Pontificia Universidad
JAVERIANA
Cali

SISTEMA DE RECOMENDACIÓN DE COLABORADORES ACADÉMICOS BASADO EN CLUSTERING DE PERFILES DE INVESTIGACIÓN

Carolina de la Espriella Alvarez

Código 9.024.710

Proyecto Aplicado para optar al título de

Magister en Ciencia de Datos

Director

Diego Fernando Mosquera Valencia

FACULTAD DE INGENIERÍA Y CIENCIAS

MAESTRÍA EN CIENCIA DE DATOS

SANTIAGO DE CALI, JUNIO DE 2026

TABLA DE CONTENIDO

LISTA DE FIGURAS

LISTA DE TABLAS

LISTA DE ANEXOS

INTRODUCCIÓN

La colaboración científica es un pilar fundamental para el avance del conocimiento y la innovación. Sin embargo, en el ecosistema académico actual, los investigadores enfrentan una dificultad recurrente y significativa: la identificación y conexión efectiva con colegas que compartan áreas de conocimiento, trayectorias o intereses investigativos afines. Este problema se manifiesta en la fragmentación de los sistemas de información existentes, como CvLAC y las plataformas institucionales, y la ausencia de mecanismos inteligentes de análisis que permitan un aprovechamiento óptimo de los datos disponibles sobre el capital humano científico del país. Como resultado, los procesos de búsqueda de colaboradores suelen ser manuales, lentos, ineficientes y propensos a sesgos, limitando la formación de redes académicas robustas, la formulación de proyectos interdisciplinarios de alto impacto y, en última instancia, debilitando la producción científica nacional y la capacidad de abordar retos complejos de manera colaborativa.

La persistencia de esta desconexión no solo subutiliza el potencial investigativo individual y colectivo, sino que también impacta negativamente la colaboración interinstitucional e intrainstitucional. Esto dificulta la conformación de equipos de investigación diversificados y con la masa crítica necesaria para competir en convocatorias nacionales e internacionales, además de frenar la transferencia de conocimiento y la alineación con los principios de la ciencia abierta. La optimización de la búsqueda de colaboradores permitiría reducir la duplicidad de esfuerzos, maximizar el uso de recursos y fomentar la creación de sinergias que impulsen la investigación de frontera.

Para solucionar este desafío, el presente proyecto presenta el diseño y desarrollo de un "Sistema de Recomendación de Colaboradores Académicos Basado en Clustering de Perfiles de Investigación". La solución se fundamentó en la aplicación de técnicas de ciencia de datos, específicamente el aprendizaje no supervisado (clustering), sobre perfiles de investigadores extraídos de fuentes públicas como las bases de datos de Minciencias (SNCTI) y CvLAC. El proyecto se desarrolló siguiendo una adaptación del ciclo de vida CRISP-DM, caracterizado por su naturaleza iterativa y cíclica. Este enfoque metodológico se compuso de las siguientes fases clave:

1. Comprensión del Negocio: Definición del problema, delimitación del alcance, identificación de los objetivos del proyecto y establecimiento de los criterios de éxito desde la perspectiva de la colaboración académica y la recomendación de perfiles compatibles.
2. Comprensión de los Datos: Revisión de la fuente de información disponible, análisis de la estructura del conjunto de datos, exploración inicial de variables, identificación de patrones relevantes y verificación preliminar de calidad para determinar su pertinencia frente al objetivo del proyecto.
3. Preparación de Datos: Recolección, limpieza, transformación e ingeniería de características para construir una base de datos consolidada y estructurada de perfiles de investigadores.
4. Modelado: Selección, aplicación y calibración de algoritmos de clustering para identificar agrupaciones naturales y significativas de investigadores con perfiles similares.
5. Evaluación: Validación rigurosa del modelo de segmentación mediante métricas cuantitativas y cualitativas para asegurar su coherencia y robustez.
6. Despliegue: Desarrollo de un prototipo funcional de herramienta interactiva que integre la visualización exploratoria de los segmentos identificados y permita el ingreso dinámico de perfiles para obtener recomendaciones personalizadas.

Como resultados principales de este proyecto, se generó una base de datos depurada de investigadores nacionales como insumo para el análisis, se validó el modelo computacional de segmentación que identificó

grupos de investigadores con características afines y, de manera fundamental, se desarrolló un prototipo funcional de la herramienta interactiva. Este prototipo permite a los usuarios explorar los segmentos de la comunidad científica, filtrar su perfil de interés y recibir recomendaciones automatizadas de potenciales colaboradores. De esta manera, se desarrolló una herramienta con el potencial de crear conexión entre investigadores y facilitar la colaboración interdisciplinaria, fortaleciendo el ecosistema científico nacional.

1. CONTEXTUALIZACIÓN DEL PROYECTO

1.1. Definición del problema

1.2. Planteamiento del problema

En el ecosistema científico para la generación de conocimiento, los investigadores enfrentan dificultades recurrentes para identificar y establecer contacto con colegas que compartan áreas de conocimiento, trayectorias académicas o intereses investigativos comunes [1], [2]. Esta desconexión impide el fortalecimiento de redes académicas, limita la eficiencia en la formulación de proyectos interdisciplinarios y debilita la producción científica nacional [3].

A pesar de que se cuenta con la existencia de bases de datos institucionales [4], [5], [6], [7], [8], [9] y plataformas como el Sistema Nacional de Ciencia, Tecnología e Innovación (SNCTI) de Minciencias, la fragmentación de los sistemas y la ausencia de mecanismos inteligentes de análisis para la caracterización de investigadores dificultan el aprovechamiento efectivo de los datos disponibles [10]. Como resultado, los investigadores suelen recurrir a estrategias manuales para la búsqueda de aliados académicos. Estos métodos, que incluyen la revisión de publicaciones, la asistencia a eventos o el uso de redes de contacto personales, son a menudo poco eficientes, consumen tiempo valioso y pueden estar sesgados por la visibilidad o el reconocimiento previo de ciertos individuos, dejando de lado a potenciales colaboradores valiosos, pero menos conocidos. Este panorama impacta negativamente la colaboración tanto interinstitucional como intra-institucional, dificultando la conformación de equipos de investigación diversificados y con la masa crítica necesaria para abordar retos científicos complejos [11].

A nivel internacional, diversos estudios han evidenciado que los sistemas de recomendación de perfiles académicos, basados en técnicas de minería de datos y aprendizaje automático, pueden facilitar la detección de comunidades científicas con intereses comunes y aumentar la producción colaborativa [12], [13]. Estos enfoques permiten aplicar modelos de aprendizaje no supervisado, como el clustering, para identificar grupos latentes dentro de un conjunto complejo de trayectorias académicas [14].

Además, estudios recientes han mostrado que la representación visual de redes colaborativas y la exploración de afinidades tienen efectos positivos en la promoción de proyectos interdisciplinarios [15]. Por otro lado, herramientas basadas en sistemas de recomendación personalizados han sido empleadas con éxito en universidades como es el caso de la *Graduate University for Advanced Studies* en Japón, donde en el 2017 se utilizaron variables como la producción académica, redes de coautoría e intereses declarados [16].

Por otra parte, en América Latina, aún son escasos los desarrollos tecnológicos orientados a caracterizar dinámicamente a las comunidades científicas locales con base en datos abiertos y enfoques automatizados. Esta brecha tecnológica ha sido resaltada por investigaciones recientes que proponen estrategias basadas en análisis de texto y agrupamiento no supervisado para generar redes académicas emergentes [17].

En este contexto, la aplicación de la ciencia de datos se vuelve fundamental para abordar esta problemática a través del diseño de una herramienta que vaya más allá de la simple identificación de grupos de investigadores, sino que, mediante la aplicación de métodos computacionales, específicamente el aprendizaje no supervisado sobre los datos disponibles de los investigadores identifique segmentos de investigadores con perfiles y características profesionales similares. Además, de contar con la capacidad de proporcionar a cada usuario recomendaciones precisas, personalizadas y contextualizadas de perfiles académicos potencialmente compatibles. Esta solución se alinea con los principios de la ciencia abierta y

fortalece los ecosistemas científicos nacionales al facilitar conexiones significativas entre actores con trayectorias afines [18].

1.2.1. Formulación del problema

Pregunta general de investigación

¿Cómo se puede, con las herramientas de la ciencia de datos, construir un sistema inteligente no supervisado que permita a los investigadores del territorio nacional vinculados al Sistema Nacional de Ciencia, Tecnología e Innovación (SNCTI), crear agrupaciones automáticas de perfiles basadas en similitudes de trayectorias académicas y áreas de actuación mediante un sistema interactivo de segmentación y recomendación de perfiles?

1.2.2. Sistematización

1. ¿Cuáles son las variables y fuentes de datos más relevantes y accesibles para caracterizar los perfiles de investigación (formación académica, experiencia profesional, áreas de actuación, producción científica, afiliación institucional) de los investigadores adscritos al SNCTI a través de CvLAC, y cómo pueden ser estas preprocesadas y estructuradas para su análisis computacional?
2. ¿Qué técnicas de aprendizaje no supervisado (*clustering*) son más efectivas para identificar agrupaciones naturales y significativas de perfiles de investigadores con características similares, dadas las particularidades de los datos disponibles, y cómo se pueden validar la coherencia, robustez y significancia de dichas agrupaciones?
3. ¿Cómo se puede diseñar y desarrollar una herramienta de visualización interactiva que permita a los usuarios explorar de manera intuitiva los segmentos de investigadores identificados, sus características distintivas y las posibles relaciones entre ellos, facilitando la comprensión de la estructura de la comunidad científica modelada?
4. ¿Qué arquitectura y funcionalidades debe poseer un prototipo de sistema de información visual e interactivo para que los investigadores puedan ingresar su propio perfil y recibir recomendaciones personalizadas de colaboradores potencialmente compatibles, basadas en el modelo de segmentación validado?

1.3. Objetivos

1.3.1. Objetivo General

Construir una herramienta de aprendizaje no supervisado que le permita a los investigadores registrar su trayectoria académica, encontrar colaboradores potenciales y visualizar los segmentos identificados de perfiles compatibles.

1.3.2. Objetivos Específicos

1. Construir una base de datos de investigadores del territorio nacional, incluyendo variables como formación académica, experiencia profesional, áreas de actuación y afiliación institucional, con los procesos necesarios de limpieza y transformación.
2. Aplicar técnicas de aprendizaje no supervisado para identificar agrupaciones naturales de perfiles con características similares.

3. Validar el modelo computacional de segmentación obtenido, utilizando métricas de evaluación para asegurar la significancia de las agrupaciones generadas.
4. Construir una herramienta interactiva que integre la visualización exploratoria de los segmentos identificados de investigadores y permita el ingreso dinámico de datos para obtener recomendaciones de perfiles compatibles.

1.4. Marco de Referencia

El marco de referencia organiza los fundamentos conceptuales que sostienen el sistema de recomendación de colaboradores académicos basado en clustering de perfiles de investigación.

En este proyecto, la articulación teórica se construye sobre cuatro ejes. El primero corresponde a la colaboración científica y a las dificultades de búsqueda de aliados académicos en ecosistemas de investigación complejos. El segundo se relaciona con el SNCTI, CvLAC y las fuentes públicas que hacen posible representar perfiles de investigadores de manera estructurada. El tercero corresponde a los sistemas de recomendación y a los enfoques de similitud que permiten ordenar candidatos potenciales. El cuarto se refiere al aprendizaje no supervisado, la validación interna de agrupamientos y la metodología CRISP-DM como marco de organización del proyecto aplicado.

1.4.1. Marco Teórico

1.4.1.1. Colaboración científica y búsqueda de colaboradores académicos

La colaboración científica es un componente central de la producción contemporánea de conocimiento, porque permite combinar capacidades, trayectorias, recursos e intereses de investigación que difícilmente se concentran en un único investigador o grupo. En las redes universitarias, la colaboración también opera como una forma de capital social académico: conecta actores, amplifica oportunidades de investigación, facilita la formulación de proyectos interdisciplinarios y mejora la circulación de capacidades entre instituciones [2].

Sin embargo, la identificación de colaboradores nuevos no es un proceso sencillo. En ecosistemas amplios, la información sobre investigadores suele encontrarse distribuida entre repositorios institucionales, hojas de vida académicas, sistemas de información científica, publicaciones y redes de coautoría. Esta dispersión aumenta el costo de búsqueda y favorece estrategias manuales basadas en contactos previos, reconocimiento visible o revisión parcial de repositorios [19], [20]. En consecuencia, perfiles pertinentes pero menos visibles pueden quedar por fuera de las decisiones de articulación, mientras que los investigadores tienden a reproducir redes conocidas.

Desde esta perspectiva, problemas como la baja formulación de proyectos interdisciplinarios, la subutilización de capacidades científicas y la menor competitividad en convocatorias no son hechos aislados, sino manifestaciones de una misma dificultad de descubrimiento académico: el ecosistema no siempre convierte capacidades dispersas en oportunidades visibles de colaboración. Esta lectura conceptual debe quedar en el marco teórico; la comprensión del negocio solo debe registrar como esas manifestaciones delimitan el caso del proyecto.

La recomendación de colaboradores académicos busca intervenir este problema mediante mecanismos sistemáticos de descubrimiento. La literatura ha abordado la selección de colaboradores a partir de redes de coautoría, predicción de enlaces, intereses de investigación, similitud de contenido y modelos integrados de selección académica [10], [11], [13], [16]. Estos enfoques comparten una idea de fondo: la afinidad entre

investigadores puede inferirse a partir de rasgos observables de su trayectoria, producción, área, colaboraciones o posición en redes científicas. Para este proyecto, dicha afinidad se aborda desde la caracterización de perfiles y la búsqueda de semejanzas dentro de segmentos de investigadores.

La colaboración científica, por tanto, no se entiende aquí como una simple relación administrativa entre personas, sino como un fenómeno susceptible de análisis computacional cuando existen datos públicos suficientes para representar perfiles, comparar trayectorias y descubrir patrones latentes de afinidad. Esta lectura permite conectar el problema académico con técnicas de ciencia de datos sin perder de vista que la recomendación final no reemplaza el juicio del investigador: solo organiza candidatos potencialmente compatibles para apoyar el descubrimiento.

1.4.1.2. SNCTI, CvLAC y datos públicos de perfiles de investigación

El Sistema Nacional de Ciencia, Tecnología e Innovación (SNCTI) constituye el entorno institucional en el que se ubica el problema de colaboración abordado por la tesis. Minciencias, como entidad rectora del sistema, articula políticas, instrumentos y mecanismos de reconocimiento orientados al fortalecimiento de la investigación, el desarrollo tecnológico y la innovación en Colombia [21]. En este contexto, los registros asociados a investigadores, grupos e instituciones no solo cumplen una función administrativa, sino que también ofrecen una base informacional para analizar capacidades científicas y dinámicas de articulación.

Dentro del SNCTI, plataformas como Scienti, CvLAC, GrupLAC e InstituLAC concentran información relevante sobre hojas de vida, grupos de investigación e instituciones [21]. CvLAC, en particular, funciona como una fuente para caracterizar trayectorias académicas, formación, filiación, áreas de conocimiento y otros atributos que permiten construir representaciones comparables de investigadores. La disponibilidad de datos abiertos de Minciencias y la orientación de ciencia abierta refuerzan la pertinencia de utilizar fuentes públicas para construir análisis replicables y auditables [22], [23].

La existencia de estos registros no elimina los problemas de acceso, integración y búsqueda. La literatura y el diagnostico institucional han señalado limitaciones en la consulta semántica, la actualización manual de información, la fragmentación de plataformas y la ausencia de mecanismos analíticos que faciliten recomendaciones sistemáticas de colaboradores [21]. Por ello, el valor teórico de estas fuentes no reside solo en su disponibilidad, sino en la posibilidad de transformarlas en perfiles computacionales consistentes.

En este documento, el capítulo de comprensión del negocio registra cifras, requisitos, restricciones y decisiones especificas del proyecto. En el marco teórico, en cambio, el SNCTI y CvLAC deben tratarse como el ecosistema conceptual y de datos que hace viable el sistema de recomendación. Esta separación evita trasladar al marco la totalidad del diagnóstico operativo y conserva en el capítulo 2 la función CRISP-DM de delimitar recursos, supuestos, criterios de éxito y restricciones.

1.4.1.3. Sistemas de recomendación para colaboración académica

Los sistemas de recomendación son técnicas y herramientas orientadas a sugerir elementos potencialmente relevantes para un usuario a partir de información disponible sobre usuarios, ítems, preferencias, contenido o relaciones [24]. En dominios académicos, los ítems recomendados pueden ser artículos, cursos, eventos, revisores, fuentes de financiación o colaboradores potenciales. En todos los casos, el problema consiste en reducir la sobrecarga de información y priorizar alternativas que resulten pertinentes para una necesidad de búsqueda [25], [26], [27].

En este sentido, el paso de estrategias reactivas de búsqueda a mecanismos proactivos, sistemáticos y

basados en evidencia pertenece a la fundamentación conceptual de los sistemas de recomendación académica. El aporte del sistema no consiste en sustituir la decisión del investigador, sino en ampliar el universo visible de candidatos y ordenar alternativas mediante criterios explícitos de representación y similitud.

La literatura distingue varias familias de recomendación. El filtrado basado en contenido recomienda elementos similares a aquellos que comparten atributos con el perfil o las preferencias conocidas del usuario [24], [28]. El filtrado colaborativo utiliza patrones de comportamiento o valoraciones de usuarios similares para inferir nuevas sugerencias [24], [28]. Los enfoques híbridos combinan fuentes y métodos para mitigar limitaciones de cada familia, como el cold start, la sobreespecialización o la escasez de datos [29], [30], [31]. también existen sistemas basados en conocimiento y enfoques basados en grafos, útiles cuando el dominio requiere reglas explicitas o representación de relaciones complejas [32].

En recomendación académica, los enfoques de redes, grafos heterogéneos, análisis de intereses y contenido de publicaciones han mostrado utilidad para descubrir artículos, temas o colaboradores potenciales [11], [33], [34], [35]. No obstante, estos sistemas enfrentan desafíos propios del dominio: sparsity, cold start, trayectorias académicas heterogéneas, diferencias disciplinares y necesidad de interpretar la recomendación más allá de una puntuación numérica. En colaboración científica, una sugerencia no es simplemente un ítem consumible, sino una posible relación académica que debe ser comprensible y justificable.

Este proyecto adopta una aproximación compatible con recomendación basada en contenido y similitud de perfiles. El sistema no parte de valoraciones históricas de usuarios ni de una variable objetivo de colaboraciones exitosas, sino de atributos observables de investigadores. En este sentido, la recomendación se construye a partir de la representación vectorial de perfiles, la segmentación previa mediante clustering y el ordenamiento de candidatos por proximidad dentro de un universo filtrado. Esta decisión es coherente con un escenario en el que se dispone de datos públicos de perfiles, pero no de etiquetas externas que permitan entrenar un modelo supervisado de afinidad.

1.4.1.4. Representación de perfiles, similitud y recomendación basada en contenido/k-NN

La representación de perfiles es el puente entre el problema académico y el sistema computacional. Un perfil de investigación puede entenderse como una caracterización estructurada de un investigador a partir de dimensiones como formación, trayectoria, área de conocimiento, filiación, producción, movilidad o participación en convocatorias. Para que estos perfiles puedan compararse, deben transformarse en atributos medibles, codificados y normalizados que permitan calcular semejanzas entre investigadores.

En los sistemas basados en contenido, la recomendación depende de la comparación entre representaciones de ítems o usuarios. En este proyecto, el ítem recomendado es un investigador potencialmente compatible y la comparación se realiza sobre variables que sintetizan su trayectoria y características académicas. La literatura sobre recomendación científica y métodos basados en contenido respalda el uso de atributos observables para ordenar elementos relevantes cuando no existen valoraciones previas suficientes [36], [37], [38].

La similitud vectorial permite ubicar perfiles en un espacio de características y estimar la cercanía entre ellos. Por otro lado, la distancia euclidiana y la similitud coseno son medidas frecuentes para comparar vectores, aunque responden a propiedades distintas: la distancia euclidiana mide separación geométrica y la similitud coseno mide cercanía de orientación entre vectores [36], [37], [39].

En el marco de este proyecto, el método k-Nearest Neighbors (k-NN) se utiliza como estrategia de recuperación: dado un perfil de consulta, identifica los candidatos más cercanos en el espacio de similitud. Asimismo, la eficiencia de esta búsqueda puede apoyarse en estructuras espaciales como Ball Tree, adecuadas para consultas de vecinos más cercanos en espacios métricos de dimensión moderada [40], [41].

1.4.1.5. Aprendizaje no supervisado y clustering de perfiles de investigación

El aprendizaje no supervisado agrupa técnicas orientadas a descubrir estructura en datos sin contar con etiquetas de salida predefinidas. Dentro de este campo, el clustering o análisis de conglomerados busca organizar observaciones en grupos de manera que los elementos de un mismo clúster sean más similares entre sí que respecto a elementos de otros grupos [42]. En el contexto de perfiles de investigación, esta técnica permite identificar segmentos de investigadores con trayectorias, características o patrones académicos afines.

La aplicación de clustering a perfiles de investigadores ha sido explorada en estudios sobre redes de colaboración, comunidades científicas, agrupación de investigadores por producción o intereses, y recomendación académica [3], [10], [17], [39], [43]. Estos trabajos muestran que el agrupamiento puede servir para construir tipologías, detectar comunidades o apoyar decisiones de recomendación. Sin embargo, también evidencian que la calidad de la segmentación depende de la representación de los perfiles, el preprocesamiento, la selección de variables y la interpretabilidad de los grupos resultantes.

Existen distintas familias de algoritmos de clustering. Los métodos particionales, como K-Means y K-Means++, dividen el conjunto de datos en un numero predefinido de grupos y buscan minimizar la variación interna [39], [42], [43]. Los métodos probabilísticos, como los modelos de mezclas gaussianas, permiten representar pertenencias suaves y estructuras elipsoidales. Los métodos jerárquicos construyen agrupaciones anidadas que pueden analizarse a distintos niveles de granularidad, mientras que los métodos basados en densidad identifican regiones densas separadas por zonas de baja densidad y pueden detectar ruido [42], [43], [44].

La selección de un algoritmo no debe responder solo a su rendimiento numérico. también importan la escala del dataset, la estabilidad, la interpretabilidad de los grupos, la cobertura de la población objetivo y la coherencia con los objetivos de negocio o investigación. Por ello, una estrategia multiparadigma permite comparar supuestos geométricos distintos y reducir la dependencia de un único enfoque.

Para este proyecto, se combinan cuatro enfoques complementarios de clustering. Dado que cada paradigma opera bajo supuestos geométricos y estadísticos distintos, la comparación sistemática de sus resultados permite evaluar la robustez de las agrupaciones identificadas y seleccionar la configuración óptima para representar la heterogeneidad del ecosistema científico colombiano. De esta manera, se evita la dependencia de un único enfoque y se fortalece la validez de las conclusiones derivadas del modelado.

En la Tabla 1 se presenta la comparación de los cuatro algoritmos candidatos, donde se detalla el paradigma al que pertenecen, su principio de funcionamiento y su ventaja principal en este contexto.

Criterio	K-Means++	GMM	Agglomerative	HDBSCAN
Paradigma	Basado en centroides (partitional)	Basado en distribución (probabilístico)	Basado en conectividad (jerárquico)	Basado en densidad (jerárquico-densidad)
Principio	Minimización de	Maximización de	Fusión	Jerarquía de densidad

Criterio	K-Means++	GMM	Agglomerative	HDBSCAN
	varianza intra-clúster	verosimilitud de mezcla de k Gaussianas multivariadas (EM)	aglomerativa: inicia con n clústeres y fusiona los más cercanos	basada en mutual reachability distance; extracción de clústeres estables
Tipo de asignación	Cada punto a un solo clúster	Probabilidades de pertenencia a cada clúster	Corte del dendrograma a una altura definida	Con etiqueta de ruido (−1) para baja densidad
Supuesto geométrico	Clústeres esféricos de varianza comparable	Clústeres elipsoidales con covarianza propia	Sin supuesto; depende del enlace	Sin supuesto; detecta densidad variable
Ventaja principal	Escalabilidad; centroides interpretables	Cuantificación de incertidumbre; BIC/AIC	Dendrograma multi-granularidad	No requiere predefinir k; robusto ante ruido

Tabla 1. Comparación de paradigmas de clustering seleccionados para la segmentación de perfiles del SNCTI [42], [45], [46]

1.4.1.6. Validación de clustering sin variable objetivo

La validación de clustering plantea un reto especifico porque, al no existir una etiqueta verdadera de referencia, no es posible evaluar el modelo con métricas supervisadas convencionales. La calidad del agrupamiento debe analizarse mediante evidencias internas, estabilidad, interpretabilidad y coherencia con el dominio. En consecuencia, las métricas internas no prueban por si solas que los grupos sean "verdaderos" en sentido sustantivo, pero si permiten examinar si la estructura encontrada es consistente, separable y técnicamente defendible [47], [48], [49].

Inicialmente, como paso previo a la aplicación del clustering, identificar la tendencia de los datos al agrupamiento permite evaluar si éstos presentan una estructura distinta de una distribución aleatoria. Para esto, el estadístico de Hopkins es una herramienta que permite estimar esta tendencia y apoyar la decisión de si el agrupamiento resulta metodológicamente pertinente [50], [51].

Ahora, posterior a la aplicación de las técnicas de clustering, es indispensable realizar una la validación interna de la segmentación que permite definir la calidad de las agrupaciones generadas. El Coeficiente de Silueta evalúa la relación entre cohesión intra-clúster y separación respecto al clúster vecino [47], el índice de Calinski-Harabasz compara dispersión entre grupos frente a dispersión interna [43], el índice de Davies-Bouldin penaliza configuraciones con clusters poco compactos o muy similares entre sí [52],y y por último, el método del codo analiza el rendimiento marginal de aumentar el número de clusters y ayuda a identificar puntos de inflexión en la inercia [53]. Es importante tener en cuenta que ninguna de estas métricas debe ser considerada de forma aislada, su valor aumenta cuando convergen con estabilidad, interpretabilidad y coherencia con el objetivo del sistema.

La literatura reciente sobre índices internos y protocolos de validación muestra que la elección de métricas depende de la forma, densidad, separación y dimensionalidad de los datos [44], [45], [54], [55], [56], [57], [58], [59]. Esto es especialmente relevante para perfiles de investigadores, donde pueden coexistir dimensiones académicas, institucionales, territoriales y temporales. Por ello, la validación debe presentarse como una acumulación de evidencias técnicas, no como demostración definitiva de utilidad práctica.

En este proyecto, la ausencia de variable objetivo y de validación con investigadores reales delimita el alcance de las conclusiones. Las métricas internas permiten evaluar la estructura del modelo y su coherencia técnica; las pruebas de cordura y análisis posteriores pueden apoyar la interpretabilidad; pero la validación

practica con usuarios pertenece a una etapa futura. Esta distinción es fundamental para evitar sobreafirmar los resultados del sistema.

1.4.1.7. CRISP-DM como marco metodológico del proyecto aplicado

CRISP-DM (Cross-Industry Standard Process for Data Mining) es un modelo de proceso para proyectos de minería de datos y ciencia de datos que organiza el ciclo de vida en fases: comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación y despliegue [60]. Su aporte principal es ofrecer una estructura iterativa que conecta objetivos del proyecto, datos disponibles, decisiones de modelado, evaluación de resultados y condiciones de uso o despliegue.

La naturaleza iterativa de CRISP-DM resulta pertinente para proyectos aplicados en los que la comprensión del problema, la disponibilidad de datos y las decisiones técnicas se ajustan progresivamente [60], [61]. En lugar de concebir el análisis como una secuencia lineal, el modelo permite volver sobre fases anteriores cuando aparecen restricciones, hallazgos de calidad de datos, limitaciones metodológicas o necesidades de reinterpretación. Esta flexibilidad es relevante para el desarrollo de un sistema de recomendación basado en datos públicos, dado que la fuente, la preparación, la segmentación y el prototipo se condicionan mutuamente.

1.4.1.8. Alcance de prototipos aplicados y límites de validación en sistemas de recomendación

Un sistema de recomendación aplicado puede tener distintos niveles de madurez: formulación conceptual, modelo entrenado, motor funcional, prototipo interactivo, uso controlado y validación con usuarios reales. Estos niveles no son equivalentes. Un prototipo funcional demuestra que la arquitectura puede operacionalizar el modelo y permitir consultas; una validación técnica interna muestra coherencia, estabilidad o rendimiento bajo datos y supuestos definidos; una validación practica con usuarios exige evidencia de uso, utilidad percibida, pertinencia de las sugerencias y condiciones reales de adopción.

Esta distinción es especialmente importante en sistemas de recomendación académica. Las métricas técnicas pueden indicar que los perfiles cercanos son recuperados de manera consistente, pero no prueban por sí mismas que un investigador considere pertinente iniciar una colaboración. La utilidad académica depende de factores adicionales: interpretabilidad de la recomendación, confianza en los datos, vigencia de los perfiles, contexto disciplinar, objetivos del usuario y restricciones institucionales. La literatura sobre calidad de métricas en sistemas de recomendación advierte que las medidas deben calcularse e interpretarse de forma consistente con el objetivo evaluado [38].

En el marco de CRISP-DM, la fase de evaluación debe diferenciar entre evidencias técnicas del modelo y decisión de avance hacia uso real [60], [61]. Para este proyecto, el prototipo constituye un artefacto aplicado relevante porque integra segmentación, filtros y recuperación de perfiles en una interfaz consultable. No obstante, la validación con investigadores reales debe mantenerse como trabajo futuro si no existe evidencia empírica de pruebas con usuarios. Esta delimitación no debilita el aporte del proyecto; al contrario, precisa su alcance académico y evita atribuir al sistema una validación que todavía no se ha realizado.

1.4.2. Antecedentes

1.4.2.1. El SNCTI y su Contexto en Colombia: Marco Relevante para la Caracterización de Perfiles

El SNCTI de Colombia conforma el entorno operativo y la principal fuente de datos para el sistema de recomendación de colaboradores académicos propuesto en esta tesis. Comprender su estructura, evolución,

actores y herramientas de información es crucial para contextualizar la problemática y el alcance de la investigación.

Minciencias, como ente rector del SNCTI, heredó funciones de Colciencias y ha buscado ampliar su capacidad de coordinación intersectorial y la formulación de políticas de CTI a largo plazo [21]. Su misión es formular, orientar, dirigir, coordinar, ejecutar, controlar y evaluar la política estatal en CTI, con el fin de promover la sostenibilidad y el fortalecimiento de la investigación, el desarrollo tecnológico y la innovación, articulando estos esfuerzos con diversos sectores para generar desarrollo productivo y bienestar social [21]

Finalmente, el concepto de Ciencia Abierta ha comenzado a permear el SNCTI, con la Política Nacional de Ciencia Abierta (PNCA) 2022-2031 de Colombia [22]. Esta política busca aumentar el acceso, visibilidad, reproducibilidad y utilidad de los resultados de CTI, y valora diversas formas de ciencia, incluyendo saberes tradicionales [22]. Este contexto de apertura y la disponibilidad de datos abiertos oficiales, como los de Minciencias [23], son fundamentales para la viabilidad y pertinencia de la presente propuesta de investigación.

1.4.2.2. Plataformas Nacionales de Información: Limitaciones y Oportunidades

Dentro del SNCTI, existen diversas plataformas nacionales de información que actúan como fuentes primarias de datos sobre investigadores en Colombia. Entre ellas, destacan Scienti, que consolida la información de CvLAC (hojas de vida de investigadores), GrupLAC (información de grupos de investigación) e InstituLAC (consolida información de las instituciones) [21]. Si bien estas plataformas son esenciales, la herramienta de consulta principal, Scienti, presenta limitaciones significativas que obstaculizan la colaboración científica efectiva y la identificación precisa de perfiles afines.

Una evaluación exhaustiva del SNCTI [21] indica que Scienti carece de una búsqueda semántica sofisticada y de un análisis colaborativo de redes. La identificación de expertos específicos es un desafío, y la calidad de los datos se ve comprometida por el ingreso manual. Además, se ha observado la falta de mecanismos eficaces de recomendación de los colaboradores. Esta ausencia de herramientas analíticas y de interconectividad sistemáticas representa un desafío importante, a pesar de la disponibilidad de registros estatales.

1.4.2.3. Sistemas de Recomendación en el Dominio Académico: Avances y Limitaciones

Los sistemas de recomendación (SR) han emergido como herramientas cruciales en el ámbito académico para mitigar la sobrecarga de información y facilitar conexiones relevantes. La literatura existente revela un creciente interés en aplicar estos sistemas para diversos fines, desde la recomendación de artículos científicos hasta la identificación de posibles colaboradores, revisores o incluso sedes para publicación.

Una revisión sistemática de quince años de investigación sobre SR en educación superior (2007-2021) [26] muestra un aumento constante en las publicaciones, con un pico en 2017. Este estudio identifica dieciséis temas principales, siendo el *e-learning* y la selección de cursos los más frecuentes. Clasifica los SR en enfoques basados en contenido, colaborativos e híbridos, y destaca la utilidad de herramientas como VOSviewer para el análisis bibliométrico y la identificación de tendencias. Otro estudio que abarca un período similar (2015-2020) y se centra en SR que apoyan prácticas educativas [27], corrobora la orientación hacia la educación formal universitaria y los estudiantes como principales usuarios. Confirma la prevalencia

del filtrado colaborativo, basado en contenido e híbrido, con una tendencia al uso de *machine learning* desde 2019. Ambas revisiones señalan la necesidad de integrar datos de múltiples fuentes y explorar enfoques semánticos para mejorar las recomendaciones.

En el contexto específico de la recomendación académica, se han propuesto diversos métodos. Un enfoque novedoso para la recomendación de artículos científicos, denominado CARE, explota las relaciones de autor común y las preferencias históricas de los investigadores en un grafo heterogéneo [33]. Este método identifica patrones de búsqueda basados en autores e incorpora estas relaciones para mejorar las recomendaciones, demostrando su efectividad en el dataset CiteULike. Otros trabajos proponen modelos de *embedding* de redes heterogéneas (PR-HNE) para la recomendación de artículos, buscando superar limitaciones de enfoques previos que ignoran factores como citaciones, colaboración, temas y etiquetas, y abordando problemas de *cold start* y dispersión de datos [34].

Los sistemas de recomendación basados en intereses también han sido explorados en redes académicas, utilizando Análisis de Redes Sociales (SNA) para recomendar nuevos colaboradores y temas de investigación, demostrando ser efectivos para abordar problemas de *sparsity* y *cold start* [35]. La explotación de redes de colaboración y contenido de publicaciones para la recomendación de colaboradores ha sido implementada en modelos híbridos como CCRec, que utiliza clustering de temas (Word2vec con K-means) y random walk para calcular la influencia, mostrando ser efectivo en el descubrimiento de nuevos colaboradores [11].

A pesar de estos avances, la literatura también identifica limitaciones. Muchos sistemas tradicionales de recomendación en el ámbito académico, aunque útiles, pueden no capturar la complejidad completa de las afinidades interdisciplinarias o las trayectorias evolutivas de los investigadores [34].

1.4.2.4. Aplicación de Técnicas de Agrupamiento a Perfiles de Investigación: Técnicas y Aplicaciones Previas

La caracterización y representación de perfiles de investigadores es un paso fundamental para cualquier sistema de recomendación de colaboradores. Las técnicas de agrupamiento, una forma de aprendizaje no supervisado, se han empleado para identificar comunidades científicas, tipologías de investigadores o afinidades latentes basadas en sus perfiles.

Un estudio de Ezugwu et al. [42] ofrece una revisión amplia de algoritmos de clustering y sus aplicaciones, incluyendo la recuperación de información y el clustering de texto, relevantes para el análisis de perfiles basados en producción científica. Se destaca la importancia del preprocesamiento de datos y se mencionan algoritmos como K-Means y jerárquicos. Rodríguez-Bárcenas (2022) [39] propone un método integral basado en minería de texto y análisis de datos para perfilar investigadores y agruparlos por áreas de conocimiento e intereses, demostrando su efectividad en casos de estudio en Cuba y Ecuador para formar Comunidades Colectivas de Conocimientos.

Los estudios comparativos de algoritmos de *clustering* son valiosos para seleccionar la técnica más adecuada. Un trabajo que compara cinco algoritmos de clustering (*meta-learning, self-supervised learning, reinforcement learning, explainable clustering* y *deep clustering*) para perfilar investigadores usando datos de Google Scholar [43] encontró que *meta-learning for clustering* tuvo el mejor rendimiento.

La validación del *clustering* es un aspecto crítico. La literatura sobre índices de validez de *clustering* es

extensa y aborda tanto la validación interna como la externa. Trabajos como el de Dudek del 2020 [47] evalúan el Índice *Silhouette*, mientras que Eustáquio et al. [44] proponen protocolos para evaluar índices internos en datos de alta dimensión utilizando *Fuzzy Cluster Validity Indices* (FCVIs) con técnicas de *Soft Subspace Clustering (SSC).* Karanikola et al. [48] y otros [49] comparan empíricamente múltiples índices de validez para estimar el número óptimo de clústeres en diversos datasets. La elección del índice de validez y del algoritmo de *clustering* puede depender significativamente de la naturaleza de los datos y de las propiedades deseadas de los clústeres [54], [45]. Trabajos recientes proponen nuevos índices de validación interna [55], [56], [57] o marcos para su evaluación [58], [59], destacando la importancia de considerar la forma, densidad y separación de los clústeres.

A pesar de la diversidad de técnicas y aplicaciones, la agrupación de perfiles de investigación enfrenta desafíos particulares, como la alta dimensionalidad de los datos textuales, la naturaleza evolutiva de los intereses de investigación y la necesidad de métricas de evaluación que capturen la coherencia y significancia de las comunidades identificadas.

2. COMPRENSIÓN DEL NEGOCIO

En CRISP-DM, la comprensión del negocio traduce el problema del dominio en objetivos, condiciones de ejecución y criterios que orientan las fases posteriores de datos, modelado, evaluación y despliegue [60]. En esta tesis, el marco de referencia explica el fundamento conceptual del problema y de la solución; este capítulo delimita su aplicación al caso del SNCTI y deja definido el alcance operativo del proyecto.

2.1. Delimitación del caso y problema de negocio

El proyecto se orienta a la comunidad de investigadores del territorio nacional como beneficiaria final de la solución. El ecosistema colombiano de ciencia, tecnología e innovación ha crecido de forma sostenida: los investigadores reconocidos por Minciencias pasaron de 8016 en 2013 a 25478 en los resultados preliminares de 2024, y los grupos de investigación reconocidos aumentaron de 4304 a 6087 en el mismo periodo [23], [62], [63].

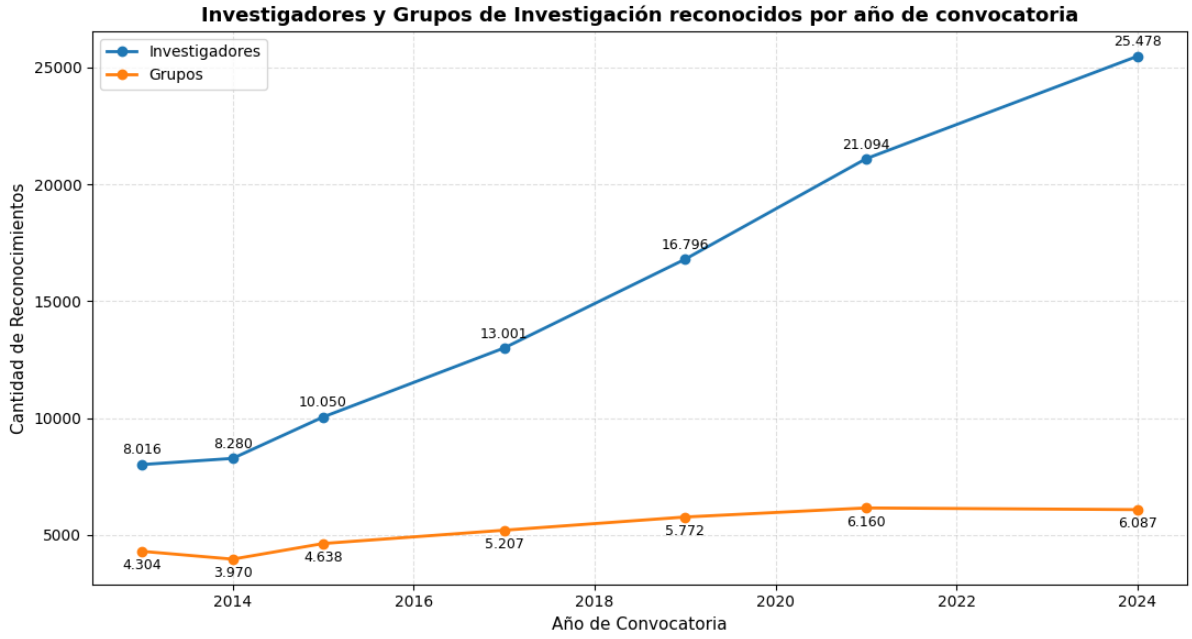

Figura 1. Investigadores y grupos de investigación reconocidos por categoría

Aunque existen cifras generales de 2024, los datos detallados de esa convocatoria no estaban disponibles para las fases de análisis desarrolladas en la tesis. Por esta razón, el proyecto trabaja con la última versión consolidada y estable disponible para ejecución: la Convocatoria 2021. Esta decisión delimita el corte temporal del sistema, sin modificar el diseño metodológico ni impedir actualizaciones posteriores cuando se publique una nueva base oficial.

2.1.1. Problema de negocio

Para efectos de la fase de comprensión del negocio, el problema se delimita como la necesidad de contar con un mecanismo sistemático que permita identificar investigadores con perfiles potencialmente complementarios dentro del ecosistema científico nacional. Esta necesidad operativa afecta la eficiencia de

los procesos de búsqueda de colaboradores, la visibilidad de perfiles afines y la capacidad de conformar redes académicas con base en criterios explícitos de similitud, trayectoria y afinidad investigativa.

En la Figura 2 se presenta el árbol del problema de negocio, el cual permite organizar sus principales causas, manifestaciones y efectos, y sirve como insumo para traducir la necesidad identificada en objetivos de negocio y objetivos de minería de datos:

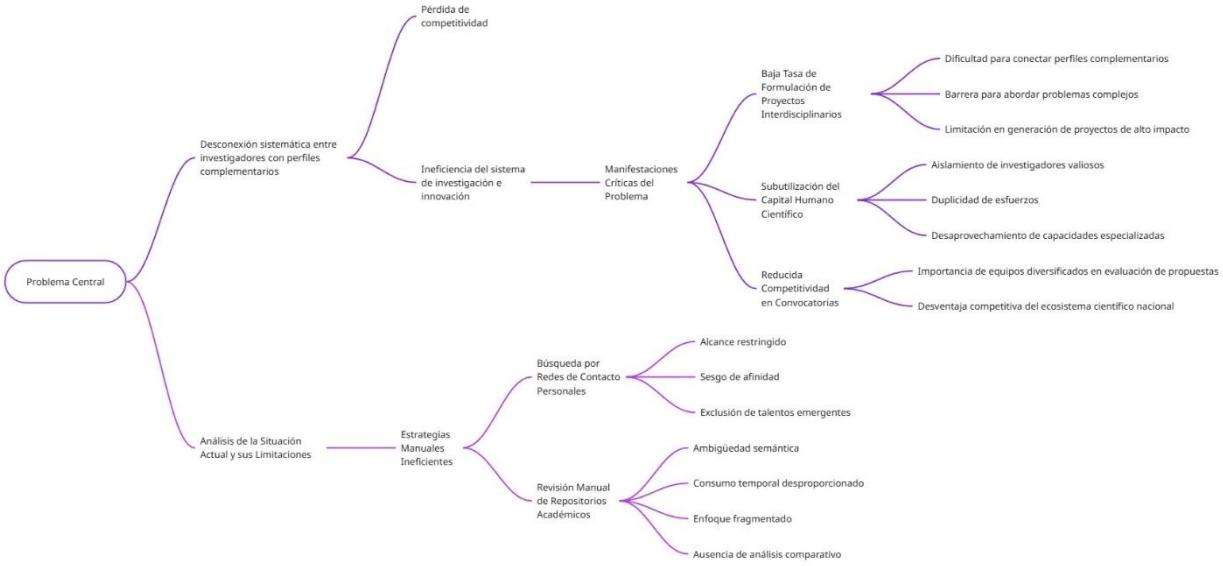

Figura 2. Árbol del problema de negocio del sistema de recomendación

2.1.2. Objetivos del negocio

2.1.2.1. Objetivo general del negocio

Desarrollar e implementar un sistema inteligente de recomendación que facilite la identificación y conexión de investigadores con perfiles complementarios, contribuyendo al fortalecimiento del ecosistema científico nacional mediante la optimización de la formación de redes de colaboración académica.

2.1.2.2. Objetivos Específicos de Negocio

Objetivo de Eficiencia Operacional: Reducir significativamente el tiempo y esfuerzo requerido para la identificación de potenciales colaboradores académicos.

Objetivo de Descubrimiento: Facilitar el descubrimiento de colaboradores altamente compatibles que los investigadores no habrían identificado mediante métodos tradicionales.

Objetivo de Comprensión del Ecosistema: Proporcionar una visión estructurada y analítica del ecosistema científico nacional que permita identificar nichos, fortalezas y oportunidades de colaboración.

Objetivo de Democratización del Acceso: Mejorar la visibilidad y accesibilidad de investigadores con perfiles valiosos, pero de menor reconocimiento en el ecosistema científico.

2.2. Evaluación de la situación actual

2.2.1. Inventario de Recursos Disponibles

En esta sección se identifican los recursos disponibles para el desarrollo del proyecto, considerando los componentes humanos, de datos y tecnológicos que soportan su ejecución. Este inventario permite establecer las capacidades iniciales, así como delimitar las condiciones operativas necesarias para avanzar en las fases de preparación, modelado, evaluación y despliegue del sistema de recomendación propuesto.

2.2.1.1. Recursos de Datos

Las fuentes de datos primarias del proyecto corresponden a las bases de datos de Minciencias disponibles en el Portal de Datos Abiertos de Colombia, las cuales contienen información oficial sobre los resultados de las convocatorias de reconocimiento y medición de investigadores del SNCTI.

Características de los Datos Disponibles:

- Volumen: Cobertura estimada de más del 80% de investigadores activos registrados en el sistema nacional de ciencia y tecnología.
- Variedad: Datos estructurados (identificadores, fechas, clasificaciones) y semi-estructurados (áreas de actuación, palabras clave).
- Accesibilidad: Datos de dominio público disponibles a través de portales web oficiales.

2.2.1.2. Recursos Tecnológicos

Infraestructura Computacional:

- Computador personal y máquina virtual universitaria: Se cuenta con una máquina virtual universitaria proporcionada por la Pontificia Universidad Javeriana Cali, junto con un computador personal con las siguientes características técnicas: LENOVO 82HU con AMD Ryzen 7 5700U, 16 GB de RAM, SSD NVMe y Windows 11 Pro de 64 bits..
- Software de Código Abierto: Utilización exclusiva de herramientas open source incluyendo Python, R, bibliotecas de machine learning y herramientas de visualización.

2.2.2. Análisis de Requisitos, Supuestos y Restricciones

2.2.2.1. Requisitos del Proyecto

- Funcionalidad del Sistema: Capacidad de generar recomendaciones de colaboradores en tiempo real.
- Cobertura de Datos: Procesamiento de al menos el 80% de perfiles de investigadores activos disponibles en las fuentes oficiales.
- Interpretabilidad: Los clústeres resultantes deben ser cualitativamente interpretables y coherentes desde la perspectiva del dominio académico.

2.2.2.2. Supuestos del Proyecto

Supuestos sobre Calidad de Datos:

- Completitud Suficiente: Se asume que los perfiles en CvLAC contienen información suficiente para caracterizar adecuadamente a los investigadores, aunque se reconoce la existencia de datos faltantes.

- Actualización Periódica: Se supone que los investigadores mantienen actualizados sus perfiles con una frecuencia que permite capturar su estado actual de investigación.
- Representatividad: Se asume que la muestra de investigadores registrados en las fuentes oficiales es representativa del ecosistema científico nacional.

Supuestos Metodológicos:

- Efectividad del Clustering: Se supone que las técnicas de clustering no supervisado pueden identificar patrones significativos de afinidad entre perfiles de investigación.

2.2.2.3. Restricciones del Proyecto

Restricciones Técnicas:

- Software Open Source: Utilización exclusiva de herramientas de código abierto, excluyendo licencias comerciales.
- Capacidad Computacional: Limitaciones en el procesamiento de datasets extremadamente grandes que podrían requerir recursos de computación distribuida.

Restricciones Éticas y Legales:

- Fuentes Públicas Exclusivamente: Manejo únicamente de datos de dominio público, excluyendo información privada o confidencial.
- Anonimización: Protección de información personal sensible no relevante para el perfilamiento académico.

2.2.3. Evaluación de Riesgos y Planes de Contingencia

2.2.3.1. Riesgos Identificados y Análisis

- **Calidad y Granularidad de Datos**
  Descripción del Riesgo: Los datos disponibles en las fuentes públicas pueden no contener información de calidad adecuada para diferenciar perfiles de manera significativa, comprometiendo la efectividad del *clustering*.

  Acción: Implementación de *web scraping* dirigido a perfiles públicos de CvLAC para enriquecer el dataset con variables específicas.

- **Limitaciones de Infraestructura**
  Descripción del Riesgo: La infraestructura computacional disponible puede resultar insuficiente para procesar *datasets* de gran tamaño o implementar algoritmos computacionalmente intensivos.

  Acción: Implementación de versiones optimizadas de algoritmos de *clustering* con menor complejidad computacional.

3. COMPRENSIÓN Y PREPARACIÓN DE LOS DATOS

Este capítulo integra dos fases consecutivas de CRISP-DM: la comprensión de los datos y la preparación de los datos. La integración responde a la necesidad de presentar el tratamiento de datos como un bloque metodológico continuo, sin separar el diagnóstico del dataset y las transformaciones que permiten llegar a la matriz de modelado. La primera fase verifica procedencia, estructura, cobertura, patrones descriptivos y calidad del conjunto de información; la segunda transforma ese insumo en una base depurada, integrada, enriquecida con indicadores y formateada para el modelado.

3.1. Alcance de la fase integrada y fuente de datos

En esta sección se presenta la primera fase del proceso, orientada a verificar que el archivo fuente esté disponible, íntegro y legible, y que su estructura sea consistente con lo esperado para avanzar. El objetivo es dejar trazabilidad sobre el insumo, registrar la versión utilizada y justificar que los análisis posteriores se apoyan en un dataset estable.

El insumo principal para el desarrollo de este proyecto fue extraído del Portal de Datos Abiertos de Colombia. Esta fuente soporta la trazabilidad de versiones y asegura disponibilidad para auditoria y replicación. Durante la formulación del proyecto se había previsto trabajar con la versión más actualizada que incorporara los resultados de la Convocatoria 957. Sin embargo, debido al aplazamiento de los resultados definitivos de dicha convocatoria, se desarrolló el proyecto con la última versión consolidada y estable disponible antes del aplazamiento: la Convocatoria nacional para el reconocimiento y medición de grupos de investigación, desarrollo tecnológico o de innovación y para el reconocimiento de investigadores del SNCTI 2021.

Esta decisión garantiza continuidad analítica, permite producir hallazgos útiles y deja explicita la posibilidad de actualizar los resultados cuando se disponga de una nueva versión oficial. Metodológicamente, el cambio afecta el corte temporal de la base, pero no altera el diseño de variables, las definiciones operativas ni los procedimientos de control implementados.

El archivo fuente utilizado presenta una estructura tabular compuesta por 77.237 registros y 30 variables, correspondientes a información de identificación de convocatoria, datos demográficos, ubicación geográfica, área de conocimiento, nivel de formación, clasificación del investigador, condiciones poblacionales e institución de filiación. La revisión inicial permitió establecer que el conjunto combina variables numéricas discretas, variables numéricas continuas, campos temporales y variables categóricas nominales u ordinales. Esta composición es consistente con el objetivo del proyecto, dado que permite caracterizar a los investigadores desde dimensiones académicas, disciplinares, territoriales e institucionales.

Para efectos de trazabilidad metodológica, el diccionario completo de variables debe conservarse en el Anexo 1. En el cuerpo se conserva la descripción estructural porque justifica la viabilidad de construir perfiles comparables de investigadores.

3.2. Descripción inicial, coherencia y valores faltantes

A partir del perfilamiento descriptivo, se validó la coherencia de los encabezados, se confirmó la ausencia de filas duplicadas o vacías y se verifico que no existieran columnas constantes. Estos resultados indican un punto de partida estructuralmente limpio, sin artefactos evidentes de carga que pudieran introducir sesgos en los análisis posteriores.

Asimismo, se identificó cobertura completa en variables relevantes como edad, año de convocatoria y

clasificación del investigador, lo cual favorece la comparabilidad longitudinal y por estratos analíticos. No obstante, se observaron faltantes en INST_FILIA y nombres de institución con longitudes considerablemente altas, aspecto que justifica las estrategias de tratamiento de nulos y estandarización textual desarrolladas durante la preparación de datos.

En la Tabla 2 se contrastan reglas mínimas de plausibilidad para identificar valores incoherentes de acuerdo con la definición y tipo de dato de las variables críticas.

Regla	Total no nulos	N° Fuera de rango	% Fuera de rango	N° Datos interpretables	% Datos interpretables	Min fecha	Max fecha
EDAD_ANOS_PR en [18,100]	77.236	27,0	0,03	NaN	NaN	NaT	NaT
ANO_CONVO fechas válidas	77.237	NaN	NaN	77.237	100,0	2013-10-31	2021-02-25

Tabla 2. Revisión de coherencia

En primer lugar, se evaluó EDAD_ANOS_PR con un rango de 18 a 100 años. Este umbral responde a dos criterios: el proceso de reconocimiento exige trayectoria verificable y, en la práctica, casi la totalidad de investigadores activos se concentra dentro de ese intervalo. Se identificaron 27 casos fuera de rango, equivalentes a 0,03 % del total, lo que sugiere una ocurrencia marginal que puede corregirse o excluirse sin afectar las conclusiones globales.

En segundo lugar, se verifico que ANIO_CONVO se comporte como fecha valida, con día, mes y año interpretables. El conjunto evaluado abarca las cohortes 2013, 2014, 2015, 2017, 2019 y 2021, y no se encontraron infracciones, lo que confirma la fiabilidad temporal de la serie.

Como parte del perfilamiento inicial de calidad, se revisó la presencia de valores nulos y de valores especiales. El diagnostico mostro que INST_FILIA concentra la mayor proporción de valores nulos, con 12,07 %, seguida por los códigos geográficos COD_DANE_NAC_PR y COD_DANE_RES_PR, con 5,94 % y 3,17 %, respectivamente. Estos resultados indican que la principal atención debe centrarse en la información institucional y geográfica, dado que ambas dimensiones son relevantes para la caracterización de perfiles y para la posterior generación de recomendaciones.

En cuanto a valores especiales, se identificó una alta concentración en ID_VICTIMA_CONFLICTO, con 73,56 %. Esta variable debe tratarse con cautela y no interpretarse como un campo de completitud convencional. Adicionalmente, las variables asociadas al área de conocimiento presentan una proporción baja de valores especiales, cercana al 1,44 %, por lo que su impacto sobre el análisis general es acotado. El detalle completo debe permanecer en el Anexo 2.

3.3. Exploración de los datos

La exploración de los datos caracteriza el comportamiento de las variables clave, permite descubrir patrones iniciales y detecta indicios de problemas de calidad que pueden afectar los análisis posteriores. Se priorizan preguntas orientadas al negocio: perfil de los investigadores, evolución por cohortes, composición por áreas del conocimiento y distribución territorial.

3.3.1. Distribución de edad y valores atípicos

Se describió la variable EDAD_ANOS_PR para identificar su rango efectivo y la presencia de valores

extremos. Esto ayuda a entender el sesgo etario del dataset analizado y a estimar el posible impacto de registros erróneos

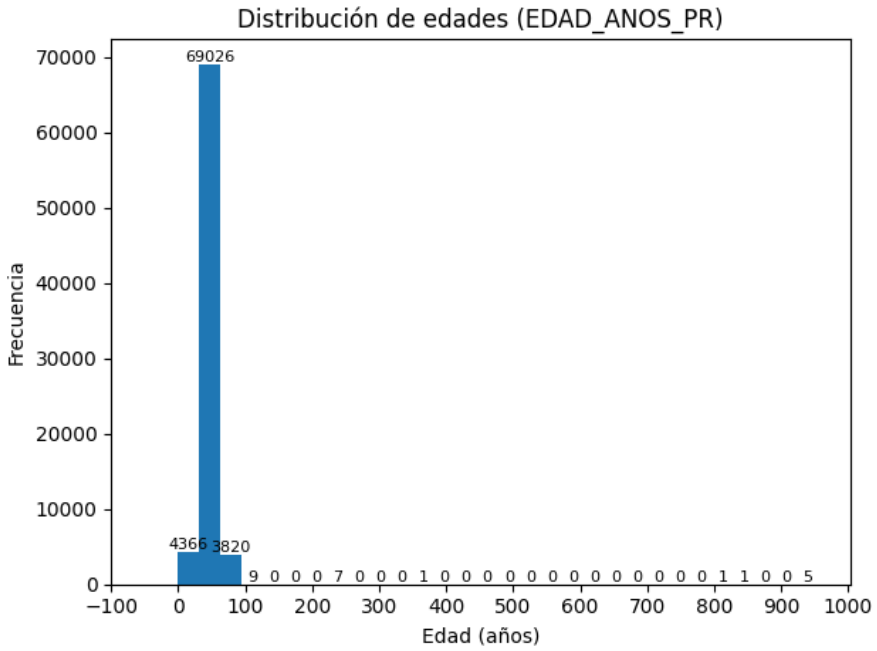

Figura 3. Distribución edades

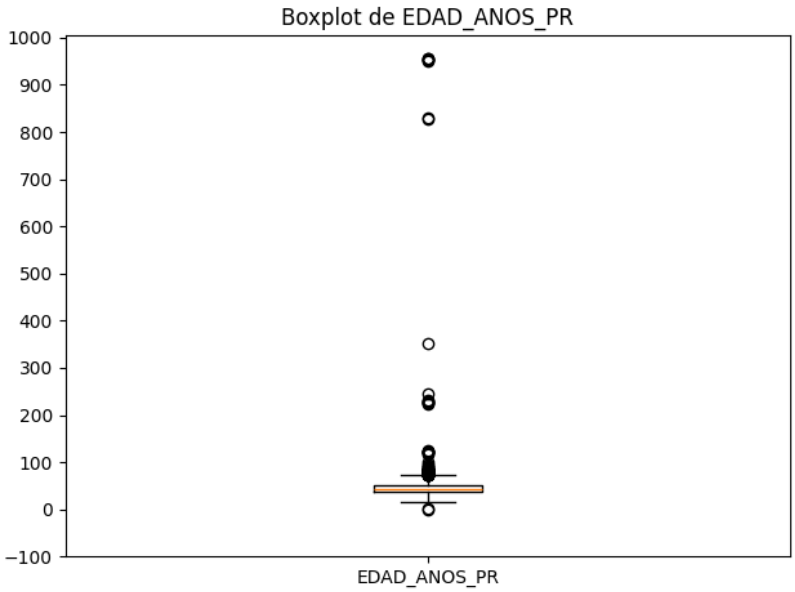

Figura 4. Boxplot edades

De acuerdo con las Figuras 3 y 4, la edad media se sitúa cerca de 44-45 años con mediana en 43 años, lo que sugiere un perfil de investigadores de media carrera. Se observan casos aislados fuera de un rango razonable de edad para los investigadores que corresponden a ≈0,33 % del total. Por lo tanto, al tener una frecuencia

baja no hay una afectación en las medidas de centralidad, pero conviene marcarlos para corrección o exclusión en próximos análisis.

3.3.2. Tamaño y edad por cohorte

En la Tabla 3 se cuantificó el volumen de registros por año de convocatoria y se identificó la edad media y mediana por cohorte. El propósito es observar crecimiento del sistema y cambios demográficos a través del tiempo.

Año convocatoria	N° registros	Media (edad)	Mediana (edad)
2013	8016	46,92	46,05
2014	8280	44,50	43,05
2015	10050	44,56	43,04
2017	13001	44,63	42,99
2019	16796	44,60	42,91
2021	21093	44,63	43,08

Tabla 3. Características por cohorte

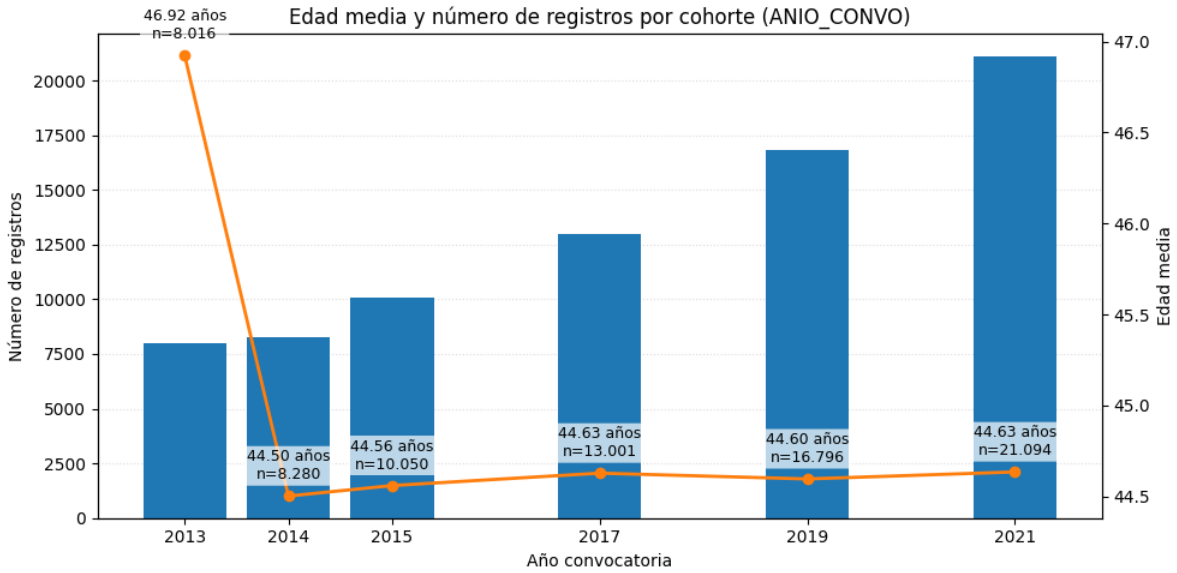

Figura 5. Edad media por cohorte

De acuerdo con la Tabla 3 y la Figura 5, se evidencia que el número de investigadores presenta un crecimiento sostenido entre 2013 y 2021, con incrementos destacados en 2019 y 2021. En términos agregados, este comportamiento se resume en una tasa de crecimiento anual compuesta (CAGR) [64] de 12,86% para el periodo 2013–2021, lo que confirma una expansión consistente de los registros a lo largo de las convocatorias analizadas.

$$CAGR = \left(\frac{N_{final}}{N_{inicial}}\right)^{\frac{1}{Años}} - 1 \qquad (1)$$

Por su parte, la edad media por cohorte, representada por la línea naranja, es estable (44-45 años) con variaciones marginales, lo que indica renovación moderada y ausencia de cambios bruscos en la composición etaria entre convocatorias.

3.3.3. Estructura de categorización de investigadores

En la Figura 6 se presenta la distribución de investigadores por clasificación (Junior, Asociado, Sénior y Emérito), reportando tanto el número absoluto como la proporción de cada categoría, con el fin de caracterizar el punto de partida del universo analizado y detectar posibles desbalances estructurales.

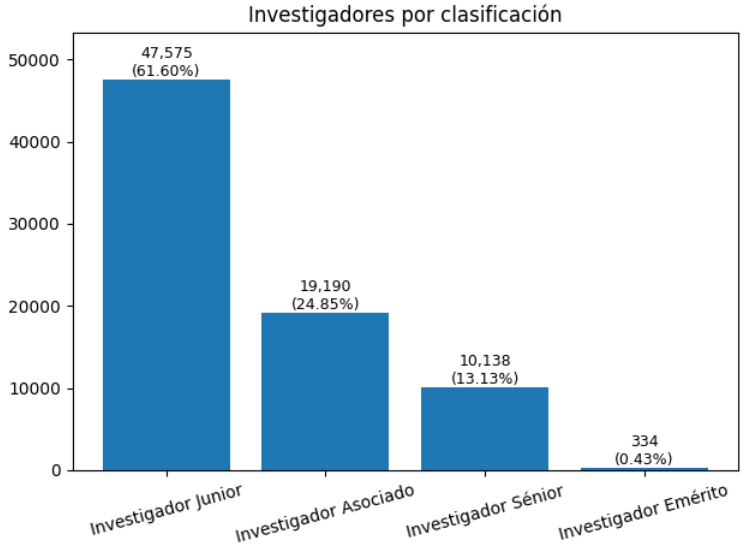

Figura 6. Distribución categorías

A partir de los resultados, se puede decir que la categoría Investigador Junior (61%) es predominante, seguido de Asociado (24%), Sénior (13%) y, debido a los criterios y características de esta categoría, Emérito representa una fracción muy reducida (<1 %). Esta composición es consistente con sistemas de investigación donde la base está conformada por investigadores en etapas iniciales y medias, y respalda que los análisis subsecuentes focalicen las transiciones.

También se revisó la evolución de la mezcla por clasificación dentro de cada cohorte.

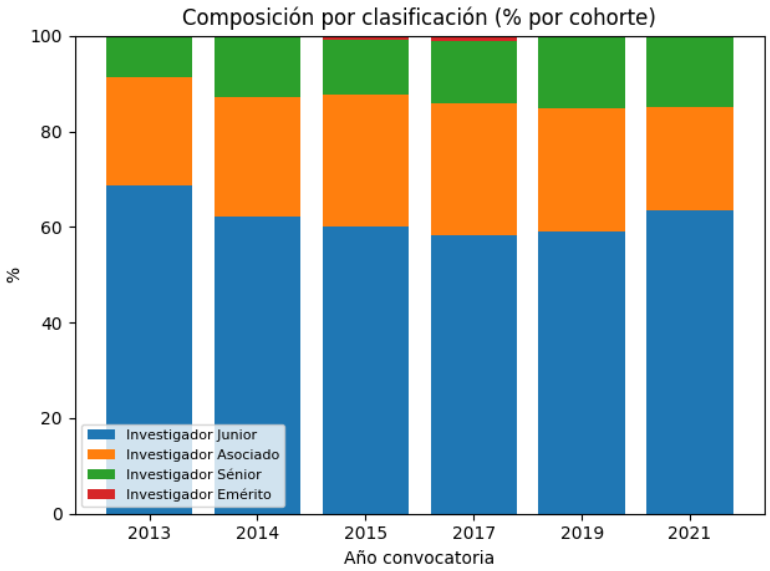

Figura 7. Composición categorías por cohorte

De acuerdo con la Figura 7, se puede resaltar que la participación de Junior se mantiene alta pero ligeramente decreciente en algunos cortes, mientras que Sénior muestra una tendencia leve al alza (8,7 % en 2013 a >14 % en 2021). Este patrón sugiere acumulación de trayectoria y refuerzo de capas superiores, congruente con un sistema que madura entre cohortes.

3.3.4. Áreas de conocimiento

Se exploró la composición por gran área (`NME_GRAN_AREA_PR`) para ubicar los principales focos de actividad, y se profundizó en los Top-10 de áreas específicas por cada gran área de mayor concentración de registros.

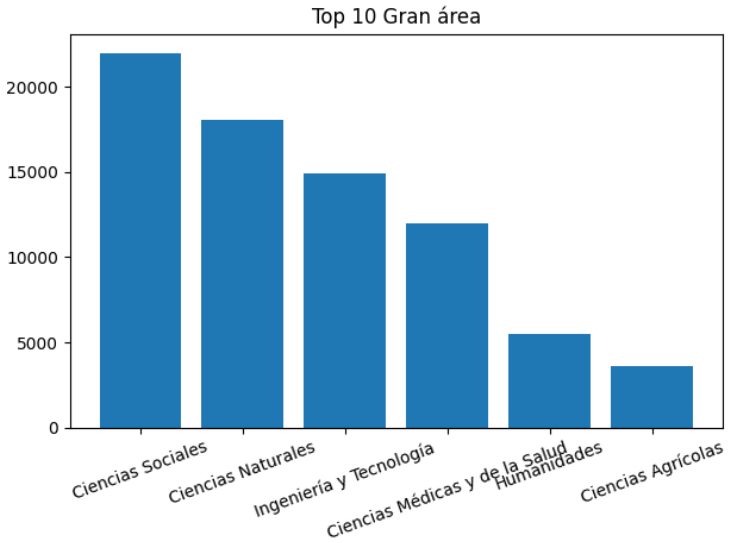

Figura 8. Top 10 grandes áreas de conocimiento

De acuerdo con la Figura 8, el volumen de investigadores se concentra principalmente en Ciencias Sociales, Ciencias Naturales e Ingeniería y Tecnología, que agrupan la mayor proporción de registros del conjunto analizado. Esta distribución evidencia una estructura disciplinar con polos claramente definidos, lo cual resulta relevante para el proyecto porque permite anticipar diferencias en la densidad de perfiles, oportunidades de colaboración y posibles sesgos de representación entre áreas.

La revisión detallada de áreas específicas, presentada en el Anexo 3, muestra que los mayores volúmenes se ubican en campos como Educación General, Negocios y Management, Ingeniería Eléctrica y Electrónica, Derecho, Economía y Psicología. Estos núcleos son coherentes con la composición agregada observada por gran área y sugieren espacios estratégicos para el análisis comparativo de perfiles y para la identificación de posibles comunidades académicas.

Por otro lado, se estimó la participación porcentual de cada clasificación dentro de cada gran área con el fin de detectar diferencias estructurales disciplinares.

Categoría / Gran área	Investigador Junior (%)	Investigador Asociado (%)	Investigador Sénior (%)	Investigador Emérito (%)
Ciencias Agrícolas	63,85	23,02	12,52	0,61
Ciencias Médicas y de la Salud	59,91	23,46	16,28	0,35
Ciencias Naturales	59,37	21,41	18,65	0,57
Ciencias Sociales	64,39	27,52	7,63	0,46
Humanidades	71,86	23,17	4,35	0,62
Ingeniería y Tecnología	55,56	28,33	15,90	0,21
Consolidado de toda las grandes áreas	61,60	28,85	13,13	0,43

Tabla 4. Categorías por área de conocimiento

De acuerdo con la Tabla 4, las Humanidades exhiben la mayor proporción de Junior (>70 %), mientras que Ciencias Naturales y Ciencias Médicas y de la Salud muestran cuotas comparativamente más altas de Sénior con 18,65% y 16,28% respectivamente. Estas asimetrías son esperables por trayectorias formativas y dinámicas de financiación específicas, y conviene considerarlas cuando se comparen desempeños entre campos.

3.3.5. Correlación entre variables numéricas

Con el propósito de explorar relaciones bivariadas entre las variables disponibles, se construyó un subconjunto conformado únicamente por variables numéricas, descartando los campos de identificación y códigos administrativos como los IDs y códigos geográficos, dado que su carácter referencial no aporta información sustantiva para el análisis y podría sesgar la interpretación de asociaciones. Sobre este conjunto depurado se estimó la matriz de correlación utilizando el coeficiente ρ de *Spearman*, apropiado para evaluar asociaciones monotónicas sin requerir supuestos estrictos de normalidad y permitiendo la inclusión de variables de tipo ordinal, obteniendo como resultado el *heatmap* presentado en la Figura 9, donde se aprecia que las correlaciones entre los pares analizados presentan magnitudes bajas, sin evidencia de relaciones fuertes.

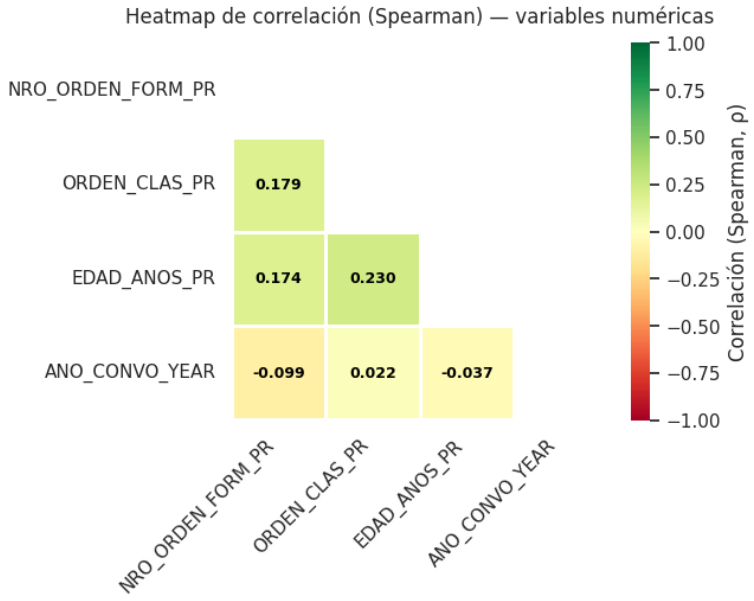

Figura 9. Heatmap de correlación

3.3.6. Geografía: residencia y diferenciales con lugar de nacimiento

Se analizó la distribución territorial de los investigadores a partir del departamento de residencia, con el propósito de identificar las principales concentraciones geográficas del ecosistema científico nacional. Adicionalmente, se comparó la residencia con el lugar de nacimiento para reconocer diferencias netas que sugieren dinámicas de atracción, permanencia o movilidad territorial del talento científico.

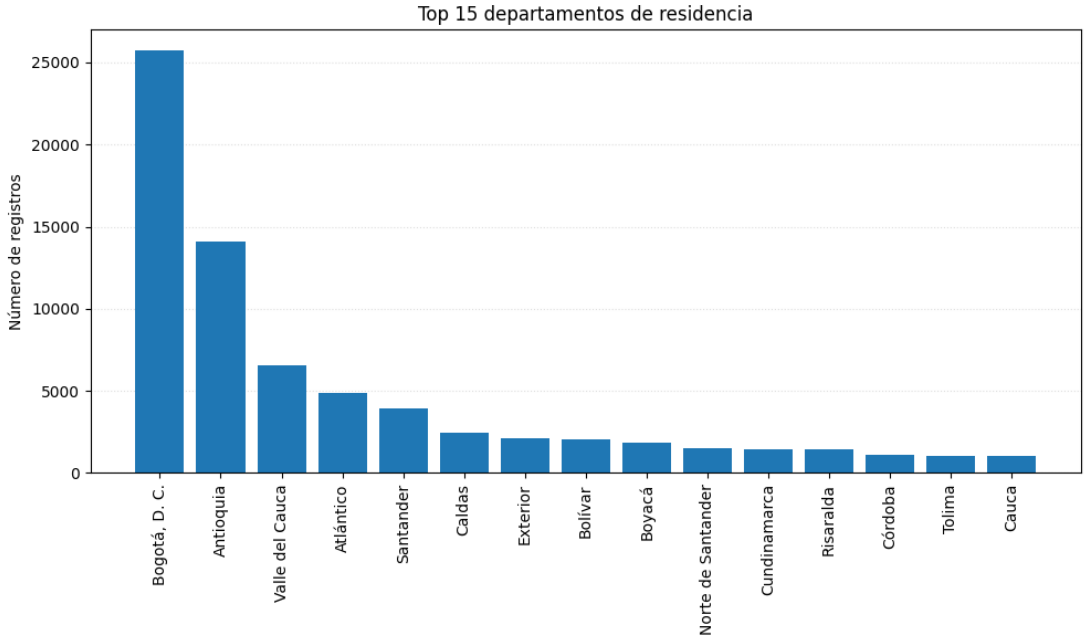

Figura 10. Top 15 departamentos de residencia

De acuerdo con la Figura 10, Bogotá, D. C. concentra el mayor número de investigadores residentes, con 33,28 % del total de registros, seguida de Antioquia con 18,21 % y Valle del Cauca con 8,49 %. Esta distribución evidencia una alta concentración territorial de capacidades científicas en los principales centros académicos e institucionales del país. El detalle de la comparación entre departamento de residencia, departamento de nacimiento y diferencia neta de registros se conserva en el Anexo 4.

En conjunto, la exploración confirma que el sistema evaluado crece en tamaño, mantiene estabilidad etaria, presenta fuerte presencia de perfiles Junior y exhibe un progreso moderado hacia categorías Senior. Las áreas con mayor masa crítica se alinean con las principales disciplinas del país y la geografía muestra polos urbanos de atracción de talento. Estas evidencias respaldan los objetivos del proyecto al proporcionar segmentaciones operativas sobre las que se pueden construir indicadores, modelar transiciones y priorizar recomendaciones.

3.4. Verificación de calidad

La verificación de calidad documenta la evaluación sistemática del conjunto de datos y las decisiones adoptadas para asegurar su uso confiable en análisis posteriores. Esta sección debe conservarse porque demuestra que el paso a preparación y modelado no se hizo sobre una base asumida como valida, sino sobre un diagnóstico explicito.

Se fijaron variables críticas y reglas de evaluación para edad, cohorte temporal, clasificación, filiación institucional y columnas geográficas. Sobre estos campos se definieron controles de completitud, coherencia básica y cobertura. Para edad se estableció el intervalo de plausibilidad [18, 100] años; para completitud se definió un umbral operativo de 10 % de ausencias por variable; y para cobertura geográfica se optó por un enfoque de reporte, no de sanción, para no descartar información útil sin análisis posterior.

Columna	No nulos	% no nulos	N° nulos	% nulos
EDAD_ANOS_PR	77236	99,99	1	0,00001
ANIO_CONVO	77237	100,00	0	0,00
NME_CLASIFICACION_PR	77237	100,00	0	0,00
COD_DANE_RES_PR	74789	96,83	2448	3,17
COD_DANE_NAC_PR	72646	94,06	4591	5,94
INST_FILIA	67911	87,93	9326	12,07

Tabla 5. Completitud

De acuerdo con la Tabla 5, la cobertura es aproximadamente del 100 % en edad, año y clasificación, lo que garantiza estabilidad para análisis comparativos por cohorte y estrato. INST_FILIA presenta alrededor de 12 % de nulos, por lo que los análisis por institución requieren políticas explicitas de tratamiento. Los códigos DANE alcanzan coberturas altas, cercanas a 94-97 %, suficientes para análisis agregados.

También se evaluó la combinación (ID_CONVOCATORIA, ID_PERSONA_PR) como llave de fila, verificando ausencia de nulos y duplicados. No se encontró ningún registro repetido bajo esta llave, lo que confirma la unicidad de la participación por persona y convocatoria y habilita la construcción de indicadores longitudinales sin pérdida o duplicación de información.

Para sintetizar la aptitud del conjunto de datos, se construyó un índice global de calidad. Este índice combina cuatro componentes: completitud del núcleo, plausibilidad de reglas, solidez de la llave compuesta y sesgos de cobertura por grupo. A continuación, se relacionan los criterios y forma de cálculo del índice:

Para la definición de los ponderados del índice de calidad (0,40 para completitud, 0,25 para plausibilidad de reglas, 0,20 para llave compuesta y 0,15 para sesgos de cobertura) se partió de la guía conceptual que ofrece Wand y Wang sobre las dimensiones intrínsecas de calidad de datos. En su trabajo, los autores muestran que la literatura y su modelo coinciden en destacar la completitud y la corrección/fiabilidad como dimensiones centrales, a partir de las cuales se derivan muchas de las demás (consistencia, precisión, relevancia), y evidencian además, que ciertas dimensiones aparecen con mayor frecuencia y peso en los estudios empíricos [65]. Aunque el artículo no propone pesos numéricos explícitos, este énfasis relativo sirvió como criterio para asignar la mayor ponderación a la completitud y a la plausibilidad de reglas, dejando ponderaciones menores, a la integridad de la llave y a la estabilidad de la cobertura por segmentos, de forma coherente con el marco teórico y con las necesidades específicas del proyecto.

A continuación, se relacionan las dimensiones evaluadas y pesos. Se consideraron cuatro componentes con los siguientes pesos:

1. Completitud (0,40): mide el porcentaje promedio de no nulos en las columnas críticas para el análisis (C: `EDAD_ANOS_PR`, `ANIO_CONVO`, `NME_CLASIFICACION_PR`, `INST_FILIA`, `COD_DANE_RES_PR`, `COD_DANE_NAC_PR`).

$$\sum_{c \in C} \frac{N° \, registros(c) - Nulos(c)}{N° \, registros(c)} \quad (2)$$

2. Plausibilidad de reglas (0,25): refleja el cumplimiento de reglas mínimas de sentido común acordadas con el proyecto (p. ej., edad en [18, 100] años y que `ANIO_CONVO` sea interpretable como fecha válida). Se calcula como:

$$(1 - \% \, Violación) * 100 \quad (3)$$

3. Llave compuesta (0,20): evalúa la integridad y unicidad de la clave (`ID_CONVOCATORIA`, `ID_PERSONA_PR`). El sub-índice se define como:

$$1 - (\% \text{ de nulos en la clave} + \% \text{ de duplicados por clave}). \quad (4)$$

4. Sesgos de cobertura (0,15): estima la estabilidad de la completitud por segmentos relevantes; en este caso, por año de convocatoria y por clasificación de investigador. Se define como:

$$(1 - \% \, faltantes \, por \, grupo) * 100 \quad (5)$$

De esta manera se obtiene la siguiente fórmula general:

$$Calidad = 0{,}40 \cdot Completitud + 0{,}25 \cdot Plausibilidad + 0{,}20 \cdot Llave + 0{,}15 \cdot Sesgos \quad (6)$$

Como resultado de los subíndices, en la Tabla 6, se obtuvieron los siguientes resultados:

Índice	**Resultado**
Completitud	0,965
Plausibilidad	1,000
Llave	1,000
Sesgos	0,918

Tabla 6. Resultados índice de calidad

El valor de completitud indica que 96,5 % de los campos principales no presenta vacíos; las excepciones se concentran en INST_FILIA y, en menor medida, en códigos DANE. La plausibilidad igual a 1,000 sugiere que las reglas mínimas se cumplen casi por completo, dado que las 24 observaciones fuera del rango de edad no afectan el redondeo del subíndice. La llave compuesta alcanza 1,000 porque no se hallaron nulos ni duplicados, y el subíndice de sesgos, 0,918, evidencia asimetrías de completitud por grupos. Como puntaje global se obtuvo 0,973, lo que ubica al dataset en un nivel alto de aptitud para análisis descriptivos, comparaciones entre cohortes y cruces por clasificación y grandes áreas.

3.5. Selección de los datos

La selección de los datos constituye un paso critico de preparación, porque de ella depende la relevancia y estabilidad del modelo posterior. El proceso buscó garantizar que las variables incorporadas aportaran valor explicativo al análisis, evitar redundancias y minimizar sesgos en las fases posteriores de limpieza, construcción y modelado.

De acuerdo con el Anexo 5, donde se detalla el análisis realizado por cada una de las variables, se incluyeron variables que permiten describir dimensiones sociodemográficas, formativas, institucionales y geográficas de los investigadores, así como aquellas necesarias para construir indicadores derivados: madurez investigativa, movilidad, diversidad, trayectoria y recencia. En contraste, se excluyeron identificadores internos, códigos administrativos redundantes, variables jerárquicas demasiado agregadas y campos textuales o descriptivos que no aportaban información adicional frente a variables interpretables.

La selección privilegio variables medibles, estandarizadas y disponibles de forma consistente en el tiempo. También incluyo variables geográficas y sociales con finalidad analítica y ética: no para inducir segmentación discriminatoria, sino para evaluar sesgos potenciales y monitorear equidad durante la evaluación.

Bloque temático	Variables	Uso previsto
1. Identificación y trazabilidad	ID_PERSONA_PR, ID_CONVOCATORIA, ANO_CONVO	Identificadores y referencia temporal para consolidación y trazabilidad.
2. Formación y clasificación	NME_NIV_FORMACION_PR, NRO_ORDEN_FORM_PR, NME_CLASIFICACION_PR, ORDEN_CLAS_PR	Nivel educativo y categoría Minciencias; base para el índice de madurez investigativa.
3. Áreas de conocimiento	NME_GRAN_AREA_PR, NME_AREA_PR, NME_ESP_AREA_PR	Clasificación temática y especialidad disciplinar; insumo para medir diversidad y segmentar perfiles.
4. Información geográfica	NME_DEPARTAMENTO_NAC_PR, NME_DEPARTAMENTO_RES_PR, NME_MUNICIPIO_RES_PR, NME_MUNICIPIO_NAC_PR	Análisis de movilidad territorial (nacimiento-residencia) y caracterización regional.
5. Datos demográficos y sociales	EDAD_ANOS_PR, NME_GENERO_PR, TXT_GRUPO_ETNICO, TXT_POBLACION_DISCA, ID_VICTIMA_CONFLICTO	Variables de equidad y diversidad; soporte para auditorías y análisis descriptivo.
6. Información institucional	INST_FILIA	Institución de filiación; base para movilidad institucional y redes de colaboración.

Tabla 7. Listado de variables seleccionadas

De acuerdo con la Tabla 7, el conjunto final quedo conformado por 20 variables agrupadas en seis bloques temáticos: identificación y trazabilidad, formación y clasificación, áreas de conocimiento, información geográfica, datos demográficos y sociales, e información institucional. Este esquema garantiza una visión integral del perfil de los investigadores y mantiene trazabilidad para limpieza, integración y construcción de indicadores derivados.

La estructura final se define como una vista consolidada por persona, con una única fila por `ID_PERSONA_PR`. Esta vista integra la información más reciente y valida de cada investigador y se complementa con variables derivadas como madurez investigativa, antigüedad, diversidad temática, movilidad geográfica e institucional y trayectoria.

3.6. Limpieza de los datos

En esta sección se describen las acciones orientadas a garantizar la calidad, coherencia y confiabilidad del conjunto de datos seleccionado, mediante la identificación y corrección de errores, inconsistencias y valores faltantes. El propósito de esta fase es asegurar que la información empleada en las etapas posteriores de construcción y modelado sea completa, consistente y libre de distorsiones que puedan afectar la validez de los resultados analíticos [60], [61].

La limpieza de los datos abarca desde el tratamiento de valores atípicos y la imputación de ausencias, hasta la estandarización de textos y la verificación de duplicados, contemplando además el tratamiento especial de la variable institución debido a que en su estructura se puede presentar más de una institución separada por "|". De esta manera, esta etapa sienta las bases técnicas para disponer de un conjunto de datos depurado y estructurado que permita avanzar hacia la construcción de indicadores derivados y la aplicación del modelo de segmentación con un alto nivel de integridad y trazabilidad.

Como resultado de la etapa anterior, el dataset pasó de 30 variables y 77.237 registros a 20 variables y 77.230 registros. Las 20 variables resultantes fueron definidas para la preparación y construcción de datos, teniendo en cuenta que algunas de ellas se usarán como insumos para derivar indicadores y para análisis de equidad. En la Figura 11, se puede observar de manera general cada una de las fases del proceso de limpieza de los datos, que se desarrollarán a continuación junto con su análisis y resultados. Esta depuración deja un maestro limpio, consistente y listo para la integración.

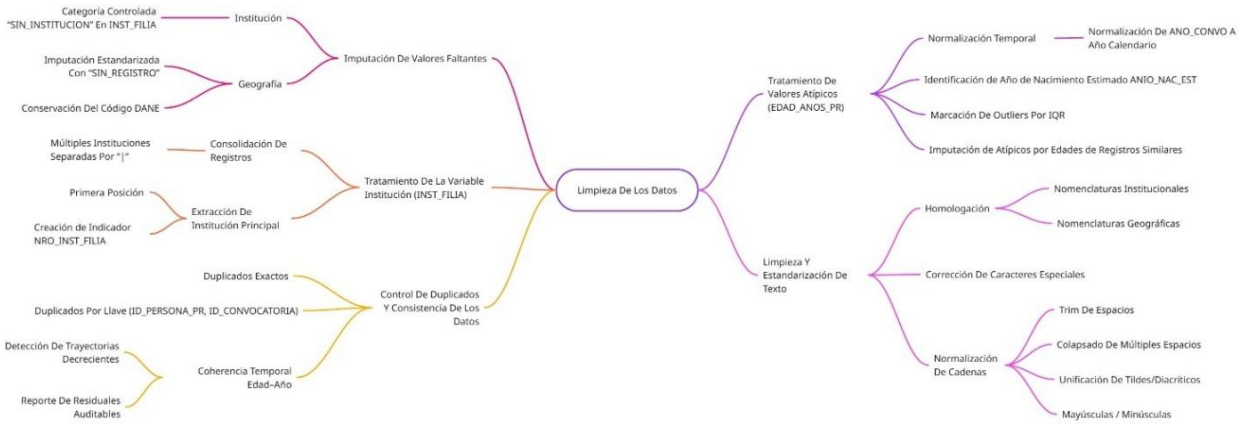

Figura 11. Limpieza de los datos

3.6.1. Tratamiento de valores atípicos

Con el fin de evitar correcciones genéricas y asegurar consistencia longitudinal, se depuró la variable EDAD_ANOS_PR en dos fases con trazabilidad. Primero, se normalizó ANO_CONVO a año calendario y, para cada ID_PERSONA_PR con historial en varias convocatorias, se calculó su año de nacimiento estimado como $\widehat{ANIO\_NAC\_EST} = mediana(ANO_{CONVO} - EDAD_{ANOS_{PR}})$. Con este año de nacimiento aproximado se recalcularon las edades únicamente en los registros atípicos o incoherentes, sustituyéndolas por el año de convocatoria: $ANO\_CONVO - \widehat{ANIO\_NAC\_EST}$. Segundo, sobre el conjunto resultante se identificaron atípicos estadísticos por la regla IQR (Q1 ± 1,5·IQR), y los casos resultantes se imputaron con la mediana de su cohorte definida por nivel de formación × clasificación.

Este procedimiento no modifica edades ya válidas y deja trazabilidad mediante dos campos: EDAD_FUENTE_CORR (UNCHANGED/ANCLA/CLIPPING/NA) y ANIO_NAC_EST (año de nacimiento estimado). Este enfoque corrige errores evidentes, armoniza la trayectoria temporal de cada investigador y minimiza sesgos, preservando la información sustantiva para el modelado.

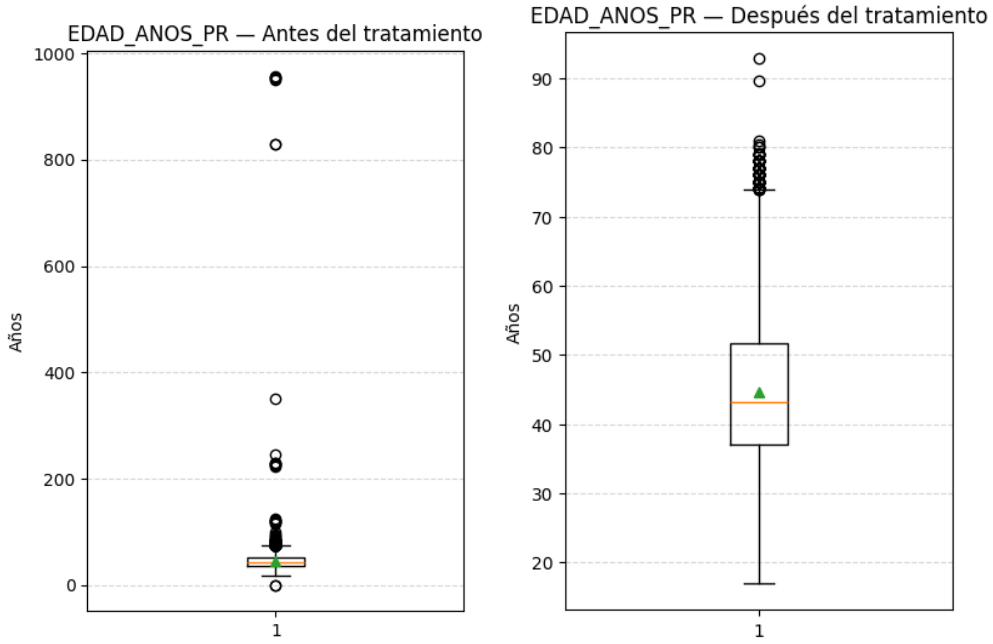

Figura 12. Variable edad antes y después de tratamiento

En la Figura 12 se observa que, previo al tratamiento, la variable EDAD_ANOS_PR presentaba valores incoherentes (negativos y muy altos), con una dispersión que distorsionaba la escala del gráfico. Tras el tratamiento aplicado, la distribución queda centrada en una mediana ≈ 43 años, con Q1 ≈ 36,98 y Q3 ≈ 51,76 (IQR ≈ 14,78). Los bigotes del boxplot se sitúan aproximadamente en LOW ≈ 14,81 y HIGH ≈ 73,93, reflejando una variabilidad coherente con trayectorias académicas reales. Sin embargo, persisten observaciones atípicas en el gráfico que se sitúan principalmente entre 74 y 93 años, las cuales representan 110 registros sobre 77.237 (≈ 0,14 %), los cuales corresponden al comportamiento real de los datos. Por lo tanto, se realizó una validación adicional donde se evidenció que estos datos atípicos remanentes son mayores en las clasificaciones Investigador Sénior/Emérito y en el nivel de formación de Doctorado/Postdoctorado, lo que

sustenta que corresponden al comportamiento normal de la población y no a errores, motivo por el cual no se aplicaron correcciones adicionales.

3.6.2. Imputación de valores faltantes

Se ejecutó la imputación sin inferir valores para evitar sesgos que no representen la realidad de los datos, reemplazando vacíos y registros equivalentes, como null o sin dato, por categorías unificadas:

- `INST_FILIA`: "`SIN_INSTITUCION`".
- Variables geográficas (nacimiento y residencia en país/departamento/municipio): "`SIN_REGISTRO`".

Este proceso se aplicó sobre el maestro limpio obteniendo los siguientes resultados sobre los 77.237 registros:

- `INST_FILIA` imputados: 9.326 (12,075%).
- `NME_DEPARTAMENTO_NAC_PR`: 222 (0,287%); `NME_MUNICIPIO_NAC_PR`: 494 (0,640%).
- `NME_DEPARTAMENTO_RES_PR`: 316 (0,409%); `NME_MUNICIPIO_RES_PR`: 388 (0,502%).

Con esta decisión se estandariza la ausencia de información como una categoría explícita, se mejora la consistencia del dataset para codificación y agregaciones posteriores, y se preserva la neutralidad analítica.

3.6.3. Limpieza y estandarización de texto

Se aplicó una normalización sistemática de cadenas en las variables textuales clave (institución, áreas de conocimiento, geografía y descriptores de formación/clasificación) para eliminar ruido tipográfico y garantizar categorías consistentes. Las acciones incluyeron: (i) pasar a mayúsculas, (ii) eliminar tildes/diacríticos (normalización Unicode) y (iii) trim de espacios y colapso de espacios múltiples.

Como resultado, se resalta que no se alteró el número de registros, debido a que las transformaciones fueron puramente de formato y no semánticas. Y, de esta manera se reducen duplicidades, se mejora la calidad de las llaves para uniones y consolidaciones, y se garantiza una codificación estable para etapas posteriores.

3.6.4. Tratamiento de la variable institución

Debido a que la variable institución presenta 569 registros (≈ 0,74% del total) en los que el investigador presenta más de una institución separada por "|", se realizó un proceso de consolidación sin expansión de filas, con dos acciones principales:

1. Creación del indicador `NRO_INST_FILIA`, que contabiliza el número de instituciones informadas por registro (separador |).
2. Definición de una institución principal tomando la primera del listado normalizado.

Esta estrategia preserva la información sobre vínculos múltiples sin duplicar observaciones, evitando sobreponderar investigadores con varias afiliaciones, y deja un insumo directo para los indicadores de movilidad institucional y para el índice de madurez investigativa.

3.6.5. Control de duplicados y consistencia de los datos

Se ejecutó la verificación de duplicados y la coherencia edad-año por persona. Primero, se evaluaron duplicados exactos (fila completa) y por llave lógica (`ID_PERSONA_PR`, `ID_CONVOCATORIA`), obteniendo como resultado 0 eliminaciones, manteniéndose 77.237 registros. Segundo, tras la

reconstrucción de edad, se ordenaron las convocatorias por `ANIO_CONVO` para cada investigador y se calculó la variación de edad entre observaciones consecutivas. De esta manera, se detectaron 11 personas con edad decreciente, en los cuales, se aplicó un ajuste dirigido solo a los casos en los que se evidenció una reducción mayor a un año, debido a que en los casos en los que la diferencia corresponde a un año podría ser causado por la fecha en la que se tomaron los datos de la convocatoria sin incurrir en un error real. Este ajuste consistió en calcular la edad aproximada para esa convocatoria con base en el año de nacimiento calculado anteriormente. Con este procedimiento se corrigieron 5 eventos de los 11 registros detectados inicialmente y se eliminaron los 6 restantes quedando con 77.231 observaciones.

Finalmente, se identificó y excluyó un registro que carecía simultáneamente de los campos críticos requeridos para el análisis (edad, formación, clasificación, área e institución), debido a que el registro no aporta información verificable ni es imputable sin introducir sesgo, su exclusión no afecta la representatividad del conjunto y mejora la calidad y trazabilidad del dataset maestro quedando con **77.230** registros válidos.

En la Figura 13, se presenta de manera resumida el proceso llevado a cabo para contar con la versión limpia y consistente del dataset:

Figura 13. Control de duplicados y consistencia de los datos

3.7. Integración de los datos

La integración consolida la versión final de limpieza en una estructura orientada al análisis y al despliegue del modelo. El trabajo comprende la conformación de una tabla persona-convocatoria con orden temporal y llaves primarias/foráneas, la generación de dimensiones mínimas y la construcción de una vista analítica consolidada para modelado y recomendación.

En la Figura 14 se definió un esquema estrella para parametrizar los datos limpios en consultas y alimentar el pipeline de modelado. El hecho central es fact_participacion, cuya llave primaria queda establecida como

(ID_PERSONA_PR, ANIO_CONVO). Sobre este hecho se declararon claves foráneas hacia las dimensiones de persona, institución, área de conocimiento y tiempo.

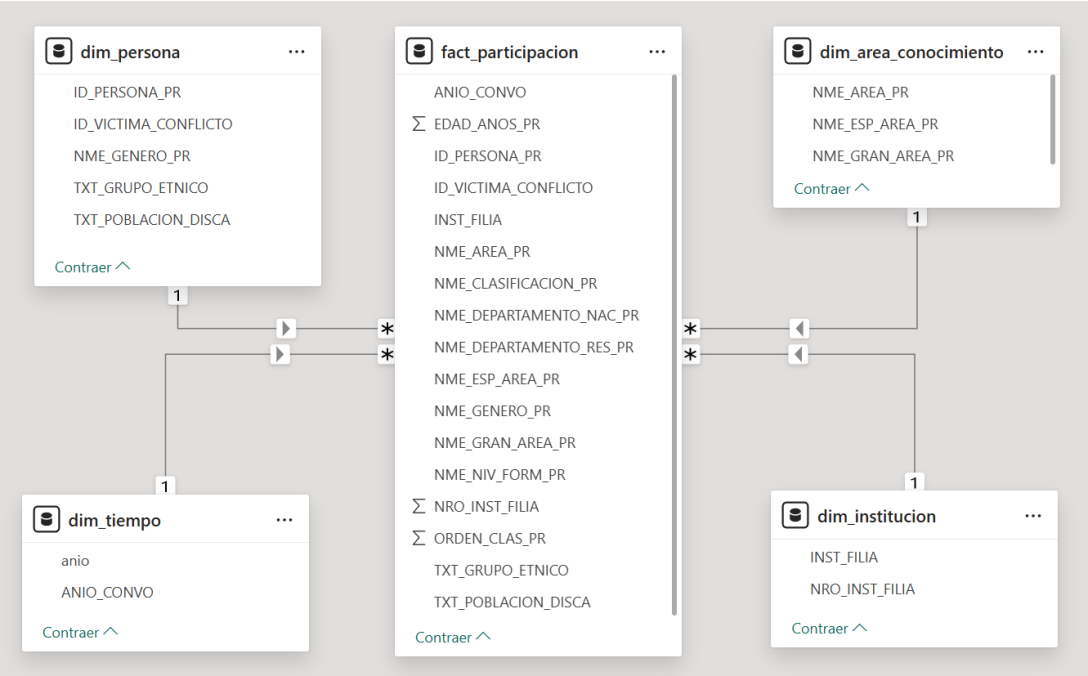

Figura 14. Esquema de estrella

El resultado del armado y la validación fue el siguiente: el hecho contiene 77.230 filas, que representan 30.085 personas a lo largo de seis años de convocatoria sin registros duplicados. Las claves foráneas verificaron 100 % de cobertura contra sus dimensiones, y la dimensión tiempo quedo con seis años válidos. La vista consolidada aporta 30.085 filas, exactamente una por persona, y la vista detallada replica el histórico completo. Con estas piezas se armó view_consolidado_entrenamiento, compuesto por una fila por persona y anclado en su última convocatoria, con agregados necesarios para indicadores, entrenamiento y recomendación.

3.8. Construcción de datos

Esta sección documenta la construcción de datos necesaria para calcular el Índice de Madurez Investigativa (IMI) y el índice de movilidad geográfica, a partir del clean master de investigadores reconocidos. El alcance contempla: (i) una fase preliminar de construcción y normalización de métricas base, que incluye trayectoria, clasificación vigente con ponderación por recencia, diversidad temática, movilidad institucional, nivel de formación y movilidad geográfica (nacimiento vs. residencia); (ii) la síntesis del IMI mediante transformaciones y agregación ponderada de los componentes principales; y (iii) una revisión final de calidad y equidad, que verifica cobertura, rangos [0,1], estabilidad de la distribución por año ancla, correlaciones esperadas con trayectoria, cortes por variables sociales y resultados del índice de movilidad geográfica, antes de publicar las salidas finales (IMI, subíndices, índice de movilidad y diccionario de variables) en las vistas consolidadas para entrenamiento.

3.8.1. Fase preliminar y construcción de métricas base

A partir del data mart integrado, se construyeron las métricas base necesarias para alimentar el Índice de Madurez Investigativa (IMI) y el índice de movilidad geográfica. Todas las métricas se calculan a nivel de persona, utilizando el historial de participación por convocatoria y el perfil más reciente disponible en el período de análisis.

La fase se organiza en cuatro grupos de indicadores: (i) trayectoria y antigüedad, (ii) movilidad institucional, (iii) diversidad temática, (iv) formación y clasificación vigente, complementados con métricas de recencia y movilidad geográfica. A continuación se describen cada una de ellas y su formulación.

3.8.1.1. Trayectoria y antigüedad

Número de convocatorias (`N_CONVOCATORIAS`)

Conteo de convocatorias en las que la persona ha sido reconocida, a partir de la tabla de hechos fact_participacion:

$$N_{\text{Convocatorias}(i)} = |\ \{\text{ANIO\_CONVO}: (\text{ID\_PERSONA\_PR} = i)\}\ | \quad (7)$$

Años mínimo y máximo de participación (`ANIO_MIN, ANIO_MAX`)

Se calcula, para cada persona, el primer y último año en el que aparece en el histórico:

$$ANIO_{\text{MIN}(i)} = \min\{\text{ANIO\_CONVO}_i\}, ANIO_{\text{MAX}(i)} = \max\{\text{ANIO\_CONVO}_i\} \quad (8)$$

Antigüedad en años (`ANTIGUEDAD_ANIOS`)

Longitud de la trayectoria observada en el período:

$$\text{ANTIGUEDAD\_ANIOS}(i) = ANIO_{\text{MAX}(i)} - ANIO_{\text{MIN}(i)} \quad (9)$$

3.8.1.2. Movilidad institucional

Número de instituciones distintas (`N_INST`)

Número de instituciones de filiación distintas en las que la persona ha participado:

$$N_{\text{INST}(i)} = |\ \{\text{INST\_FILIA}: (\text{ID\_PERSONA\_PR} = i)\}\ | \quad (10)$$

Tasa de movilidad institucional (`TASA_MOV_INST`)

Mide el promedio de "cambios de institución" por año de trayectoria. Se considera que cada vez que una persona ha estado en más de una institución, ha realizado $N_\text{INST} - 1$cambios. Para evitar división por cero se introduce un término pequeño ε:

$$\text{TASA\_MOV\_INST}(i) = \frac{\max\{N\_\text{INST}(i) - 1, 0\}}{\text{ANTIGUEDAD\_ANIOS}(i) + \varepsilon} \quad (11)$$

Componente de movilidad institucional (MOB)

A partir de la tasa anterior se construye un subíndice acotado en $[0,1]$ mediante una función sigmoide centrada en un parámetro b(tasa "típica") y con pendiente controlada por a. Antes se aplica un *cap* en el percentil 95 para reducir el efecto de atípicos:

$$\text{MOB}(i) = \frac{1}{1 + \exp\left(-a(\text{TASA\_MOV\_INST\_c}(i) - b)\right)} \quad (12)$$

Donde `TASA_MOV_INST_c` es la tasa recortada en $p95$.

3.8.1.3. Diversidad temática

Entropía normalizada por área de conocimiento específica (DIV_ENTROPY_N)

A partir de la distribución de la persona en las diferentes áreas (`NME_ESP_AREA_PR`), se calcula una entropía de Shannon normalizada por el número de categorías distintas $K(i)$, de forma que 0 indica ausencia de diversidad y valores cercanos a 1 reflejan una distribución uniforme:

$$H(i) = -\sum_k p_{ik} \log(p_{ik} + \varepsilon), \text{DIV\_ENTROPY\_N}(i) = \begin{cases} \frac{H(i)}{\log K(i)} & \text{si } K(i) > 1 \\ 0 & \text{en otro caso} \end{cases} \quad (13)$$

donde p_{ik}es la proporción de participaciones de la persona $i$en la especialidad k.

Componente de diversidad (DIV)

La entropía normalizada se transforma luego en un subíndice en $[0,1]$ con un área alrededor de un valor objetivo μ, dado que una diversidad temática muy baja indica especialización excesiva, pero una diversidad demasiado alta puede reflejar dispersión y baja profundidad en las líneas de trabajo. Se usa una función gaussiana que penaliza tanto la baja diversidad como la diversificación extrema fuera de la banda:

$$\text{DIV}(i) = \exp\left(-\frac{(\text{DIV\_ENTROPY\_N}(i) - \mu)^2}{2\sigma^2 + \varepsilon}\right) \quad (14)$$

3.8.1.4. Formación y clasificación vigente

Nivel de formación (EDU)

A partir del nivel de formación (`NIVEL_MAP`), se aplica un mapeo ordinal a la escala $[0,1]$según la taxonomía establecida (técnico, pregrado, maestría, doctorado, postdoctorado, etc.):

$$\mathrm{EDU}(i) = \mathrm{NIVEL\_MAP}\big(\mathrm{NIVEL\_ANCLA}(i)\big) \qquad (15)$$

donde `NIVEL_MAP` es el diccionario que asigna un valor entre 0 y 1 a cada nivel.

Clasificación (`CLS_raw`)

Si se dispone de un orden numérico de la clasificación (`ORDEN_CLAS_ANCLA`), se escala a $[0,1]$por min-máx:

$$\mathrm{CLS\_raw}(i) = \frac{\mathrm{ORDEN\_CLAS\_ANCLA}(i) - \min\,(\mathrm{ORDEN\_CLAS\_ANCLA})}{\max\,(\mathrm{ORDEN\_CLAS\_ANCLA}) - \min\,(\mathrm{ORDEN\_CLAS\_ANCLA}) + \varepsilon} \qquad (16)$$

En ausencia del orden numérico, se utiliza un mapeo ordinal de textos (`CLS_MAP`).

Año de última clasificación (`ANIO_ULT_CLS`)

Se calcula la cantidad de años que han pasado desde la última clasificación de la persona con respecto a la última clasificación vigente. La recencia de clasificación se expresa como:

$$\Delta_{\mathrm{CLS}(i)} = \max\{\,\mathrm{ANIO\_ANCLA}(i) - \mathrm{ANIO\_ULT\_CLS}(i), 0\} \qquad (17)$$

Clasificación con vigencia (`CLS_VIG`)

Para incorporar el efecto de la recencia, esta ecuación ajusta la clasificación cruda `CLS_raw` con base en qué tan antigua es la última clasificación. Cuando $\Delta_{\mathrm{CLS}} = 0$ (clasificación reciente), el factor exponencial es 1 y el puntaje se mantiene; a medida que aumentan los años desde la última clasificación, $\exp\,(-\lambda_{\mathrm{CLS}}\Delta_{\mathrm{CLS}})$toma valores cada vez menores entre 0 y 1, reduciendo progresivamente el aporte de esa clasificación al IMI. El parámetro λ_{CLS}controla la velocidad de este decaimiento: valores más altos implican penalizaciones más fuertes a clasificaciones desactualizadas:

$$\mathrm{CLS\_VIG}(i) = \mathrm{CLS\_raw}(i)\ \exp\left(-\lambda_{\mathrm{CLS}\,\Delta_{\mathrm{CLS}(i)}}\right) \qquad (18)$$

3.8.1.5. Métricas de experiencia y recencia

Componente de experiencia (`EXP`)

La experiencia investigativa se aproxima a partir del número de convocatorias. Para capturar rendimientos decrecientes, se utiliza una función cóncava $1 - e^{-kx}$sobre el conteo recortado:

$$N\_\mathrm{CONVOCATORIAS\_c}(i) = \min\,\{N\_\mathrm{CONVOCATORIAS}(i), p95\} \qquad (19)$$

$$\mathrm{EXP}(i) = 1 - \exp\left(-k_{\mathrm{EXP}\,N_{\mathrm{CONVOCATORIAS\_c}(i)}}\right) \qquad (20)$$

Recencia de participación (REC_S)

Se construye una medida de qué tan reciente es el perfil ancla respecto al año más actual observado en los datos (CURRENT_YEAR) . Se utiliza un decaimiento lineal en una ventana de recencia W:

$$\text{gap}(i) = \text{CURRENT\_YEAR} - \text{ANIO\_ANCLA}(i) \quad (21)$$

$$\text{REC\_S}(i) = \max\left\{0,\ 1 - \frac{\text{gap}(i)}{W}\right\} \quad (22)$$

con $\text{REC\_S} \in [0,1]$, donde 1 indica perfiles muy recientes.

3.8.1.6. Estandarización de componentes

Para facilitar la agregación posterior en el IMI, se generan versiones **estandarizadas en** $[0,1]$de las principales métricas, añadiendo el sufijo _S y aplicando *clip* de seguridad:

- CLS_VIG_S: versión recortada de CLS_VIG en $[0,1]$.
- EXP_S: versión estandarizada de EXP.
- DIV_S: versión estandarizada de DIV.
- MOB_S: versión estandarizada de MOB.
- REC_S: recencia ya definida en $[0,1]$.
- EDU_S: versión recortada de EDU en $[0,1]$.

Estas columnas _S constituyen los **componentes formales del IMI** que se combinarán en la siguiente sección mediante una agregación ponderada.

3.8.1.7. Índice base de movilidad geográfica

Finalmente, como insumo para el análisis de movilidad territorial, se define un índice simple a partir de los departamentos de nacimiento y residencia:

Categoría de movilidad geográfica (MOB_GEO_CAT)

A partir de NME_DEPARTAMENTO_NAC_PR y NME_DEPARTAMENTO_RES_PR se comparan los departamentos normalizados:

- Sin movilidad: nacimiento y residencia en el mismo departamento.
- Inter-departamental: nacimiento y residencia en departamentos diferentes.

Componente de movilidad geográfica (MOB_GEO_S)

Se mapea la categoría anterior a una escala $[0,1]$para su uso analítico:

$$\text{MOB\_GEO\_S}(i) = \begin{cases} 0{,}0 & \text{si MOB\_GEO\_CAT} = \text{“Sin movilidad”} \\ 1{,}0 & \text{si MOB\_GEO\_CAT} = \text{“Inter-departamental”} \end{cases} \quad (23)$$

3.8.2. Construcción de los índices

A partir de las métricas base descritas en la sección anterior, se define un índice de madurez investigativa por persona, $IMI(i)$, como combinación lineal ponderada de subíndices normalizados en $[0,1]$:

$$IMI(i) = 0{,}35\text{CLS\_VIG\_S}(i) + 0{,}25\text{EXP\_S}(i) + 0{,}10\text{DIV\_S}(i) + 0{,}05\text{MOB\_S}(i) + 0{,}15\text{REC\_S}(i) + 0{,}10\text{EDU\_S}(i)$$

En la implementación se asignan pesos mayores a la clasificación con vigencia (`CLS_VIG_S`) y a la trayectoria (`EXP_S`), pesos intermedios a la diversidad temática (`DIV_S`), la movilidad institucional (`MOB_S`) y la recencia del perfil (`REC_S`), y un peso menor al nivel de formación (`EDU_S`).

A partir del valor continuo de $IMI(i)$, se calculan percentiles y se definen cinco niveles ordinales (`IMI_TIER`) ("Muy Bajo"-0, "Bajo"≤0,2, "Medio"≤0,4, "Alto"≤0,8 y "Muy Alto"=1), lo que facilita su uso en análisis descriptivo, segmentación y definición de reglas de negocio.

3.8.3. Revisión, sensibilidad y publicación

Antes de liberar el Índice de Madurez Investigativa y los subíndices asociados al data mart analítico, se implementó una capa de control que verifica estabilidad, coherencia interna y posibles sesgos, además de dejar parametrizada la sensibilidad del modelo.

En primer lugar, se realiza un análisis de calidad técnico del IMI sobre el dataset completo. Se verifica: (i) cobertura total del índice obteniendo como resultado 100 %, (ii) coherencia de signo mediante las correlaciones de IMI con `CLS_VIG_S` y `EXP_S`, validando que el índice debe crecer con mejor clasificación vigente y mayor trayectoria, (iii) distribución de los quintiles (`IMI_TIER`) para asegurar cortes razonablemente balanceados, y (iv) sanidad de rangos en todos los subíndices, confirmando que `CLS_VIG_S`, `EXP_S`, `DIV_S`, `MOB_S`, `REC_S` y `EDU_S` se mantienen en la escala [0,1] sin valores atípicos fuera de banda.

En segundo lugar, se evalúa la estabilidad del IMI por cohorte temporal. Se calcula la media, desviación estándar y proporción de perfiles "Muy Alto" y "Muy Bajo" por año ancla. Este corte permite detectar rupturas de escala entre promociones (por ejemplo, cohortes con IMI sistemáticamente más alto por cambios de política o de datos). Cuando la información está disponible, el mismo análisis se puede extender por grandes áreas de conocimiento, como verificación adicional de consistencia entre dominios disciplinares:

Año convocatoria	n	Promedio IMI	Desviación estándar IMI	Muy alto (%)	Muy bajo (%)
2013	1,691	0,164845	0,059888	0,000000	93,258427
2014	585	0,173186	0,086875	0,000000	84,444444
2015	1,148	0,233847	0,125210	6,968641	58,797909
2017	2,127	0,273878	0,130147	7,522332	34,179596
2019	3,443	0,310997	0,115682	8,103398	25,210572
2021	21,091	0,410220	0,152505	26,561092	0,000000

Tabla 8. Revisión IMI

En la Tabla 8, se observa un incremento progresivo del promedio de IMI entre 2013 y 2021 (de 0,16 a 0,41), lo que indica que, en conjunto, las cohortes más recientes presentan trayectorias de mayor madurez investigativa, lo cual es algo esperable en un índice acumulativo que incorpora experiencia y clasificación. Adicionalmente, la proporción de perfiles "Muy bajos" cae drásticamente: pasa de más del 90 % en 2013 a

84 % en 2014, 59 % en 2015 y continúa descendiendo hasta desaparecer en 2021, mientras que la proporción de perfiles “Muy altos” crece desde 0 % en 2013-2014 hasta alrededor del 26 % en 2021. Asimismo, las desviaciones estándar por año se mantienen en rangos adecuados (0,06-0,15), sin cambios drásticos, lo que sugiere que no hay saltos de escala entre convocatorias ni comportamientos anómalos en alguna cohorte específica. En conjunto, la tabla respalda que el IMI es sensible a las diferencias de madurez entre años, pero al mismo tiempo muestra una evolución temporal suave y consistente con la dinámica esperada del sistema de investigación.

Como tercer componente, con ayuda de la Figura 15, se implementa un monitoreo de equidad. El IMI por persona se enriquece con variables sociales provenientes del consolidado (`NME_GENERO_PR`, `TXT_GRUPO_ETNICO`, `TXT_POBLACION_DISCA`, `ID_VICTIMA_CONFLICTO`) y, para cada eje disponible, se genera un resumen por grupo: tamaño de muestra, media y dispersión del IMI, y porcentaje de perfiles “Muy Alto” y “Muy Bajo”. Estos cortes no se utilizan para ajustar el score, sino como radar de vigilancia para identificar brechas sistemáticas entre géneros, grupos étnicos, población con discapacidad o víctimas del conflicto armado, y documentar hallazgos en futuras iteraciones del modelo:

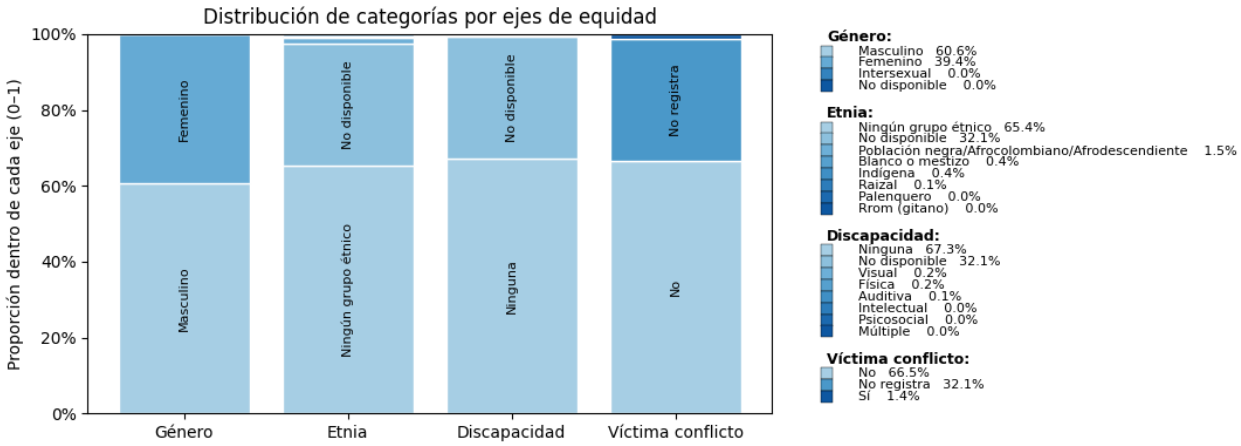

Figura 15. Distribución de IMI por categorías de equidad

3.9. Formateo de datos

3.9.1. Conjunto base de trabajo

A partir del dataset obtenido en la etapa de construcción de datos, se realizó un análisis variable por variable para definir su rol dentro de la fase de modelado. Esta revisión consideró la pertinencia de cada campo frente al objetivo del estudio, su naturaleza estadística, su posible redundancia con variables derivadas, el riesgo de introducir sesgos no deseados y su utilidad para la interpretación de los resultados.

Como resultado, las variables fueron clasificadas en cuatro grupos: variables utilizadas directamente por el algoritmo, variables de monitoreo e interpretación, variables de trazabilidad y variables excluidas del modelado. Esta clasificación se detalla en el Anexo 6 y permitió separar los campos que aportan señal analítica para la segmentación de aquellos que cumplen una función contextual, descriptiva o de control.

3.9.2. Diagnóstico para modelado de variables numéricas y categóricas

En esta sección se presenta el inventario general de las variables consideradas para la etapa de modelado, donde, si bien la totalidad de las variables que conforman el conjunto de entrada son de naturaleza numérica, su origen es heterogéneo: un primer grupo corresponde a variables discretas derivadas directamente de los datos originales (conteos y medidas de trayectoria), mientras que un segundo grupo está compuesto por índices que fueron construidos en la etapa de preparación de datos a partir de atributos categóricos u ordinales, tales como el nivel educativo, la clasificación en el escalafón y la diversidad temática.

El propósito de esta sección es documentar de manera sintética el tipo de variable y su rango observado, con el fin de identificar posibles requerimientos de normalización, estandarización de escala o tratamiento específico previo al entrenamiento del modelo. En la Tabla 9 se relacionan estas variables numéricas junto con su origen y respectivos rangos, información que servirá de base para las decisiones posteriores de preprocesamiento y balanceo.

Variable	Descripción	Tipo de numérica	Rango observado
EDAD_ANCLA	Edad de la persona (en años) en la convocatoria más reciente.	Discreta.	17,58 - 92,83
N_CONVOCATORIAS	Número total de convocatorias en las que ha participado la persona.	Discreta.	1 - 6
N_INST	Número de instituciones distintas en las que ha participado.	Discreta.	1 - 5
N_AREAS	Número de áreas de conocimiento diferentes en la trayectoria.	Discreta.	1 - 4
ANTIGUEDAD_ANIOS	Años transcurridos entre la primera y la última participación.	Discreta.	0 - 8
EDU	Índice normalizado de nivel educativo alcanzado.	Índice continuo [0,1] (ordinal).	0,55 - 1,00
CLS_raw	Puntaje bruto de clasificación en el escalafón (Junior/Asociado/Senior) en la situación actual.	Índice continuo [0,1].	0,00 - 1,00
CLS_VIG	Puntaje de clasificación vigente/ajustada utilizado en el índice de madurez.	Índice continuo [0,1].	0,00 - 1,00
TASA_MOV_INST	Tasa de movilidad institucional (proporción de años con cambio de institución).	Índice continuo [0,1].	0,00 - 1,00
DIV_ENTROPY_N	Entropía normalizada de la distribución de instituciones a lo largo de la trayectoria.	Índice continuo ≈[0,1].	-0,00 - 1,00
EXP	Indicador de exposición/experiencia acumulada en el sistema de convocatorias.	Índice continuo [0,1].	0,259 - 0,835
DIV	Indicador complementario de diversidad de trayectorias (institucional/temática).	Índice continuo [0,1].	0,044 - 0,755
MOB	Índice sintético de movilidad institucional acumulada en convocatorias	Índice continuo [0,1].	0,119 - 0,953

Tabla 9. Inventario variables para el modelo

3.9.3. Escalado y normalización de variables numéricas

3.9.3.1. Criterios de escalado

En esta subsección se definen los criterios de escalado aplicados a las variables numéricas que ingresan al modelo de clustering. De esta forma, se ha decidido implementar una estandarización tipo z-score (media cero y desviación estándar uno) para todas las variables del modelo, de manera que queden en órdenes de magnitud comparables y ninguna domine el cálculo de distancias únicamente por su escala, dado que en contextos de clustering las variables con mayor varianza absoluta tienden a dominar el cálculo de distancias si no se estandarizan previamente [66]. Aunque una parte de los indicadores (EDU, CLS_raw, CLS_VIG, TASA_MOV_INST, DIV_ENTROPY_N, EXP, DIV y MOB) había sido construida previamente en la etapa 4.4 sobre una escala normalizada en el rango [0,1], se consideró que mantener criterios de escalado distintos podría afectar el comportamiento del modelo. Por esto, se optó por aplicar el mismo esquema de estandarización también a estos índices, mitigando el riesgo de que diferencias residuales de varianza influyan de forma desproporcionada en la formación de los clusters.

3.9.3.2. Aplicación del escalado sobre el conjunto de entrenamiento.

A continuación, en la Tabla 10 se presentan los estadísticos básicos de cada variable antes y después de la estandarización tipo z-score, con el fin de evidenciar el efecto del escalado sobre la distribución de los datos. En la sección izquierda de la tabla se observan las escalas originales, donde variables como EDAD_ANCLA presentan una media de 45,29 y una desviación estándar de 10,36, mientras que los índices construidos se concentran en el rango [0,1] con desviaciones inferiores a 0,30. Esta disparidad confirma la necesidad de la estandarización señalada en el criterio anterior, dado que sin ella las variables con mayor dispersión absoluta dominarían el cálculo de distancias. En la sección derecha de la tabla se verifica que, tras el escalado, todas las variables alcanzan media cero y desviación estándar uno, lo que garantiza su participación equitativa en el modelo. Asimismo, los valores mínimos y máximos estandarizados permiten identificar la presencia de trayectorias particularmente extremas, como en el caso de N_AREAS, la cual alcanza un máximo de 9,47 desviaciones estándar, lo que indica investigadores con una diversidad de áreas de conocimiento muy por encima de la tendencia central de la distribución.

Variable	Antes del escalado				Después del escalado			
	Media	Desv estándar	Mínimo	Máximo	Media	Desv estándar	Mínimo	Máximo
EDAD_ANCLA	45,29	10,36	17,58	92,83	0,00	1,00	-2,67	4,58
N_CONVOCATORIAS	2,56	1,65	1,00	6,00	0,00	1,00	-0,94	2,07
N_INST	1,31	0,56	1,00	5,00	0,00	1,00	-0,55	6,55
N_AREAS	1,09	0,30	1,00	4,00	0,00	1,00	-0,32	9,47
ANTIGUEDAD_ANIOS	3,03	3,01	0,00	8,00	0,00	1,00	-1,00	1,64
EDU	0,85	0,17	0,10	10,00	0,00	1,00	-4,73	0,21
CLS_raw	0,14	0,24	0,00	10,00	0,00	1,00	-0,60	3,50
CLS_VIG	0,14	0,24	0,00	10,00	0,00	1,00	-0,60	3,50
TASA_MOV_INST	0,07	0,14	0,00	10,00	0,00	1,00	-0,49	6,25
DIV_ENTROPY_N	0,07	0,23	0,00	10,00	0,00	1,00	-0,32	3,87
EXP	0,48	0,20	0,25	0,83	0,00	1,00	-1,09	1,69
DIV	0,07	0,14	0,04	0,75	0,00	1,00	-0,24	4,78
MOB	0,24	0,24	0,11	0,95	0,00	1,00	-0,51	2,88

Tabla 10. Estadísticos básicos antes y después del escalado tipo z-score

3.9.4. Dataset final formateado

En esta sección se construyó la matriz de características definitiva para el modelo de clustering, denominada X_{Modelo}. A partir del dataset formateado y escalado, se extrajeron exclusivamente las trece variables numéricas estandarizadas que se definieron como insumo del algoritmo (`EDAD_ANCLA`, `N_CONVOCATORIAS`, `N_INST`, `N_AREAS`, `ANTIGUEDAD_ANIOS`, `EDU_z`, `CLS_raw_z`, `CLS_VIG_z`, `TASA_MOV_INST_z`, `DIV_ENTROPY_N_z`, `EXP_z`, `DIV_z` y `MOB_z`), conformando una matriz de dimensión $N \times 13$, donde N corresponde al número de investigadores con información completa en estas variables. El propósito de este paso es separar explícitamente el vector de entrada utilizado por el modelo de clustering del resto de atributos de monitoreo, dejando claramente delimitado el conjunto de datos que será utilizado en la fase de modelado y garantizando que todas las variables que intervienen en el cálculo de distancias han pasado por el mismo proceso de tratamiento y estandarización.

4. MODELADO

4.1. Selección de la técnica de modelado

En la fase de Modelado de la metodología CRISP-DM, la selección de la técnica constituye el primer paso analítico, dado que de esta decisión depende la capacidad del algoritmo para capturar la estructura latente en los datos [60]. A partir de la matriz de características X_Modelo, conformada por 30.067 investigadores y 13 variables numéricas estandarizadas mediante *z-score*, se procedió a evaluar la aptitud del dataset para el análisis de clustering y a seleccionar los algoritmos candidatos que permitan identificar agrupaciones naturales de perfiles de investigación.

4.1.1. Validación de la tendencia al agrupamiento

Como paso previo a la selección de algoritmos, se verificó la tendencia al agrupamiento del dataset mediante el estadístico de Hopkins (*Hopkins Statistic*), que contrasta la distribución espacial observada contra una distribución uniforme aleatoria [51]. Se obtuvo un valor de $H = 0{,}9907 \pm 0{,}0005$ sobre 10 iteraciones independientes con muestras de 3.006 puntos cada una, resultado que supera ampliamente el umbral de referencia $H > 0{,}75$ establecido en la literatura para confirmar una tendencia fuerte al clustering. Este hallazgo indica que la matriz X_Modelo presenta una estructura interna altamente diferenciada, con regiones de alta densidad separadas por zonas de baja densidad, lo que valida la pertinencia de aplicar técnicas de segmentación no supervisada sobre los datos preparados en el capítulo anterior.

4.1.2. Técnicas candidatas

Teniendo en cuenta la estrategia multiparadigma adoptada previamente, se define que las técnicas que serán implementadas son:

- K-Means++
- GMM
- Agglomerative
- HDBSCAN

4.2. Diseño del plan de prueba

4.2.1. Protocolo de iteración

El plan de prueba se estructuró como un barrido sistemático de configuraciones de $k = 2$ a $k = 10$ clústeres para cada uno de los cuatro algoritmos seleccionados, dado que este rango abarca desde la dicotomía mínima hasta un nivel de granularidad compatible con la interpretabilidad del dominio académico. En cada iteración, se registró el resultado del coeficiente de silueta, el índice de Calinsk-Harabasz y el método del codo, junto con el conteo de observaciones por clúster, lo que generó un *log* de optimización completo y auditable.

Adicionalmente, se incorporaron variantes de hiperparámetros específicos por algoritmo como el tipo de

covarianza en GMM y método de enlace en Agglomerative Clustering, con lo cual el espacio de búsqueda total comprendió múltiples combinaciones evaluadas de forma exhaustiva. Para cada configuración, se verificó que ningún clúster concentrara menos del 5 % de la población, criterio que responde al requisito de robustez estadística y asegura que los grupos resultantes posean suficiente masa para sustentar inferencias válidas sobre los perfiles de investigación del SNCTI.

En síntesis, el plan de prueba establece un protocolo de validación tridimensional compuesto por cohesión (Silueta), separación (Calinski-Harabasz) e inercia marginal (Codo), aplicado de forma sistemática sobre el rango $k \in [2, 10]$ para los cuatro paradigmas de *clustering*, con lo que se asegura que la selección del modelo en la siguiente se sustente en evidencia cuantitativa convergente y no en el resultado de un único criterio.

4.3. Construcción del modelo

A partir del plan de prueba definido, se procedió con la construcción de los modelos de *clustering* sobre la matriz X_{Modelo} (30.067 × 13), con el fin de identificar la configuración de hiperparámetros que maximice la calidad de la segmentación de perfiles del SNCTI. En esta subsección se documentan los resultados del barrido experimental para K-Means++, algoritmo que obtuvo el mejor desempeño global entre los tres paradigmas evaluados, se presentan las curvas de inercia y de Silueta que sustentan la elección de *k* = 4 clústeres, y se describe la configuración final del modelo seleccionado.

4.3.1. Ejecución del barrido experimental

El protocolo descrito en la sección anterior se aplicó de forma exhaustiva a las 151 combinaciones de algoritmo e hiperparámetros: 14 configuraciones de K-Means++ ($k \in [2, 15]$), 56 de GMM (4 tipos de covarianza × 14 valores de *k*), 56 de Agglomerative Clustering (4 métodos de enlace × 14 valores de *k*) y 25 de HDBSCAN (5 valores de `min_cluster_size` × 5 de `min_samples`). Para cada configuración se registraron el Coeficiente de Silueta, el Índice de Calinski-Harabasz (CH) y el Índice de Davies-Bouldin (DB), junto con el tamaño mínimo y máximo de clúster y el tiempo de ejecución, lo que generó un *log* de optimización completo y reproducible.

El mejor resultado correspondió a K-Means++ con k = 4, con un ranking promedio de 2,7, seguido por configuraciones de GMM con cuatro componentes. Por su parte, HDBSCAN generó soluciones excesivamente fragmentadas, con un número elevado de clústeres según la combinación de hiperparámetros, mientras que Agglomerative Clustering presentó evidencia de encadenamiento, concentrando una proporción mayoritaria de observaciones en un solo grupo. Estos resultados orientaron el análisis posterior hacia K-Means++, cuya estabilidad y capacidad de segmentación resultaron más adecuadas para el propósito de recomendación académica, y se detallan en el Anexo 7.

4.3.2. Comparación multi-criterio de los modelos candidatos

En la Tabla 11 se consolidan las métricas de desempeño técnico de la mejor configuración de cada paradigma. Para los tres primeros algoritmos, la evaluación se realizó con *k* = 4 clústeres según la convergencia observada, pero en el caso de HDBSCAN, el número de clústeres fue determinado automáticamente por el algoritmo a partir de la estructura de densidad del dataset, por lo que se reporta el valor de *k* resultante de la mejor configuración. La prueba de estabilidad consistió en ejecutar cada algoritmo con cinco semillas aleatorias distintas (42, 123, 456, 789 y 2024) y calcular el ARI promedio entre todos los pares de asignaciones resultantes, donde ARI = 1,0 indica concordancia perfecta [59].

Criterio	K-Means++	GMM (esférica)	Agglomerative (Ward)	HDBSCAN (mcs=300, ms=30)
Clústeres (k)	4 (predefinido)	4 (predefinido)	4 (predefinido)	24 (auto-detectado)
Coeficiente de Silueta ($\bar{S}$)	**0,3986**	0,3829	0,3394	0,2760
Calinski-Harabasz (CH)	**13.592,7**	13.115,4	12.067,6	4.082,7
Davies-Bouldin (DB)	**1,1517**	1,1724	1,2960	1,4254
ARI (estabilidad)	**0,9999 ± 0,0000**	0,5688 ± 0,2909	1,0000 (det.)	1,0000 (det.)
Puntos de ruido	0 (0,0 %)	0 (0,0 %)	0 (0,0 %)	1.040 (3,5 %)
Ranking promedio (de 151)	**2,7**	5,0	28,0	90,3
Tiempo de ejecución (s)	7,15	7,14	40,07	11,70

Tabla 11. Comparación de métricas de desempeño técnico de los modelos candidatos

Los resultados evidencian que K-Means++ con *k* = 4 domina en las tres métricas de validación interna de forma simultánea: obtiene el Coeficiente de Silueta más alto ($\bar{S} = 0{,}3986$), el Índice de Calinski-Harabasz más elevado ($\mathrm{CH} = 13.592{,}7$) y el Índice de Davies-Bouldin más bajo ($\mathrm{DB} = 1{,}1517$). GMM con covarianza esférica ocupa la segunda posición en las tres métricas, mientras que Agglomerative Clustering con enlace Ward se ubica en tercer lugar con diferencias sustanciales en Silueta (-0,0592) y en Davies-Bouldin (+0,1443) respecto al líder.

HDBSCAN, por su parte, se ubica en la posición 90 del *leaderboard* de 151 configuraciones, con un desempeño significativamente inferior en las tres métricas intrínsecas: Silueta de 0,2760 (-0,1226 respecto al líder), Calinski-Harabasz de 4.082,7 (-69,9 %) y Davies-Bouldin de 1,4254 (+23,8 %). No obstante, la limitación más relevante no reside en las métricas individuales, sino en la fragmentación excesiva del espacio de perfiles: la mejor configuración (*min_cluster_size* = 300, *min_samples* = 30) genera 24 clústeres, cifra que se eleva hasta 172 con parámetros más granulares (*min_cluster_size* = 50).

Adicionalmente, HDBSCAN clasifica como ruido 1.040 investigadores (3,5 % del total), puntos que quedarían excluidos del sistema de recomendación al carecer de asignación a un grupo. Esta exclusión contradice el requisito funcional de cobertura universal definido previamente, donde se estableció que el prototipo debe generar recomendaciones para la totalidad de los perfiles del SNCTI. Por estas razones, HDBSCAN se descarta del listado de selección del modelo.

En síntesis, la comparación multi-criterio presentada en la Tabla 11 evidencia que K-Means++ con k = 4 supera de manera consistente a los demás paradigmas evaluados en las tres métricas de validación interna,

al tiempo que alcanza la mayor estabilidad (ARI = 0,9999 ± 0,0000) y el menor tiempo de ejecución (7,15 s). A partir de estos resultados, se selecciona K-Means++ como el algoritmo definitivo para las siguientes etapas del proyecto, dado que su combinación de desempeño estadístico, reproducibilidad y eficiencia computacional satisface los criterios establecidos en el plan de prueba. De esta manera, los centroides y las asignaciones de clúster producidos por este modelo constituirán el insumo directo tanto para la caracterización de perfiles investigativos como para el cálculo de similitud en el motor de recomendación de colaboradores académicos.

4.3.3. Análisis del Método del Codo del algoritmo seleccionado

En la Figura 16 se presenta la curva de inercia (SSE) de K-Means++ para $k \in [2, 10]$, donde se puede observar un decrecimiento pronunciado de la inercia entre $k = 2$ (y $k = 4$ (165.847,6), seguido de una atenuación gradual a partir de $k = 5$. En términos porcentuales, la transición de $k = 3$ a $k = 4$ produce una reducción de inercia del 20,2 %, mientras que el paso de $k = 4$ a $k = 5$ reduce solo un 10,2 %, lo que representa una caída a la mitad en el rendimiento marginal. Este cambio de pendiente identifica a $k = 4$ como el punto de inflexión de la curva, donde adicionar clústeres deja de aportar una mejora sustancial en la compactación de los grupos.

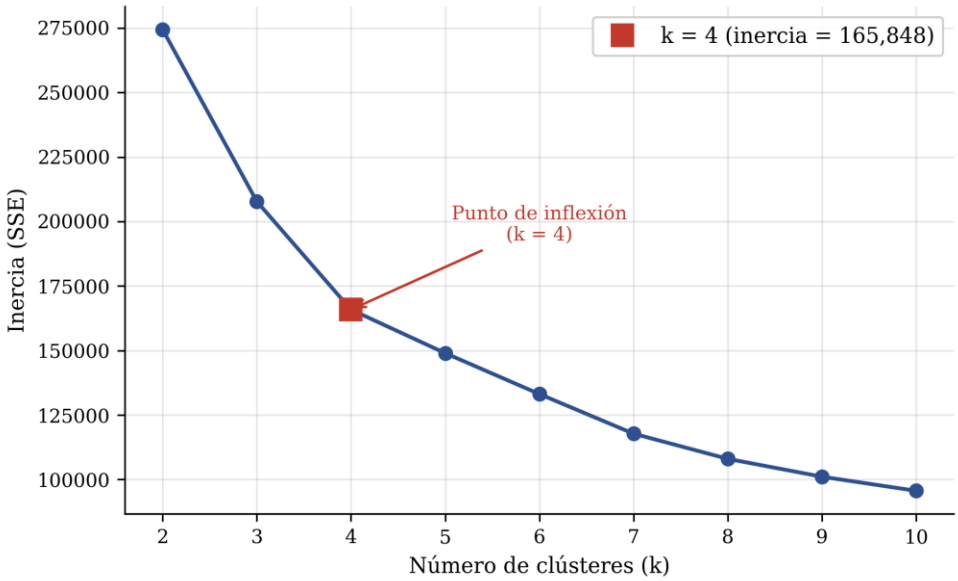

Figura 16. Curva de inercia de K-Means++ para k ∈ [2, 10].

4.3.4. Análisis del Coeficiente de Silueta del algoritmo seleccionado

De forma complementaria, en la Figura 17 se muestra el Coeficiente de Silueta promedio ($\bar{S}$) para cada valor de k, donde se evidencia que $k = 4$ alcanza el máximo absoluto con $\bar{S} = 0,3986$. A partir de $k = 5$, el coeficiente desciende a $\bar{S} = 0,3382$ y se mantiene por debajo de $\bar{S} = 0,35$ para todos los valores subsiguientes, lo que indica que incrementar el número de grupos más allá de cuatro deteriora la cohesión interna sin mejorar la separación entre clústeres.

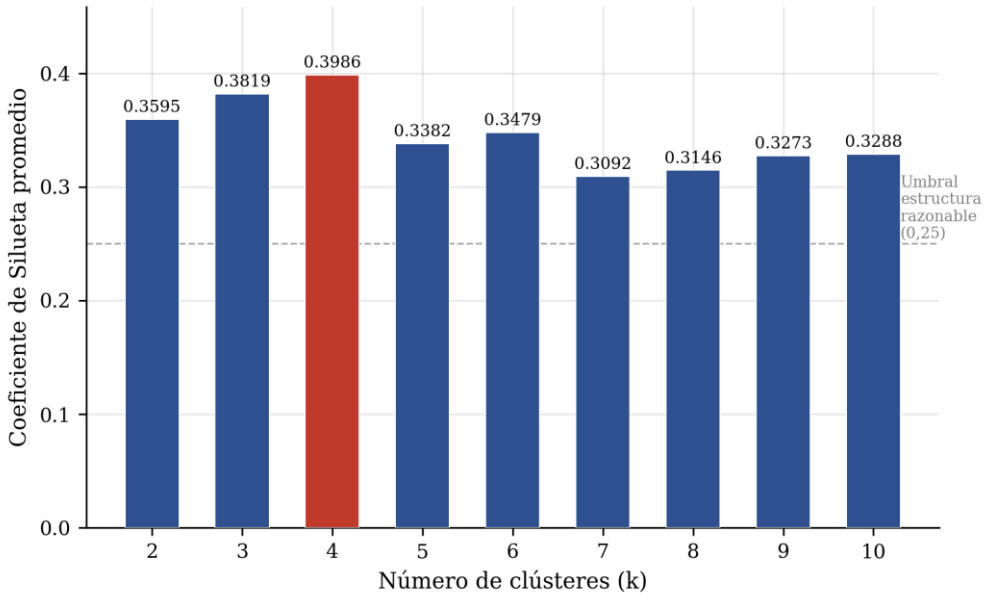

Figura 17. Coeficiente de Silueta promedio de K-Means++ para k ∈ [2, 10].

Asimismo, el índice de Calinski-Harabasz corrobora este resultado: k = 4 obtiene el valor más alto (CH = 13.592,7), lo que refleja la mayor razón de dispersión inter-clúster sobre dispersión intra-clúster dentro del rango evaluado. De forma complementaria, la inercia (SSE) permite observar la suma de las distancias cuadráticas de las observaciones respecto al centroide de su respectivo clúster, por lo tanto, valores menores indican mayor compactación interna de los grupos. Sin embargo, su disminución debe analizarse junto con el porcentaje de reducción de inercia, ya que este muestra cuánto mejora la compactación al aumentar k. En la Tabla 12 se observa que, al pasar de k = 3 a k = 4, la inercia disminuye de 207.784,9 a 165.847,6, con una reducción del 20,2 %, mientras que para valores posteriores las reducciones son progresivamente menores. Esto indica que k = 4 representa un punto de mejora relevante antes de que los beneficios marginales de agregar nuevos clústeres empiecen a disminuir. De manera convergente, el índice de Davies-Bouldin alcanza su mínimo en k = 4 (DB = 1,1517), lo que confirma que esta configuración produce los clústeres más compactos y mejor separados entre sí.

k	Inercia (SSE)	Δ Inercia (%)	Silueta ($\bar{S}$)	Calinski-Harabasz	Davies-Bouldin
2	274.408,7	-	0,3595	12.752,9	1,4305
3	207.784,9	-24,3 %	0,3819	13.240,6	1,2777
4	**165.847,6**	**-20,2 %**	**0,3986**	**13.592,7**	**1,1517**
5	148.983,1	-10,2 %	0,3382	12.198,9	1,2024
6	133.129,0	-10,6 %	0,3479	11.636,9	1,2228
7	117.763,2	-11,5 %	0,3092	11.616,1	1,2317
8	108.014,0	-8,3 %	0,3146	11.242,6	1,2275
9	101.083,2	-6,4 %	0,3273	10.769,0	1,1875
10	95.625,8	-5,4 %	0,3288	10.309,0	1,2037

Tabla 12. Métricas de validación interna de K-Means++ por número de clústeres (k)

4.3.5. Configuración del modelo final

Con base en la convergencia de las tres métricas de validación interna en $k = 4$, se configuró el modelo final de K-Means++ con los siguientes hiperparámetros: método de inicialización *k-means++* para seleccionar centroides iniciales dispersos y reducir el riesgo de mínimos locales: `n_init` = 10 ejecuciones independientes con selección automática de la mejor inercia, `max_iter` = 300 iteraciones como criterio de convergencia, y `random_state` = 42 para garantizar la reproducibilidad del experimento. El entrenamiento del modelo completo sobre los 30.067 registros se completó en 7,15 segundos, lo que confirma la eficiencia computacional y garantiza escalabilidad en caso de requerirse en el futuro.

La segmentación resultante distribuyó los investigadores en cuatro grupos de tamaños diferenciados: el Clúster 0 con 6.480 investigadores (21,6 %), el Clúster 1 con 4.119 (13,7 %), el Clúster 2 con 16.694 (55,5 %) y el Clúster 3 con 2.774 (9,2 %). Si bien el Clúster 2 concentra la mayoría de la población, todos los grupos superan el umbral mínimo del 5 % definido, y el grupo más pequeño (Clúster 3, $n = 2.774$) posee suficiente masa estadística para sustentar análisis descriptivos e inferencias válidas sobre los perfiles de investigación.

En síntesis, la construcción del modelo confirmó que K-Means++ con $k = 4$ constituye la configuración óptima entre las 151 evaluadas, dado que maximiza simultáneamente el Coeficiente de Silueta ($\bar{S} = 0,3986$), el Índice de Calinski-Harabasz (CH = 13.592,7) y minimiza el Índice de Davies-Bouldin (DB = 1,1517), con un punto de inflexión claro en la curva de inercia.

4.4. Caracterización de los clústeres

Una vez seleccionado K-Means++ con $k = 4$ como modelo , se procedió a la caracterización semántica de los clústeres con el propósito de traducir los centroides matemáticos en perfiles investigativos comprensibles para los *stakeholders* del SNCTI. En esta subsección se analizan las variables que mayor poder discriminante aportan a la segmentación, se describen los rasgos distintivos de cada grupo mediante las 13 variables de entrada y se propone una taxonomía de arquetipos que orienta las estrategias de recomendación del sistema.

4.4.1. Variables discriminantes

Para identificar qué dimensiones del perfil investigativo contribuyen en mayor medida a la separación de los clústeres, se calculó la importancia relativa de cada variable mediante la razón de varianza inter-clúster sobre varianza total, ponderada por la distancia a los centroides. Los cinco *features* con mayor poder discriminante concentran el 65,5 % de la importancia acumulada: `ANTIGUEDAD_ANIOS` (14,3 %), `EXP` (12,4 %), `TASA_MOV_INST` (11,9 %), `MOB` (11,5 %) y `N_CONVOCATORIAS` (11,4 %). Este resultado indica que la segmentación se estructura principalmente en torno a dos ejes conceptuales: la profundidad de la trayectoria en el sistema (antigüedad, experiencia y participación en convocatorias) y la movilidad institucional (tasa de cambio entre instituciones e índice sintético de movilidad). En contraste, las variables `EDU` (0,4 %) y `EDAD_ANCLA` (1,3 %) presentan la menor contribución, lo que sugiere que el nivel educativo y la edad son relativamente homogéneos entre los investigadores reconocidos y, por tanto, no constituyen

ejes de diferenciación relevantes para la segmentación.

4.4.2. Perfiles de los clústeres

En esta sección se presentan los perfiles resultantes de la segmentación, con el propósito de traducir los centroides del modelo en arquetipos interpretables para el análisis de colaboración académica. Para ello, se utiliza un diagrama de radar que permite comparar, de forma relativa, el comportamiento de los cuatro clústeres frente a las variables con mayor capacidad discriminante en la caracterización de los investigadores.

En la Figura 18 se presenta el diagrama de radar de los centroides normalizados de los cuatro clústeres. Cada eje representa una variable seleccionada para la interpretación del modelo, entre ellas experiencia, educación, diversidad temática, movilidad, número de convocatorias, número de instituciones, número de áreas de conocimiento, antigüedad, entropía temática y tasa de movilidad institucional. Esta visualización permite observar qué dimensiones tienen mayor o menor intensidad relativa en cada clúster, facilitando la comparación entre perfiles.

Para construir el radar, los valores de los centroides fueron normalizados al rango [0,1] mediante una transformación min-max aplicada por variable, tomando como universo de referencia únicamente los cuatro centroides y no la totalidad de investigadores del dataset. La normalización se expresa de la siguiente manera:

$$Z_{kj} = \frac{c_{kj} - \min(c_{\cdot j})}{\max(c_{\cdot j}) - \min(c_{\cdot j})} \qquad (24)$$

Donde:

- c_{kj}: Representa el valor del centroide del clúster $k$en la variable j.
- $\min(c_{\cdot j})$: Corresponde al menor valor observado para la variable $j$entre los cuatro centroides.
- $\max(c_{\cdot j})$: Corresponde al mayor valor observado para la variable $j$entre los cuatro centroides.

Esta transformación se utilizó exclusivamente con fines de visualización comparativa y no corresponde al procedimiento de escalado empleado para entrenar el modelo de clustering. En consecuencia, el radar no debe interpretarse como una escala absoluta del investigador promedio ni como un percentil respecto al dataset completo. Un valor cercano a 1 indica que ese clúster presenta el mayor centroide relativo en esa variable frente a los demás clústeres, mientras que un valor cercano a 0 indica que presenta el menor valor relativo entre los centroides comparados.

Esta precisión es especialmente importante para interpretar el clúster C2 - Emergentes, el cual aparece visualmente cerca del centro del radar en varias dimensiones porque registra valores mínimos o empatados en el mínimo frente a los demás clústeres. Esto no significa ausencia de información ni error en la visualización, sino una posición relativa baja respecto a los otros perfiles. Los valores reales de los centroides se conservan en el Anexo 8 como soporte de trazabilidad y replicabilidad del análisis.

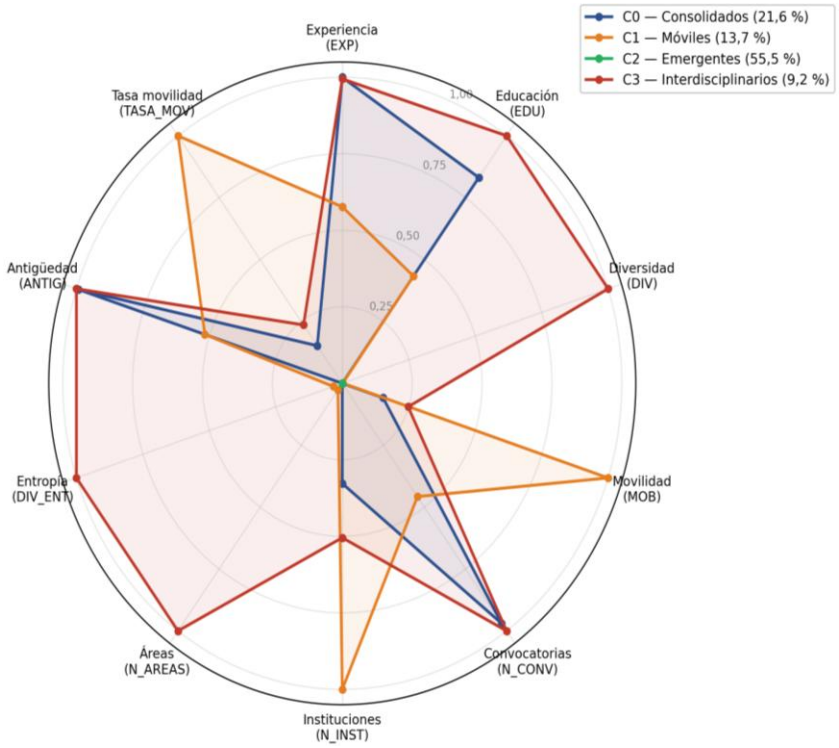

Figura 18. Diagrama de radar de los centroides normalizados de los cuatro clústeres.

A partir de la lectura del radar y de los valores reales de los centroides documentados en el Anexo 8, se construyó una síntesis interpretativa de los cuatro perfiles identificados. Esta síntesis no reemplaza los valores técnicos de los centroides, sino que los traduce en arquetipos comprensibles para el análisis del dominio y para la posterior configuración del sistema de recomendación. En la Tabla 13 se presenta el nombre asignado a cada clúster, su tamaño relativo, los rasgos dominantes, las variables que explican su diferenciación y la implicación operativa para la recomendación de colaboradores.

Clúster	Arquetipo	n (%)	Perfil dominante	Variables definitorias	Implicación para la recomendación
C0	**Consolidados**	6.480 (21,6 %)	Investigadores sénior, alta experiencia, una sola área, baja movilidad	EXP ↑, ANTIG ↑, N_CONV ↑, MOB ↓, DIV ↓	Recomendar colaboradores complementarios en áreas adyacentes
C1	**Móviles**	4.119 (13,7 %)	Investigadores en crecimiento con alta rotación institucional	MOB ↑↑, TASA_MOV ↑↑, N_INST ↑, EXP ↕	Sugerir investigadores de instituciones destino para facilitar inserción
C2	**Emergentes**	16.694 (55,5 %)	Recién ingresados, mínima trayectoria, una institución, un área	EXP ↓, ANTIG ↓, MOB = 0, N_CONV ↓	Conectar con Consolidados o Interdisciplinarios como mentores
C3	**Interdiscipl.**	2.774 (9,2 %)	Investigadores sénior con diversidad temática alta y entropía elevada	DIV ↑↑, DIV_ENT ↑↑, N_AREAS ↑, EXP ↑	Recomendar como puentes entre disciplinas para proyectos multi-área

Tabla 13. Taxonomía de arquetipos por clúster y su implicación para el sistema de recomendación

Clúster 0 - Consolidados (*n* = 6.480, 21,6 %). Este grupo reúne investigadores con alta experiencia acumulada (EXP = 0,72), prolongada antigüedad en el sistema (ANTIG = 6,53 años) y elevada participación en convocatorias (N_CONV = 4,46). Su nivel educativo es alto (EDU = 0,92), con mediana en el máximo de la escala, lo que indica predominio de formación doctoral o posdoctoral. Se caracterizan por una movilidad institucional baja (MOB = 0,22; TASA_MOV = 0,06), lo que refleja una trayectoria estable vinculada mayoritariamente a una sola institución (N_INST = 1,40). La diversidad temática es mínima (DIV = 0,04; N_AREAS = 1,0), lo que configura un perfil de especialista disciplinar. La edad promedio de 50,0 años confirma que se trata de investigadores en etapa de madurez plena de su carrera académica.

Clúster 1 - Móviles (*n* = 4.119, 13,7 %). El rasgo definitorio de este grupo es la alta movilidad institucional: MOB = 0,77 y TASA_MOV = 0,38, los valores más elevados de toda la segmentación, acompañados de un promedio de N_INST = 2,23 instituciones. Su experiencia es intermedia (EXP = 0,56) con una antigüedad de 3,83 años y una participación de 2,86 convocatorias, lo que sugiere investigadores en fase de crecimiento activo que han transitado entre distintas entidades para consolidar su trayectoria. El nivel educativo permanece alto (EDU = 0,87), aunque ligeramente inferior al de los Consolidados. La clasificación vigente es la más baja de los cuatro grupos (CLS_VIG = 0,10), lo que podría reflejar que el cambio frecuente de institución dificulta mantener una clasificación actualizada ante Minciencias. Con una edad promedio de 43,0 años, este clúster representa una cohorte de investigadores en etapa intermedia que priorizan la diversificación de redes institucionales.

Clúster 2 - Emergentes (*n* = 16.694, 55,5 %). Se trata del grupo más numeroso y homogéneo de la segmentación, compuesto por investigadores con baja experiencia (EXP = 0,34), antigüedad cercana a cero (ANTIG = 0,89 años) y participación limitada a 1,43 convocatorias en promedio. La movilidad institucional es nula (MOB = 0,12, valor mínimo de la escala; TASA_MOV = 0,00) y los investigadores se vinculan a una sola institución (N_INST = 1,0) en una sola área de conocimiento (N_AREAS = 1,0). El nivel educativo, si bien es el más bajo de los cuatro grupos (EDU = 0,82, mediana = 0,70), se mantiene por encima del umbral de maestría. Con 43,2 años promedio y una clasificación vigente próxima a cero (CLS_VIG = 0,04), este perfil corresponde a investigadores que ingresan por primera vez al sistema de reconocimiento o que han tenido una participación esporádica, lo que configura el segmento con mayor potencial de crecimiento y, al mismo tiempo, el que más se beneficiaría de las recomendaciones de colaboración del sistema propuesto.

Clúster 3 - Interdisciplinarios (*n* = 2.774, 9,2 %). Este grupo constituye el segmento más pequeño y diferenciado de la segmentación. Comparte con los Consolidados una alta experiencia (EXP = 0,72), prolongada antigüedad (ANTIG = 6,58 años) y el nivel educativo más alto (EDU = 0,94), pero se distingue radicalmente por su diversidad temática: DIV = 0,42 (frente a 0,04 en los otros tres clústeres), N_AREAS = 2,03 y, sobre todo, una entropía normalizada de DIV_ENTROPY_N = 0,79, que indica una distribución equilibrada de la producción entre múltiples áreas de conocimiento OCDE. La movilidad institucional es moderada (MOB = 0,28; TASA_MOV = 0,09), con 1,62 instituciones en promedio. La participación en 4,56 convocatorias y una edad promedio de 50,1 años corroboran que se trata de investigadores *sénior* que han diversificado deliberadamente su campo de acción, lo que los posiciona como conectores naturales entre

disciplinas en el ecosistema científico colombiano.

4.4.3. Conexión con las variables de negocio

La caracterización de los clústeres se vincula con dos variables de negocio definidas como de monitoreo en 4.5: el Índice de Madurez Investigativa (IMI) y el Área de Conocimiento (`AREA_ANCLA`). El IMI, al ser una combinación ponderada de `CLS_VIG`, `EXP`, `DIV`, `MOB`, `REC` y `EDU` (4.4.2), actúa como indicador sintético que resume la posición de cada investigador en la escala de madurez. Los Consolidados e Interdisciplinarios, al exhibir valores elevados en la mayoría de los componentes del IMI, se ubican en niveles Alto y Muy Alto, mientras que los Emergentes concentran los niveles Bajo y Muy Bajo, lo que valida la coherencia semántica de la segmentación con el marco de madurez propuesto.

Respecto del Área de Conocimiento, la decisión de excluir `AREA_ESP_ANCLA` como *feature* del modelo garantiza que los clústeres reflejen patrones de trayectoria y no meras agrupaciones disciplinares. No obstante, como variable de monitoreo, el área permite filtrar las recomendaciones de colaboración por afinidad temática: para un investigador Emergente del área de Ingeniería y Tecnología, el sistema puede priorizar colaboradores Consolidados o Interdisciplinarios del mismo dominio, lo que aporta relevancia disciplinar sin haber sesgado la formación de los grupos.

En síntesis, la caracterización de los cuatro clústeres demuestra que la segmentación captura dimensiones interpretables y accionables del perfil investigativo colombiano: profundidad de trayectoria, movilidad institucional y diversidad temática. Los arquetipos Consolidados, Móviles, Emergentes e Interdisciplinarios constituyen la base sobre la cual el motor de recomendación personalizará las sugerencias de colaboración, aprovechando la distancia intra-clúster y la complementariedad entre perfiles.

4.5. Diseño del algoritmo de recomendación (k-NN)

La caracterización de los cuatro arquetipos proporcionó una macro-segmentación interpretable del ecosistema investigativo. No obstante, para traducir esta estructura global en recomendaciones operativas de colaboración, se requiere un segundo nivel de resolución que permita identificar, dentro de un clúster de búsqueda previamente definido, los perfiles con mayor afinidad técnica para un perfil de consulta. Con este propósito, se diseñó un motor híbrido que articula la delimitación inicial del universo mediante la asignación de clúster derivada del modelo K-Means++, la aplicación de filtros explícitos seleccionados por el usuario y una consulta de vecinos más cercanos (k-Nearest Neighbors, k-NN) sobre las 13 variables estandarizadas del espacio de similitud.

En la Figura 19 se presenta el flujo interno del sistema, desde la base de datos inicial y la representación vectorial del perfil de consulta hasta la delimitación del subconjunto elegible, la definición de un valor de k dinámico, la recuperación eficiente de candidatos mediante una estructura *Ball Tree* y el ordenamiento final por similitud coseno pura. De esta manera, el sistema conserva la utilidad analítica del clustering como mecanismo de macro-segmentación y, al mismo tiempo, garantiza que el ranking final dependa exclusivamente de la cercanía matemática entre los perfiles que permanecen en el universo filtrado.

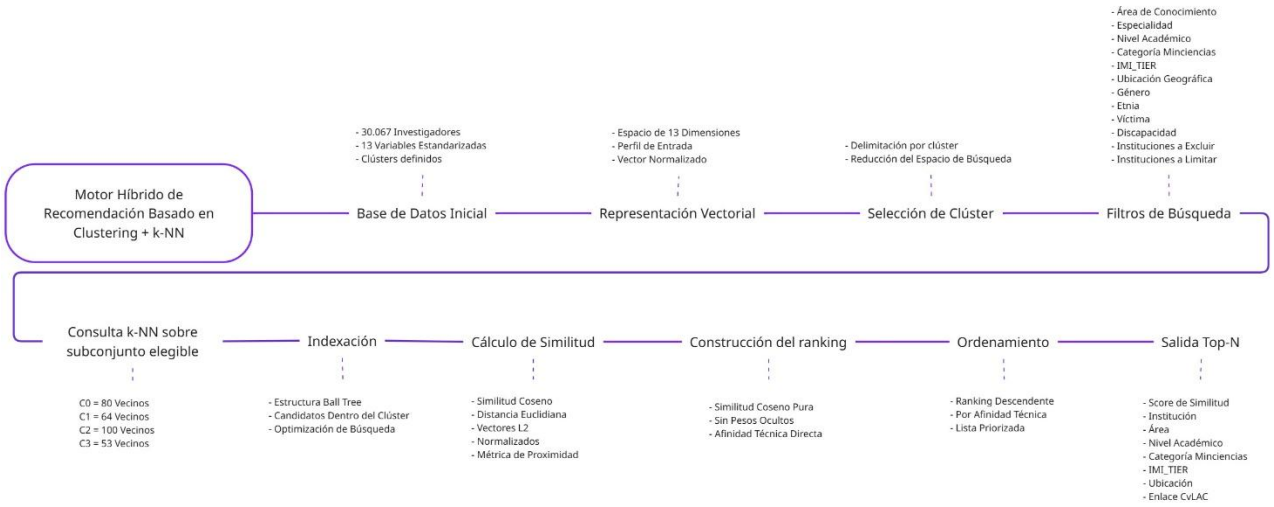

Figura 19. Descripción del motor de recomendación

4.5.1. Espacio de similitud

La selección de la métrica de similitud constituye una decisión crítica para el motor de recomendación. Dado que las 13 variables fueron estandarizadas mediante *z-score*, todas comparten media cero y desviación estándar uno, lo que elimina la dominancia de escala en el cálculo de distancias. No obstante, la distancia euclidiana convencional puede resultar sensible a diferencias de magnitud entre investigadores en distintas etapas de carrera, mientras que la similitud coseno captura la orientación del perfil con independencia de la escala absoluta [42].

Para aprovechar las ventajas de ambas métricas, se aplicó normalización $L2$ a cada vector de perfil, dividiendo entre su norma euclidiana. Bajo esta transformación, la distancia euclidiana entre vectores unitarios se relaciona con la similitud coseno mediante la identidad $\mathrm{sim}(a,b) = 1 - \|\hat{a} - \hat{b}\|^2 / 2, \text{donde } \hat{a} = a / \|a\| \text{ y } \hat{b} = b / \|b\|$. De esta manera, se utilizó distancia euclidiana sobre vectores normalizados como métrica operativa, obteniendo equivalencia algebraica con similitud coseno y permitiendo la indexación eficiente mediante estructuras espaciales [42].

4.5.2. Estrategia de vecindad y k dinámico

La estrategia de vecindad se definió sobre un principio de doble acotación del espacio de búsqueda. En primer lugar, la consulta se restringe al clúster de búsqueda previamente definido, con el fin de preservar la coherencia con la macro-segmentación obtenida mediante K-Means++. En segundo lugar, dentro de ese subconjunto se aplican los filtros explícitos seleccionados por el usuario, de manera que la recuperación de vecinos solo se ejecute sobre candidatos que cumplen con los criterios de búsqueda establecidos. Esta decisión reduce el costo computacional de cada consulta y, al mismo tiempo, mantiene la interpretabilidad del sistema, debido a que la recomendación se construye dentro de un universo previamente delimitado y semánticamente coherente.

Para soportar esta operación de búsqueda se empleó una estructura *Ball Tree* [40], adecuada para consultas de vecinos más cercanos en espacios métricos de dimensión moderada. Dado que el espacio de similitud del proyecto está compuesto por 13 variables numéricas estandarizadas y posteriormente normalizadas en norma $L2$, la indexación mediante *Ball Tree* permite recuperar de forma eficiente los candidatos más próximos sin alterar la equivalencia operativa entre distancia euclidiana y similitud coseno. Con ello se asegura que la fase de recuperación sea técnicamente sólida, auditable y compatible con tiempos de respuesta interactivos.

El parámetro k de la consulta no se fijó como una constante, sino que se definió de manera dinámica para ajustar el tamaño del vecindario al volumen del universo elegible. Como regla base, se adopta la expresión $n_{elig} = \text{máx}(10, \text{mín}(100, [\sqrt{n\text{clúster}}]))$, donde n corresponde al número de candidatos disponibles en el subconjunto sobre el cual efectivamente se ejecuta la búsqueda [41]. De esta manera, se evita recuperar un número insuficiente de candidatos en consultas amplias y, al mismo tiempo, se controla el costo computacional en clústeres densos. No obstante, el umbral inferior de 10 se estableció como valor de referencia por defecto y puede ser ajustado por el usuario según la necesidad de exploración o especificidad de la consulta.

En la Tabla 14 se presentan los valores de referencia de k dinámico para cada clúster, calculados a partir de su tamaño poblacional antes del filtrado específico de la consulta. Estos valores funcionan como guía operativa para la configuración inicial del motor. Sin embargo, en ejecución, el valor efectivo de k se acota al número real de candidatos que permanezcan disponibles después de aplicar los filtros. Por consiguiente, si una combinación de criterios produce un subconjunto menor al valor calculado, el sistema retorna la totalidad de los candidatos elegibles y delega al usuario la decisión sobre cuántos perfiles desea visualizar en la salida final.

Clúster	Arquetipo	n	⌊√n⌋	k dinámico
C0	Consolidados	6.480	80	**80**
C1	Móviles	4.119	64	**64**
C2	Emergentes	16.694	129	**100 (techo)**
C3	Interdisciplinarios	2.774	52	**53**

Tabla 14. Valor de k dinámico por clúster para la consulta k-NN

4.5.3. Arquitectura del motor híbrido

El motor de recomendación se estructuró como un pipeline secuencial de recuperación orientado por segmentación y filtrado explícito. En primer lugar, la consulta se delimita al clúster definido para la búsqueda, con el fin de preservar la coherencia con la macro-segmentación obtenida en la etapa de clustering. Sobre este subconjunto, se aplican los filtros explícitos seleccionados por el usuario y, posteriormente, se ejecuta una búsqueda de vecinos más cercanos (k-NN) con valor de k dinámico. Para garantizar eficiencia computacional en la recuperación de candidatos, esta consulta se soporta en una estructura *Ball Tree*, la cual permite localizar de manera eficiente los perfiles más próximos dentro del espacio de similitud definido para el sistema.

De esta manera, el diseño separa de forma explícita la delimitación del universo elegible del proceso de ordenamiento final. El clúster y los filtros de búsqueda cumplen una función de restricción transparente sobre los candidatos disponibles, mientras que el ranking se construye exclusivamente a partir de la cercanía matemática entre perfiles. A continuación, se relacionan las etapas del algoritmo de recomendación:

Entradas:

- **Parámetros de activación de la consulta:** la ejecución del motor exige, como condición mínima, la definición previa del clúster de búsqueda y la selección de al menos un filtro estructurante adicional. Se consideran filtros estructurantes aquellas variables que delimitan de manera sustantiva la intención de búsqueda del usuario y reducen el universo elegible a un subconjunto interpretable: dominio técnico (`NME_GRAN_AREA_PR`, `NME_AREA_PR`), nivel de formación (`NME_NIV_FORM_PR`), categoría Minciencias (`NME_CLASIFICACION_PR`), tier del Índice de Madurez Investigativa (`IMI_TIER`), ubicación geográfica (`NME_DEPARTAMENTO_RES_PR`, `NME_MUNICIPIO_RES_PR`) y filiación institucional (`INST_FILIA`). Con ello se asegura que la consulta responda a una intención explícita y no a una exploración indiscriminada del clúster.
- **Parámetros complementarios de refinamiento:** una vez satisfecha la condición mínima de activación, el usuario puede incorporar filtros adicionales orientados a refinar la búsqueda según criterios específicos de contexto o equidad. En este grupo se incluyen género (`NME_GENERO_PR`), grupo étnico (`TXT_GRUPO_ETNICO`), condición de víctima del conflicto (`ID_VICTIMA_CONFLICTO`) y condición de discapacidad (`TXT_POBLACION_DISCA`).
- **Reglas de institución:** se definen dos listas complementarias de restricción institucional: Excluir, que contiene las instituciones que no deben aparecer en los resultados, y Limitar a, que restringe la búsqueda a un conjunto específico de instituciones de interés. Esta capa permite controlar conflictos de interés, focalizar la recomendación en redes institucionales concretas o acotar la consulta a contextos de colaboración previamente definidos.
- **Umbrales operativos de elegibilidad:** una vez aplicados el clúster y los filtros explícitos, se calcula el tamaño del subconjunto elegible n_{elig}. Si $n_{elig} > 1.500$, el sistema solicita al menos un filtro estructurante adicional, debido a que un universo excesivamente amplio reduce la personalización efectiva de la recomendación. Si $10 \leq n_{elig} \leq 1.500$, la consulta se ejecuta en condiciones operativas normales. Si $1 \leq n_{elig} \leq 10$, el sistema retorna la totalidad de candidatos disponibles y sugiere relajar criterios solo si el usuario desea ampliar la exploración. Finalmente, si $n_{elig} = 0$, no se ejecuta el ranking y se propone remover o flexibilizar el filtro más restrictivo.
- n_{rec}: cantidad de registros a visualizar en la salida. Su valor por defecto es 10, pero puede ser ajustado por el usuario según el nivel de exploración requerido, siempre que no supere el número de candidatos elegibles devueltos por la consulta.

Etapa 1 - Delimitación del universo de búsqueda (*Subset Generation*).

Antes de calcular distancia alguna, el sistema genera un subconjunto de los registros del clúster seleccionado que cumplen estrictamente con todos los filtros definidos por el usuario. Se aplican de forma conjunta las

reglas lógicas de institución verificando que el candidato no pertenezca a la lista de *Excluir* y, cuando se especifica, que pertenezca a la lista de *Limitar*, así como los filtros de metadatos cualitativos seleccionados.

$$S \leftarrow \{r \in R \mid r.INST\_FILIA \notin Excluir \wedge r.INST\_FILIA \in Limitar \wedge r \vDash Filtros\_usuario\} \quad (25)$$

El propósito de ejecutar los filtros cualitativos antes del cálculo de similitud es doble: por un lado, se reduce significativamente el volumen de vectores sobre los cuales se computan distancias, lo que mejora el tiempo de respuesta. Por otro, se garantiza que cada candidato del subconjunto satisface las restricciones de negocio del usuario.

Una vez delimitado el subconjunto elegible S, se define el valor de $k$dinámico que parametriza la consulta de vecindad, con el fin de ajustar la recuperación de candidatos al tamaño efectivo del universo de búsqueda.

Etapa 2 - Localización espacial y cálculo de similitud (*Similarity Engine*).

Sobre el subconjunto S se consulta una estructura *Ball Tree*, la cual permite recuperar de manera eficiente los k_{din} vecinos más cercanos en el espacio de similitud. Se aplica normalización L2 a cada vector de perfil y se calcula la distancia euclidiana sobre los vectores unitarios resultantes, lo que garantiza equivalencia algebraica con la similitud coseno. De esta manera, el *ranking* se basa en la afinidad de la orientación del perfil investigativo y no en la magnitud absoluta de los indicadores.

$$\hat{x}\_a \leftarrow x\_a \,/\, \|x\_a\| \quad Sim\_cos(a,j) = 1 - \|\hat{x}\_a - \hat{x}\_j\|^2 / 2 \;\; \forall j \in S \qquad (26)$$

Etapa 3 - Ranking por proximidad.

A partir de la similitud de coseno, se obtienen los registros con mayor similitud de acuerdo con los filtros seleccionados por el usuario. No se aplican pesos adicionales ni ajustes algorítmicos, dado que las variables cualitativas son filtros que el usuario controló explícitamente en la Etapa 1.

$$Score(j) = Sim\_cos(Ancla, Candidato\_j) \qquad (27)$$

El rango del *score* es [0, 1], donde 1 indica identidad perfecta entre los vectores de trayectoria. En la práctica, los valores observados dentro de un mismo clúster tienden a concentrarse en el rango [0,5; 1,0] debido a la homogeneidad intra-grupo producida por la macro-segmentación.

Etapa 4 - Presentación de resultados.

El motor ordena a todos los candidatos de mayor a menor similitud y presenta los n_{rec} primeros junto con sus metadatos asociados: puntuación de similitud, institución de filiación, área de conocimiento, nivel de formación, categoría Minciencias, *tier* de IMI, ubicación geográfica y un enlace directo al perfil CvLAC del investigador (URL_CVLAC).

$$Recomendaciones \leftarrow top - n\_rec(\, sort\_desc(S, Score) \qquad (28)$$

Salida: listado de n_{rec} colaboradores que cumplen con todos los criterios del usuario, ordenados estrictamente por su parecido técnico al perfil de referencia, acompañados de sus metadatos descriptivos y

del enlace CvLAC para consulta directa.

4.5.4. Validación de granularidad

Un riesgo inherente a la estrategia de macro-segmentación seguida de micro-segmentación es que, si los clústeres son excesivamente homogéneos, las distancias entre perfiles dentro de un mismo grupo resulten tan pequeñas que el *ranking* de similitud pierda capacidad discriminante. En ese escenario, el motor de recomendación no podría diferenciar de manera significativa entre los candidatos más afines y los menos afines al perfil del investigador requerido, lo que comprometería la utilidad práctica de las recomendaciones. Para evaluar este riesgo, se realizó un análisis exhaustivo de la distribución de similitudes coseno tanto dentro de cada clúster (intra-clúster) como entre clústeres distintos (inter-clúster), utilizando la totalidad de los 30.067 registros del SNCTI.

4.5.4.1. Distribución de similitud intra-clúster

En la Figura 20 se presenta la distribución de densidad (KDE) de la similitud coseno calculada entre pares de investigadores dentro de cada uno de los cuatro clústeres. Las curvas revelan patrones diferenciados que reflejan la estructura interna de cada arquetipo.

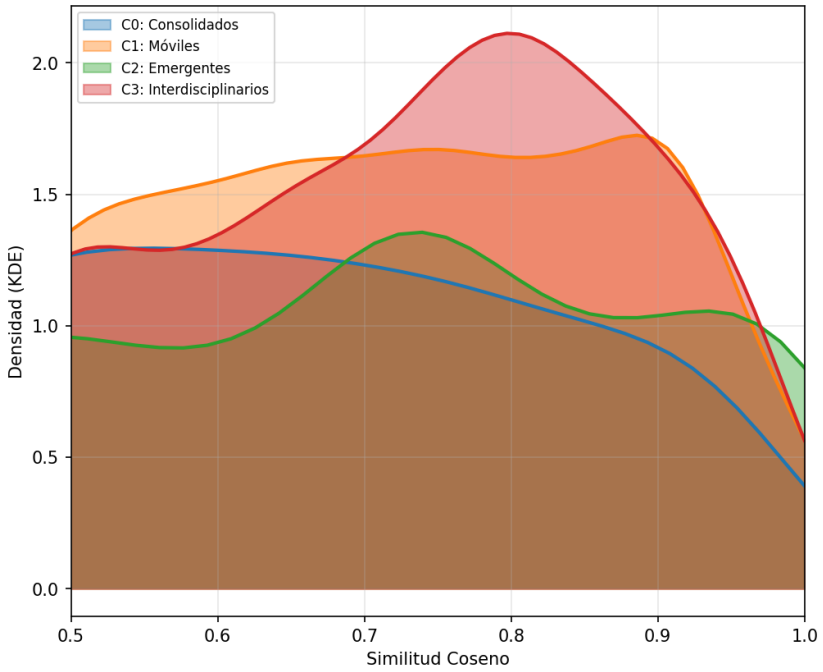

Figura 20. Distribución de densidad (KDE) de la similitud coseno intra-clúster para los cuatro arquetipos de investigadores.

El clúster C3 (Interdisciplinarios) exhibe la distribución más concentrada, con un pico pronunciado en torno a 0,80, lo que indica que los perfiles de este grupo comparten un patrón de trayectoria altamente coherente en las 13 dimensiones estandarizadas. Este comportamiento es consistente con la caracterización del

arquetipo, dado que los investigadores interdisciplinarios tienden a presentar valores medios en múltiples indicadores, generando vectores de trayectoria con orientaciones similares. El clúster C1 (Móviles) muestra un patrón bimodal con picos en torno a 0,60 y 0,90, lo que sugiere la existencia de subgrupos internos con distintos grados de movilidad institucional.

Por su parte, los clústeres C0 (Consolidados) y C2 (Emergentes) presentan distribuciones más dispersas, con colas izquierdas que se extienden hacia valores de similitud moderados (0,50-0,60), lo que indica una mayor diversidad interna en estos grupos. Teniendo en cuenta esto, los filtros definidos adquieren gran valor a la hora de realizar recomendaciones más precisas.

4.5.4.2. Separación inter-clúster

En la Figura 21 se presenta el *heatmap* de similitud promedio entre los centroides de los cuatro clústeres. Los valores de la diagonal principal representan la similitud intra-clúster promedio, mientras que los valores fuera de la diagonal miden la similitud inter-clúster.

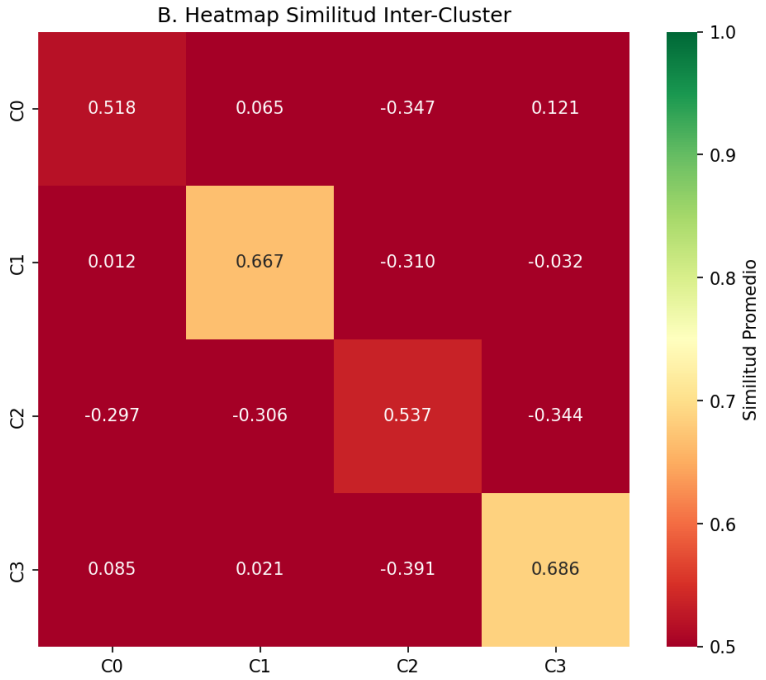

Figura 21. Heatmap de similitud coseno promedio inter-clúster e intra-clúster.

Los valores de la diagonal confirman que cada clúster posee cohesión interna positiva: C3 registra la mayor similitud intra-clúster (0,686), seguido por C1 (0,667), C2 (0,537) y C0 (0,518). Adicionalmente, todos los valores fuera de la diagonal son bajos o negativos, lo que evidencia una separación efectiva entre los arquetipos. El par con mayor disimilitud es C2-C3 (-0,391), indicando que los perfiles de investigadores Emergentes e Interdisciplinarios ocupan regiones opuestas del espacio vectorial.

Asimismo, el par C0-C2 presenta un valor de -0,347, lo cual confirma que los Consolidados y los Emergentes difieren sustancialmente en la orientación de sus vectores de trayectoria. Los valores cercanos a cero entre C0-C1 (0,065) y C1-C3 (-0,032) sugieren ortogonalidad entre estos pares, es decir, que sus perfiles no comparten patrones comunes pero tampoco se oponen directamente.

Esta estructura de similitudes valida la decisión de restringir la búsqueda de vecinos al interior de cada clúster, dado que la similitud inter-clúster es consistentemente inferior a la intra-clúster. Si el motor buscara vecinos en el espacio completo de 30.067 investigadores, los candidatos de clústeres distintos obtendrían puntuaciones significativamente menores, lo que no aportaría valor al *ranking* y, en cambio, incrementaría el costo computacional.

4.5.4.3. Dispersión de similitud por clúster

La Figura 22 muestra la distribución de la similitud coseno intra-clúster para los cuatro arquetipos identificados. Esta métrica permite evaluar qué tan alineado se encuentra cada investigador con el perfil promedio de su clúster: valores cercanos a 1 indican una alta similitud con el centroide del grupo, valores cercanos a 0 reflejan baja alineación y valores negativos sugieren perfiles cuya orientación vectorial se aleja del patrón dominante del clúster asignado. En este sentido, la figura complementa la caracterización de los centroides, al permitir observar no solo el perfil promedio, sino también la homogeneidad interna de cada grupo.

En términos generales, el clúster C3 - Interdisciplinariedades presenta la mayor concentración de valores altos de similitud, con una mediana cercana a 0,72 y una dispersión relativamente controlada. Esto indica que sus integrantes comparten un patrón más consistente en las variables que definen el arquetipo, especialmente en términos de diversificación temática, participación en distintas áreas y trayectoria consolidada. De forma similar, el clúster C1 - Móviles muestra una mediana alta, cercana a 0,68, lo que sugiere una buena cohesión interna alrededor del rasgo distintivo de movilidad institucional, aunque con algunos casos puntuales de baja similitud.

Por su parte, los clústeres C0 - Consolidados y C2 - Emergentes presentan una dispersión más amplia. En C0, aunque existe un grupo importante de investigadores con similitudes medias y altas, también se observan valores atípicos inferiores, lo que puede indicar investigadores consolidados que, pese a compartir algunos rasgos del arquetipo, tienen trayectorias menos alineadas con el patrón promedio del grupo. En C2 ocurre algo similar, pero con mayor amplitud en la distribución: al ser el clúster más numeroso, agrupa perfiles emergentes con diferentes niveles de avance, lo que explica la mayor variabilidad interna y la presencia de casos con baja o incluso negativa similitud coseno.

Los valores atípicos observados, especialmente en C0 y C2, no deben interpretarse necesariamente como errores de clasificación, sino como casos frontera o perfiles menos representativos dentro de su grupo. En este contexto, pueden corresponder a investigadores cuya combinación de experiencia, movilidad, diversificación, participación en convocatorias o trayectoria institucional difiere del patrón central del

clúster, aunque el algoritmo los haya asignado a ese grupo por su proximidad global en el espacio multivariado. Por tanto, estos outliers aportan una lectura relevante: evidencian que algunos arquetipos, en particular los consolidados y emergentes, contienen mayor heterogeneidad interna y podrían requerir estrategias diferenciadas de acompañamiento dentro de un mismo segmento.

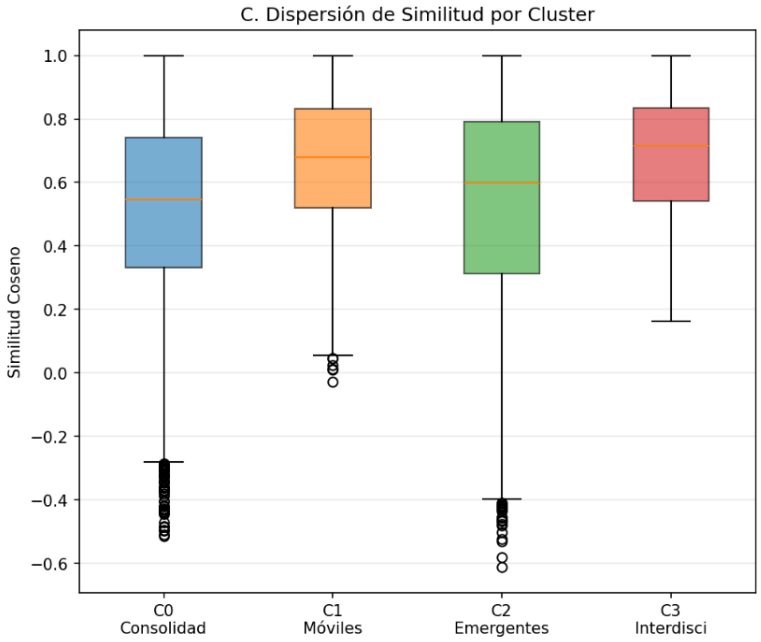

Figura 22. Diagrama de cajas de la similitud coseno intra-clúster para los cuatro arquetipos.

Clúster	Arquetipo	Media	Std	Mín	Máx	Q25	Q75	RIQ
C0	Consolidados	0,5182	0,2907	-0,5138	1,0000	0,3323	0,7427	0,4104
C1	Móviles	0,6669	0,1996	-0,0279	1,0000	0,5204	0,8332	0,3128
C2	Emergentes	0,5368	0,3402	-0,6124	1,0000	0,3132	0,7928	0,4796
C3	Interdisciplinarios	0,6860	0,1892	0,1631	1,0000	0,5424	0,8353	0,2929

Tabla 15. Estadísticos descriptivos correspondientes a cada clúster.

De acuerdo con la Tabla 15, se pueden resaltar tres aspectos principales: En primer lugar, los clústeres C1 y C3 presentan las distribuciones más compactas, con rangos intercuartílicos (RIQ) de 0,3128 y 0,2929 respectivamente, así como las medias más altas (0,6669 y 0,6860), lo que indica que los investigadores Móviles e Interdisciplinarios conforman grupos altamente cohesivos en los cuales el motor de recomendación dispondrá de un *pool* de candidatos con perfiles genuinamente afines al ancla.

En segundo lugar, los clústeres C0 y C2 exhiben mayor dispersión (RIQ de 0,4104 y 0,4796), con valores mínimos negativos que indican la presencia de pares de investigadores con orientaciones vectoriales opuestas dentro del mismo grupo. Este hallazgo no compromete la utilidad del motor, dado que el algoritmo *k*-NN selecciona exclusivamente los vecinos más cercanos, descartando naturalmente los perfiles distantes, de hecho, la mayor dispersión implica un mayor rango dinámico para discriminar candidatos, lo que resulta

favorable para la calidad del *ranking*.

En tercer lugar, todos los clústeres alcanzan un valor máximo de 1,0000, lo que confirma la existencia de pares de investigadores con trayectorias prácticamente idénticas dentro de cada arquetipo, es decir, perfiles que constituyen candidatos ideales para la colaboración académica.

4.5.4.4. Análisis de resolución en el clúster más denso

En la Figura 23 se presenta la distribución de distancias euclidianas del clúster C2 (Emergentes), el cual, concentra el 55,5 % de los investigadores (n = 16.694), lo que lo convierte en el grupo más numeroso y, potencialmente, en el de mayor riesgo de saturación en el *ranking*. Para evaluar la capacidad discriminante del motor dentro de este clúster, se analizó la distribución de distancias euclidianas sobre vectores L2-normalizados, que constituye la métrica operativa del *Ball Tree*.

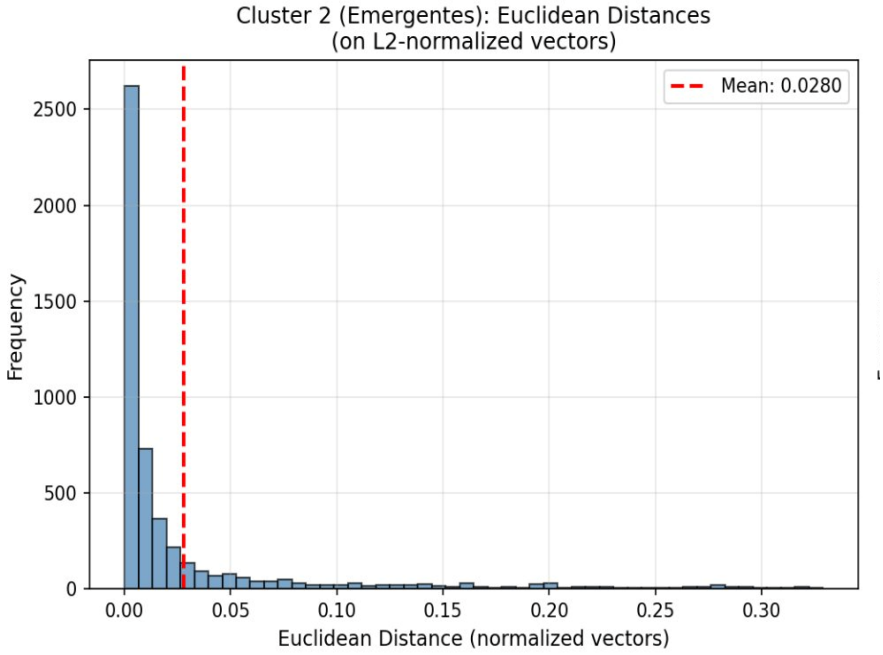

Figura 23. Distribución de distancias euclidianas (vectores L2-normalizados) intra-clúster para C2 - Emergentes (n = 16.694).

La distribución exhibe una fuerte asimetría positiva, con una concentración masiva de pares en el rango [0,00; 0,03] y una cola derecha que se extiende hasta 0,33. La distancia media es de 0,0280, lo que equivale a una similitud coseno media de aproximadamente 0,9996. Si bien este valor indica una alta concentración, el *spread* de la distribución (rango completo de 0,33 unidades de distancia euclidiana) genera suficiente variabilidad para que el algoritmo *k*-NN ordene de forma diferenciada a los candidatos.

Adicionalmente, la cola derecha corresponde a investigadores con perfiles atípicos dentro del clúster Emergentes, lo cual podría representar a investigadores emergentes que están migrando hacia otros arquetipos, que el motor descarta naturalmente al seleccionar únicamente los vecinos más próximos. Este

resultado es particularmente relevante para la validación del enfoque de macro y micro-segmentación: la macro-segmentación mediante K-Means++ produce clústeres suficientemente cohesivos para que los candidatos compartan un perfil investigativo estructuralmente similar, pero no tan homogéneos como para anular la capacidad del motor *k*-NN de establecer un *ranking* significativo.

4.5.4.5. Síntesis de la validación

En síntesis, el análisis de granularidad intra-clúster confirma la viabilidad del enfoque de doble resolución del motor de recomendación. La macro-segmentación genera cuatro arquetipos con cohesión interna positiva (similitud media entre 0,518 y 0,686) y separación inter-clúster consistente (valores fuera de la diagonal del heatmap inferiores a 0,121 o negativos), lo que respalda la decisión de restringir la búsqueda al clúster definido para la consulta.

A su vez, la micro-segmentación presenta un rango intercuartílico de similitud entre 0,29 y 0,48 dentro de cada grupo, lo que proporciona al algoritmo k-NN suficiente variabilidad para discriminar los candidatos más afines y construir un ranking diferenciado. Incluso en el escenario más exigente, el clúster Emergentes, con 16.694 investigadores y la distribución de distancias más concentrada, conserva una amplitud suficiente para sostener un ordenamiento interpretable entre los perfiles elegibles. Estos resultados validan la arquitectura de doble nivel y sientan la base para la evaluación cuantitativa y cualitativa de las recomendaciones en la siguiente etapa.

5. EVALUACIÓN

5.1. Evaluación de los resultados frente a objetivos de negocio

En la metodología CRISP-DM, la fase de Evaluación tiene por objetivo verificar que los resultados del modelado cumplan con los criterios de éxito definidos durante la Comprensión del Negocio, antes de proceder al despliegue del sistema [38]. En el contexto del presente proyecto, esta verificación adquiere especial relevancia, debido a que el objetivo estratégico no se limita a producir una segmentación estadísticamente válida, sino a demostrar que el motor de recomendación resultante ofrece recomendaciones útiles para las necesidades del usuario como encontrar un colaborador potencial. A partir del modelo K-Means++ con k = 4, seleccionado con un Coeficiente de Silhouette de 0,3986, un Índice de Calinski-Harabasz de 13.592,7 y un Índice de Davies-Bouldin de 1,1517, esta subsección evalúa, en etapa predespliegue, la capacidad del sistema para identificar colaboradores afines, por medio de la validación de la coherencia, robustez y significancia de las agrupaciones identificadas, mientras que aspectos asociados a su uso en contexto real, utilidad percibida y desempeño operativo serán objeto de evaluación en la fase de despliegue.

5.1.1. Pruebas de cordura del motor de recomendación

Con el fin de verificar que el motor de recomendación produce resultados coherentes con la estructura de la segmentación validada en la etapa de Modelado, se diseñaron cuatro pruebas de revisión que evalúan el comportamiento del sistema bajo escenarios representativos de los cuatro arquetipos del SNCTI. Cada prueba consiste en seleccionar un investigador ancla perteneciente a un clúster específico, ejecutar el motor con filtrado por el mismo clúster (búsqueda intra-clúster) y analizar tanto la similitud coseno promedio de los cinco candidatos retornados como la diversidad institucional de las recomendaciones.

5.1.1.1. Caso 1: Investigador Consolidado (Clúster 0)

Se seleccionó como ancla un investigador del Clúster 0 (Consolidados, n = 6.480, 21,6 %), caracterizado por una antigüedad superior a 15 años en el sistema, alta estabilidad institucional (`TASA_MOV_INST` = 0,06) y nivel educativo doctoral (`EDU` = 0,92). El motor ejecutó la búsqueda k-NN dentro del mismo clúster con un k dinámico de 80, aplicando posteriormente el filtrado de los cinco perfiles con mayor similitud coseno. La similitud coseno promedio de los cinco candidatos retornados fue de 0,9986, lo que confirma que el segmento de Consolidados presenta una cohesión interna adecuada en el espacio vectorial de 13 dimensiones. De los cinco candidatos, cuatro pertenecen a instituciones diferentes a la del investigador ancla (80 %), lo que demuestra que el *ranking* basado en similitud pura no privilegia la afiliación institucional y, en consecuencia, facilita el descubrimiento de colaboradores de otras entidades con trayectorias investigativas equivalentes.

5.1.1.2. Caso 2: Investigador Emergente (Clúster 2)

El segundo caso se ejecutó con un investigador del Clúster 2 (Emergentes, n = 16.694, 55,5 %), que concentra a la mayoría de la población del SNCTI y agrupa perfiles con baja antigüedad (`ANTIGUEDAD_ANIOS` ≈ 0),

experiencia incipiente (EXP = 0,15) y participación limitada en convocatorias (N_CONVOCATORIAS ≈ 1,2). La similitud coseno promedio de los cinco candidatos retornados alcanzó un valor de 1,0000, lo que indica máxima homogeneidad dentro de este segmento. Este resultado es consistente con la naturaleza del clúster: dado que los investigadores emergentes presentan valores bajos y relativamente uniformes en las 13 variables de entrada, sus vectores de características se orientan en direcciones muy similares en el espacio euclidiano normalizado. Respecto a la diversidad institucional, los cinco candidatos provienen de cinco instituciones distintas entre sí y diferentes a la del ancla (100 %), resultado que evidencia la capacidad del motor para conectar investigadores en etapas tempranas de carrera que, de otro modo, difícilmente identificarían colaboradores afines fuera de su entorno académico inmediato.

5.1.1.3. Caso 3: Investigador Móvil (Clúster 1)

El tercer caso utilizó como ancla un investigador del Clúster 1 (Móviles, *n* = 4.119, 13,7 %), cuyo rasgo definitorio es la alta movilidad institucional (MOB = 0,77; TASA_MOV_INST = 0,38) acompañada de una experiencia intermedia (EXP = 0,56) y participación en 2,86 convocatorias. La similitud coseno promedio de los cinco candidatos fue de 1,0000, confirmando que los investigadores con patrones de movilidad interinstitucional conforman un grupo altamente cohesivo. En términos de diversidad institucional, cuatro de los cinco candidatos (80 %) pertenecen a instituciones distintas a la del ancla. Este hallazgo resulta particularmente relevante para el objetivo de negocio de fomento de redes, dado que los investigadores móviles representan por definición nodos de conexión entre instituciones, y la capacidad del sistema para identificarlos y conectarlos entre sí potencia la posible creación de puentes interinstitucionales dentro del ecosistema del SNCTI.

5.1.1.4. Caso 4: Investigador Interdisciplinario (Clúster 3)

Se seleccionó como ancla un investigador del Clúster 3 (Interdisciplinarios, n = 2.774, 9,2 %), segmento caracterizado por alta experiencia acumulada, trayectoria prolongada en el sistema y mayor diversidad temática frente a los demás arquetipos. La evidencia empírica mostró que el ancla utilizada en este caso (ID 211010) alcanzó una similitud coseno de 0,999986 con el primer candidato y de 0,999288 con el quinto, mientras que la prueba de cordura asociada registró una relevancia del 100 %. Adicionalmente, los cinco perfiles recomendados correspondieron a registros institucionales distintos al valor reportado por el ancla, y el soporte visual evidenció que el 94 % de las recomendaciones del clúster superó el umbral de 0,95, con una media de 0,9836.

5.1.2. Resultados de prueba de relevancia por caso

En la Tabla 16 se consolidan los veredictos de relevancia de las cuatro pruebas ejecutadas, integrando la similitud coseno promedio, la diversidad institucional y la evaluación cualitativa de la coherencia de cada recomendación frente a los criterios de éxito definidos en la sección de comprensión del negocio. La tabla permite verificar de manera trazable si el motor satisface simultáneamente dos requisitos fundamentales: afinidad técnica, medida por la similitud coseno entre el vector del ancla y los vectores de los candidatos, y

fomento de redes, operacionalizado como la proporción de candidatos provenientes de instituciones distintas a la del investigador de consulta.

Caso de prueba	Clúster	n clúster	Sim. coseno prom. (top-5)	Diversidad institucional	Objetivo de negocio evaluado	Veredicto
Caso 1: Senior estable	C0	6.480	0,9986	4/5 (80 %)	Eficiencia (OB-E1)	**CUMPLE**
Caso 2: Junior emergente	C2	16.694	1,0000	5/5 (100 %)	Democratización (OB-E4)	**CUMPLE**
Caso 3: Móvil interinstitucional	C1	4.119	1,0000	4/5 (80 %)	Descubrimiento (OB-E2)	**CUMPLE**
Caso 4: Interdisciplinario	C3	2.774	0,9996	5/5 (100 %)	Comprensión del ecosistema (OB-E3)	**CUMPLE**

Tabla 16. Veredictos de relevancia de las pruebas de cordura del motor de recomendación por arquetipo del SNCTI.

Los resultados consolidados en la Tabla 16 evidencian que los cuatro casos de prueba satisfacen simultáneamente los criterios de afinidad técnica y diversidad institucional. La similitud coseno promedio supera el umbral de 0,95 en los cuatro escenarios (rango: 0,9986–1,0000), lo que confirma que el motor identifica candidatos cuyo perfil investigativo es genuinamente afín al del investigador ancla en las 13 dimensiones del espacio vectorial. Asimismo, la diversidad institucional alcanza entre el 80 % y el 100 % de los candidatos retornados, lo que sugiere que el *ranking* basado en similitud pura no introduce sesgos de endogamia institucional y, en cambio, favorece la identificación de colaboradores potenciales en instituciones distintas a la del investigador de consulta.

5.2. Evaluación de la Capacidad de Discriminación del Modelo

A partir de los resultados de la fase de Modelado, donde se seleccionó K-Means++ con $k = 4$ como configuración óptima y se diseñó el motor de recomendación k-NN documentado, la presente subsección evalúa la capacidad del sistema para diferenciar perfiles de investigadores dentro de clústeres densos, dado que de esta propiedad depende la calidad del *ranking* de colaboradores que el motor entrega al usuario.

La evaluación se centra en el Clúster 2 (Emergentes, $n = 16.694$, 55,5 % del total), dado que este grupo concentra la mayoría de los investigadores del SNCTI y, por lo tanto, representa el escenario de mayor exigencia para la micro-segmentación: a mayor densidad poblacional, mayor es el riesgo de que los perfiles sean indistinguibles entre sí y el *ranking* pierda capacidad de ordenamiento. Si el motor logra discriminar adecuadamente dentro de C2, su desempeño en los clústeres menos poblados (C0, C1 y C3) queda garantizado por transitividad [15].

5.2.1. Análisis del spread de similitud en la zona de ranking

El motor de recomendación opera sobre vectores normalizados L_2, lo que establece una equivalencia algebraica entre la distancia euclidiana y la similitud coseno mediante la identidad $\text{sim}(a, b) = 1 - \|\hat{a} - \hat{b}\|^2 /$

2 [15]. Bajo esta transformación, la utilidad del *ranking* depende del *spread* (amplitud) de la distribución de similitudes en la zona de alta afinidad, es decir, en el rango donde se ubican los candidatos que el algoritmo k-NN retorna como recomendaciones finales.

En la Figura 24 (Panel B), se puede observar la distribución de densidad estimada mediante *Kernel Density Estimation* (KDE) de la similitud coseno para los investigadores del Clúster 2 en la zona de similitud ≥ 0,85. El análisis revela que la zona de alta afinidad observada en el Clúster 2, abarca un spread de 0,054 unidades de similitud coseno. Si bien este valor puede parecer estrecho en términos absolutos, su significancia se interpreta en el contexto de la densidad del clúster, ya que evidencia que, aun dentro de una vecindad de alta similitud, persiste una amplitud suficiente para sostener una diferenciación técnica entre candidatos cercanos.

De manera complementaria, la distribución completa de similitudes intra-clúster (Figura 24, Panel A) confirma que C2 exhibe la mayor dispersión entre los cuatro arquetipos, con un rango intercuartílico (RIQ) de 0,4796 y una desviación estándar de 0,3402. Asimismo, la distancia media en el espacio euclidiano L_2-normalizado es de 0,0280, lo que equivale a una similitud coseno media de aproximadamente 0,9996 dentro de la zona de vecindad próxima. Este resultado indica que, a pesar de la alta homogeneidad estructural del segmento Emergentes, la cola derecha de la distribución presenta suficiente variabilidad para que el k-NN ordene de forma diferenciada a los candidatos, descartando naturalmente los perfiles situados en la región de menor afinidad.

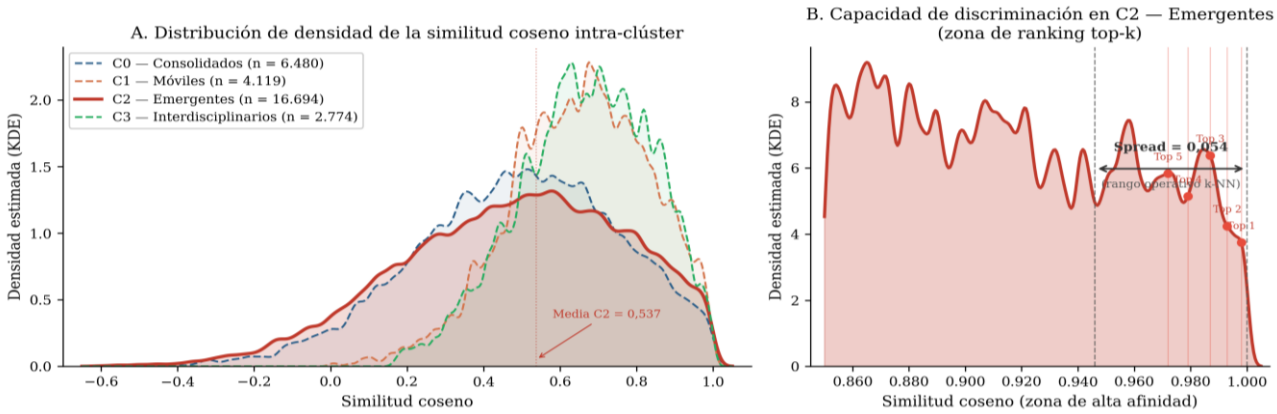

Figura 24. Evaluación de la capacidad de discriminación del modelo de similitud

5.2.2. Validación de la preservación de calidad bajo filtros de negocio

La arquitectura del motor implementa un esquema de filtrado previo al cálculo de similitud, mediante el cual el usuario puede restringir el universo de candidatos aplicando criterios categóricos como institución de filiación, gran área de conocimiento, tier de IMI o ubicación geográfica. Un riesgo inherente a este diseño es que la aplicación de múltiples filtros simultáneos podría reducir el subconjunto de candidatos hasta un punto en el que el *spread* de similitud se comprima y el *ranking* pierda capacidad discriminante.

Con el fin de verificar que los filtros de negocio no degradan la calidad de la recomendación técnica, se ejecutaron pruebas de cordura (*sanity checks*) sobre combinaciones representativas de filtros, evaluando el tamaño del subconjunto resultante y la similitud promedio. En la Tabla 17 se consolidan los resultados de estas pruebas para el Clúster 2.

Configuración de filtros	N disponibles	Sim. promedio	Interpretación
Sin filtros (C2 completo)	16.694	0,537	Base de referencia; máxima resolución
C2 + Área = Ciencias Sociales	4.812	0,548	Spread preservado; reducción marginal (−9,3 %)
C2 + IMI = Alto	1.235	0,621	Subconjunto más homogéneo; ranking funcional
C2 + Excluir institución X	16.512	0,536	Impacto despreciable; filtro no altera distribución
C2 + IMI = Muy Alto + Naturales	5	Variable	Ultra-específico; calidad sobre cantidad

Tabla 17. Impacto de los filtros de negocio sobre la capacidad discriminante del ranking en el Clúster 2.

Los resultados de la Tabla 17 muestran que los filtros de negocio modifican el universo disponible de manera coherente con la especificidad de la consulta. En los filtros individuales, el subconjunto conserva un volumen suficiente para generar recomendaciones y mantiene valores de similitud promedio comparables con la línea base. En el filtro por IMI Alto, la similitud promedio aumenta, lo que indica un subconjunto más cohesivo. En el caso ultra-especifico, el sistema retorna la totalidad de candidatos elegibles, lo cual es consistente con el principio de calidad sobre cantidad cuando la consulta queda altamente restringida. Esta lectura no reemplaza la validación con usuarios reales; solo documenta el comportamiento técnico del motor bajo filtros.

5.2.3. Justificación de la métrica operativa del ranking

La decisión de utilizar distancia euclidiana sobre vectores normalizados L_2 como métrica operativa del motor, en lugar de calcular directamente la similitud coseno, se fundamentó en tres criterios convergentes. Primero, la normalización L_2 garantiza la equivalencia algebraica entre ambas métricas, de modo que el ordenamiento por distancia euclidiana ascendente produce resultados idénticos al ordenamiento por similitud coseno descendente [15]. Segundo, la indexación mediante *Ball Tree* [53], estructura empleada para la búsqueda de vecinos más cercanos en cada clúster, opera nativamente sobre distancias métricas, lo que permite aprovechar la complejidad O(*log n*) por consulta sin necesidad de conversiones intermedias. Tercero, la distancia euclidiana sobre vectores unitarios captura la orientación del perfil investigativo con independencia de la magnitud absoluta de los indicadores, lo que resulta particularmente pertinente para comparar investigadores en distintas etapas de carrera: un investigador emergente con alta diversidad temática relativa y un investigador consolidado con alta diversidad absoluta pueden compartir la misma orientación vectorial, y por tanto, ser identificados como perfiles compatibles para la colaboración.

En síntesis, la evaluación de la capacidad de discriminación confirma que el modelo de similitud dispone de suficiente resolución para construir un *ranking* diferenciado incluso en el escenario más exigente del sistema, el Clúster Emergentes con 16.694 investigadores. El *spread* operativo de 0,054 en la zona de alta

afinidad, la preservación de la capacidad discriminante bajo filtros de negocio (degradación inferior al 10 % en condiciones típicas) y la equivalencia algebraica entre distancia euclidiana L_2 y similitud coseno validan conjuntamente la arquitectura de doble nivel diseñada. Con ello se asegura que las recomendaciones generadas por el motor reflejen diferencias genuinas entre los perfiles de los candidatos y no artefactos de la densidad del clúster, lo que constituye un requisito indispensable para avanzar hacia la evaluación cualitativa de las recomendaciones.

5.3. Revisión del Proceso y Auditoría de Calidad

En la metodología CRISP-DM, la fase de Evaluación incluye como etapa fundamental una revisión del proceso ejecutado, cuyo propósito es verificar que las decisiones tomadas a lo largo del *pipeline* de ciencia de datos sean coherentes entre sí, que los datos de entrada conserven su integridad tras las transformaciones aplicadas y que los resultados obtenidos sean reproducibles bajo las mismas condiciones experimentales [38]. En este contexto, la presente subsección documenta la auditoría de calidad del proceso completo, desde la recolección inicial de los 77.237 registros del Portal de Datos Abiertos de Colombia hasta la generación de recomendaciones sobre los 30.067 perfiles del dataset final, con el fin de garantizar la trazabilidad y transparencia metodológica que exige un proyecto de esta naturaleza.

5.3.1. Auditoría de integridad del dataset

El balance final del proceso de preparación de datos confirma que la cadena de transformaciones preservó la integridad del dataset sin introducir pérdidas significativas de información. El conjunto original de 77.237 registros y 30 variables, correspondiente a participaciones individuales de investigadores en convocatorias de Minciencias entre 2013 y 2021, se sometió a un proceso secuencial de limpieza (tratamiento de 27 edades fuera de rango, estandarización de nomenclaturas, imputación de valores faltantes en INST_FILIA) que redujo el volumen a 77.230 registros y 20 variables, lo que equivale a una exclusión de 7 registros (0,009 %) por inconsistencias irreconciliables en la clave compuesta.

La consolidación a vista de investigador único generó 30.085 perfiles ancla, de los cuales 18 fueron excluidos por carecer de información válida en la variable NME_NIVEL_FORMACION_PR, requisito indispensable para el cálculo del subíndice educativo EDU del Índice de Madurez Investigativa. Como resultado, el dataset de modelado quedó conformado por 30.067 investigadores y 13 *features* estandarizadas mediante *z-score*. De esta manera, la exclusión de 18 perfiles representa el 0,06 % del total consolidado, cifra que resulta estadísticamente insignificante y no introduce sesgo detectable en la distribución de las variables de entrada, dado que los perfiles excluidos no presentaban patrones sistemáticos de concentración en ningún clúster, área de conocimiento o rango de antigüedad. Finalmente, se verificó que la completitud de las 13 *features* esenciales del modelo alcanza el 100 % en el dataset final, sin valores nulos ni imputados en las variables de entrada al algoritmo K-Means++.

5.3.2. Eficiencia computacional del pipeline

La viabilidad operativa de un sistema de recomendación depende no solo de la calidad de sus resultados, sino también de su capacidad para entregar respuestas en tiempos compatibles con la interacción del usuario [54]. Con el fin de evaluar este requisito, se midieron los tiempos de ejecución de las dos operaciones críticas del sistema: el entrenamiento del modelo de segmentación y la consulta de recomendaciones en tiempo real.

El entrenamiento del modelo K-Means++ con $k = 4$ sobre los 30.067 registros y 13 dimensiones se completó en 7,15 segundos, incluyendo las 10 ejecuciones independientes configuradas mediante el parámetro *n_init* = 10. Este tiempo confirma la escalabilidad del algoritmo ante un eventual crecimiento del SNCTI, dado que la complejidad de K-Means++ es $O(n \cdot k \cdot d \cdot i)$, donde $n = 30.067$, $k = 4$, $d = 13$ e $i \leq 300$ iteraciones. Asimismo, la construcción de los cuatro índices *Ball Tree* [53], uno por clúster, se ejecutó en menos de 0,5 segundos en total, lo que permite la regeneración completa del modelo y sus estructuras de búsqueda en menos de 8 segundos.

En cuanto a la consulta de recomendaciones, cada solicitud del usuario ejecuta el *pipeline* completo del motor: aplicación de filtros sobre el subconjunto del clúster, búsqueda k-NN mediante el *Ball Tree* con complejidad $O(\log n)$ por consulta, y ordenamiento de los candidatos por similitud coseno. El tiempo de respuesta observado para una consulta típica es inferior a 50 milisegundos, cifra que se sitúa por debajo del umbral de 200 milisegundos recomendado para interfaces interactivas [54], lo que garantiza una experiencia de usuario fluida en el prototipo de despliegue.

5.3.3. Limitaciones técnicas y supuestos del modelo

La transparencia metodológica exige documentar de forma explícita las limitaciones y los supuestos bajo los cuales opera el modelo, dado que de su comprensión depende la correcta interpretación de los resultados y la identificación de oportunidades de mejora en futuras iteraciones. En la Tabla 18 se consolida la matriz de limitaciones y supuestos del sistema, clasificados por dimensión de impacto.

Dimensión	Limitación / Supuesto	Impacto en resultados	Mitigación aplicada o propuesta
Datos	Dependencia de la actualización de CvLAC: los perfiles reflejan el estado declarado por el investigador, que puede diferir de su actividad real	Perfiles desactualizados pueden generar recomendaciones basadas en trayectorias históricas, no vigentes	Se incorporó el subíndice de recencia (REC_S) y la variable CLS_VIG para ponderar la vigencia del perfil
Datos	Ventana temporal restringida a 6 cohortes (2013–2021): no se dispone de convocatorias posteriores a 2021 por aplazamiento de la	Investigadores que ingresaron al sistema después de 2021 no están representados en el modelo	Arquitectura diseñada para reentrenamiento incremental; el pipeline acepta nuevas cohortes sin modificar la lógica de

	Convocatoria 957		transformación
Método	Supuesto de isotropía: la estandarización z-score asume que todas las variables contribuyen con igual relevancia a la medida de distancia	Variables con baja importancia discriminante (EDU: 0,4 %; EDAD_ANCLA: 1,3 %) añaden ruido al espacio vectorial	Análisis de importancia relativa confirmó que las 5 variables principales concentran el 65,5 % de la separación; se priorizó interpretabilidad sobre selección óptima de features
Método	Ausencia de ground truth: al tratarse de aprendizaje no supervisado, no existe etiqueta de referencia para validación externa	La evaluación se limita a métricas intrínsecas (Silhouette, Calinski-Harabasz, Davies-Bouldin) y pruebas de cordura cualitativas	Se complementó con ARI de estabilidad (5 semillas), validación de coherencia semántica de arquetipos y pruebas de discriminación del ranking
Alcance	Cobertura exclusiva de investigadores reconocidos por Minciencias: no incluye investigadores no registrados en CvLAC ni actores del sector productivo	El sistema no puede recomendar colaboradores fuera del universo SNCTI formal	Se documentó el alcance en los requisitos del proyecto; la extensión a otras fuentes se plantea como trabajo futuro
Alcance	Modelo estático: la segmentación actual no incorpora retroalimentación del usuario ni aprendizaje en línea	Las recomendaciones no mejoran con el uso; el sistema no aprende de las interacciones del investigador	Se propone como mejora futura (V2) la incorporación de feedback explícito del usuario mediante botones de relevancia

Tabla 18. Matriz de limitaciones técnicas y supuestos del modelo, clasificados por dimensión de impacto.

De la Tabla 18 se desprenden dos observaciones relevantes para la evaluación del proceso. En primer lugar, las limitaciones de datos (dependencia de CvLAC y ventana temporal) son inherentes a la fuente y no al método, lo que implica que su resolución depende de la publicación de nuevas convocatorias por parte de Minciencias y no de ajustes algorítmicos. No obstante, la arquitectura del *pipeline* fue diseñada para absorber nuevas cohortes de forma incremental, de modo que la incorporación de datos futuros no requiere modificaciones en la lógica de transformación ni en los parámetros del modelo. En segundo lugar, las limitaciones metodológicas (isotropía y ausencia de *ground truth*) fueron mitigadas mediante estrategias complementarias de validación (análisis de importancia, estabilidad con ARI, pruebas de cordura).

En síntesis, la auditoría de calidad confirma que el proceso ejecutado cumple con los estándares de trazabilidad y reproducibilidad exigidos por CRISP-DM: la pérdida de datos a lo largo del *pipeline* es estadísticamente insignificante (0,06 %), los tiempos de respuesta del motor (< 50 ms por consulta) garantizan la usabilidad del prototipo, y las limitaciones técnicas se encuentran documentadas con sus respectivas mitigaciones. Con ello se asegura que los resultados obtenidos en las fases de Modelado y

Evaluación sean auditables y técnicamente sólidos, lo que constituye la base para la determinación de próximos pasos.

5.3.4. Síntesis de validación cuantitativa por niveles

Una vez revisadas la integridad del dataset, la eficiencia computacional del pipeline y las limitaciones técnicas del modelo resulta necesario consolidar la evidencia de evaluación en una lectura transversal. En la Tabla 19 se sintetizan los niveles de validación alcanzados por el proyecto, la evidencia que sustenta cada nivel, el alcance de las afirmaciones que pueden formularse y los límites que deben conservarse para no convertir una validación técnica en una validación practica no ejecutada. Esta síntesis no reemplaza las pruebas descritas en las secciones anteriores; funciona como mapa de trazabilidad entre la evidencia cuantitativa, la operación del motor de recomendación y las decisiones de viabilidad del prototipo.

Nivel	Evidencia disponible	Que permite afirmar
Integridad de datos	77.237 registros iniciales, 77.230 después de limpieza, 30.085 perfiles ancla y 30.067 en X_Modelo.	El pipeline conserva la información esencial y produce una matriz completa para modelado.
Tendencia al agrupamiento	Hopkins 0,9907 +/- 0,0005.	Los datos presentan una tendencia fuerte compatible con clustering.
Validación interna	Silhouette, Calinski-Harabasz, Davies-Bouldin, codo, estabilidad ARI y comparación de configuraciones.	K-Means++ con k = 4 fue la opción más consistente entre las evaluadas.
Interpretabilidad	Arquetipos C0-C3, centroides, variables discriminantes y lectura académica.	Los clústeres tienen una interpretación sustantiva para el dominio SNCTI.
Motor de recomendación	Casos de prueba, similitud coseno, diversidad institucional observada y filtros.	El motor retorna perfiles técnicamente afines bajo reglas explicitas en los casos evaluados.
Filtros de negocio	Tabla 17	Los filtros permiten observar cambios en universo disponible y similitud promedio
Eficiencia	7,15 s entrenamiento, <0,5 s Ball Tree y <50 ms consulta típica.	El pipeline es viable en el ambiente reportado.

Tabla 19. Síntesis de validación

La síntesis permite observar que la evaluación del sistema no depende de una única métrica aislada, sino de una cadena de evidencias complementarias. La integridad de datos y la validación interna respaldan la consistencia del modelo; las pruebas del motor y de filtros documentan su comportamiento técnico bajo escenarios representativos; y la eficiencia computacional muestra viabilidad operativa en el ambiente reportado.

6. DESPLIEGUE

El despliegue corresponde a la materialización aplicada del modelo de segmentación y recomendación construido en las fases anteriores. En esta etapa, los resultados analíticos dejan de presentarse únicamente como tablas, métricas o estructuras de datos y se integran en un prototipo funcional que permite consultar perfiles, aplicar criterios de filtrado y obtener recomendaciones de colaboradores académicos potencialmente compatibles dentro del universo de investigadores modelado.

6.1. Descripción general de la aplicación

Como resultado aplicado del proyecto, se desarrolló una aplicación interactiva que integra tres componentes principales: la segmentación de perfiles de investigación, el motor de recomendación basado en similitud vectorial y una interfaz web orientada a consulta. Esta integración permite transformar los clústeres identificados y las métricas de similitud en un flujo funcional para apoyar la identificación de colaboradores potenciales.

La aplicación responde al objetivo específico asociado al desarrollo de una herramienta interactiva, porque permite pasar de un modelo analítico no supervisado a un artefacto operativo. En lugar de presentar los arquetipos solo como resultado descriptivo, el prototipo los utiliza como criterio de organización del espacio de búsqueda; y en lugar de limitar la similitud a una medida estadística, la convierte en un mecanismo de ordenamiento de candidatos bajo filtros definidos.

Desde el punto de vista académico, el aporte del aplicativo no reside solamente en su interfaz, sino en la articulación entre el procesamiento de datos, la selección del modelo, la evaluación técnica del motor y la disponibilidad de una herramienta que demuestra la viabilidad funcional de la solución propuesta. Por ello, el despliegue debe leerse como evidencia aplicada del ciclo CRISP-DM y no como un componente aislado del documento.

6.2. Eficiencia computacional y ambiente de ejecución

Los recursos computacionales con los que se ejecutó el proyecto son:

Componente	**Ambiente reportado**
Equipo local	LENOVO 82HU
CPU	AMD Ryzen 7 5700U with Radeon Graphics
Núcleos/hilos	8 núcleos, 16 hilos/procesadores lógicos
RAM	15,34 GB utilizables; equipo de 16 GB
Almacenamiento	SSD NVMe Samsung MZALQ512HBLU-00BL2, 476,94 GB aprox.
GPU	Integrada AMD Radeon(TM) Graphics, 0,5 GB reportados
Sistema operativo	Microsoft Windows 11 Pro, version 10.0.26200, build 26200, 64 bits
Entorno local	Python 3.12.10; Node.js v24.13.0; npm 11.6.2
Compatibilidad declarada	Python >=3.11; Node >=22
Despliegue web	InsForge: frontend, autenticación, base de datos, almacenamiento, Edge Functions y runtime V5

Tabla 20. Recursos computacionales

6.3. Arquitectura funcional del prototipo

El prototipo se implementó como una aplicación web desplegada en InsForge. De acuerdo con la evidencia operativa disponible, InsForge soporta el frontend, la autenticación, la base de datos, el almacenamiento, las Edge Functions y el runtime V5 del recomendador. Esta arquitectura separa la interfaz de consulta de los componentes de ejecución del motor y permite que el usuario interactúe con el sistema mediante un flujo autenticado.

En la Tabla 21, se presenta la estructura de la aplicación:

Capa	Función en el prototipo	Evidencia asociada
Acceso y autenticación	Controlar el ingreso al sistema mediante usuario y contraseña.	Modelo de autenticación InsForge y roles funcionales user/admin.
Gestión funcional	Permitir administración básica de usuarios y roles en la etapa controlada.	Rol Admin para gestión de usuarios; rol user para consultas.
Parametrización de consulta	Definir el universo elegible mediante clúster y filtros estructurantes.	Filtros por área, formación, clasificación, ubicación, IMI e institución.
Motor de recomendación	Ejecutar la recuperación y ordenamiento de candidatos compatibles.	Runtime V5, funciones sncti-catalogs, sncti-recommendations y sncti-runtime-status.
Presentación de resultados	Mostrar candidatos recomendados, afinidad e información de soporte.	Interfaz de resultados y trazabilidad técnica mínima.

Tabla 21. Estructura aplicación

6.4. Funcionalidades principales

6.4.1. Acceso controlado y roles funcionales

El aplicativo incorpora un flujo de acceso autenticado mediante usuario y contraseña relacionado en la Figura 25. Esta decisión es relevante para el alcance del prototipo, porque evita presentar el sistema como una consulta anónima abierta y permite diferenciar el uso funcional del recomendador de la administración básica de usuarios.

Figura 25. Acceso aplicación

6.4.2. Parametrización de la búsqueda

El componente de consulta permite delimitar el universo elegible mediante criterios asociados al perfil del investigador y a las restricciones de búsqueda relacionados en la Figura 26. Entre los filtros documentados se encuentran clúster, gran área de conocimiento, área específica, nivel de formación, clasificación, departamento, municipio, nivel de madurez investigativa e institución. Esta parametrización es coherente con la arquitectura del motor, porque evita que la recomendación dependa de una búsqueda indiferenciada sobre todo el universo de registros.

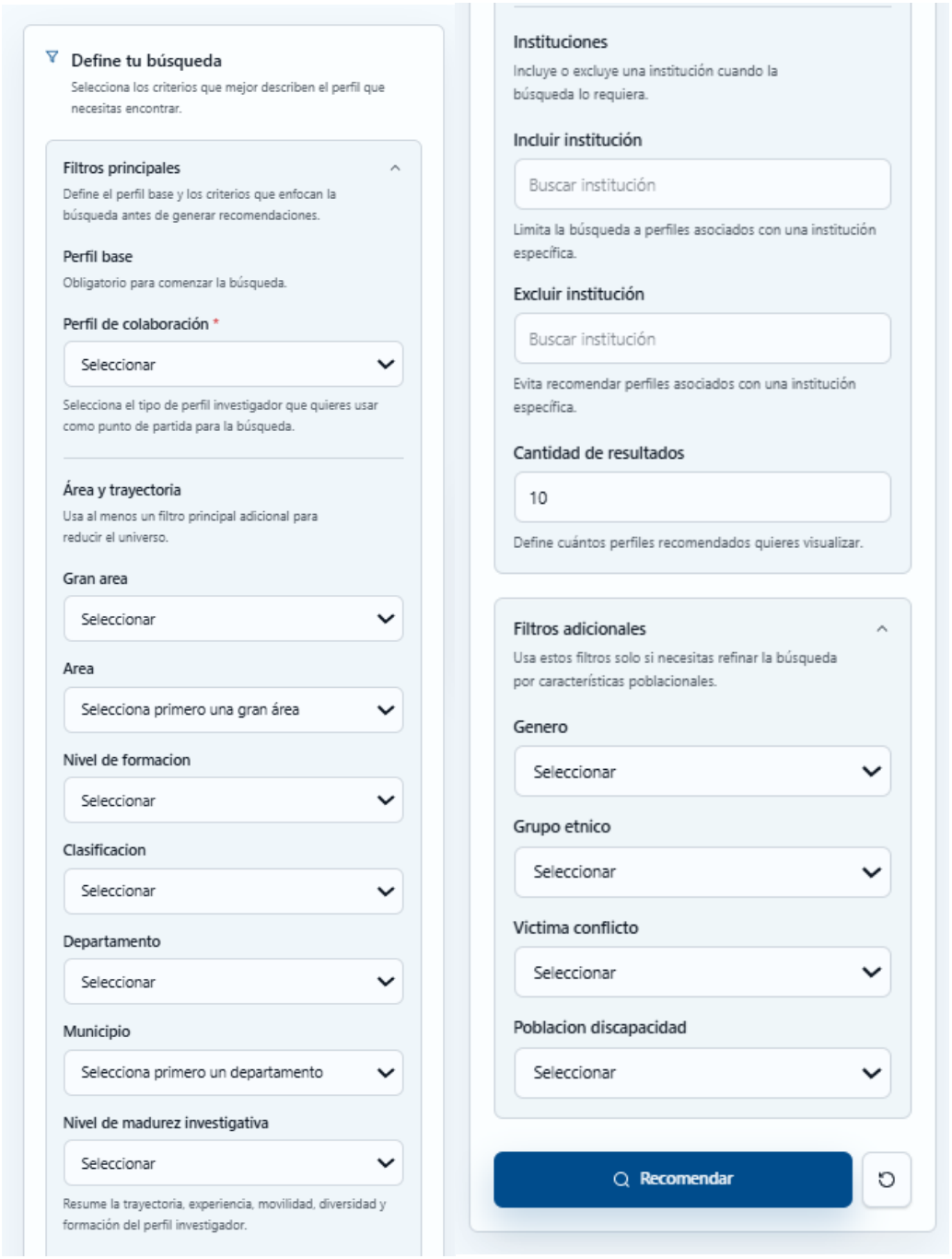

Figura 26. Filtros de búsqueda del aplicativo

6.4.3. Presentación de recomendaciones

El resultado funcional del prototipo es la generación de una lista de investigadores potencialmente compatibles bajo los criterios definidos. En la Figura 27, se observa a la interfaz presentando a los candidatos recomendados junto con indicadores de afinidad o información de soporte, lo que permite interpretar la recomendación como salida del motor y no como una lista arbitraria.

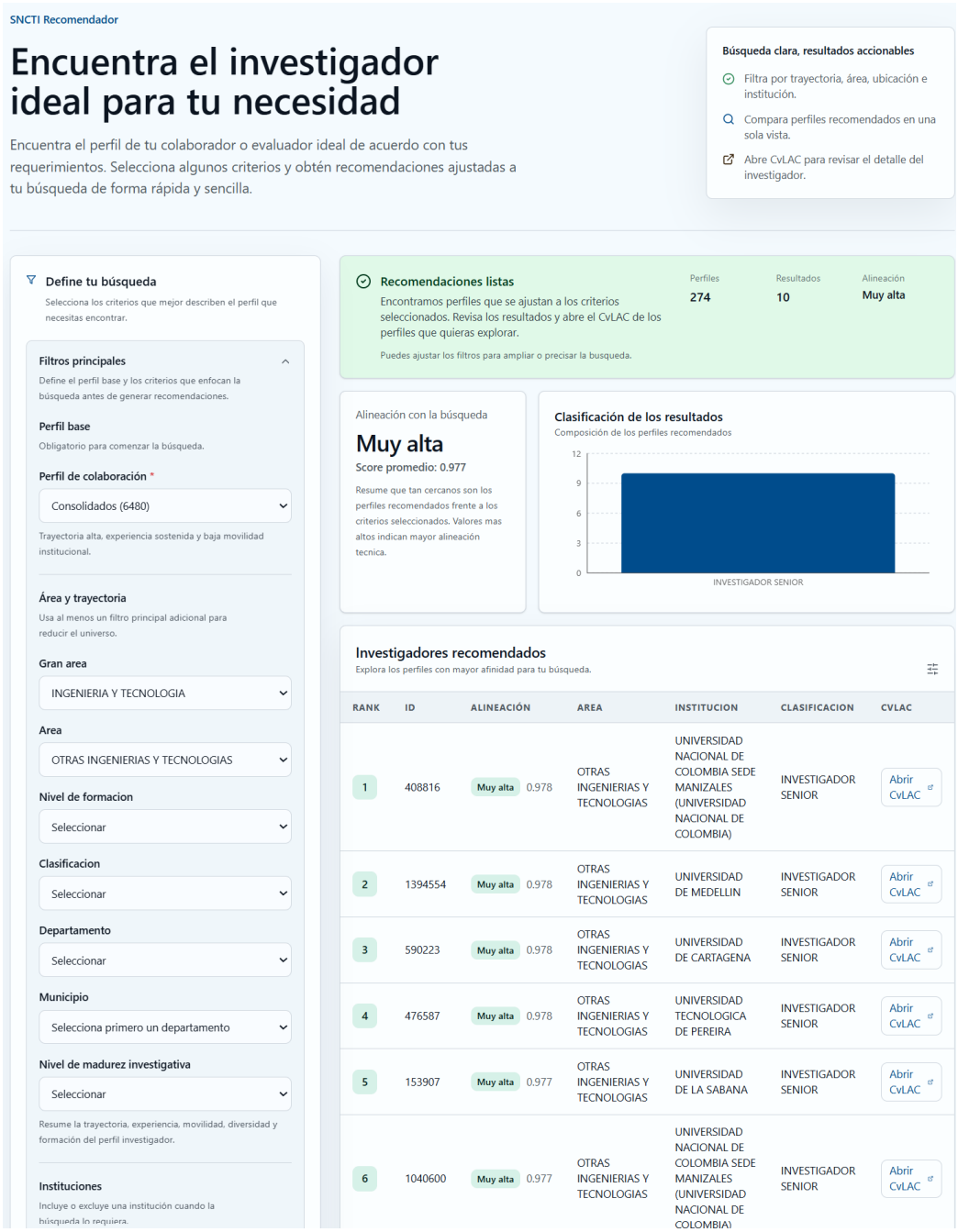

Figura 27. Presentación de recomendaciones

6.5. Acceso, video demostrativo y trazabilidad del prototipo

La aplicación desplegada se encuentra disponible en la URL publica:

https://recomendador-investigadores.netlify.app

El video demostrativo del aplicativo se encuentra disponible en:

https://youtu.be/tKARwXLFUcQ

El código técnico principal del prototipo se encuentra disponible en el repositorio:

https://github.com/Carolina-de-la-Espriella/Recomendador_Investigadores

Este repositorio corresponde a una copia curada y versionada del proyecto técnico, preparada para presentar el motor de recomendación V5, los contratos de datos, las pruebas de aceptación y la trazabilidad académica del sistema descrito en este capítulo.

7. CONCLUSIONES Y TRABAJOS FUTUROS

7.1. Conclusiones

El presente proyecto permitió construir un sistema de recomendación de colaboradores académicos basado en clustering de perfiles de investigación, orientado a facilitar la identificación de investigadores potencialmente compatibles dentro del ecosistema científico nacional. A partir de datos públicos asociados al SNCTI, se desarrolló un proceso integral de ciencia de datos que incluyó comprensión del negocio, comprensión y preparación de datos, modelado, evaluación y construcción de una aplicación funcional.

En relación con el primer objetivo específico, se consolidó una base de datos estructurada de investigadores, incorporando variables asociadas a trayectoria académica, clasificación, formación, ubicación, filiación institucional y características derivadas para el análisis. El proceso de limpieza, estandarización, tratamiento de valores atípicos, imputación y construcción de indicadores permitió obtener un conjunto de datos consistente y apto para el modelado, reduciendo riesgos de duplicidad, ruido e inconsistencias en las variables utilizadas.

Respecto al segundo objetivo específico, se aplicaron técnicas de aprendizaje no supervisado para identificar agrupaciones naturales de perfiles de investigadores. La evaluación comparativa de distintos enfoques de clustering permitió seleccionar K-Means++ con k = 4 como configuración final, al presentar un equilibrio adecuado entre desempeño técnico, estabilidad, interpretabilidad y utilidad para el propósito del sistema. Este resultado permitió estructurar cuatro arquetipos de investigadores que sintetizan diferencias relevantes en madurez investigativa, movilidad, trayectoria y perfil académico.

En cuanto al tercer objetivo específico, el modelo fue validado mediante métricas internas, análisis de estabilidad, revisión de granularidad y pruebas de cordura del motor de recomendación. Los resultados evidenciaron que la segmentación obtenida es coherente con la estructura del conjunto de datos y que el motor logra generar recomendaciones técnicamente afines dentro de cada clúster. Además, la evaluación mostró que las recomendaciones no se limitan a reproducir cercanía institucional, sino que conservan diversidad entre instituciones, aspecto relevante para fomentar nuevas conexiones académicas.

Frente al cuarto objetivo específico, se desarrolló una aplicación interactiva que materializa el modelo construido y permite consultar perfiles, identificar su arquetipo asociado y generar recomendaciones de colaboradores potenciales. Esta aplicación constituye el principal resultado aplicado del proyecto, en la medida en que transforma el análisis no supervisado en una herramienta funcional de apoyo a la colaboración académica.

En términos generales, el proyecto demuestra que es viable utilizar técnicas de ciencia de datos y aprendizaje no supervisado para caracterizar perfiles de investigadores y apoyar procesos de recomendación académica en el contexto del SNCTI. Además, la solución propuesta aporta una alternativa sistemática frente a los mecanismos tradicionales de búsqueda de colaboradores, usualmente basados en redes personales, visibilidad previa o revisión manual de información dispersa.

No obstante, los resultados deben interpretarse dentro del alcance definido para esta primera versión. La validación realizada corresponde a una etapa técnica, por lo que aspectos como utilidad percibida y generación efectiva de nuevas colaboraciones deberán ser evaluados en fases posteriores. Asimismo, el sistema depende de la actualización y cobertura de las fuentes públicas disponibles, por lo que su desempeño futuro estará condicionado por la incorporación de nuevas cohortes y fuentes complementarias de información.

En síntesis, el proyecto cumple su propósito al entregar una base de datos depurada, un modelo de segmentación validado, un motor de recomendación funcional y una aplicación interactiva que evidencia la viabilidad técnica de la solución. Con ello, se establece una base metodológica y tecnológica para avanzar hacia sistemas más robustos de recomendación académica, capaces de fortalecer la colaboración científica y mejorar la visibilidad de investigadores con perfiles compatibles dentro del ecosistema nacional.

7.2. Trabajos futuros

El trabajo futuro prioritario consiste en actualizar el sistema con nuevas cohortes oficiales de información antes de avanzar hacia la validación con usuarios reales. La versión desarrollada se apoya en el corte disponible utilizado durante la ejecución del proyecto. Sin embargo, entre dicho corte y el momento actual ha transcurrido un periodo aproximado de cinco años, durante el cual pueden haberse presentado cambios sustantivos en la clasificación de investigadores, filiaciones institucionales, trayectorias académicas, áreas de actuación, movilidad territorial, producción científica y composición general del ecosistema científico nacional. Estos cambios podrían modificar la estructura de los perfiles, la conformación de los clústeres, los arquetipos identificados y, en consecuencia, las recomendaciones generadas por el sistema. Por tanto, realizar pruebas con usuarios sobre una base desactualizada podría afectar la percepción de pertinencia, utilidad y confianza en las recomendaciones, no necesariamente por debilidades del modelo, sino por diferencias entre la información usada por el sistema y la situación actual de los investigadores.

La actualización del sistema debería contemplar la incorporación de bases posteriores, la ampliación de cobertura, la inclusión de nuevos investigadores reconocidos y la revisión de investigadores cuya categoría, institución, ubicación o trayectoria haya cambiado. Este proceso permitiría evaluar la estabilidad de los clústeres a través del tiempo y verificar si los arquetipos construidos conservan su capacidad descriptiva frente a una versión más reciente del SNCTI. La actualización debería incluir reentrenamiento periódico, comparación de centroides, análisis de cambios en los arquetipos, revisión de desplazamientos en la composición del ecosistema científico y control de variaciones que puedan alterar el comportamiento del motor de recomendación.

Una vez actualizado el sistema, el siguiente frente prioritario corresponde a realizar una validación con usuarios reales. Esta fase debería incluir investigadores, lideres de grupos de investigación y gestores de ciencia, tecnología e innovación, con el fin de evaluar la pertinencia de las recomendaciones, la claridad de la interfaz, la utilidad percibida, la comprensión de los arquetipos y la capacidad del sistema para apoyar la identificación de colaboradores no evidentes. Esta validación podría combinar escenarios de uso, cuestionarios, entrevistas y comparación cualitativa entre recomendaciones generadas y expectativas de los participantes.

Para que la validación con usuarios no sea solo una prueba informal de uso, se deben definir criterios de evaluación antes de ejecutarla. Entre ellos pueden considerarse: pertinencia percibida de los candidatos recomendados, claridad de los filtros, facilidad para interpretar la afinidad, tiempo requerido para identificar un colaborador potencial, confianza en la explicación del sistema, diversidad institucional o disciplinar de los resultados y disposición a usar la herramienta en procesos reales de formulación de proyectos. Estos criterios deberían aplicarse sobre una versión actualizada del sistema, de manera que la evaluación refleje tanto el desempeño funcional del prototipo como la vigencia de la información utilizada para generar las recomendaciones.

También se tiene como trabajo futuro enriquecer las fuentes de datos utilizadas. La integración de información proveniente de CvLAC, GrupLAC, producción científica, proyectos de investigación, redes de

coautoría o bases bibliométricas permitiría fortalecer la caracterización de perfiles y capturar dimensiones que no quedan completamente representadas en la fuente inicial. Este enriquecimiento podría mejorar la detección de afinidades temáticas, trayectorias colaborativas y complementariedades disciplinares, además de aumentar la capacidad del sistema para generar recomendaciones más contextualizadas y explicables.

Finalmente, se recomienda avanzar hacia la incorporación de mecanismos de retroalimentación explicita por parte de los usuarios. Permitir que los investigadores valoren recomendaciones como pertinentes, parcialmente pertinentes o no pertinentes generaría información útil para ajustar progresivamente el ranking. Esta evolución permitiría pasar de un sistema basado principalmente en segmentación y similitud vectorial a un esquema más adaptativo, capaz de aprender de la interacción de los usuarios sin perder trazabilidad metodológica. Esta retroalimentación sería especialmente valiosa después de actualizar las cohortes de información y ejecutar las primeras pruebas con usuarios reales, ya que permitiría comparar la coherencia técnica del sistema con la percepción práctica de quienes podrían utilizarlo.

8. REFERENCIAS BIBLIOGRÁFICAS

[1] N. Lorenzo-Escolar and F. Pastor-Ruiz, “Un análisis de los principales sistemas de identificación y perfil para el personal investigador,” *Aula Abierta*, no. 40, pp. 107–118, 2012.

[2] J. M. S. Gil, F. H. Hernández, T. G. Ramírez, A. G. Barujel, and V. M. H. Rivero, “University research networks as spaces for collaboration and social capital: The case of REUNI+D,” *Educ. Policy Anal. Arch.*, vol. 30, 2022, doi: 10.14507/epaa.30.7084.

[3] M. W. Rodrigues, M. A. J. Song, and L. E. Zárate, “Effectively clustering researchers in scientific collaboration networks: case study on ResearchGate,” *Soc. Netw. Anal. Min.*, vol. 11, no. 1, p. 71, 2021, doi: 10.1007/s13278-021-00781-9.

[4] Universidad del Valle, “Investigación en Cifras – Sistema de Indicadores de Ciencia, Tecnología e Innovación.” [Online]. Available: https://viceinvestigaciones.univalle.edu.co/investigacion-en-cifras

[5] Instituto Tecnológico Metropolitano, “AMCTI – Aplicativo para la Medición de Capacidades en Ciencia, Tecnología e Innovación.” [Online]. Available: https://amcti.itm.edu.co/

[6] Universidad Distrital Francisco José de Caldas, “SICIUD – Sistema de Información Científica.” [Online]. Available: https://siciud.udistrital.edu.co/

[7] Universidad de Antioquia, “SIIU – Sistema de Información para la Gestión de la Investigación.” [Online]. Available: https://www.udea.edu.co/wps/portal/udea/web/inicio/investigacion/gestion-investigacion/!ut/p/z1/...

[8] Universidad de La Sabana, “OLIS – Observatorio de Ciencia, Tecnología e Innovación.” [Online]. Available: https://olis.unisabana.edu.co/olis/

[9] Universidad Nacional de Colombia, “Hermes – Sistema de Información de la Investigación.” [Online]. Available: http://www.hermes.unal.edu.co/

[10] Y. Ma, G. Cheng, Z. Liu, and X. Liang, “Clustering-based link prediction in scientific coauthorship networks,” *International Journal of Modern Physics C*, vol. 28, no. 06, p. 1750082, May 2017, doi: 10.1142/S0129183117500826.

[11] X. Kong, H. Jiang, Z. Yang, Z. Xu, F. Xia, and A. Tolba, “Exploiting Publication Contents and Collaboration Networks for Collaborator Recommendation,” *PLoS One*, vol. 11, no. 2, pp. e0148492-, Feb. 2016, [Online]. Available: https://doi.org/10.1371/journal.pone.0148492

[12] X. Sun, W. Zhang, X. Guo, and W. Lu, “Unsupervised Graph Neural Network with Self-Expressive Attention for Community Detection,” in *2023 26th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, 2023, pp. 1890–1895. doi: 10.1109/CSCWD57460.2023.10152715.

[13] O. Husain *et al.*, “Modeling Academic Research Collaborator Selection Using an Integrated Model,” *IEEE Access*, vol. 9, pp. 102397–102421, 2021, doi: 10.1109/ACCESS.2021.3096250.

[14] J. Zhu, X. Li, C. Gao, Z. Wang, and J. Kurths, “Unsupervised community detection in attributed networks based on mutual information maximization,” *New J. Phys.*, vol. 23, no. 11, p. 113016, 2021, doi: 10.1088/1367-2630/ac2fbd.

[15] D. He *et al.*, “Community-Centric Graph Convolutional Network for Unsupervised Community Detection,” in *Proceedings of the Twenty-Ninth International Joint Conference on Artificial*

Intelligence, IJCAI-20, C. Bessiere, Ed., International Joint Conferences on Artificial Intelligence Organization, May 2020, pp. 3515–3521. doi: 10.24963/ijcai.2020/486.

[16] M. ARAKI, M. KATSURAI, I. OHMUKAI, and H. TAKEDA, "Interdisciplinary Collaborator Recommendation Based on Research Content Similarity," *IEICE Trans. Inf. Syst.*, vol. E100.D, no. 4, pp. 785–792, 2017, doi: 10.1587/transinf.2016DAP0030.

[17] P. C. Jr and C. F. de S. Araújo, "Scientific Production on Clusters in South America," *Journal of Scientometric Research*, vol. 11, no. 3, pp. 400–408, Jan. 2023, doi: 10.5530/jscires.11.3.43.

[18] UNESCO, "UNESCO Recommendation on Open Science," Paris, 2021. doi: 10.54677/MNMH8546.

[19] S. Gunawardena and R. Weber, "Discovering Patterns of Collaboration for Recommendation," 2009. [Online]. Available: www.aaai.org

[20] G. Halevi, H. Moed, and J. Bar-Ilan, "Suitability of Google Scholar as a source of scientific information and as a source of data for scientific evaluation—Review of the Literature," Aug. 01, 2017, *Elsevier Ltd*. doi: 10.1016/j.joi.2017.06.005.

[21] C. H. González Campo, G. Murillo Vargas, and M. García Solarte, *Sistema Nacional de Ciencia Tecnología e Innovación: un análisis desde la percepción de actores*. Programa editorial Universidad del Valle, 2023. doi: 10.25100/peu.863.

[22] E. P. V. Dann, C. M. G. Pavão, and F. C. C. da Silva, "A Política Nacional de Ciencia Abierta 2022-2031 da Colômbia: reflexões e perspectivas para o contexto brasileiro," *InCID: Revista de Ciência da Informação e Documentação*, vol. 14, no. 2, pp. 105–125, Dec. 2023, doi: 10.11606/issn.2178-2075.v14i2p105-125.

[23] T. e I. (Minciencias) Ministerio de Ciencia, "Datos abiertos." Accessed: May 25, 2025. [Online]. Available: https://minciencias.gov.co/ciudadano/datosabiertos

[24] J. Idakwo, Joshua Babatunde Agbogun, and Taiwo Kolajo, "A Survey on Recommendation System Techniques," *UMYU Scientifica*, vol. 2, no. 2, pp. 112–119, Jun. 2023, doi: 10.56919/usci.2322.012.

[25] H. Yang, H. Zhou, and Y. Li, "A Review of Academic Recommendation Systems Based on Intelligent Recommendation Algorithms," in *2022 7th International Conference on Image, Vision and Computing, ICIVC 2022*, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 958–962. doi: 10.1109/ICIVC55077.2022.9886104.

[26] V. Maphosa and M. Maphosa, "Fifteen Years of Recommender Systems Research in Higher Education: Current Trends and Future Direction," 2023, *Taylor and Francis Ltd.* doi: 10.1080/08839514.2023.2175106.

[27] M. C. Urdaneta-Ponte, A. Mendez-Zorrilla, and I. Oleagordia-Ruiz, "Recommendation systems for education: Systematic review," Jul. 02, 2021, *MDPI AG*. doi: 10.3390/electronics10141611.

[28] K. Najmani, E. habib Benlahmar, N. Sael, and A. Zellou, "A comparative study on recommender systems approaches," in *ACM International Conference Proceeding Series*, Association for Computing Machinery, Oct. 2019. doi: 10.1145/3372938.3372941.

[29] R. Widayanti, M. Heru, R. Chakim, C. Lukita, U. Rahardja, and N. Lutfiani, "Improving Recommender Systems using Hybrid Techniques of Collaborative Filtering and Content-Based Filtering," *Journal of Applied Data Sciences*, vol. 4, no. 3, pp. 289–302, 2023.

[30] A. Amangeldieva and C. Kharmyssov, “A hybrid approach for a movie recommender system using Content-Based, Collaborative and Knowledge-Based Filtering methods,” in *SIST 2024 - 2024 IEEE 4th International Conference on Smart Information Systems and Technologies, Proceedings*, Institute of Electrical and Electronics Engineers Inc., 2024, pp. 93–99. doi: 10.1109/SIST61555.2024.10629294.

[31] N. Sharma, A. Vaidya, S. Borate, A. Shinde, D. Khurge, and S. Tade, “Hybridization Unleashed: Advancing Movie Recommendations Through Collaborative, Content Based and Hybrid Filtering,” in *2023 IEEE Pune Section International Conference, PuneCon 2023*, Institute of Electrical and Electronics Engineers Inc., 2023. doi: 10.1109/PuneCon58714.2023.10450097.

[32] M. Uta *et al.*, “Knowledge-based recommender systems: overview and research directions,” 2024, *Frontiers Media SA*. doi: 10.3389/fdata.2024.1304439.

[33] F. Xia, H. Liu, I. Lee, and L. Cao, “Scientific Article Recommendation: Exploiting Common Author Relations and Historical Preferences,” *IEEE Trans. Big Data*, vol. 2, no. 2, pp. 101–112, Apr. 2016, doi: 10.1109/tbdata.2016.2555318.

[34] Z. Ali, G. Qi, K. Muhammad, B. Ali, and W. A. Abro, “Paper recommendation based on heterogeneous network embedding,” *Knowl. Based. Syst.*, vol. 210, Dec. 2020, doi: 10.1016/j.knosys.2020.106438.

[35] P. Toreini, M. A. Chatti, H. Thüs, and U. Schroeder, “Interest-based Recommendation in Academic Networks using Social Network Analysis,” Hrsg, 2016.

[36] X. Bai, M. Wang, I. Lee, Z. Yang, X. Kong, and F. Xia, “Scientific paper recommendation: A survey,” 2019, *Institute of Electrical and Electronics Engineers Inc.* doi: 10.1109/ACCESS.2018.2890388.

[37] M. Kolli, L. D. Mukil, N. Kartthikeyan, M. Pranavkrishnan, N. Sampath, and P. C. Nair, “Addressing the Cold Start Challenge through the Implementation of Content-based Filtering and Hybrid Filtering for Movie Recommendation,” in *2024 15th International Conference on Computing Communication and Networking Technologies, ICCCNT 2024*, Institute of Electrical and Electronics Engineers Inc., 2024. doi: 10.1109/ICCCNT61001.2024.10725528.

[38] Y. M. Tamm, R. Damdinov, and A. Vasilev, “Quality metrics in recommender systems: Do We calculate metrics consistently?,” in *RecSys 2021 - 15th ACM Conference on Recommender Systems*, Association for Computing Machinery, Inc, Sep. 2021, pp. 708–713. doi: 10.1145/3460231.3478848.

[39] G. Rodríguez Bárcenas, “Método de algoritmo de clúster para el análisis del perfil de investigadores científicos,” *e-Ciencias de la Información*, Jul. 2022, doi: 10.15517/eci.v12i2.50456.

[40] S. M. Omohundro, “Five Balltree Construction Algorithms,” 1989.

[41] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. New York, N.Y.: Wiley, 2001. [Online]. Available: https://www.wiley.com/en-us/Pattern+Classification%2C+2nd+Edition-p-9781118586006

[42] A. E. Ezugwu *et al.*, “A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects,” Apr. 01, 2022, *Elsevier Ltd*. doi: 10.1016/j.engappai.2022.104743.

[43] D. M. Ahmed Al-Kerboly and D. Z. F. Al-Kerboly, “A Comparative Study of Clustering Algorithms for Profiling Researchers in Universities Through Google Scholar,” in *2024 IEEE 3rd International Conference on Computing and Machine Intelligence, ICMI 2024 - Proceedings*, Institute of Electrical

and Electronics Engineers Inc., 2024. doi: 10.1109/ICMI60790.2024.10585953.

[44] F. S. Eustaquio, R. A. Rios, and T. N. Rios, “A Cluster Validity Protocol to Evaluate Internal Indices for Clustering of High-Dimensional Datasets,” in *IEEE International Conference on Fuzzy Systems*, Institute of Electrical and Electronics Engineers Inc., 2024. doi: 10.1109/FUZZ-IEEE60900.2024.10611943.

[45] A. Dudek, “Silhouette index as clustering evaluation tool,” in *Studies in Classification, Data Analysis, and Knowledge Organization*, Springer Science and Business Media Deutschland GmbH, 2020, pp. 19–33. doi: 10.1007/978-3-030-52348-0_2.

[46] A. Karanikola, C. M. Liapis, and S. Kotsiantis, “A comparative study of validity indices on estimating the optimal number of clusters,” in *IISA 2021 - 12th International Conference on Information, Intelligence, Systems and Applications*, Institute of Electrical and Electronics Engineers Inc., Jul. 2021. doi: 10.1109/IISA52424.2021.9555497.

[47] H. Chouikhi, M. Charrad, and N. Ghazzali, “A comparison study of clustering validity indices,” in *2015 Global Summit on Computer & Information Technology (GSCIT)*, 2015, pp. 1–4. doi: 10.1109/GSCIT.2015.7353330.

[48] R. G. Lawson and P. C. Jurs, “New index for clustering tendency and its application to chemical problems,” *J. Chem. Inf. Comput. Sci.*, vol. 30, no. 1, pp. 36–41, Feb. 1990, doi: 10.1021/ci00065a010.

[49] A. Neme and A. Nido, “Exploratory Data Analysis through the Inspection of the Probability Density Function of the Number of Neighbors,” in *Advances in Intelligent Data Analysis XII*, A. Tucker, F. Höppner, A. Siebes, and S. Swift, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 310–321.

[50] M. Capó, A. Pérez, and J. A. Lozano, “Fast Computation of Cluster Validity Measures for Bregman Divergences and Benefits.” [Online]. Available: https://ssrn.com/abstract=4086603

[51] C. Sanitphonklang and B. Chunngam, “A machine learning framework for interdisciplinary research recommendation via researcher clustering,” *International Journal of Innovative Research and Scientific Studies*, vol. 8, no. 6, pp. 2376–2388, Sep. 2025, doi: 10.53894/ijirss.v8i6.10119.

[52] K. Yang, Y. Shi, Z. Yu, Z. Zhong, J. Bi, and M. Wang, “Hybrid Clustering Solutions Fusion based on Gated Three-way Decision,” in *Proceedings of the International Joint Conference on Neural Networks*, Institute of Electrical and Electronics Engineers Inc., 2023. doi: 10.1109/IJCNN54540.2023.10191633.

[53] J. C. Rojas-Thomas and M. Santos, “New internal clustering validation measure for contiguous arbitrary-shape clusters,” *International Journal of Intelligent Systems*, vol. 36, no. 10, pp. 5506–5529, Oct. 2021, doi: 10.1002/int.22521.

[54] M. H. A. Khan, “Internal-Cluster-Validation-Based Model Selection for k-Means Clustering,” in *International Computer Science and Engineering Conference*, Institute of Electrical and Electronics Engineers Inc., 2024. doi: 10.1109/ICSEC62781.2024.10770636.

[55] B. Cao, C. Yang, K. He, J. Fan, H. Gao, and P. Qian, “Internal Purity: A Differential Entropy based Internal Validation Index for Crisp and Fuzzy Clustering Validation,” *IEEE Transactions on Fuzzy Systems*, 2024, doi: 10.1109/TFUZZ.2024.3424479.

[56] N. Pakgohar, A. Lengyel, and Z. Botta-Dukát, “Quantitative evaluation of internal cluster validation indices using binary data sets,” *Journal of Vegetation Science*, vol. 35, no. 5, Sep. 2024, doi:

10.1111/jvs.13310.

[57] F. Kächele and N. Schneider, “Cluster Validation Based on Fisher’s Linear Discriminant Analysis,” *J. Classif.*, 2024, doi: 10.1007/s00357-024-09481-3.

[58] M. Gagolewski, “Normalised Clustering Accuracy: An Asymmetric External Cluster Validity Measure,” *J. Classif.*, 2024, doi: 10.1007/s00357-024-09482-2.

[59] P. Chapman *et al.*, “CRISP-DM 1.0 Step-by-step data mining guide,” DaimlerChrysler, 1999.

[60] IBM Corporation, “IBM SPSS Modeler CRISP-DM Guide,” 2023.

[61] T. e I. (Minciencias) Ministerio de Ciencia, “Resultados preliminares corregidos - Investigadores. Convocatoria Nacional de Actualización y Transición para el Reconocimiento de Investigadores.,” Minciencias, Bogotá, Colombia, 2025. [Online]. Available: https://minciencias.gov.co/convocatorias/investigacion/convocatoria-nacional-actualizacion-y-transicion-para-el-reconocimiento

[62] T. e I. (Minciencias) Ministerio de Ciencia, “Resultados preliminares corregidos - Grupos de investigación. Convocatoria Nacional de Actualización y Transición para el Reconocimiento de Investigadores y Grupos de Investigación.,” Minciencias, Bogotá, Colombia, 2025. [Online]. Available: https://minciencias.gov.co/convocatorias/investigacion/convocatoria-nacional-actualizacion-y-transicion-para-el-reconocimiento

[63] B. Heidecke and M. Hübscher, “Planzahlen, Wachstumsraten und der Terminal Value,” in *Funktionsverlagerung und Verrechnungspreise: Rechtsgrundlagen, Bewertungen, Praxistipps*, B. Heidecke, R. Schmidtke, and J. Wilmanns, Eds., Wiesbaden: Springer Fachmedien Wiesbaden, 2017, pp. 255–264. doi: 10.1007/978-3-658-09026-5_8.

[64] Y. Wand and R. Y. Wang, “Anchoring data quality dimensions in ontological foundations,” *Commun. ACM*, vol. 39, no. 11, pp. 86–95, Nov. 1996, doi: 10.1145/240455.240479.

[65] M. Sarstedt and E. Mooi, “Cluster Analysis,” in *A Concise Guide to Market Research: The Process, Data, and Methods Using IBM SPSS Statistics*, M. Sarstedt and E. Mooi, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2019, pp. 301–354. doi: 10.1007/978-3-662-56707-4_9.

[66] O. Aydin, C. Osorio-Murillo, and C.-C. Huang, *30th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (ACM SIGSPATIAL GIS 2022) : Tuesday November 1-Friday November 4, 2022, Seattle, WA, USA*. The Association for Computing Machinery, 2022.

ANEXOS

Anexo 1. Estructura de los datos

Posición	Columna	Tipo de dato		Descripción
0	ID_CONVOCATORIA	int64	Numérica discreta	Identificador de la convocatoria
1	NME_CONVOCATORIA	object	Categórica nominal	Nombre de la convocatoria
2	ANO_CONVO	object	Temporal (fecha discreta)	Año de realización de la convocatoria
3	ID_PERSONA_PR	int64	Numérica discreta	Identificador de la persona
4	ID_AREA_CON_PR	object	Categórica nominal	Identificador de área de conocimiento OCDE principal del investigador
5	NME_ESP_AREA_PR	object	Categórica nominal	Especialidad del área de conocimiento OCDE principal del investigador
6	NME_AREA_PR	object	Categórica nominal	Área de conocimiento OCDE principal del investigador
7	NME_GRAN_AREA_PR	object	Categórica nominal	Gran área de conocimiento OCDE principal del investigador
8	NME_GENERO_PR	object	Categórica nominal	Género del investigador
9	NME_PAIS_NAC_PR	object	Categórica nominal	País de nacimiento del investigador
10	NME_REGION_NAC_PR	object	Categórica nominal	Región de nacimiento del investigador
11	NME_DEPARTAMENTO_NAC_PR	object	Categórica nominal	Departamento de nacimiento del investigador
12	NME_MUNICIPIO_NAC_PR	object	Categórica nominal	Municipio de nacimiento del investigador
13	COD_DANE_NAC_PR	float64	Numérica discreta	Código de homologación DANE para el municipio de nacimiento
14	ID_NIV_FORMACION_PR	object	Categórica ordinal	Nivel de formación máximo alcanzado
15	NME_NIV_FORM_PR	object	Categórica ordinal	Nombre del nivel de formación máximo alcanzado
16	NRO_ORDEN_FORM_PR	int64	Numérica discreta	Importancia del nivel de formación, siendo 0 la de menor valor
17	ID_CLAS_PR	object	Categórica ordinal	Categoría alcanzada por el investigador
18	NME_CLASIFICACION_PR	object	Categórica ordinal	Nombre de la categoría alcanzada por el investigador
19	ORDEN_CLAS_PR	int64	Numérica discreta	Orden de importancia de la categoría alcanzada por el investigador
20	EDAD_ANOS_PR	float64	Numérica continua	Edad en años desde la fecha de nacimiento hasta el último mes de la ventana de la convocatoria

21	NME_PAIS_RES_PR	object	Categórica nominal	País de ubicación del investigador
22	NME_REGION_RES_PR	object	Categórica nominal	Región de ubicación del investigador
23	NME_DEPARTAMENTO_RES_PR	object	Categórica nominal	Departamento de ubicación del investigador
24	NME_MUNICIPIO_RES_PR	object	Categórica nominal	Municipio de ubicación del investigador
25	COD_DANE_RES_PR	float64	Numérica discreta	Código DANE para el municipio de ubicación del investigador.
26	ID_VICTIMA_CONFLICTO	object	Categórica nominal	Población víctima del conflicto armado
27	TXT_GRUPO_ETNICO	object	Categórica nominal	Grupo étnico del investigador
28	TXT_POBLACION_DISCA	object	Categórica nominal	Población en situación de discapacidad
29	INST_FILIA	object	Categórica nominal	Lista de instituciones que el investigador indica tener filiación separados por PIPE (|)

Anexo 2. Nulos y valores especiales

Tipo de revisión	Columna/Campo	Porcentaje (%)	Observación
Valores nulos	INST_FILIA	12,07	Variable con mayor proporción de datos faltantes. Requiere tratamiento específico por su importancia para la caracterización institucional de los investigadores.
Valores nulos	COD_DANE_NAC_PR	5,94	Presenta faltantes asociados al código geográfico de nacimiento. Su impacto debe considerarse en análisis territoriales.
Valores nulos	COD_DANE_RES_PR	3,17	Presenta faltantes asociados al código geográfico de residencia. Requiere revisión si se utiliza para análisis geográficos o segmentación territorial.
Valores especiales	ID_VICTIMA_CONFLICTO	73,56	Alta proporción de registros bajo categorías especiales. No debe interpretarse como completitud plena, dado que puede representar ausencia funcional de información.
Valores especiales	NME_AREA_PR	1,44	Baja proporción de valores especiales en el área de conocimiento principal. Su impacto sobre el análisis general es acotado.
Valores especiales	NME_ESP_AREA_PR	1,44	Baja proporción de valores especiales en la especialidad del área de conocimiento. Debe monitorearse en análisis disciplinares detallados.
Valores especiales	NME_GRAN_AREA_PR	1,44	Baja proporción de valores especiales en la gran área de conocimiento. No representa una afectación significativa para la caracterización general.

Anexo 3. Top 20 áreas de conocimiento específicas

Gran área	Área específica	N° registros
Ciencias Sociales	Educación General (Incluye Capacitación, Pedagogía, ...)	4981
Ciencias Sociales	Negocios y Management	3462
Ingeniería y Tecnología	Ingeniería Eléctrica y Electrónica	3002
Ciencias Sociales	Derecho	2908
Ciencias Sociales	Economía	2248
Ciencias Sociales	Psicología (Incluye Terapias de Aprendizaje, ...)	2108
Ciencias Médicas y de la Salud	Salud Pública	1671
Ciencias Naturales	Ciencias de la Computación	1613
Ciencias Naturales	Biología Celular y Microbiología	1394
Ingeniería y Tecnología	Ingeniería Mecánica	1376
Ciencias Naturales	Bioquímica y Biología Molecular	1294
Ciencias Naturales	Ecología	1235
Humanidades	Filosofía	1231
Ciencias Agrícolas	Agronomía	1177
Ciencias Sociales	Ciencias Políticas	1174
Ciencias Naturales	Química Orgánica	1131
Ingeniería y Tecnología	Ingeniería Civil	1038
Ingeniería y Tecnología	Ingeniería Química (Plantas y Productos)	1003
Ciencias Médicas y de la Salud	Epidemiología	973
Ciencias Agrícolas	Ciencias Veterinarias	921

Anexo 4. Comparación residencia vs nacimiento

Departamento	N° registros residencia	N° registros nacimiento	Diferencia
Bogotá, D. C.	25705	20094	5611
Antioquia	14061	11671	2390
Valle del Cauca	6556	6633	-77
Atlántico	4892	3587	1305
Santander	3928	4503	-575
Caldas	2417	2824	-407
Exterior	2132	4369	-2237
Bolívar	2047	2066	-19
Boyacá	1817	3336	-1519
Norte de Santander	1510	2043	-533
Cundinamarca	1439	1293	146
Risaralda	1408	1304	104
Córdoba	1079	1394	-315
Tolima	1047	1958	-911
Cauca	1024	1363	-339

Anexo 5. Justificación de variables seleccionadas y descartadas

Variable	Descripción	Decisión	Criterio
ID_CONVOCATORIA	Identificador numérico de la convocatoria	Incluida	Identificador clave para construir el histórico y trazar registros por convocatoria.
NME_CONVOCATORIA	Nombre de la convocatoria	Descartada	Texto descriptivo del ID Convocatoria y redundante con este ID.
ANO_CONVO	Año de realización de la convocatoria	Incluida	Referencia temporal necesaria para construir históricos y definir año ancla.
ID_PERSONA_PR	Identificador del investigador	Incluida	Clave primaria del investigador.
ID_AREA_CON_PR	Identificador del área de conocimiento	Descartada	ID redundante con NME_AREA_PR.
NME_ESP_AREA_PR	Especialidad del área de conocimiento OCDE	Incluida	Variable temática de mayor detalle, la cual se usará como base para diversidad temática, clasificación disciplinar y caracterización de clústeres.
NME_AREA_PR	Área de conocimiento OCDE	Incluida	Nivel intermedio de la jerarquía, que se usará para el análisis descriptivo y apoyo en la construcción de diversidad.
NME_GRAN_AREA_PR	Gran área de conocimiento OCDE	Incluida	Nivel agregado de área que servirá de insumo para el modelo de segmentación.
NME_GENERO_PR	Género del investigador	Incluida	Variable social de equidad, no se usará como insumo para el modelo, pero servirá para el análisis de brechas y monitoreo de equidad en los clústeres.
NME_PAIS_NAC_PR	País de nacimiento del investigador	Descartada	Nivel agregado jerárquico redundante con departamento. Se prioriza usar departamento para mayor detalle.
NME_REGION_NAC_PR	Región de nacimiento	Descartada	Nivel agregado jerárquico redundante con departamento. Se prioriza usar departamento para mayor detalle.
NME_DEPARTAMENTO_NAC_PR	Departamento de nacimiento	Incluida	Nivel territorial clave para movilidad geográfica y caracterización regional de origen.
NME_MUNICIPIO_NAC_PR	Municipio de nacimiento	Incluida	Máximo detalle territorial de origen. No se usará como insumo para el modelo pero servirá para el monitoreo y trazabilidad

			geográfica de los clústers.
COD_DANE_NAC_PR	Código DANE del municipio de nacimiento	Descartada	Código numérico correlacionado con NME_MUNICIPIO_NAC_PR. Es un identificador administrativo que no aporta valor analítico adicional.
ID_NIV_FORMACION_PR	Identificador del nivel de formación	Descartada	ID redundante con NME_NIV_FORM_PR y NRO_ORDEN_FORM_PR, que capturan el nivel y su orden jerárquico de manera interpretable.
NME_NIV_FORM_PR	Nombre del nivel máximo de formación	Incluida	Variable educativa central; base para el componente Nivel de formación (EDU) del Índice de Madurez Investigativa (IMI) y para describir los perfiles de formación en los clústeres.
NRO_ORDEN_FORM_PR	Orden numérico del nivel de formación	Incluida	Codifica la jerarquía de niveles (técnico, pregrado, maestría, doctorado, etc.) y permite mapear el nivel a una escala [0,1] en el componente EDU.
ID_CLAS_PR	Identificador de la clasificación MinCiencias	Descartada	Código interno redundante con NME_CLASIFICACION_PR y ORDEN_CLAS_PR. No agrega información más allá de la categoría y su orden.
NME_CLASIFICACION_PR	Nombre de la categoría de clasificación	Incluida	Base nominal de la clasificación del investigador; se utiliza para derivar el componente de clasificación del IMI y para interpretar la trayectoria.
ORDEN_CLAS_PR	Orden de importancia de la clasificación	Incluida	Escala ordinal que permite transformar la clasificación en una variable numérica (CLS_raw y CLS_VIG).
EDAD_ANOS_PR	Edad del investigador en años	Incluida	Variable demográfica clave; se usará como EDAD_ANCLA (Edad más reciente en la última convocatoria registrada por el investigador) en el consolidado y puede aportar contexto a los clústeres.
NME_PAIS_RES_PR	País de residencia / ubicación del investigador	Descartada	Nivel agregado jerárquico redundante con departamento. Se prioriza usar departamento para mayor detalle.
NME_REGION_RES_PR	Región de residencia	Descartada	Nivel agregado jerárquico redundante con departamento. Se prioriza usar departamento para mayor detalle.
NME_DEPARTAMENTO_RES_PR	Departamento de residencia	Incluida	Nivel territorial clave en destino; se usa en indicadores de movilidad y análisis regional.

NME_MUNICIPIO_RES_PR	Municipio de residencia	Incluida	Máximo detalle territorial de residencia, útil para análisis de concentración geográfica y movilidad fina.
COD_DANE_RES_PR	Código DANE del municipio de residencia	Descartada	Código numérico redundante con NME_MUNICIPIO_RES_PR. Es un identificador administrativo que no aporta valor analítico adicional.
ID_VICTIMA_CONFLICTO	Condición de víctima del conflicto armado	Incluida	Variable social sensible utilizada para análisis de equidad y brechas en el IMI y en los clústeres.
TXT_GRUPO_ETNICO	Grupo étnico del investigador	Incluida	Variable de equidad y diversidad; se usa para cortes de IMI y de segmentación por pertenencia étnica.
TXT_POBLACION_DISCA	Condición de discapacidad	Incluida	Variable social de equidad; permite analizar posibles brechas en resultados del IMI y en la participación.
INST_FILIA	Instituciones de filiación declaradas	Incluida	Base para caracterizar la afiliación institucional, movilidad institucional y redes; además alimenta el componente de trayectoria (N_INST, movilidad, etc.).

Anexo 6. Listado de variables para el modelo

Variable	Descripción	Uso en modelado	Justificación
ID_PERSONA_PR	Identificador único de la persona investigadora.	Sí - Monitoreo	Necesario para trazabilidad, joins y recomendaciones puntuales. No se usa como feature para evitar que el modelo aprenda patrones sobre investigadores particulares.
ANIO_ANCLA	Año de la convocatoria más reciente (perfil ancla).	Sí - Monitoreo	Útil para análisis de cohorte y lectura temporal de clústers. No se usa como feature para no mezclar efectos de calendario en la distancia.
EDAD_ANCLA	Edad de la persona en el año de la convocatoria ancla.	Sí - Algoritmo	Variable numérica clave para caracterizar etapa de carrera y trayectoria, se incorpora como input del clustering.
CLAS_ANCLA	Clasificación textual (Junior, Asociado, Senior, etc.) en el año ancla.	Sí - Monitoreo	Sirve para interpretar los clústers y para reportes. Su contenido numérico ya está capturado en CLS_raw/CLS_VIG.
CLAS_ORDEN_ANCLA	Orden numérico asociado a CLAS_ANCLA.	No	Redundante frente a CLS_raw y CLS_VIG, que son las variables de clasificación que alimentan el índice de madurez. Se puede excluir del dataset analítico.
NIVEL_ANCLA	Nivel de formación textual en el año ancla.	Sí - Monitoreo	Facilita la lectura de resultados por nivel educativo. Su versión numérica está en EDU, que entra al algoritmo.
AREA_ESP_ANCLA	Especialidad de área de conocimiento de la última convocatoria.	Sí - Monitoreo	Alta cardinalidad, se usa para filtrar y describir clústers, no para calcular distancias. Evita que el clustering se reduzca a “agrupación por disciplina”.
AREA_ANCLA	Área de conocimiento (nivel intermedio) de la última convocatoria.	Sí - Monitoreo	Se usa para lectura temática y filtros de recomendación (ej. área requerida por el usuario). No se usa como feature para que los clústers reflejen patrones de madurez/trayectoria y no sólo disciplina.

GRAN_AREA_ANCLA	Gran área de conocimiento (≈ 7 categorías).	Sí - Monitoreo	Variable temática agregada. Se usará como filtro (usuario pide una gran área específica) y para caracterizar clústers, pero no se incluye en el algoritmo para evitar que domine la segmentación.
INST_ANCLA	Institución de la última convocatoria (nombre).	Sí - Monitoreo	Permite analizar y filtrar clústers por institución o tipo de institución. No entra al algoritmo por su alta cardinalidad y naturaleza nominal.
NRO_INST_ANCLA	Código numérico de la institución ancla.	Sí - Monitoreo	Identificador técnico por si se requiere cruzar con otras bases institucionales. No es adecuado como feature ya que el código no tiene significado ordinal.
NME_DEPARTAMENTO_NAC_PR	Departamento de nacimiento de la persona.	Sí - Monitoreo	Variable de origen territorial para análisis de equidad/región. No se usa en el algoritmo para evitar que la distancia agrupe primordialmente por territorio.
NME_DEPARTAMENTO_RES_PR	Departamento de residencia actual de la persona.	Sí - Monitoreo	Clave para análisis territorial y posibles índices de movilidad geográfica. Se mantiene para lectura y filtros, no como feature numérico/categórico del clustering.
NME_GENERO_PR	Género declarado de la persona.	Sí - Monitoreo	Variable sensible de equidad. Se usa para analizar brechas y composición de clústers, pero no se incluye como feature para evitar segmentación directa por género.
TXT_POBLACION_DISCA	Indicador textual de población con discapacidad.	Sí - Monitoreo	Se utiliza para análisis de inclusión y equidad; queda fuera del algoritmo para no inducir clústers dominados por condición de discapacidad.
TXT_GRUPO_ETNICO	Grupo étnico de la persona.	Sí - Monitoreo	Variable central para análisis de equidad étnica; uso exclusivamente analítico/monitoreo, no como insumo del clustering.
ID_VICTIMA_CONFLICTO	Indicador de víctima del conflicto armado.	Sí - Monitoreo	Se usa para seguimiento de inclusión de víctimas. No entra al algoritmo por ser una condición social sensible.
ANIO_MIN	Primer año de participación registrada en convocatorias.	Sí - Monitoreo	Ayuda a identificar cohortes de ingreso y “antigüedad bruta”. Como

			feature se prefiere la variable resumida ANTIGUEDAD_ANIOS.
ANIO_MAX	Último año de participación registrada.	Sí - Monitoreo	Complementa la lectura temporal de la trayectoria; no se utiliza directamente en el algoritmo por redundancia con ANIO_ANCLA y ANTIGUEDAD_ANIOS.
N_CONVOCATORIAS	Número total de convocatorias en las que ha participado la persona.	Sí - Algoritmo	Representa la intensidad de participación en el sistema; es una de las variables base para perfilar trayectoria y madurez.
N_INST	Número de instituciones distintas en las que ha participado.	Sí - Algoritmo	Indicador cuantitativo de movilidad institucional; se incluye como feature clave del modelo.
N_AREAS	Número de áreas de conocimiento diferentes en la trayectoria.	Sí - Algoritmo	Captura la diversificación temática de la carrera investigativa (especialización vs diversificación).
ANTIGUEDAD_ANIOS	Diferencia en años entre la primera y la última participación (antigüedad en el sistema).	Sí - Algoritmo	Resume el "tiempo activo" en el ecosistema; relevante para entender patrones de madurez y movilidad.
EDU	Índice/escala numérica de nivel educativo alcanzado.	Sí - Algoritmo	Versión numérica estandarizada del nivel de formación; se incorpora directamente al modelo.
CLS_raw	Clasificación "cruda" del investigador en el escalafón (escala numérica).	Sí - Algoritmo	Mide el nivel actual de reconocimiento en el sistema (Junior/Asociado/Senior) en formato numérico.
CLS_VIG	Clasificación vigente/ajustada usada en el índice de madurez.	Sí - Algoritmo	Complementa a CLS_raw como medida consolidada de estatus actual; se incluye como feature.
TASA_MOV_INST	Tasa de movilidad entre instituciones (cambios relativos en la trayectoria).	Sí - Algoritmo	Variable central de movilidad institucional; entra directamente al cálculo de distancias del clustering.
DIV_ENTROPY_N	Entropía normalizada de la distribución de instituciones en la trayectoria.	Sí - Algoritmo	Captura diversidad institucional de forma robusta; discrimina perfiles con concentraciones vs distribuciones más dispersas.
EXP	Indicador de exposición/experiencia	Sí - Algoritmo	Resume el "volumen" de participación; contribuye a

	acumulada (según definición del índice).		diferenciar trayectorias más intensas de las incipientes.
DIV	Indicador complementario de diversidad en la trayectoria (no necesariamente entropía).	Sí - Algoritmo	Aporta otra dimensión de diversidad (p. ej. combinaciones institucionales/temáticas) que complementa DIV_ENTROPY_N.
MOB	Índice sintético de movilidad institucional acumulada en convocatorias	Sí - Algoritmo	Integra información de movilidad en una dimensión numérica adicional; se incluye como feature siempre que el índice compuesto (IMI) no entre al modelo.
IMI	Índice de Madurez Investigativa (combinación de educación, clasificación, experiencia, movilidad, etc.).	Sí - Monitoreo	Índice compuesto que se usa para validar y caracterizar clústers y para ordenar recomendaciones; no entra al algoritmo para evitar doble conteo de sus componentes.
IMI_PCTL	Percentil del IMI en la distribución de investigadores.	Sí - Monitoreo	Permite ubicar cada persona en la distribución relativa de madurez; útil para análisis y cortes de política, no como feature.
IMI_TIER	Tier/categoría del IMI (p. ej. baja, media, alta madurez).	Sí - Monitoreo	Se usa para describir los clústers (qué proporción de cada tier hay en cada perfil) y para reglas de recomendación, no para entrenar el modelo.

Anexo 7. Rank de barrido experimental

Rank	Rank prom	Rank Silhouette	Rank Calinski-Harabasz	Rank Davies-Bouldin	Algoritmo	Parámetros	N° clústers	Silhouette score	Calinski-Harabasz	Davies-Bouldin	Min clúster size	Max clúster size	Runtime
1	2,7	3	1	4	KMeans	n_clústers = 4	4	0,39859	13592,71	1,151733	2774	16694	7,145696
2	5	6	3	6	GMM	n_components = 4, covariance = spherical	4	0,382851	13115,35	1,172395	2915	15804	7,143389
3	11	4	14	15	GMM	n_components = 4, covariance = tied	4	0,384966	11844,24	1,210062	2886	17571	8,956681
4	11,3	9	9	16	GMM	n_components = 4, covariance = full	4	0,37394	12132,6	1,21536	2697	17358	9,602678
5	12,3	8	12	17	GMM	n_components = 3, covariance = diag	3	0,381083	11917,12	1,215423	2886	20686	6,879781
6	14	22	8	12	KMeans	n_clústers = 5	5	0,338195	12198,89	1,202399	2768	15281	7,049029
7	15	1	41	3	Agglomerative	n_clústers = 2, linkage = single	2	0,468441	9512,499	1,097475	2886	27181	18,26219
8	17,7	7	2	44	KMeans	n_clusters = 3	3	0,38185	13240,61	1,277705	2886	17495	7,334039
9	17,7	17	16	20	KMeans	n_clusters = 6	6	0,34793	11636,92	1,222821	1363	15283	7,25234
10	19,3	26	23	9	KMeans	n_clusters = 9	9	0,327337	10769	1,187516	1360	7267	6,7891

Anexo 8. Valores de los centroides reales por clúster

Variable	C0 - Consolidados (n = 6.480; 21,6 %)	C1 - Móviles (n = 4.119; 13,7 %)	C2 - Emergentes (n = 16.694; 55,5 %)	C3 - Interdisciplinariedades (n = 2.774; 9,2 %)
EXP	0,7203	0,5583	0,3376	0,7183
EDU	0,9193	0,8709	0,8182	0,9399
DIV	0,0439	0,0447	0,0439	0,4239
MOB	0,2195	0,7721	0,1192	0,2811
N_CONVOCATORIAS	4,4620	2,8602	1,4293	4,5570
N_INST	1,4011	2,2258	1,0000	1,6182
N_AREAS	1,0000	1,0272	1,0000	2,0263
ANTIGUEDAD_ANIOS	6,5319	3,8286	0,8868	6,5789
DIV_ENTROPY_N	0,0000	0,0264	0,0000	0,7899
TASA_MOV_INST	0,0578	0,3799	0,0000	0,0897
EDAD_ANCLA	49,9875	43,0102	43,2376	50,0824
CLS_raw	0,3724	0,1029	0,0413	0,3319
CLS_VIG	0,3724	0,1029	0,0413	0,3319