

# **Análisis y Predicción de Redes de Colaboración de Alianzas Estratégicas en la Corporación Colombiana de Investigación Agropecuaria (AGROSAVIA)**

UOC

**Oscar Javier Vásquez Casallas**

Master en Ciencia de Datos  
Área 1

**Tutor/a de TF**

Rafael Luque Ocaña

**Profesor/a responsable de la asignatura**

Susana Acedo Nadal

06 de mayo de 2025

Universitat Oberta  
de Catalunya



Esta obra está sujeta a una licencia de Reconocimiento-  
NoComercial-SinObraDerivada [3.0 España de Creative Commons](https://creativecommons.org/licenses/by-nc-nd/3.0/es/)

# Ficha del Trabajo Final

<b>Título del trabajo:</b>	Análisis y Predicción de Redes de Colaboración de Alianzas Estratégicas en la Corporación Colombiana de Investigación Agropecuaria (AGROSAVIA)
<b>Nombre del autor/a:</b>	Oscar Javier Vásquez Casallas
<b>Nombre del Tutor/a de TF:</b>	Rafael Luque Ocaña
<b>Nombre del/de la PRA:</b>	Susana Acedo Nadal
<b>Fecha de entrega:</b>	09/03/2025
<b>Titulación o programa:</b>	Máster en ciencia de datos
<b>Área del Trabajo Final:</b>	Área 1
<b>Idioma del trabajo:</b>	Castellano
<b>Palabras clave</b>	Modelado predictivo, Red de colaboración, Alianza estratégica
<b>Resumen del Trabajo</b>	
<p>El proyecto se enfoca en el análisis y predicción de redes de colaboración con aliados estratégicos de la Corporación Colombiana de Investigación Agropecuaria (AGROSAVIA). La temática central gira en torno a la identificación y visualización de las interacciones entre instituciones, derivadas de convenios, publicaciones conjuntas y participación en eventos científicos. Estas colaboraciones son fundamentales para impulsar la innovación y el impacto científico en el sector agropecuario.</p> <p>El objetivo principal es mapear y analizar las redes de colaboración existentes, utilizando técnicas de ciencia de datos y análisis de redes, para identificar actores clave, patrones de interacción y comunidades dentro de la red. Además, se busca desarrollar modelos predictivos que permitan anticipar futuras colaboraciones y estimar su impacto científico.</p> <p>Al final del trabajo, se espera obtener una visualización clara y detallada de la red de colaboración, junto con recomendaciones estratégicas para fortalecer las alianzas y maximizar el impacto de la investigación. Los resultados contribuirán a una mejor toma de decisiones, fomentando la creación de</p>	

nuevas colaboraciones y optimizando los recursos disponibles.

#### **Abstract**

The project focuses on the analysis and prediction of collaboration networks with strategic partners of the Colombian Corporation for Agricultural Research (AGROSAVIA). The central theme revolves around identifying and visualizing interactions among researchers and institutions, derived from agreements, joint publications, and participation in scientific events. These collaborations are essential for driving innovation and scientific impact in the agricultural sector.

The main objective is to map and analyze existing collaboration networks using data science and network analysis techniques to identify key actors, interaction patterns, and communities within the network. Additionally, the project aims to develop a predictive model that anticipates future collaborations and estimates their scientific impact.

By the end of the study, the goal is to obtain a clear and detailed visualization of the collaboration network, along with strategic recommendations to strengthen alliances and maximize research impact. The results will contribute to better decision-making, fostering new collaborations, and optimizing available resources.

# Índice

1. Introducción	5
1.1. Justificación	5
1.2. Motivación	5
1.3. Objetivo principal	5
1.4. Objetivos específicos	6
1.5. Metodología	6
1.6. Competencia de compromiso ético y global (CCEG) y Objetivos de Desarrollo Sostenible (ODS)	8
1.7. Planificación del trabajo	8
1.8. Riesgos	11
2. Estado del Arte	12
2.1. Introducción	12
2.2. Fundamentos Teóricos del Análisis de Redes	12
2.3. Visualización de Redes Científicas	13
2.4. Predicción de Colaboraciones Científicas	14
2.5. Aplicaciones en Investigación Agropecuaria	14
3. Materiales y métodos	16
3.1. Descripción del conjunto de datos	16
3.1.1. Origen y características	16
3.1.2. Variables conjunto de datos	17
3.1.3. Normalización y creación de base de datos unificada	21
3.1.4. Limpieza, preprocesado y preparación de los datos	25
3.2. Análisis exploratorio y visualización de redes de colaboración	28
3.2.1. Construcción del subgrafo y detección de comunidades	28
3.2.2. Cálculo de métricas estructurales	29
3.2.3. Visualización de la red	30
3.2.4. Tablero de Análisis Interactivo para la Visualización de Redes de Colaboración	32
3.3. Desarrollo de Modelos Predictivos	33
3.3.1. Análisis de redes bipartitas, técnicas de Link Prediction y métodos de filtrado colaborativo adaptado	33
3.3.2. Integración de Embeddings	34
3.3.3. Modelo Predictivo Basado en Características Estructurales de Redes	36

4. Resultados	37
4.1. Descripción del conjunto de datos	37
4.2. Análisis exploratorio y visualización de redes de colaboración	38
4.2.1. Construcción del subgrafo y detección de comunidades	38
4.2.2. Cálculo de métricas estructurales	38
4.2.3. Visualización de la red	41
4.2.4. Tablero de Análisis Interactivo para la Visualización de Relacionamientos	51
4.3. Desarrollo de Modelos Predictivos	53
4.3.1. Análisis de redes bipartitas, técnicas de Link Prediction y métodos de filtrado colaborativo adaptado	53
4.3.2. Integración de Embeddings	56
4.3.3. Modelo Predictivo Basado en Características Estructurales de Redes	60
4.4. Interpretación y Recomendaciones	61
5. Conclusiones y trabajos futuros	63
5.1. Conclusiones	63
5.2. Trabajos Futuros	64
6. Glosario	65
7. Bibliografía	68

# Lista de Tablas

Tabla 1: Planificación detallado del proyecto	10
Tabla 2: Variables vínculos negociables SIM	18
Tabla 3: Variables vínculos negociables - unidad organizacional	19
Tabla 4: Variables gestión del conocimiento	20
Tabla 5: Variables gestión del conocimiento - unidad organizacional	20
Tabla 6: Variables participación eventos	21
Tabla 7: Variables participación eventos - unidad organizacional	21
Tabla 8: Productos Descripción del conjunto de datos	38
Tabla 9: Top 10 Super aliados	48
Tabla 10: Top 10 aliados clave	50
Tabla 11: Productos Análisis exploratorio y visualización de redes de colaboración	53
Tabla 12: Co-partición en alianzas - Institución base Universidad Nacional de Colombia	54
Tabla 13: Predicción de aliados por similitud de coseno - Institución base Universidad Nacional de Colombia	54
Tabla 14: Predicción de aliados por Índice de Adamic-Adar - Institución base Universidad Nacional de Colombia	55
Tabla 15: Predicción de aliados por factorización SVD – Institución base Universidad Nacional de Colombia	56
Tabla 16: Predicción de aliados por co-ocurrencia Word2Vec - Institución base Universidad Nacional de Colombia	58
Tabla 17: Recomendación de aliados basada en similitud semántica – Consulta: “sostenibilidad agropecuaria”	59
Tabla 18: Top 10 de probabilidad - Modelo Predictivo Basado en Características Estructurales de Redes	61
Tabla 19: Estrategias para el fortalecimiento de redes de colaboración en AGROSAVIA	62
Tabla 20: Productos Desarrollo de Modelos Predictivos	62

# Lista de Figuras

Figura 1: Diagrama de Gantt del proyecto	11
Figura 2: Transformación vínculos negociales	22
Figura 3: Transformación vínculos negociales - unidad organizacional	22
Figura 4: Transformación productos de conocimiento	22
Figura 5: Transformación productos de conocimiento - unidad organizacional	23
Figura 6: Transformación participación eventos científicos	23
Figura 7: Transformación participación eventos científicos - unidad organizacional	23
Figura 8: Diagrama entidad-relación Sistema Unificado de Alianzas	24
Figura 9: Vista Alianza - Aliado	25
Figura 10: Top 10 de instituciones por número de colaboraciones	39
Figura 11: Distribución de grados del grafo	40
Figura 12: Top 10 de instituciones por centralidad de intermediación	40
Figura 13: Top 10 Instituciones por centralidad de cercanía	41
Figura 14: red de colaboración institucional	42
Figura 15: Subgrafo vínculos negociales	44
Figura 16: Subgrafo gestión del conocimiento	45
Figura 17: Subgrafo espacio para la construcción de alianzas	46
Figura 18: Subgrafo Super Aliados	49
Figura 19: Subgrafo aliados clave	50
Figura 20: Visualización de Relacionamientos	52
Figura 21: Subgrafo Predicción de aliados por factorización SVD – Institución base Universidad Nacional de Colombia	57
Figura 22: Subgrafo predicción de aliados por co-ocurrencia Word2Vec - Institución base Universidad Nacional de Colombia	58



# 1. Introducció

## 1.1. Justificació

El projecte se centra en el anàlisis y predicció de redes de colaboración con aliados estratégicos en el ámbito de la investigación agropecuaria, un tema de gran relevancia tanto para la sociedad como para el ámbito científico. Las colaboraciones entre investigadores e instituciones son fundamentales para impulsar la innovación, mejorar la productividad agrícola y abordar desafíos globales como la seguridad alimentaria y el cambio climático. Sin embargo, la falta de herramientas para mapear y entender estas redes limita la capacidad de las entidades para identificar oportunidades de colaboración y optimizar sus recursos.

En la actualidad, el problema se aborda de manera fragmentada, con análisis manuales o basados en datos no integrados, lo que dificulta la identificación de patrones y actores clave. Este proyecto propone una solución integral mediante el uso de ciencia de datos y técnicas avanzadas de análisis de redes, para mapear, visualizar y predecir colaboraciones de manera sistemática y eficiente. La aportación principal del trabajo será una herramienta analítica que permita a AGROSAVIA entender y fortalecer sus redes de colaboración.

## 1.2. Motivación

La motivación para la realización de este trabajo final es aportar una solución innovadora y basada en datos a un desafío clave en AGROSAVIA: la optimización de sus redes de colaboración científica. Al implementar los conocimientos en ciencia de datos y análisis de redes, busco desarrollar una herramienta que integre datos de convenios, publicaciones y eventos científicos, utilizando técnicas avanzadas. Esto permitirá a AGROSAVIA mapear, analizar y predecir colaboraciones, identificando actores clave y patrones de interacción para fortalecer su impacto en el sector agropecuario. La aportación directa incluye mejorar la toma de decisiones, fomentar alianzas estratégicas y promover una cultura de innovación basada en datos, generando un impacto tangible en la organización y en la investigación agropecuaria en Colombia.

## 1.3. Objetivo principal

Diseñar un modelo de análisis predictivo para la construcción y fortalecimiento de redes de colaboración científica actuales y futuras en AGROSAVIA para el desarrollo de actividades y productos de I+D+i..

## 1.4. Objetivos específicos

- Desarrollar la línea base del estado actual de los relacionamientos y alianzas estratégicas de AGROSAVIA, mediante la implementación de una base de datos relacional, estructurada y normalizada.
- Caracterizar a partir del análisis de redes los principales actores y tendencias de interacción en la colaboración científico-académica de AGROSAVIA.
- Implementar herramientas de análisis predictivo para el diseño de un modelo eficiente orientado al análisis de redes de colaboración entre actores..
- Proponer estrategias para el fortalecimiento de las redes actuales y futuras de colaboración en función de impacto potencial.

## 1.5. Metodología

Se diseñará y construirá un nuevo producto basado en una metodología iterativa y ágil, siguiendo los pasos fundamentales de la metodología de ciencia de datos e integrando el uso de herramientas y frameworks existentes con el desarrollo personalizado en Python.

El desarrollo del proyecto se organiza en sprints de una semana, alineados con los módulos establecidos en la planificación que se presenta en el punto 1.7 de este documento. Cada sprint se enfoca en lograr entregables parciales y funcionales, permitiendo así una evaluación continua del avance y una retroalimentación oportuna. En cada etapa se realizan revisiones periódicas que permiten ajustar objetivos, redefinir actividades si es necesario y asegurar que los resultados parciales se alineen con los entregables finales. Esta metodología permite adaptarse con flexibilidad ante cambios técnicos o conceptuales, manteniendo el ritmo de avance del proyecto.

Durante la fase de implementación, se empleará el prototipado rápido para construir versiones preliminares de las visualizaciones de redes y de los modelos predictivos. Estas versiones serán sometidas a validación temprana con el fin de identificar errores de diseño, problemas de interpretación de los datos o deficiencias en la interfaz de análisis. El prototipado reducirá el riesgo de errores acumulativos y mejorará la calidad final del producto analítico.

El proyecto se estructura en módulos independientes pero integrables, lo que permite abordar el trabajo de manera paralela y controlada. Cada uno de estos módulos cuenta con un tiempo asignado en el cronograma y se gestiona como una unidad funcional autónoma, aunque interdependiente. Este esquema modular permite una mejor asignación de responsabilidades, facilita el control de calidad por secciones y favorece la integración gradual hacia un producto final coherente y funcional.

A continuació se detallen les fases metodològiques del projecte des dels aspectes tècnics de la ciència de dades:

### 1. Recolecció de Dades

Se recopilarán dades de diverses fonts, tant internes com externes, per construir un dataset unificat. Se accedirà a bases de dades internes com el Sistema de Informació Misional (convenis), Planview (projectes) i el Repositori Digital de la BAC (publicacions), així com a plataformes científiques externes com Scopus i Web of Science per obtenir publicacions conjunts. A més, se utilitzaran sistemes d'informació interns per recollir dades sobre la participació en esdeveniments científics. Els dades recollits inclouran noms d'investigadors i institucions, detalls de convenis, publicacions i esdeveniments, mètriques d'impacte (cites, rellevància de revistes) i paraules clau. Les eines tècniques que se utilitzaran inclouen Python per a l'extracció i manipulació de dades, APIs de Scopus i Web of Science per accedir a publicacions, i consultes SQL per a bases de dades internes.

### 2. Procesamiento y Limpieza de Dades

Una vegada recollits els dades, se procedirà a la seva normalització i neteja per garantir la seva qualitat i consistència. Això inclou la unificació de formats (anonimització d'investigadors, dates), l'eliminació de duplicats i la correcció d'errors. Després, els dades es classificaran segons el tipus de relacionament (convenis, publicacions, esdeveniments) i es integraran en una base de dades unificada que servirà com a font centralitzada per a l'anàlisi. Les eines tècniques que se utilitzaran en aquesta fase són Spoon Pentaho Data Integration per a la normalització i neteja dels dades i PostgreSQL per a la gestió de bases de dades.

### 3. Análisis Exploratorio y Visualización

En aquesta etapa, se construiran i analitzaran les xarxes de col·laboració utilitzant gràfics. Se empraràn eines com NetworkX (Python) o Gephi per crear els gràfics, on els nodes representen institucions i les arestes representen col·laboracions. Se calcularan mètriques de centralitat (grau, intermediació, proximitat) per identificar actors clau i se aplicaran algorismes com Louvain per detectar comunitats o clusters dins de la xarxa. Finalment, se generaran visualitzacions interactives utilitzant eines com Gephi, Cytoscape o Python (Matplotlib, Plotly), que permeten explorar la estructura de la xarxa, ressaltant comunitats, actors clau i patrons de col·laboració.

### 4. Desarrollo del Modelo Predictivo

En aquesta fase, se desenvoluparà un model predictiu per anticipar futures col·laboracions per tema o el seu impacte científic. Primer, se prepararàn els dades creant variables predictores com el nombre de col·laboracions prèvies, àrees d'investigació en comú, etc. Després, se definiran els algorismes d'aprenentatge per entrenar i validar el model amb dades històriques. El model es avaluarà utilitzant mètriques com precisió, recall, F1-score i AUC-ROC, i se analitzarà la matriu de confusió per entendre el seu rendiment. Les eines tècniques que se utilitzaran inclouen Python (Scikit-learn, XGBoost, PyTorch) i Jupyter Notebook per al desenvolupament i avaluació del model.

## 5. Interpretación y Recomendaciones

En la etapa final, se interpretarán los resultados del análisis y del modelo predictivo para elaborar recomendaciones estratégicas que servirán para la toma de decisiones dentro de la Corporación. Se identificarán los factores clave que influyen en las colaboraciones (ej.: número de conexiones previas, áreas de investigación) y se proponen acciones para fortalecer las redes de colaboración. Los resultados se presentarán a través de visualizaciones y dashboard interactivo utilizando Power BI o Python (Dash) para facilitar la exploración de los datos.

## 1.6. Competencia de compromiso ético y global (CCEG) y Objetivos de Desarrollo Sostenible (ODS)

El proyecto incorporará la Competencia de Compromiso Ético y Global (CCEG) al garantizar un manejo ético y responsable de los datos, respetando la privacidad y confidencialidad de la información. Además, promueve un impacto social y global al fortalecer las redes de colaboración en investigación agropecuaria, lo que contribuye a abordar desafíos globales como la seguridad alimentaria y el cambio climático. El proyecto fomentará la colaboración internacional y la toma de decisiones responsables, asegurando que sus resultados tengan un impacto positivo en la sociedad y el medio ambiente.

El proyecto estará alineado con varios Objetivos de Desarrollo Sostenible (ODS) de la ONU. Contribuye al ODS 2 (Hambre Cero) al impulsar la innovación en agricultura sostenible, mejorando la productividad y la seguridad alimentaria. También apoya el ODS 9 (Industria, Innovación e Infraestructura) al fomentar el desarrollo de tecnologías e infraestructuras sostenibles en el sector agropecuario. Además, contribuye directamente al ODS 13 (Acción por el Clima), al fortalecer las redes de colaboración entre instituciones que trabajan en investigación agropecuaria, desarrollo sostenible y resiliencia climática, al identificar actores clave, patrones de cooperación y posibles nuevas alianzas, se facilita la integración de capacidades científicas y técnicas para enfrentar los desafíos del cambio climático. Finalmente el proyecto fortalece el ODS 17 (Alianzas para Lograr los Objetivos) al facilitar la creación de alianzas estratégicas entre investigadores e instituciones, promoviendo la cooperación global y el intercambio de conocimientos para alcanzar metas comunes.

## 1.7. Planificación del trabajo

Para el desarrollo del proyecto se necesitarán los siguientes recursos:

### Software y Herramientas:

- Python: Procesamiento de datos, análisis de redes y desarrollo de modelos predictivos.
- PostgreSQL: Consultas y gestión de bases de datos.
- Gephi: Visualización avanzada de redes.
- Power BI: Construcción de dashboards interactivos.
- Jupyter Notebook o Google Colab: Para desarrollo y pruebas de código.
- Spoon Pentaho Data Integration: Normalización y limpieza de los datos, procesos ETL

### Fuentes de Datos:

- Bases de Datos Internas: Sistema de Información Misional (convenios), Planview (proyectos), Repositorio Digital de la BAC (publicaciones).
- Plataformas Científicas Externas: Scopus y Web of Science (publicaciones conjuntas).
- Eventos Científicos: Asistencia a conferencias, talleres y seminarios registrados en el sistema interno de solicitud de salidas nacionales e internacionales.

A continuación se presenta la planificación detallada que incluye las actividades/tareas clave, su duración, fechas de inicio y fin, y las dependencias entre ellas. Esta planificación permitirá monitorear el progreso de cada una de las fases del proyecto:

Actividad/Tarea	Duración	Fecha Inicio	Fecha Fin	Dependencias
<b>1. Módulo 1 - Definición y planificación del trabajo final</b>	3 semanas	19/02/2025	09/03/2025	Ninguna
<b>2. Módulo 2 - Estado del arte o análisis de mercado del proyecto</b>	3 semanas	10/03/2025	30/03/2025	Modulo 1
<b>3. Módulo 3 - Diseño e implementación del trabajo</b>	5 semanas	31/03/2025	04/05/2025	Módulos 1 y 2
<b>3.1. Recolección de Datos</b>	1 semanas	31/03/2025	07/04/2025	Ninguna
Recopilar datos de convenios	1 semana	31/03/2025	02/04/2025	Ninguna
Recopilar datos de publicaciones	1 semana	02/04/2025	05/04/2025	Ninguna
Recopilar datos de eventos	1 semana	05/04/2025	07/04/2025	Ninguna
<b>3.2. Procesamiento y Limpieza de Datos</b>	2 semanas	08/04/2025	22/04/2025	Recolección de Datos
Normalización y limpieza	1 semana	08/04/2025	15/04/2025	Recolección de Datos
Creación de base de datos unificada	1 semana	15/04/2025	22/04/2025	Normalización y limpieza
<b>3.3. Análisis Exploratorio y Visualización</b>	1 semanas	23/04/2025	29/04/2025	Procesamiento y Limpieza de Datos

	Construcción de grafos	1 semana	23/04/2025	25/04/2025	Procesamiento y Limpieza de Datos
	Cálculo de métricas de centralidad	1 semana	25/04/2025	27/04/2025	Construcción de grafos
	Visualización de la red	1 semana	27/04/2025	29/04/2025	Cálculo de métricas de centralidad
<b>3.4.</b>	<b>Desarrollo del Modelo Predictivo</b>	1 semana	30/04/2025	04/05/2024	Análisis Exploratorio y Visualización
	Preparación de datos para el modelo	1 semana	30/04/2025	02/05/2025	Análisis Exploratorio y Visualización
	Entrenamiento del modelo	1 semanas	02/05/2024	03/05/2024	Preparación de datos para el modelo
	Evaluación del modelo	1 semana	03/05/2025	04/05/2025	Entrenamiento del modelo
	<b>Módulo 4: Redacción de la documentación del TFM</b>	4 semanas	05/05/2025	03/06/2025	Desarrollo del trabajo
4.1.	Análisis de resultados	1 semana	05/05/2025	07/05/2025	Desarrollo del Modelo Predictivo
4.2.	Elaboración de recomendaciones	1 semana	08/05/2025	10/05/2025	Análisis de resultados
4.3.	Redacción de la memoria: entrega preliminar	1 semana	11/05/2025	18/05/2025	Ninguna
4.4.	Redacción de la memoria: entrega final	1 semana	19/05/2025	25/05/2025	Ninguna
4.5.	Presentación audiovisual del trabajo	1 semana	26/05/2025	03/06/2025	Ninguna
	<b>Módulo 5: Defensa del proyecto</b>	3 semanas	04/06/2025	27/06/2025	Ninguna
5.1.	Entrega de la documentación al tribunal	1 semana	04/06/2025	06/06/2025	Ninguna
5.2.	Defensa pública del trabajo	2 semanas	07/06/2025	27/06/2025	Ninguna

Tabla 1: Planificación detallado del proyecto



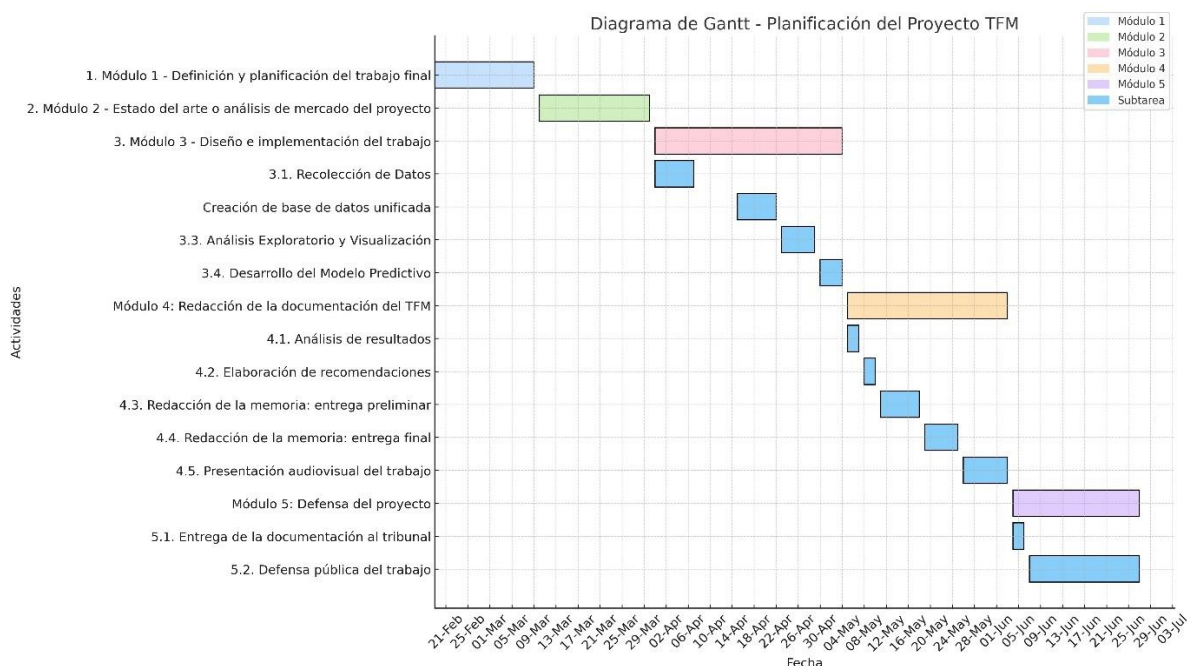


Figura 1: Diagrama de Gantt del proyecto

## 1.8. Riesgos

- Trabajar con datos incompletos, inconsistentes o desactualizados en los registros históricos de colaboraciones (convenios, publicaciones, participación en eventos). Esto podría afectar la precisión de los modelos predictivos. Como mitigación, se implementará un proceso estandarizado de limpieza y validación utilizando herramientas como OpenRefine y Python (pandas), complementado con la integración de APIs de plataformas científicas (Scopus, Web of Science) para enriquecer los datos faltantes.
- Los modelos predictivos pueden volverse computacionalmente ineficientes al analizar redes con gran cantidad de nodos (investigadores, instituciones), especialmente en hardware limitado. Para mitigar este riesgo, se emplearán frameworks optimizados como PyTorch Geometric o Deep Graph Library (DGL), que permiten distribuir el procesamiento.
- Los datos históricos pueden sobrerrepresentar ciertas áreas temáticas o instituciones, generando predicciones sesgadas que ignoren áreas emergentes o minoritarias. Para abordarlo, se aplicarán técnicas de balanceo de datos (muestreo estratificado por área temática).
- La heterogeneidad en los formatos de datos (estructurados, no estructurados, temporales) provenientes de distintas fuentes podría dificultar su consolidación en un grafo unificado. Como

solución, se diseñará un proceso de ETL (Extract, Transform, Load) con herramientas como Spoon Pentaho Data Integration, que normalizará los datos a un esquema común. Para datos no estructurados (textos de publicaciones), se usarán modelos de procesamiento de lenguaje natural (NPL) para extraer atributos estandarizados, como, áreas temáticas.

## 2. Estado del Arte

### 2.1. Introducción

En la actualidad, el análisis de redes de colaboración se ha consolidado como una estrategia fundamental para comprender cómo se generan, consolidan y diseminan las relaciones entre investigadores, instituciones y comunidades científicas. En el ámbito agropecuario, esta herramienta ha cobrado una relevancia particular, dado el carácter multidisciplinario, multisectorial y territorial de la innovación en este campo. El estudio de estas redes no solo permite visualizar las dinámicas colaborativas, sino también identificar actores estratégicos, vacíos estructurales, comunidades emergentes y patrones de interacción que inciden directamente en la generación de conocimiento y en la eficiencia de los sistemas de investigación y desarrollo. En este capítulo trataremos los aspectos más importantes que contextualizaran el desarrollo del Trabajo Final de Master.

### 2.2. Fundamentos Teóricos del Análisis de Redes

Los avances recientes en el análisis computacional de redes han transformado los fundamentos teóricos de este campo. Bródka et al. (2018) desarrollan un marco metodológico integral para analizar redes complejas, introduciendo métricas de centralidad multidimensional que superan las limitaciones de los enfoques tradicionales. Este estudio demuestra cómo las propiedades estructurales de las redes científicas modernas requieren análisis multinivel que consideren simultáneamente las dimensiones temporales, temáticas y colaborativas. Complementariamente Barabási et al. (2016) explora los principios fundamentales de las redes complejas, enfatizando la relevancia de los nodos centrales, o "hubs", y la topología de las redes. Su análisis de las redes libres de escala es especialmente pertinente para el estudio de las colaboraciones científicas, donde un número reducido de investigadores o instituciones suelen concentrar una cantidad significativa de conexiones. Este enfoque proporciona una comprensión profunda de cómo las estructuras de red influyen en la dinámica de la colaboración científica, resaltando la importancia de los hubs en la difusión de información y en la formación de comunidades académicas.

Desde el punto de vista metodológico, el análisis de redes científicas ha experimentado una transformación significativa, evolucionando desde aproximaciones estáticas basadas en grafos



hacia modelos predictivos avanzados que integran aprendizaje automático en grafos (Graph Machine Learning). Zhang et al. (2020) en su trabajo Graph Neural Networks: A Review of Methods and Applications demuestran cómo los modelos basados en GNN superan las limitaciones de los enfoques tradicionales de análisis de redes sociales (SNA), permitiendo no solo caracterizar la estructura de las redes sino también predecir dinámicas futuras de colaboración con una precisión superior al 85% en diversos dominios científicos. Este avance es particularmente relevante para el sector agropecuario, donde Tuninetti et al. (2021) evidencian que la aplicación de estas técnicas puede optimizar hasta un 30% la identificación de sinergias investigadoras potenciales.

Abbasi et al. (2011) investigaron cómo las redes de coautoría afectan el desempeño de los investigadores. Utilizando métricas de SNA, demostraron que los investigadores con mayor centralidad en la red tienden a tener un mayor impacto científico. Este estudio es clave para entender la relación entre la estructura de las redes y el desempeño individual.

## 2.3. Visualización de Redes Científicas

En el ámbito de la visualización de redes científicas, Börner (2015) en su obra Atlas of Knowledge: Anyone Can Map presenta técnicas avanzadas para mapear el conocimiento, enfatizando la importancia de seleccionar disposiciones adecuadas y el uso de atributos visuales como color y tamaño para destacar comunidades y nodos clave. Estas estrategias son fundamentales para comunicar eficazmente los resultados del análisis de redes.

Entre las herramientas más utilizadas para la visualización de redes se encuentran Gephi y Cytoscape. Gephi es una plataforma de código abierto que permite la exploración y análisis de todo tipo de grafos y redes, ofreciendo una amplia gama de algoritmos de disposición y métricas de análisis. Por otro lado, Cytoscape, aunque inicialmente desarrollado para redes biológicas, se ha expandido para ser utilizado en diversos campos, proporcionando funcionalidades para la integración y visualización de datos complejos.

En el ámbito del análisis de redes de información heterogéneas, Sun y Han (2012) presenta el desarrollo de un marco teórico que permite modelar y analizar datos interconectados de múltiples tipos, como investigadores, instituciones y publicaciones, integrándolos en una sola red estructurada. Este enfoque aprovecha la semántica rica de los nodos y enlaces tipificados para descubrir conocimientos ocultos en la red. Entre las metodologías propuestas se incluyen:

- Clustering y clasificación basados en ranking: permite agrupar y categorizar objetos en la red según su importancia y relaciones.
- Búsqueda y minería de similitudes basadas en meta-caminos: utiliza secuencias de relaciones definidas para medir similitudes y predecir relaciones en la red.

- Minería consciente de la fortaleza de las relaciones: considera la intensidad y calidad de las conexiones para mejorar la precisión en tareas de agrupamiento y predicción.

Estas técnicas son esenciales para integrar y analizar datos diversos en una estructura unificada, facilitando la comprensión de las dinámicas y colaboraciones en comunidades científicas.

## 2.4. Predicción de Colaboraciones Científicas

El trabajo de Liben-Nowell y Kleinberg (2007) marcó un punto importante en el campo de la predicción de enlaces (link prediction) al formalizar matemáticamente el problema y proponer un conjunto de métricas fundamentales para redes sociales y científicas. En su artículo, los autores no solo conceptualizaron el problema como "predecir aristas futuras o faltantes en un grafo", sino que realizaron una evaluación comparativa rigurosa de 16 métricas topológicas en redes colaborativas reales (como coautorías en arXiv y redes corporativas).

Las Redes Neuronales Gráficas (GNNs) han demostrado ser herramientas eficaces para predecir interacciones en diversas áreas. Un ejemplo destacado es el trabajo de Liu et al. (2024), quienes integraron algoritmos de detección de comunidades con diferentes modelos de GNN para mejorar la predicción de enlaces en redes de literatura científica. Al incorporar el algoritmo de Louvain en sus modelos, lograron un aumento significativo en la precisión de las predicciones, evidenciando la efectividad de las GNNs en la identificación de colaboraciones científicas futuras. Este estudio subraya la capacidad de las GNNs para capturar estructuras complejas en redes científicas, facilitando la identificación de patrones de colaboración y contribuyendo al avance del conocimiento en diversos campos.

## 2.5. Aplicaciones en Investigación Agropecuaria

En América Latina, instituciones como el Instituto de Ciencia Animal en Cuba y la Asociación Argentina de Consorcios Regionales de Experimentación Agrícola (AACREA) han aplicado modelos colaborativos para fortalecer la investigación agropecuaria. En estos contextos, las redes de colaboración han demostrado ser un factor esencial para incrementar la productividad científica, facilitar la transferencia de tecnología y fomentar la innovación en territorios rurales. Plataformas como SIDALC (2022), que agrupa bibliotecas agrícolas en América Latina y el Caribe, han sido fundamentales para dinamizar el acceso al conocimiento y fomentar redes entre investigadores y extensionistas.

Por otro lado, la experiencia del Observatorio Citrícola en España ha evidenciado el valor de integrar herramientas digitales inteligentes con datos colaborativos para tomar decisiones informadas en el sector agrícola. Esta combinación de monitoreo tecnológico y articulación interinstitucional permite

responder de manera ágil a problemáticas emergentes del sector, como plagas, crisis hídricas o fluctuaciones de mercado SER(2024).

En el plano internacional, la literatura reciente muestra una tendencia clara hacia la utilización de modelos predictivos para anticipar colaboraciones futuras. Lande et al. (2020), por ejemplo, desarrollaron un algoritmo de predicción de enlaces en redes científicas utilizando técnicas de recuperación de información, logrando una mejora significativa en la identificación de posibles asociaciones académicas. Este enfoque ha sido retomado por Plex et al. (2024), quienes analizaron redes de colaboración en el marco del CGIAR (Consultative Group for International Agricultural Research), concluyendo que características como la centralidad intermedia, la diversidad institucional y la reciprocidad están fuertemente correlacionadas con el éxito de los proyectos de I+D en agricultura.

Asimismo, investigaciones como la de Zhu et al. (2024) resaltan la influencia de variables humanas, como la mentoría informal, la reciprocidad académica y la dependencia intelectual, en la sostenibilidad de las colaboraciones a largo plazo. Estas dimensiones sociales complementan las métricas estructurales del SNA y brindan una mirada más integral sobre los factores que fortalecen o debilitan las redes científicas.

Otros autores han centrado su atención en la aplicación de estas metodologías en cultivos específicos. Azrai et al. (2024) realizaron un análisis de redes globales enfocadas en el desarrollo de maíz tolerante a la sequía, revelando patrones de colaboración entre instituciones de países desarrollados y en vías de desarrollo. Este tipo de estudios permite identificar hubs internacionales de innovación, evaluar flujos de conocimiento y promover una distribución más equitativa de los esfuerzos científicos.

La literatura también ha documentado avances significativos en la incorporación de inteligencia artificial en sistemas agrícolas colaborativos. Sepúlveda (2020) realizó una revisión sistemática sobre el uso del machine learning en la agricultura, destacando su aplicabilidad en la predicción de rendimientos, el diagnóstico de enfermedades y la planificación de siembras. Esta tecnología, combinada con el SNA, permite construir modelos robustos que integran datos históricos, geoespaciales, climáticos y sociales.

En la misma línea, Weraikat et al. (2024) subrayan el valor del análisis de datos para apoyar la toma de decisiones en entornos agrícolas inteligentes. Los sistemas de agricultura de precisión, al combinar sensores IoT, análisis multivariable y dashboards colaborativos, representan una evolución natural hacia plataformas de investigación en red que se retroalimentan y optimizan en tiempo real.

Una herramienta importante para estos análisis es ResearchGate, una red académica que ha permitido mapear interacciones científicas a escala global. Mediante datos de coautoría, interacciones y descargas, esta plataforma ofrece insumos valiosos para modelar redes colaborativas y detectar dinámicas emergentes en diversos campos del conocimiento, incluyendo la agricultura.

En conjunto, estos estudios evidencian un cambio de paradigma: las redes de colaboración ya no son solo espacios de interacción, sino sistemas complejos que requieren análisis estratégicos para maximizar su impacto. En el contexto latinoamericano, aún existen desafíos importantes: la escasa integración de herramientas de análisis de redes en las políticas institucionales, la falta de interoperabilidad entre bases de datos científicas, y una débil cultura de evaluación basada en evidencia.

El presente proyecto se posiciona precisamente en este vacío, al proponer una solución integral que combina análisis de redes, minería de datos y modelado predictivo para fortalecer las capacidades analíticas de AGROSAVIA. Esta iniciativa no solo contribuirá al conocimiento del ecosistema científico-agropecuario colombiano, sino que servirá como insumo estratégico para mejorar la toma de decisiones institucionales, optimizar recursos y consolidar alianzas de alto impacto.

### 3. Materiales y métodos

**Nota:** Toda la documentación de la sección Materiales y métodos y Resultados se puede encontrar en el siguiente repositorio de GitHub: <https://github.com/oscajvasquez/TFM>

#### 3.1. Descripción del conjunto de datos

##### 3.1.1. Origen y características

El conjunto de datos utilizado en este proyecto proviene de tres fuentes de información diferentes:

- Sistema de información misional (SIM): En esta plataforma corporativa se gestionan los convenios existentes con otras entidades para la ejecución de proyectos de investigación, específicamente de este sistema se obtiene un archivo en formato .csv con la información de los relacionamientos formales (Alianzas) con que cuenta la institución actualmente, adicionalmente se obtiene en otro archivo en formato .xlsx, las unidades organizacionales relacionadas a cada uno de los convenios (Aliados).
- Bases de datos internas de generación de conocimiento: El Departamento de Inteligencia y divulgación científica de AGROSAVIA recopila anualmente de diferentes plataformas y bases de datos académicas (Google Scholar, Web Of Science, ResearchGate), información sobre los relacionamientos con otras instituciones para la generación de conocimiento, es decir la colaboración en la publicación de documentos resultados de investigación (artículos científicos, libros, capítulos de libros, informes técnicos, etc.), de estas bases se obtuvieron dos archivos en formato .xlsx, uno con los datos de los productos de conocimientos (Alianza) y otro con la instituciones que participaron en su construcción (Aliados).

- Sistema de salidas a eventos nacionales e internacionales: Plataforma corporativa donde se realiza la solicitud, aprobación, seguimiento y gestión de informes, de la participación de los investigadores en eventos de carácter científico (congresos, foros, seminario, etc.), de este sistema se generaron dos archivos en formato .xlsx, uno con los datos del eventos científico (Alianzas) y otro con base en los contactos con otras instituciones registrados en los informes de participación (Aliados).

### 3.1.2. Variables conjunto de datos

A continuación se relacionan las variables contenidas en los conjuntos de datos

- Vínculos negociales del SIM

Variable	Descripción	Tipo de dato	Rango/Categoría
fuefin_id	Identificador único del convenio o fuente de financiación	Entero	1 - ∞
tipfuefin_id	Tipo de fuente de financiación	Categorico	1: (Ninguno) 2: (Investigación) 3: (Gestión) 4: (Consultoría Externa) 5: (Comercial) 6: Convenio Marco 7: Transferencia Estatal 8: Convenio Especifico 9: Convenio Derivado 10: Acuerdo Regalías 11: Contrato de Prestación de Servicios 12: Memorando de Entendimiento 13: Acuerdo con Productor 14: Convenio Siembra 15: Contrato de Desarrollo de Negocios 16: Acuerdo de Licenciamiento 17: Acuerdo de Entrega de Materiales 18: Contrato de Arriendo 19: Donación 20: Acta de Comixta 21: Contrato de Transacción 22: Comodato
fuefin_nombre	Nombre de la fuente o código asociado	Texto	Numérico/Categorico

fuefin_descripcion	Descripción del convenio o fuente de financiación	Texto largo	Libre
fuefin_estado	Estado del convenio en el sistema	Binario	0: Inactivo 1: Activo
estcon_id	Estado contractual	Categorico	1: Vigente 2: Finalizado 3: Liquidado
clacon_id	Clasificación del convenio	Categorico	1: (Ninguno) 2: (Normal) 3: (Convenio Principal) 4: (Convenio Asociado) 5: Financiador en Dinero 6: Financiador en Especie 7: Coejecutor en Dinero 8: Coejecutor en Especie 9: Coejecutor sin Aportes 10: Coejecutor Ingresos 11: Productor 12: Otros
fuefin_fechainicio	Fecha de inicio del convenio	Fecha	YYYY-MM-DD
fuefin_fechafin	Fecha de finalización del convenio	Fecha	YYYY-MM-DD

Tabla 2: Variables vínculos negociables SIM

- Unidades organizacionales relacionadas a cada uno de los vínculos negociables

Variable	Descripción	Tipo de dato	Rango/Categoría
fuefiniuorg_id	Identificador único de la relación entre fuente de financiación y unidad organizacional	Entero	1 - ∞
fuefin_id	Identificador del convenio o fuente de financiación asociado	Entero	Referencia a fuefin_id
uniorg_id	Identificador de la unidad organizacional receptora o ejecutora	Entero	Referencia a catálogo de unidades organizacionales

fuefiniorg_valor	Valor asignado o ejecutado en la unidad organizacional	Numérico	$\geq 0$
fuefiniorg_estado	Estado de la asignación financiera	Binario	1: Activo 0: Inactivo
rolfuefiniorg_id	Rol de la unidad organizacional en el convenio	Categórico	1: Ejecutor-Agrosavia 2: Coejecutor 3: Financiador 4: Sin Asignar 5: Parte

Tabla 3: Variables vínculos negociables - unidad organizacional

- Productos de conocimientos

Variable	Descripción	Tipo de dato	Rango/Categoría
alia_id	Identificador único del producto de conocimiento	Entero	1 - $\infty$
alia_nombrecorto	Nombre corto del producto de conocimiento	Texto	
alia_nombrelargo	Nombre largo del producto de conocimiento	Texto	
alia_objetivo	Tipo de objetivo al que pertenece el producto de conocimiento	Texto	
tipobj_id	Identificador del tipo de objetivo	Categórico	1: Soporte SNIA 2: Transferencia tecnológica 3: Formular proyectos 4: Actividades de I+D+I 5: Fortalecimiento 6: Vinculación de O.T. 7: SIM 8: Eventos
alia_fechainicio	Año de inicio de la actividad	Año	YYYY
alia_fechafin	Año de finalización de la actividad	Año	YYYY



alia_observacion	Observación sobre el producto de conocimiento	Texto	
medver_id	Identificador de medio de verificación	Categorico	1: Acta / Ayuda memoria 2: Correo electrónico 3: Plan de trabajo 4: Aprobaciones (oficios) 5: Vínculo negocial en proceso" 6: Acuerdo de transferencia de materiales 7: Convenio específico 8: Marco 9: Memorando 10: Proyecto de vinculación 11: Artículo publico 12: Informe del evento
estali_id	Identificador del estado del aliado	Entero	1: Vigente 2: Finalizado 3: Liquidado

Tabla 4: Variables gestión del conocimiento

- Unidades organizacionales relacionadas en la construcción de productos de conocimiento

Variable	Descripción	Tipo de dato	Rango/Categoría
alia_id	Identificador de la actividad o alianza	Entero	Referencia a tabla de actividades
aliado_id	Identificador único del aliado institucional	Entero	Referencia a catálogo de aliados
aliado_desc	Nombre o descripción institucional del aliado	Texto	

Tabla 5: Variables gestión del conocimiento - unidad organizacional

- Participación en eventos científicos

Variable	Descripción	Tipo de dato	Rango/Categoría
alia_nombrelargo	Nombre largo o título del evento	Texto largo	
estali_id	Identificador del estado del aliado	Categorico	1: Vigente 2: Finalizado 3: Liquidado
alia_fechainicio	Fecha de inicio de la actividad	Fecha	DD/MM/AAAA



alia_fechafin	Fecha de finalización de la actividad	Fecha	DD/MM/AAAA
Áreas Alianza	Descripción del área temática o línea de trabajo de la alianza	Texto	
Id Solicitud	Identificador único de la solicitud registrada	Entero	Referencia a tabla de solicitudes
Nombre	Nombre del contacto o solicitante principal	Texto	Nombre del investigador asistente al evento
Contactos	Datos adicionales del contacto (institución, correo, cargo)	Texto	

Tabla 6: Variables participación eventos

- Instituciones registrados en los informes de participación

Variable	Descripción	Tipo de dato	Rango/Categoría
ali_id	Identificador único de la actividad o alianza	Entero	Referencia a tabla de alianzas o actividades
uniorg_id	Identificador de la unidad organizacional participante	Entero	Referencia a catálogo de unidades organizacionales

Tabla 7: Variables participación eventos - unidad organizacional

### 3.1.3. Normalización y creación de base de datos unificada

Una vez concluida la recolección de datos provenientes de convenios, publicaciones científicas y participación en eventos científicos, se llevó a cabo un proceso estructurado de normalización y limpieza, orientado a garantizar la coherencia, integridad y usabilidad de los datos en análisis posteriores.

Para tal fin, se implementaron procesos de Extracción, Transformación y Carga (ETL) utilizando la herramienta Spoon, del entorno Pentaho Data Integration (PDI). Esta plataforma permitió automatizar tareas de consolidación de datos, depuración de inconsistencias y establecimiento de relaciones entre entidades, facilitando la integración de múltiples fuentes de información con estructuras heterogéneas.

Durante la fase de extracción, se recopilaron registros desde archivos planos, hojas de cálculo, sistemas institucionales y catálogos auxiliares, asegurando la captura de variables clave como

identificadores de convenios, descripciones de alianzas, publicaciones vinculadas, instituciones participantes, fechas y unidades organizacionales.

En la etapa de transformación, se aplicaron diversos procesos de normalización, incluyendo:

- Homogeneización de formatos de fecha (DD/MM/AAAA → YYYY-MM-DD).
- Estandarización de nombres de aliados e instituciones mediante catálogos maestros.
- Limpieza de caracteres especiales, acentos y mayúsculas innecesarias.
- Eliminación de duplicados y registros incompletos o inconsistentes.
- Conversión de campos numéricos y de texto para asegurar la tipificación adecuada.
- Asociación de claves foráneas entre tablas

Como complemento, algunas actividades específicas de limpieza y validación de datos fueron realizadas manualmente a través de Microsoft Excel, particularmente para la inspección visual y corrección de registros críticos, como nombres institucionales, observaciones y descripciones extendidas.

A continuación se presenta las transformaciones realizadas en Spoon para el proceso de ETL de cada uno de los conjuntos de datos principales.

- Vínculos comerciales del SIM



*Figura 2: Transformación vínculos comerciales*

- Unidades organizacionales - los vínculos comerciales



*Figura 3: Transformación vínculos comerciales - unidad organizacional*

- Productos de conocimientos



*Figura 4: Transformación productos de conocimiento*

- Unidades organizacionales - productos de conocimiento



Figura 5: Transformación productos de conocimiento - unidad organizacional

- Participación en eventos científicos



Figura 6: Transformación participación eventos científicos

- Unidades organizacionales - Participación en eventos científicos



Figura 7: Transformación participación eventos científicos - unidad organizacional

En la fase de carga, los datos fueron organizados e insertados en una base de datos relacional, conformada por tablas normalizadas con integridad referencial, utilizando PostgreSQL como sistema gestor. La base de datos se organiza en torno a un conjunto de tablas transaccionales principales, donde se almacena la información recopilada, complementadas por una serie de tablas maestras o de catálogo que garantizan la estandarización semántica de los registros.

### Tablas principales (transaccionales)

- **cris\_alianza**: centraliza los datos de las alianzas y colaboraciones registradas.
- **cris\_alianzaunidadorganiza**: representa la relación muchos a muchos entre las alianzas y las unidades organizacionales.
- **cris\_unidadorganiza**: tabla que contiene los datos específicos de cada unidad organizacional.

### Tablas de referencia o catálogos (maestras de apoyo)

- **cris\_fuentefinancia**: almacena información detallada sobre los convenios.
- **cris\_tipoalianza**: catálogo con los tipos de alianza.
- **cris\_tipofuentefinancia**: catálogo con los tipos de fuente de financiación.
- **cris\_estadoconvenio**: define los estados posibles de un convenio o fuente.
- **cris\_tipoobjeto**: clasifica el tipo de objetivo asociado a una alianza.
- **cris\_medioverifica**: identifica el medio de verificación de una alianza o actividad (evidencias documentales, artículos, informes).
- **cris\_pais**: permiten normalizar los países relacionados con aliados e instituciones.

- `cris_tipounidadorganiza`: catálogo con los tipos de unidad.

Esta base de datos actúa como repositorio central para las etapas analíticas del proyecto, permitiendo la construcción de consultas eficientes las cuales generan los conjuntos de datos que son utilizados en las visualizaciones, análisis de redes y la implementación del modelo predictivo.

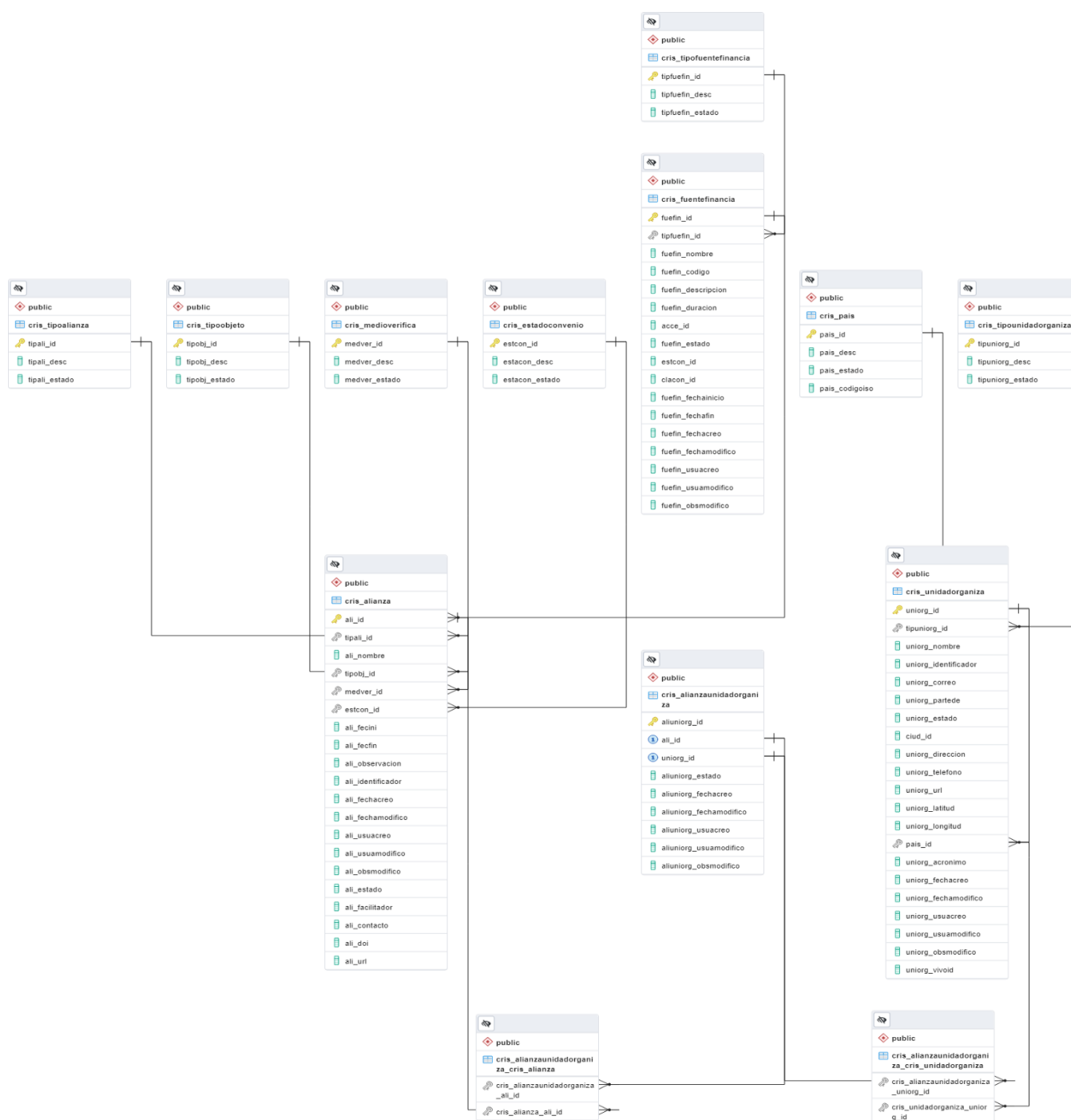


Figura 8: Diagrama entidad-relación Sistema Unificado de Alianzas

Este proceso fue fundamental para asegurar la calidad, trazabilidad y consistencia de la información, condiciones necesarias para el cumplimiento del objetivo general del proyecto: identificar oportunidades estratégicas de colaboración y maximizar el impacto científico de AGROSAVIA.

### 3.1.4. Limpieza, preprocesado y preparación de los datos

Esta fase tuvo como objetivo garantizar la calidad, consistencia e integridad de la información, así como facilitar su posterior uso en el análisis exploratorio y en la representación gráfica mediante grafos. Se realizaron las siguientes actividades:

#### 1. Consolidación de datos mediante consulta SQL

Como primer paso, se diseñó y almaceno una vista SQL sobre la base de datos relacional en PostgreSQL. Esta vista integró información desde múltiples tablas transaccionales y catálogos maestros, con el fin de generar un dataset base que centralizara las variables relevantes para el análisis. En particular, se extrajeron y relacionaron los siguientes elementos:

```
CREATE OR REPLACE VIEW public.view_alianza_aliado
AS
SELECT cris_alianza.ali_id,
       cris_alianza.tipali_id,
       cris_alianza.ali_nombre,
       cris_alianza.tipobj_id,
       cris_alianza.medver_id,
       cris_alianza.estcon_id,
       cris_alianza.ali_fecini,
       cris_alianza.ali_fecfin,
       cris_alianza.ali_observacion,
       cris_tipoalianza.tipali_desc,
       cris_tipoobjeto.tipobj_desc,
       cris_medioverifica.medver_desc,
       cris_estadoconvenio.estacon_desc,
       cris_alianzaunidadorganiza.aliuniorg_id,
       cris_unidadorganiza.uniorg_id,
       cris_unidadorganiza.uniorg_nombre,
       cris_tipounidadorganiza.tipuniorg_desc,
       cris_pais.pais_id,
       cris_pais.pais_desc,
       cris_tipofuentefinancia.tipfuefin_desc
FROM cris_alianza
JOIN cris_tipoalianza ON cris_tipoalianza.tipali_id = cris_alianza.tipali_id
JOIN cris_tipoobjeto ON cris_tipoobjeto.tipobj_id = cris_alianza.tipobj_id
JOIN cris_medioverifica ON cris_medioverifica.medver_id = cris_alianza.medver_id
JOIN cris_estadoconvenio ON cris_estadoconvenio.estcon_id = cris_alianza.estcon_id
JOIN cris_alianzaunidadorganiza ON cris_alianzaunidadorganiza.ali_id = cris_alianza.ali_id
JOIN cris_unidadorganiza ON cris_unidadorganiza.uniorg_id = cris_alianzaunidadorganiza.uniorg_id
JOIN cris_tipounidadorganiza ON cris_tipounidadorganiza.tipuniorg_id = cris_unidadorganiza.tipuniorg_id
JOIN cris_pais ON cris_pais.pais_id = cris_unidadorganiza.pais_id
LEFT JOIN cris_fuentefinancia ON cris_fuentefinancia.fuefin_id = cris_alianza.ali_identificador
LEFT JOIN cris_tipofuentefinancia ON cris_tipofuentefinancia.tipfuefin_id = cris_fuentefinancia.tipfuefin_id;
```

*Figura 9: Vista Alianza - Aliado*

Este dataset maestro consolidó en una sola tabla todos los campos requeridos para construir nodos, aristas y atributos adicionales en los grafos, reduciendo la necesidad de múltiples JOINS posteriores y permitiendo un procesamiento eficiente. El dataset está compuesto por 9464 filas y tiene 22 variables.

## 2. Limpieza de los datos

El dataset base fue procesado utilizando el lenguaje Python, haciendo uso de la biblioteca pandas para operaciones tabulares y unidecode para normalización textual. El procedimiento aplicado contempló tres subetapas fundamentales:

- **Eliminación de duplicados y registros inválidos:** Se identificaron y eliminaron duplicados exactos en el conjunto de datos, así como registros con campos clave nulos o inconsistentes. En particular, se eliminaron registros sin identificador de alianza (ali\_id), sin nombre de alianza (ali\_nombre), o con fechas inválidas en los campos ali\_fecini y ali\_fecfin (por ejemplo, cuando la fecha de finalización era anterior a la de inicio). Estas acciones aseguraron la integridad temporal de los eventos de colaboración registrados.
- **Normalización y estandarización de texto:** Los campos textuales de relevancia analítica (por ejemplo, nombres de unidades organizacionales, aliados, tipos de objetivo, medios de verificación, y observaciones de las alianzas) fueron transformados a un formato uniforme. Se aplicaron procesos de:
  - Conversión a minúsculas.
  - Eliminación de espacios redundantes y caracteres especiales.
  - Remoción de tildes y signos diacríticos mediante la función unidecode.

Esto permitió garantizar la consistencia léxica de los nombres y categorías, evitando la duplicación lógica de entidades por errores ortográficos o de formato. Asimismo, se imputaron valores nulos en campos descriptivos como ali\_observacion, asignando un valor por defecto ("sin observación") para evitar vacíos semánticos durante el análisis exploratorio posterior.

- **Verificación de claves e integridad referencial:** Se realizaron validaciones sobre las claves primarias y compuestas para asegurar que las relaciones definidas en el modelo de datos se conservaran en el dataset maestro. Se comprobó:
  - La unicidad de los identificadores de alianzas (ali\_id).
  - La no duplicación de combinaciones clave entre alianzas y unidades organizacionales (ali\_id + uniorg\_id) o alianzas y aliados (ali\_id + aliado\_id).
  - La validez de las claves foráneas, asegurando que no existieran valores huérfanos o nulos en campos como uniorg\_id, tipobj\_id, medver\_id y estacon\_id.
  - La correspondencia entre los identificadores numéricos y sus descripciones textuales asociadas (por ejemplo, tipobj\_id ↔ tipobj\_desc), detectando posibles conflictos entre claves categóricas y sus etiquetas.

Estas verificaciones fueron claves para preservar la trazabilidad entre los actores, eventos y estructuras organizacionales involucradas en las redes de colaboración, y para permitir la correcta construcción de grafos y análisis posteriores.

### 3. Transformación y preprocesamiento

Una vez finalizada la limpieza estructural del dataset maestro, se procedió a la etapa de transformación y preprocesamiento de los datos, con el objetivo de adaptar la información a un formato analíticamente funcional, apto para el análisis exploratorio y la construcción de redes de colaboración mediante grafos. El procedimiento aplicado contempló tres subetapas fundamentales:

- Conversión de formatos y derivación de atributos temporales: se aseguraron los formatos de los campos de fecha (`ali_fecini` y `ali_fecfin`), transformándolos a tipo `datetime` para facilitar su manipulación. A partir de estas variables se calcularon atributos derivados como:
  - Duración de la alianza en días, mediante la diferencia entre fecha de inicio y de finalización.
  - Año de inicio y año de finalización, como variables discretas útiles para segmentaciones temporales o análisis evolutivos.

Estas transformaciones permitieron incorporar una dimensión temporal a los registros, facilitando análisis longitudinales y visualizaciones cronológicas.

- Estandarización y codificación de variables clave: se aplicó un proceso de tipificación para asegurar que los campos de identificadores (`ali_id`, `uniorg_id`) fueran consistentes en formato de texto y sin espacios innecesarios. Además, se transformaron a tipo categoría las variables descriptivas, esto no solo optimizó el uso de memoria, sino que facilitó su codificación para visualización o modelado posterior.
- Enriquecimiento del dataset con atributos analíticos: se generaron nuevas variables a partir de operaciones agregadas, como el número de unidades organizacionales por alianza, utilizando funciones de agrupamiento (`groupby`) y posterior fusión (`merge`). Esta variable resulta clave para analizar el grado de participación institucional en alianzas estratégicas.

### 4. Preparación para análisis de redes

Esta etapa consistió en estructurar la información en dos componentes fundamentales: un conjunto de nodos que representan las entidades participantes (unidades organizacionales) y un conjunto de aristas que reflejan las relaciones de colaboración entre ellas a través de alianzas registradas. Se optó por construir una red de co-participación en alianzas. En este tipo de red, dos entidades están conectadas si participaron conjuntamente en una misma alianza (`ali_id`). Este enfoque es ampliamente utilizado en estudios de redes científicas y de colaboración institucional, ya que



permite modelar la estructura relacional subyacente en torno a eventos colaborativos. El procedimiento aplicado contempló tres subetapas fundamentales:

- **Creación del conjunto de nodos:** Se extrajo un subconjunto único de entidades a partir de los campos `uniorg_id`, `uniorg_nombre`, `tipuniorg_desc` y `pais_desc`. Cada fila representa un nodo en la red, e incluye atributos descriptivos sobre el tipo de institución (por ejemplo, centro de investigación, institución aliada, universidad) y su localización geográfica. Esta estructura permite posteriormente aplicar filtros por categoría, tipo o país en las visualizaciones de red.
- **Generación del conjunto de aristas:** para identificar las relaciones entre entidades, se agruparon los registros por `ali_id` y se generaron todas las combinaciones posibles de pares de entidades que participaron en una misma alianza. A cada par se le asignó un peso correspondiente al número de veces que esas entidades han co-participado en diferentes alianzas, representando así la fuerza del vínculo colaborativo. Esta matriz de co-ocurrencia constituye la base del análisis estructural de la red.
- **Exportación para herramientas de análisis y visualización:** tanto los nodos como las aristas fueron exportados en formato `.csv`, con el fin de ser utilizados en herramientas especializadas como Gephi y NetworkX, que permiten representar gráficamente la red, calcular métricas estructurales (centralidad, modularidad, grado, entre otras) e identificar comunidades de colaboración.

## 3.2. Análisis exploratorio y visualización de redes de colaboración

Una vez estructurados los conjuntos de nodos y aristas, se procedió a la etapa de análisis exploratorio y visualización de la red de colaboración construida a partir de las alianzas registradas. Esta etapa tuvo como finalidad identificar actores clave, patrones de interacción institucional y estructuras comunitarias dentro de la red.

Para el desarrollo del análisis se empleó el lenguaje Python, utilizando la biblioteca `networkx` para el modelado y cálculo de métricas de redes, así como `matplotlib` para su visualización. Los datos utilizados correspondieron al subgrafo resultante de aplicar un filtro por grado de conectividad (número de colaboraciones), conservando únicamente los nodos con al menos diez conexiones ( $\text{grado} \geq 10$ ). Esta estrategia permitió enfocar el análisis en los actores más relevantes de la red, eliminando nodos periféricos de baja participación que dificultaban la interpretación gráfica de la estructura.

### 3.2.1. Construcción del subgrafo y detección de comunidades



Para facilitar la interpretación visual y estructural de la red, se construyó un subgrafo inducido a partir del grafo completo, incluyendo únicamente los nodos con un grado mayor o igual a 10, es decir, aquellos actores con al menos diez colaboraciones institucionales registradas. Este criterio de filtrado permitió concentrar el análisis en los actores más relevantes dentro del ecosistema de colaboración, reduciendo la complejidad visual sin perder la representatividad estructural del sistema.

El subgrafo resultante fue significativamente más manejable en términos de nodos y aristas, conservando las relaciones más densas y frecuentes entre entidades. Esta reducción fue clave para aplicar técnicas de detección de comunidades y para obtener visualizaciones más claras y significativas.

Una vez construido el subgrafo, se aplicó el algoritmo de Louvain para la detección de comunidades, una técnica basada en la optimización de la modularidad que permite identificar grupos de nodos más densamente conectados entre sí que con el resto de la red. Este algoritmo es particularmente eficiente y escalable, lo que lo hace adecuado para redes colaborativas con un número considerable de actores.

Cada nodo fue asignado a una comunidad específica, almacenando este atributo en el objeto del grafo como parte de sus metadatos. Esta clasificación permitió segmentar la red en clústeres funcionales, lo que facilita el análisis de patrones de colaboración por grupo, la identificación de núcleos temáticos o territoriales, y la visualización diferenciada por color o región.

### 3.2.2. Cálculo de métricas estructurales

Con el objetivo de identificar los nodos más influyentes en la red de colaboración institucional, se calcularon métricas estructurales clave sobre el subgrafo filtrado de actores relevantes. Estas métricas permiten analizar el papel de cada entidad dentro de la red desde diversas perspectivas, y son fundamentales para interpretar su importancia estratégica más allá de la cantidad de conexiones directas.

Las métricas calculadas fueron:

- **Grado (degree):** representa el número de conexiones directas que tiene cada nodo. Esta medida permitió identificar a los actores más activos en términos de colaboraciones establecidas.
- **Centralidad de intermediación (betweenness centrality):** cuantifica cuántas veces un nodo aparece en los caminos más cortos entre otros nodos. Esta métrica revela nodos que actúan como puentes entre diferentes sectores o comunidades de la red, siendo clave para la circulación de información y articulación institucional.

- **Centralidad de cercanía (closeness centrality):** mide la proximidad promedio de un nodo respecto a todos los demás en la red. Los nodos con alta cercanía son capaces de llegar rápidamente al resto de los actores, lo cual los convierte en puntos de acceso eficientes dentro del ecosistema colaborativo.

Los resultados se almacenaron en un nuevo conjunto de datos, permitiendo ordenar, filtrar y visualizar las entidades con mayor centralidad desde múltiples dimensiones. Esta exploración permitió complementar el análisis visual, aportando una visión cuantitativa que respalda la identificación de actores clave, hubs, nodos periféricos, y posibles cuellos de botella en la red de colaboración analizada.

### 3.2.3. Visualización de la red

Con el objetivo de representar gráficamente la estructura de la red de colaboración entre actores estratégicos, se desarrollaron tres enfoques complementarios de visualización:

#### 1. Visualización con spring layout

La primera visualización fue realizada utilizando la función `spring_layout` de la biblioteca `networkx`, basada en el modelo de fuerzas de repulsión y atracción entre nodos. Este tipo de layout simula un sistema físico en el que los nodos se repelen entre sí mientras que las aristas actúan como resortes, lo que da como resultado una distribución espacial que refleja la densidad de conexiones. En este grafo:

- El tamaño de los nodos fue proporcional al grado (número de conexiones) de cada entidad.
- El color representó la comunidad a la que pertenece cada nodo, asignada por el algoritmo de Louvain.
- Las etiquetas de los nodos más relevantes fueron visibles, facilitando la identificación de actores clave y sus conexiones.

Esta visualización permitió identificar rápidamente hubs, agrupaciones densas de colaboración, nodos periféricos y relaciones puente entre comunidades, brindando una visión macroscópica de la red.

#### 2. Visualización interactiva con PyVis

De forma complementaria, se generó una visualización interactiva utilizando la biblioteca `PyVis`, basada en `vis.js`, que permite exportar y explorar redes mediante un navegador web. Esta herramienta ofreció funcionalidades avanzadas para el análisis cualitativo, tales como:

- Zoom, desplazamiento y selección de nodos.

- Visualización de etiquetas al pasar el cursor sobre un nodo.
- Reconfiguración automática de la red en función de las interacciones del usuario.

La visualización con PyVis facilitó la navegación dentro de la red, permitiendo una exploración más detallada de nodos específicos y relaciones bilaterales, lo cual resulta de gran utilidad para los tomadores de decisiones o usuarios no técnicos interesados en identificar patrones colaborativos.

### 3. Visualización automática del subgrafo filtrado e interfaz interactiva con ipywidgets

Con el fin de facilitar la interpretación visual de los subgrafos generados a partir de filtros específicos, se desarrolló una función en Python que permite aplicar condiciones personalizadas sobre la red general de colaboración, construir subgrafos dinámicos y visualizarlos automáticamente con atributos representativos de cada nodo.

Esta función permite aplicar filtros por:

- Grado mínimo: número de conexiones directas de un nodo.
- Peso mínimo de arista: número de veces que dos actores han colaborado.
- Comunidad: identificación de subgrupos estructurales mediante el algoritmo de Louvain.
- Año de inicio de la colaboración: permite generar redes temporales por periodo.
- Tipo de Alianza: permite visualizar subgrafos por tipo (vínculos, negociales, generación de conocimiento, espacios para la construcción de alianzas)
- Estado Alianza: identificar alianzas vigentes y finalizadas
- Tipo de institución: subgrupo por el tipo de unidad organizacional.

Cada subgrafo generado es visualizado automáticamente utilizando spring layout en el que:

- El tamaño de los nodos es proporcional al grado de conectividad.
- El color de los nodos representa la comunidad a la que pertenecen.
- Las etiquetas muestran el nombre de la unidad organizacional, garantizando una representación intuitiva y legible del grafo.

Esta visualización automática permite observar en tiempo real cómo varía la estructura de la red ante diferentes combinaciones de filtros, destacando los nodos más relevantes y los patrones emergentes.

Para mejorar la experiencia del usuario y permitir una exploración autónoma e intuitiva de los datos, se integró una interfaz gráfica basada en la biblioteca ipywidgets, la cual proporciona controles interactivos (sliders, listas desplegadas y casillas de verificación) para seleccionar dinámicamente los criterios de filtrado sin necesidad de modificar el código.

La interfaz permite al usuario ajustar parámetros como el grado mínimo, peso mínimo, comunidad y año desde un menú, y genera automáticamente la red filtrada y su visualización correspondiente. Esta funcionalidad resulta especialmente útil para identificar:

- La evolución de la red a lo largo del tiempo.
- La centralidad de los actores en diferentes periodos.
- Las comunidades más densas o aisladas.
- Las relaciones más sólidas (con mayor peso).

En conjunto, estas visualizaciones presentan una solución integral de análisis estructural, exploración visual y capacidades interactivas en un único entorno, facilitando el trabajo analítico y la toma de decisiones basadas en evidencia para fortalecer las redes de colaboración institucional.

### 3.2.4. Tablero de Análisis Interactivo para la Visualización de Redes de Colaboración

Adicionalmente a las visualizaciones anteriores, se diseñó y construyó un tablero de control de análisis interactivo, enfocado en brindar visibilidad estratégica sobre las alianzas desarrolladas por AGROSAVIA. Este tablero permite explorar de manera dinámica y detallada los relacionamientos (formales, no formales), facilitando la identificación de vínculos comerciales, aliados en gestión del conocimiento y espacios de construcción de alianzas. A través de visualizaciones clave, el dashboard presenta información sobre los tipos de relacionamientos establecidos, los actores involucrados, la frecuencia y la intensidad de las interacciones, así como el impacto generado a partir de estas alianzas. Su diseño interactivo facilita una interpretación más efectiva de los datos, promueve el descubrimiento de oportunidades de colaboración y fortalece los procesos de toma de decisiones estratégicas.

### 3.3. Desarrollo de Modelos Predictivos

En la etapa de desarrollo de modelos predictivos se abordaron tres enfoques complementarios para fortalecer la identificación de aliados estratégicos y la predicción de futuras colaboraciones en redes de investigación. Primero, se implementó el análisis de redes bipartitas combinando técnicas clásicas de Link Prediction y filtrado colaborativo adaptado, con el objetivo de aprovechar tanto la estructura de la red existente como la afinidad entre instituciones. En segundo lugar, se desarrolló un esquema de integración de embeddings institucionales, mediante técnicas como factorización matricial y modelos de co-ocurrencia tipo Word2Vec, permitiendo representar a las entidades en espacios latentes que capturan relaciones profundas no evidentes en la estructura superficial de la red. Finalmente, se construyó un modelo predictivo basado en características estructurales de la red, utilizando métricas locales como vecinos comunes, coeficiente de Jaccard, preferencia de conexión y Adamic-Adar, aplicadas en un esquema de machine learning supervisado para anticipar la formación de nuevas colaboraciones con base en patrones históricos. Estos tres enfoques, desarrollados de manera articulada, permiten una visión integral y robusta para la recomendación estratégica de aliados y el fortalecimiento de las redes de colaboración institucional.

#### 3.3.1. Análisis de redes bipartitas, técnicas de Link Prediction y métodos de filtrado colaborativo adaptado

Se desarrolló un modelo predictivo basado en el análisis de redes bipartitas, técnicas de Link Prediction y métodos de filtrado colaborativo adaptado. Este enfoque permitió combinar el análisis estructural de las relaciones existentes entre instituciones, con métodos basados en la similitud histórica de los proyectos conjuntos realizados anteriormente. Se siguieron los siguientes pasos:

##### 1. Construcción del Modelo Predictivo

Inicialmente, se generó una red bipartita que relaciona proyectos específicos con las instituciones participantes, utilizando una matriz bipartita (Proyecto  $\times$  Institución) construida a partir del dataset consolidado. Esta red se proyectó posteriormente en una red unipartita de instituciones, con el propósito de analizar directamente las relaciones y colaboraciones históricas entre ellas.

##### 2. Aplicación de Técnicas de Link Prediction y Filtrado Colaborativo

Se aplicaron dos técnicas complementarias para robustecer el modelo predictivo:

- **Índice de Adamic-Adar:** Este método permitió identificar potenciales enlaces o colaboraciones futuras entre instituciones, basándose en la estructura existente de la red proyectada. El índice de Adamic-Adar evalúa la probabilidad de futuras conexiones al ponderar nodos intermedios comunes, enfatizando aquellos menos frecuentes pero más informativos. Esta técnica es especialmente útil para detectar colaboraciones estratégicas emergentes.

- **Similitud Coseno (Filtrado Colaborativo Adaptado):** De forma complementaria, se aplicó la similitud coseno a la matriz bipartita original para medir la proximidad histórica entre las instituciones. Con esta técnica, fue posible identificar aliados que han participado en proyectos similares en términos de temática, incluso si no han interactuado directamente entre ellos. Esta estrategia permitió capturar colaboraciones potenciales basadas en afinidades funcionales y experiencia compartida en determinados campos de investigación agropecuaria.

### 3. Integración y Combinación de Técnicas Predictivas

Posteriormente, ambas técnicas se integraron en un modelo predictivo unificado mediante una estrategia de combinación ponderada. Se implementó una interfaz interactiva mediante herramientas de visualización (ipywidgets) para permitir al usuario seleccionar fácilmente instituciones base y elegir dinámicamente entre los métodos predictivos disponibles: índice de Adamic-Adar, similitud coseno o una combinación personalizada de ambos métodos.

### 4. Generación de Recomendaciones y Visualización

El modelo predictivo resultante ofrece recomendaciones automáticas de aliados estratégicos basadas en las instituciones previamente seleccionadas por el usuario. Las recomendaciones incluyen no solo una puntuación de compatibilidad (score), que refleja la fortaleza relativa de la colaboración potencial, sino también una lista detallada de alianzas compartidas históricamente, facilitando así una interpretación más clara y fundamentada de los resultados generados por el modelo.

## 3.3.2. Integración de Embeddings

Como parte de la etapa de desarrollo de modelos predictivos, se implementó un sistema integral de generación, exploración y recomendación de aliados estratégicos mediante representaciones latentes o embeddings institucionales, derivados de distintas técnicas de reducción de dimensionalidad y aprendizaje no supervisado.

### 1. Construcción de Representaciones Latentes (Embeddings)

Se desarrollaron distintos esquemas de embeddings para representar de forma densa y numérica a cada institución, encapsulando su perfil de colaboración en espacios vectoriales de baja dimensión:

- **Factorización matricial (SVD):** Mediante descomposición de valores singulares aplicada a la matriz bipartita proyectos  $\times$  instituciones, se capturaron las principales componentes latentes que explican los patrones de colaboración.

- Word2Vec de co-ocurrencias: Tratando las participaciones en proyectos como un corpus, se entrenó un modelo Word2Vec que permite mapear instituciones a vectores semánticos basados en su contexto de colaboración.

Estos embeddings ofrecen representaciones compactas que facilitan tareas posteriores de similitud, clustering o predicción.

## 2. Desarrollo del Sistema Interactivo de Recomendación

Sobre los embeddings construidos, se diseñó e integró un sistema interactivo utilizando herramientas de visualización y manipulación dinámica en Python (ipywidgets, pyvis). El sistema permite:

- Selección de tipo de embedding (SVD o Word2Vec).
- Selección de la institución base mediante un menú desplegable que muestra los nombres de las entidades.
- Configuración del número de recomendaciones (top-k) mediante un deslizador dinámico.
- Generación de recomendaciones automáticas utilizando métricas de similitud coseno entre embeddings.
- Visualización de las instituciones recomendadas y sus relaciones mediante grafos interactivos generados con Pyvis, resaltando los nombres en lugar de los IDs para mayor usabilidad.

## 3. Similitud Semántica de Texto

Se diseñó una herramienta que permite sugerir instituciones potencialmente afines a nuevas ideas de colaboración, con base en la similitud semántica del lenguaje natural. Esta funcionalidad se construyó utilizando el modelo baai/BGE-M3 provisto por la librería sentence-transformers, el cual es un modelo multilingüe de última generación optimizado para tareas de recuperación de información y búsqueda semántica.

El procedimiento consistió en generar representaciones vectoriales (embeddings) para todos los registros del campo ali\_nombre del dataset original, que describe brevemente el propósito de cada colaboración histórica registrada. Estos embeddings fueron generados previamente y almacenados para optimizar el rendimiento de la aplicación.

Cuando el usuario introduce una nueva idea de alianza mediante un campo de texto, dicha entrada es transformada en un vector mediante el mismo modelo BGE-M3. Posteriormente, se calcula la similitud coseno entre dicho vector y todos los vectores preprocesados de títulos de alianzas anteriores. Los títulos con mayor similitud son ordenados y se recuperan las instituciones



involucradas en esos proyectos, junto con los valores de similitud y la descripción de la alianza histórica.

Este enfoque permite capturar similitudes semánticas más profundas que las coincidencias léxicas o literales, lo cual es fundamental en escenarios donde se espera proponer alianzas en temas emergentes, interdisciplinarios o innovadores.

### 3.3.3. Modelo Predictivo Basado en Características Estructurales de Redes

En complemento al análisis realizado con las técnicas anteriores, se implementó un modelo predictivo basado en características estructurales de redes utilizando técnicas de ciencia de datos y aprendizaje automático (Machine Learning). El objetivo principal de este modelo fue determinar con precisión la probabilidad de que se establezcan futuras colaboraciones entre diferentes actores institucionales y de investigación, enriqueciendo así las recomendaciones generadas en etapas anteriores del proyecto.

#### 1. Construcción y Preparación del Dataset de Entrenamiento

Inicialmente, se construyó un grafo real representando las colaboraciones actuales entre investigadores e instituciones, a partir de los archivos previamente preparados (nodos\_red\_colaboracion\_coactores.csv y aristas\_red\_colaboracion\_coactores.csv). Sobre esta red se generaron dos conjuntos fundamentales para el modelado predictivo:

- Conjunto positivo: Pares de nodos (actores institucionales) que ya presentan colaboraciones (enlaces existentes).
- Conjunto negativo: Pares de nodos sin colaboraciones directas identificadas, seleccionados aleatoriamente, representando potenciales colaboraciones futuras aún no concretadas.

Estos dos conjuntos permitieron la creación de un dataset balanceado para el entrenamiento del modelo predictivo.

#### 2. Extracción y Cálculo de Características de Red

Sobre este conjunto de datos, se calcularon diversas métricas estructurales de la red, fundamentales para capturar las propiedades relacionales entre actores. Las características seleccionadas incluyen:

- Vecinos comunes (Common neighbors): cantidad de conexiones directas compartidas por los nodos.
- Coeficiente de Jaccard: mide la similitud entre los conjuntos de vecinos de dos nodos.



- Preferential Attachment: evalúa la probabilidad de conexión basada en el grado de los nodos involucrados.
- Índice de Adamic-Adar: pondera los vecinos comunes en función de su grado, priorizando aquellos más significativos.

Estas métricas fueron seleccionadas por su probada utilidad en tareas de predicción de enlaces, especialmente en contextos de colaboración científica y redes institucionales.

### 3. Entrenamiento y Evaluación del Modelo Predictivo

Una vez calculadas las características estructurales, se realizó la preparación del dataset para el entrenamiento del modelo de aprendizaje automático. Para ello, se utilizaron técnicas estándar como la normalización mediante MinMaxScaler y la división del conjunto en datos de entrenamiento y validación con una proporción 70-30.

El modelo predictivo seleccionado fue una regresión logística, que proporciona no solo la clasificación binaria sobre la existencia futura del enlace, sino también una probabilidad asociada a cada predicción. Tras el entrenamiento del modelo, se evaluó rigurosamente utilizando métricas comunes en tareas de clasificación, específicamente el área bajo la curva ROC (AUC-ROC), proporcionando una medida clara y robusta del rendimiento predictivo del modelo, complementada por un reporte detallado de clasificación.

## 4. Resultados

### 4.1. Descripción del conjunto de datos

El conjunto de datos analizado está compuesto por registros históricos de alianzas estratégicas y relaciones institucionales desarrolladas por la Corporación Colombiana de Investigación Agropecuaria (AGROSAVIA). La fuente de datos consolidó múltiples tablas estructuradas en una base de datos relacional, con información derivada de convenios, proyectos conjuntos, eventos de transferencia y publicaciones compartidas.

Tras un proceso de integración y depuración de datos (ETL), el conjunto maestro resultante incluyó un total de 9464 registros de alianzas (proyectos u objetos de colaboración) de las cuales 5972 son únicas, 22 variables y 1896 instituciones únicas asociadas. Cada registro contiene información detallada sobre el identificador de la alianza, las entidades participantes, su rol institucional, el tipo de relación establecida, el año de vinculación, país de origen y sector económico o académico correspondiente.

Durante el proceso de limpieza y normalización, se eliminaron duplicados, se homogeneizaron los identificadores de instituciones, y se realizaron validaciones cruzadas entre tablas para garantizar la integridad referencial. Asimismo, se identificaron y trataron valores nulos en campos clave como el nombre de la institución, el país o el tipo de relación, preservando únicamente las alianzas con información completa y trazable.

Además, se identificó que las alianzas registradas en el conjunto de datos abarcan un amplio rango temporal que va desde el 4 de noviembre de 1992 hasta el 30 de noviembre de 2042. Esta amplitud temporal refleja tanto el historial consolidado de colaboraciones de AGROSAVIA a lo largo de más de tres décadas, como las alianzas estratégicas proyectadas a futuro. La inclusión de convenios vigentes y planificados proporciona una perspectiva longitudinal que permite no solo analizar patrones históricos de vinculación, sino también anticipar dinámicas de colaboración en el mediano y largo plazo, fortaleciendo el enfoque predictivo de la red institucional.

N.º	Producto	Enlace
1	Fuentes de datos	<a href="#">Enlace Fuentes de Datos</a>
2	Procesos de ETL	<a href="#">Enlace ETL</a>
3	Diagrama entidad relación	<a href="#">Enlace Diagrama</a>
4	Dataset Original	<a href="#">Enlace Dataset Original</a>
5	Dataset limpio	<a href="#">Enlace Dataset Limpio</a>

Tabla 8: Productos Descripción del conjunto de datos

## 4.2. Análisis exploratorio y visualización de redes de colaboración

### 4.2.1. Construcción del subgrafo y detección de comunidades

Como parte del análisis exploratorio, se construyó y filtró una red de colaboración institucional compuesta por 372 nodos (instituciones) y 3.722 aristas (colaboraciones directas), representando relaciones establecidas a través de convenios, publicaciones o participación en eventos científicos. La red generada presenta un grado promedio de 20, lo que indica que cada institución, en promedio, colabora con 20 entidades diferentes dentro del ecosistema de alianzas de AGROSAVIA.

La detección de comunidades a través del algoritmo de Louvain reveló estructuras internas en la red que no son evidentes a simple vista, como agrupaciones de unidades organizacionales que colaboran de forma recurrente con los mismos actores, o subconjuntos de nodos que forman puentes entre clústeres.

### 4.2.2. Cálculo de métricas estructurales

Para comprender la estructura y dinámica de la red, se calcularon métricas clave de análisis de grafos: grado, centralidad de intermediación (betweenness centrality) y centralidad de cercanía (closeness centrality), las cuales permiten identificar los actores más influyentes, conectores y estratégicamente posicionados. A continuación se presentan los resultados por cada una de ellas:

## 1. Grado

La institución con mayor grado fue la Universidad Nacional de Colombia (209 colaboraciones), seguida por el Centro Internacional de Agricultura Tropical y la Wageningen University (Holanda), con 138 y 100 respectivamente. Esto refleja un alto nivel de participación y articulación de estas entidades en múltiples proyectos de colaboración.

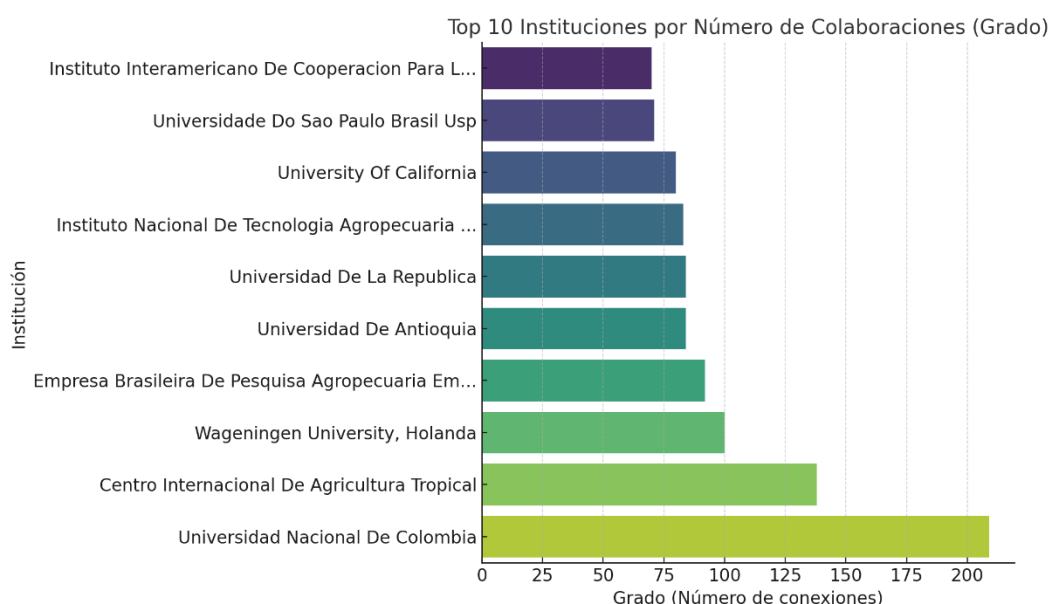


Figura 10: Top 10 de instituciones por número de colaboraciones

La distribución de grados de la red filtrada evidencia una estructura de red altamente heterogénea. Como se observa en la figura 11, la mayoría de los nodos presentan un grado entre 10 y 40, lo cual indica que una parte significativa de las instituciones mantiene colaboraciones con un número moderado de actores. Sin embargo, también se aprecian nodos con grados notablemente superiores (algunos superan los 100), lo cual revela la existencia de hubs o instituciones que actúan como nodos centrales dentro de la red de colaboración.

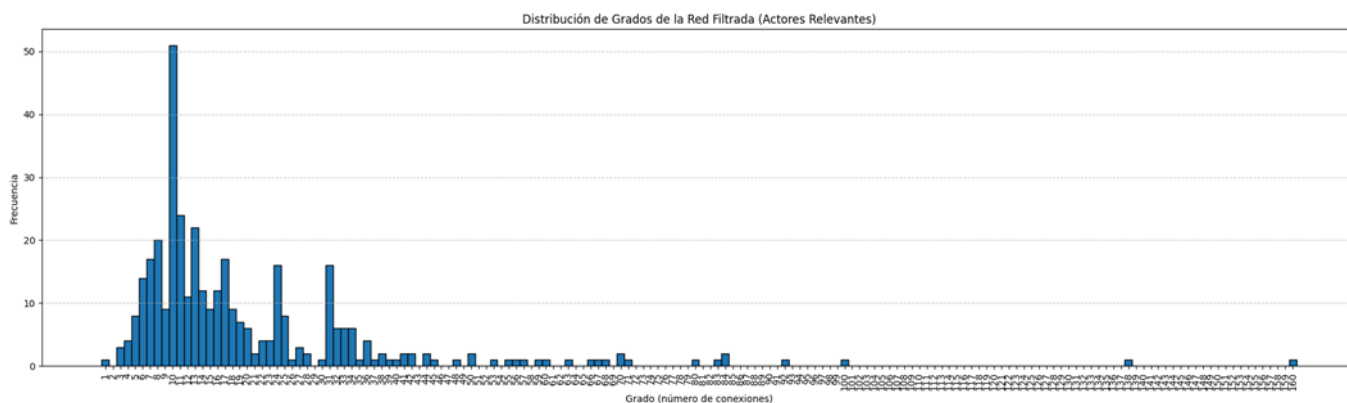


Figura 11: Distribución de grados del grafo

Este comportamiento es característico de redes con propiedades de tipo "libre de escala" (scale-free), en donde unos pocos actores concentran una gran cantidad de conexiones mientras que la mayoría mantiene un número reducido. La presencia de estos supernodos sugiere que existen instituciones con una elevada capacidad de articulación y vinculación interinstitucional, lo que puede ser clave para potenciar procesos de innovación, diseminación de conocimiento y ejecución de proyectos colaborativos de alto impacto.

## 2. Centralidad de intermediación

Nuevamente, la Universidad Nacional de Colombia destaca como el nodo con mayor intermediación (betweenness = 0.185), indicando que actúa como punto de conexión entre diversas comunidades o subredes. Le siguen el CIAT (Colombia), el IICA (Costa Rica) y la Universidad de los Andes (Colombia), los cuales ocupan posiciones clave como facilitadores de flujo de información y coordinación interinstitucional.

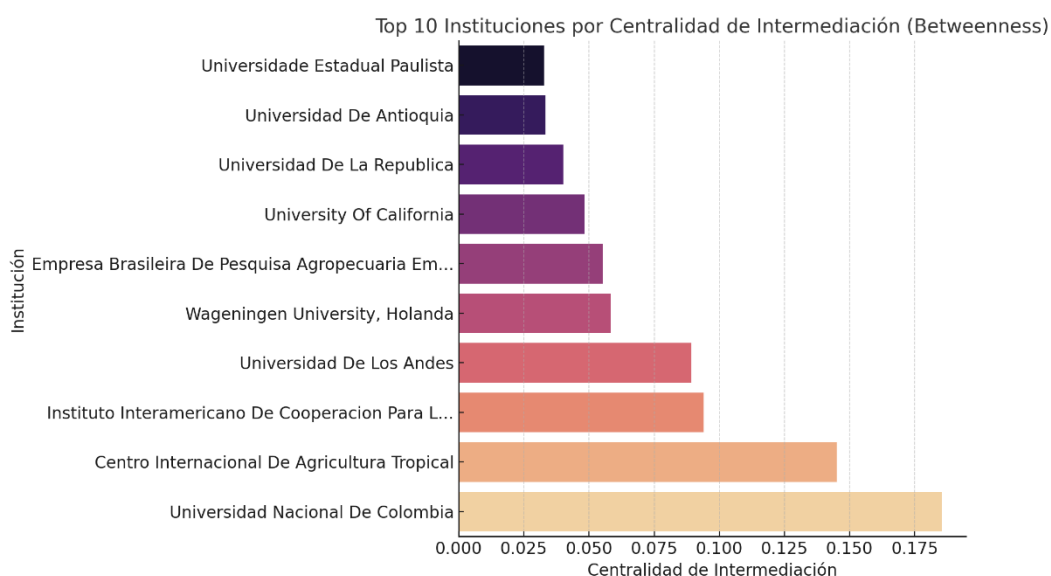


Figura 12: Top 10 de instituciones por centralidad de intermediación

### 3. Centralidad de cercanía

Las instituciones con mayor cercanía fueron aquellas que, además de tener altos grados de conexión, están estratégicamente ubicadas en el centro de la red, permitiéndoles acceder de manera eficiente a una gran parte de la red. Aparte de la Universidad Nacional de Colombia que ocupa el primer lugar, también se incluyen nuevamente a universidades de alto perfil regional e internacional como la Universidad de la República (Uruguay), la Universidade do São Paulo (Brasil) y la University of California.

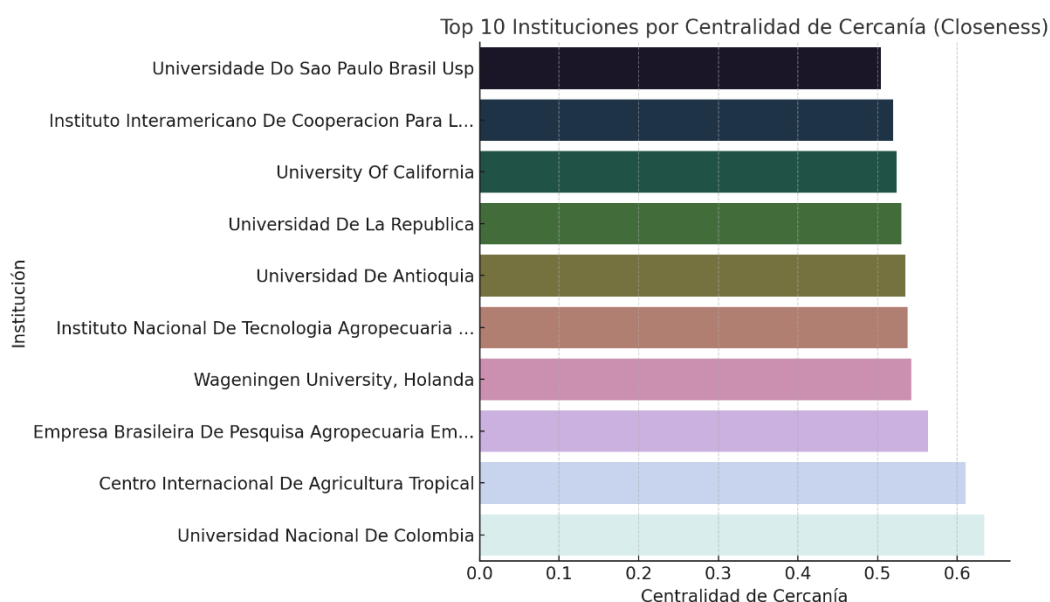


Figura 13: Top 10 Instituciones por centralidad de cercanía

Los resultados muestran una red altamente colaborativa pero heterogénea, con algunos actores centrales que concentran gran parte de las interacciones y otros muchos con posiciones periféricas. La presencia destacada de universidades colombianas, junto con centros internacionales de referencia, evidencia el papel de AGROSAVIA como articulador de conocimiento en redes multiactor y multinivel.

Por su alta intermediación, instituciones como la Universidad Nacional de Colombia y el CIAT no solo son relevantes por su volumen de colaboraciones, sino porque habilitan la conectividad entre entidades que de otro modo no estarían vinculadas, posicionándose como aliados estratégicos en futuros proyectos. Asimismo, la diversidad geográfica de los nodos más centrales —que incluye instituciones de Colombia, Brasil, Países Bajos, Costa Rica, Estados Unidos, Uruguay y Argentina— refuerza el carácter global de la red.

#### 4.2.3. Visualización de la red



Se generó una visualización de la red de colaboración institucional filtrando los nodos con grado mayor o igual a 10, con el fin de resaltar los actores más activos y conectados dentro del ecosistema de alianzas de AGROSAVIA. Esta visualización, compuesta por 372 nodos y 3.722 aristas, permite identificar patrones de interacción, agrupamientos naturales y nodos estratégicos dentro de la red.

Red de Colaboración - Actores Relevantes (Grado  $\geq 10$ )

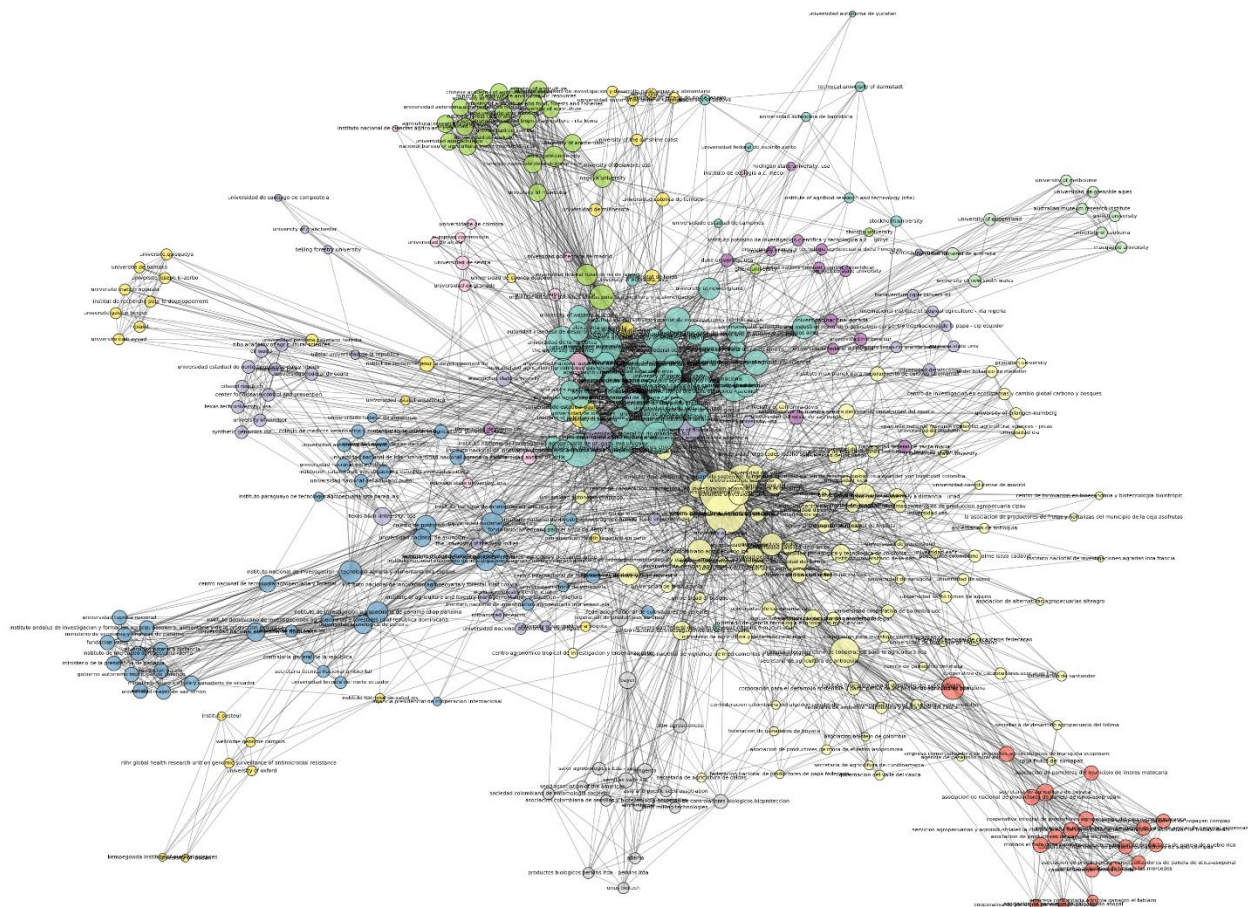


Figura 14: red de colaboración institucional

Analizando la estructura de la red, se identifica que se trata de un grafo no dirigido, ya que las relaciones de colaboración entre instituciones son bidireccionales y recíprocas por naturaleza. Es decir, si dos instituciones participan en una alianza conjunta, la relación es mutua y no se establece jerarquía o dirección entre ellas.

Asimismo, es un grafo ponderado, dado que las aristas representan no solo la existencia de una colaboración, sino también su intensidad o frecuencia, es decir, el número de alianzas compartidas entre dos instituciones (análisis de coocurrencia). Esta ponderación enriquece el análisis al permitir distinguir vínculos más significativos o repetidos.

La representación de la red revela una estructura densa y compleja, con varios nodos concentrados en la parte central que conectan múltiples comunidades de colaboración. En el núcleo de la red se destacan actores como la Universidad Nacional de Colombia, el Centro Internacional de Agricultura Tropical (CIAT) y la Wageningen University, que no solo presentan los grados más altos, sino también posiciones de alta intermediación (betweenness), actuando como nodos puente que conectan distintas regiones o áreas temáticas de la red.

Respecto a las comunidades identificadas mediante el algoritmo de detección de Louvain y representadas por colores en el grafo, se observa una segmentación natural en clústeres o grupos de instituciones con mayor densidad de conexiones internas. Estas comunidades pueden estar relacionadas con factores como:

- Proximidad geográfica o país de origen.
- Afinidad temática o tipo de investigación.
- Naturaleza institucional (universidades, centros de investigación, organismos internacionales, etc.).

La visualización evidencia también la existencia de subredes especializadas, algunas más periféricas pero cohesionadas, que podrían representar redes regionales o ámbitos de investigación específicos. El análisis de estas comunidades permite no solo identificar los nodos más influyentes dentro de cada grupo, sino también detectar oportunidades para fortalecer la integración entre comunidades mediante actores puente.

A través de la interfaz interactiva desarrollada con ipywidgets, que permite visualizar automáticamente subgrafos filtrados a partir de criterios seleccionados por el usuario, se realizó un análisis de los subgrafos por tipo de alianza:

## 1. Vínculos negociales

El subgrafo filtrado por el tipo de alianza vínculo negocial, revela una estructura de colaboración que destaca por su densidad moderada y una distribución centralizada en torno a algunos actores clave. En esta visualización, se identifican instituciones con una alta centralidad de grado, como el Instituto Interamericano de Cooperación para la Agricultura (IICA) y el Instituto Financiero para el Desarrollo del Valle del Cauca (INFIVALLE), que actúan como articuladores de múltiples relaciones institucionales en el contexto de alianzas con propósito principalmente contractual o transaccional.

Se observa una red compuesta por una mezcla de entidades públicas, universidades, organizaciones de productores, cooperativas, agencias de desarrollo y entes territoriales. La presencia significativa de asociaciones de productores, cooperativas paneleras, gremios agrícolas y secretarías departamentales de agricultura indica que el vínculo negocial es especialmente útil para estructurar proyectos de fortalecimiento productivo, gestión de recursos y cooperación técnica para el desarrollo agropecuario a nivel territorial.



Otro hallazgo importante es la existencia de comunidades temáticamente cohesionadas, especialmente alrededor de sectores como el sector panelero, la cadena de cacao y el desarrollo rural sostenible, lo cual se refleja en los agrupamientos de nodos con alta conectividad entre sí y menor conectividad con el resto de la red. Esta segmentación puede estar indicando alianzas especializadas o basadas en la localización geográfica de los actores.

Red de colaboración filtrada

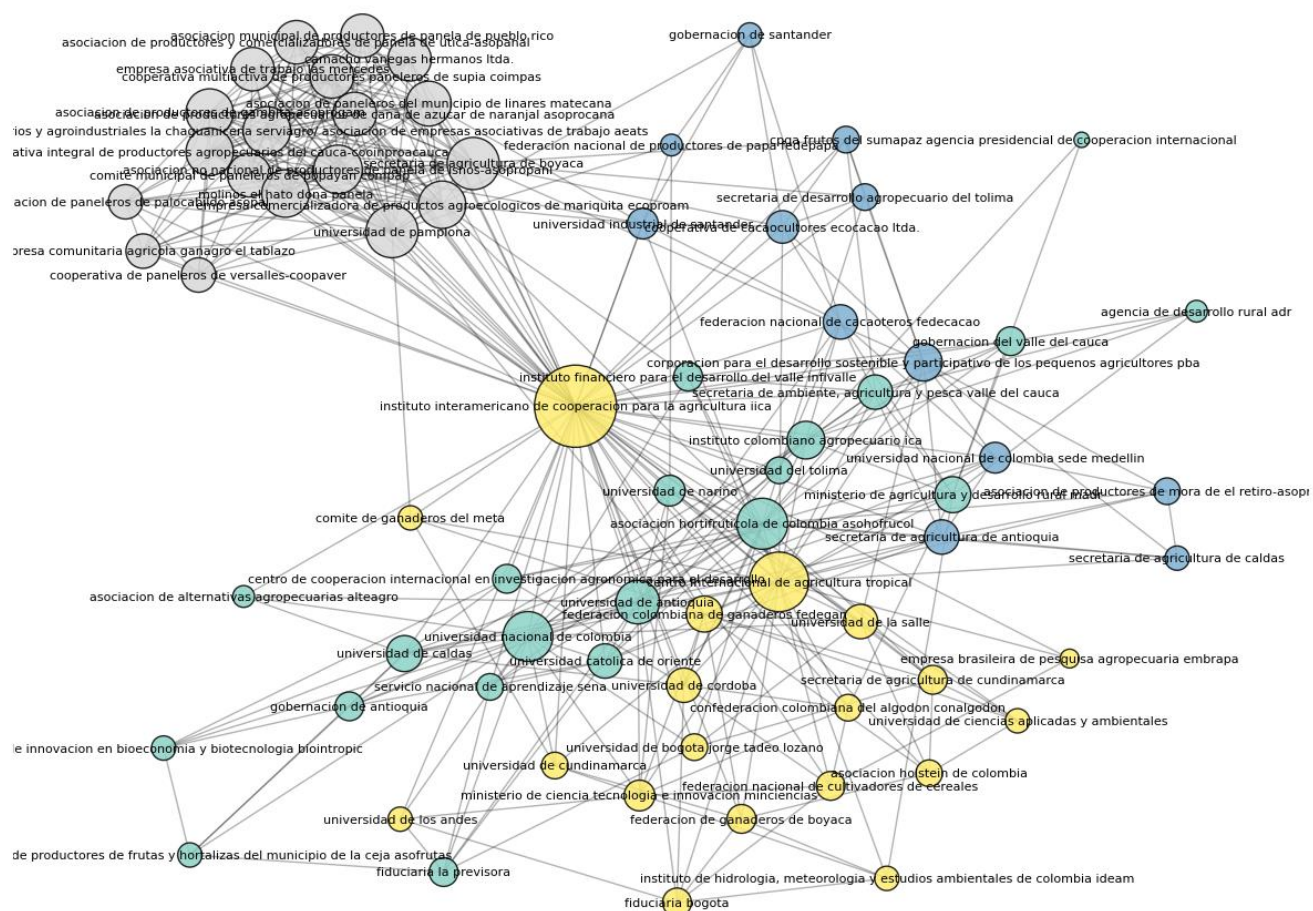


Figura 15: Subgrafo vínculos negociales

El análisis también muestra que, aunque existen algunos nodos periféricos con baja conectividad, estos tienden a estar vinculados a través de al menos una institución central, lo que sugiere una buena cohesión interna y potencial para escalar colaboraciones en futuras iniciativas de valor agregado o de fortalecimiento de capacidades.

## 2. Gestión del conocimiento

### Red de colaboración filtrada



Uno de los aspectos más destacados es la presencia de múltiples comunidades detectadas mediante el algoritmo de Louvain, las cuales están claramente representadas por agrupaciones cromáticas en el grafo. Estas comunidades parecen responder tanto a afinidades geográficas como temáticas. Por ejemplo, se identifican conglomerados bien diferenciados que agrupan a universidades y centros de investigación de América Latina, África Occidental, Asia y Europa, lo cual evidencia la diversidad global de los actores que colaboran en la producción, intercambio y aplicación del conocimiento agropecuario.

Entre los nodos más centrales, se destacan instituciones como el Centro Internacional de Mejoramiento de Maíz y Trigo (CIMMYT), el International Institute of Tropical Agriculture (IITA),



universidades como Wageningen University, University of California, y Universidad Nacional de Colombia, así como agencias de cooperación técnica como el NIHR Global Health Research Unit, entre otros. Estas instituciones actúan como nodos pivote que conectan comunidades diversas y promueven la articulación entre múltiples actores.

El subgrafo muestra una proporción considerable de enlaces entre actores de diferentes regiones, lo que indica un alto nivel de colaboración transnacional, coherente con el objetivo de este tipo de alianzas: la gestión, producción y circulación del conocimiento científico y técnico para el fortalecimiento de capacidades institucionales y la innovación colaborativa.

### 3. Espacio para la Construcción de Alianzas

El subgrafo filtrado por el tipo de alianza espacio para la construcción de alianzas representa las relaciones establecidas o fortalecidas a partir de la participación conjunta en eventos científicos, congresos, foros y encuentros académicos. La red resultante evidencia una estructura densamente conectada, con una notable diversidad institucional y una fuerte presencia de universidades, centros de investigación, ministerios, agencias técnicas y organizaciones de productores.

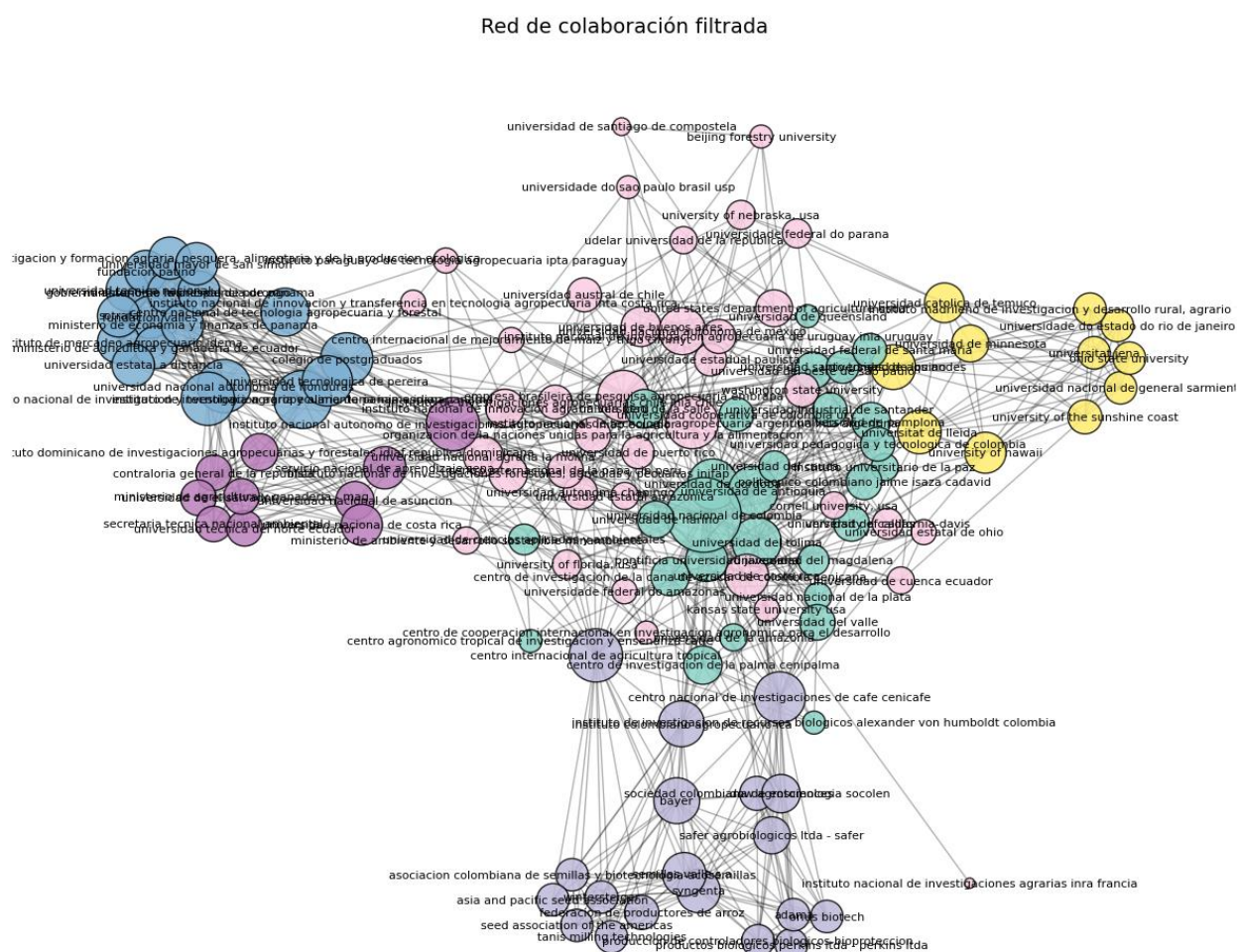


Figura 17: Subgrafo espacio para la construcción de alianzas

Desde el punto de vista topológico, la red presenta una alta conectividad global con nodos que funcionan como puentes interregionales. Instituciones como la Universidad Nacional de Colombia, el Centro Nacional de Investigaciones de Café (CENICAFÉ), y el Centro Internacional de Agricultura Tropical (CIAT) destacan como actores clave que articulan múltiples comunidades. Esto es reflejo de su activa participación en espacios colaborativos abiertos, y de su capacidad para establecer vínculos bilaterales y multilaterales de forma constante.

Se observan comunidades identificables, agrupadas tanto por afinidad temática como por proximidad geográfica. Existen clústeres claramente formados por instituciones de países andinos, del Caribe, América Central, y redes más internacionales que incluyen actores de Europa, Asia y Oceanía, por ejemplo, un grupo claramente diferenciable corresponde al clúster del Caribe y América Central, conformado por instituciones como el Instituto Dominicano de Investigaciones Agropecuarias y Forestales (IDIAF), el Ministerio de Agricultura de República Dominicana, el Ministerio de Agricultura y Ganadería de Costa Rica, el INTA de Costa Rica, y la Universidad de Puerto Rico. Este patrón de agrupamiento sugiere que los eventos científicos funcionan como entornos eficaces para promover alianzas entre instituciones que no comparten necesariamente vínculos operativos o proyectos específicos, pero que se acercan a través del diálogo técnico, la divulgación y la colaboración informal.

Otra característica destacable del grafo es la presencia de nodos periféricos que, a pesar de tener un bajo grado de conexión, participan de la red a través de relaciones puntuales con actores centrales. Esto sugiere una alta potencialidad para ampliar la red y fomentar su densificación futura si estas relaciones ocasionales se transforman en alianzas formales o proyectos conjuntos.

#### 4. Super Aliados y Aliados Clave

Como resultado del análisis estructural de la red de colaboración de AGROSAVIA, se identificaron los Top 10 de Super Aliados, definidos como aquellas instituciones que presentan una alta métrica de grado y que participan de manera activa en los tres tipos de alianza estratégicos establecidos en este estudio: vínculo negocial, gestión del conocimiento y espacios para la construcción de alianzas. Esta categoría destaca a los actores con mayor integración y transversalidad en la red, dado que no solo mantienen un volumen significativo de colaboraciones, sino que además diversifican su rol en distintos escenarios de interacción institucional.

Nº	Nombre de la Institución	Betweenness	Grado	Tipos de alianza
1	Universidad Nacional de Colombia	209	290	Gestión del conocimiento, Vínculo negocial, Construcción de alianzas
2	Centro Internacional de Agricultura Tropical (CIAT)	15	288	Gestión del conocimiento, Vínculo negocial, Construcción de alianzas
3	Empresa Brasileira de Pesquisa Agropecuária (EMBRAPA)	8	138	Gestión del conocimiento, Vínculo negocial, Construcción de alianzas
4	Universidad de Antioquia	62	124	Gestión del conocimiento, Vínculo negocial, Construcción de alianzas

5	Instituto Nacional de Tecnología Agropecuaria (INTA), Argentina	1066	110	Gestión del conocimiento, Vínculo comercial, Construcción de alianzas
6	Universidade de São Paulo (USP), Brasil	2659	93	Gestión del conocimiento, Vínculo comercial, Construcción de alianzas
7	Universidad de los Andes	555	90	Gestión del conocimiento, Vínculo comercial, Construcción de alianzas
8	Universidade Estadual Paulista (UNESP), Brasil	4336	88	Gestión del conocimiento, Vínculo comercial, Construcción de alianzas
9	United States Department of Agriculture (USDA)	169	85	Gestión del conocimiento, Vínculo comercial, Construcción de alianzas
10	Instituto Colombiano Agropecuario (ICA)	6	79	Gestión del conocimiento, Vínculo comercial, Construcción de alianzas

Tabla 9: Top 10 Super aliados

Entre los principales Super Aliados se encuentran instituciones nacionales como la Universidad Nacional de Colombia, el Centro Internacional de Agricultura Tropical (CIAT), la Universidad de Antioquia, y la Universidad de los Andes, que sobresalen tanto por su posicionamiento central como por su versatilidad institucional. A estos se suman entidades internacionales de gran trayectoria como la Empresa Brasileira de Pesquisa Agropecuária (EMBRAPA), el Instituto Nacional de Tecnología Agropecuaria de Argentina (INTA), la Universidade de São Paulo y el Departamento de Agricultura de Estados Unidos (USDA), consolidando así una red de actores de alto nivel científico, técnico y estratégico.

El hecho de que estos Super Aliados participen simultáneamente en alianzas operativas, académicas y de relacionamiento informal indica una alta capacidad de articulación interinstitucional, lo que los convierte en actores fundamentales para la sostenibilidad de la red y la generación de impactos compartidos. Además, sus altos valores en grado (entre 79 y 290 conexiones) los posicionan como puentes entre comunidades dentro de la red de colaboración, facilitando el flujo de información, conocimiento y recursos.

El subgrafo de los Super Aliados representa las interacciones entre las diez instituciones más influyentes de la red de colaboración de AGROSAVIA, este subconjunto de actores forma una estructura altamente cohesionada, donde prácticamente todos los nodos están interconectados, configurando una red casi completa (clique).

Desde una perspectiva estructural, la topología del grafo refleja un entorno colaborativo de élite, donde las instituciones no solo poseen múltiples conexiones, sino que también se relacionan directamente entre sí. Esto indica una alta densidad relacional, lo cual es una señal de confianza mutua, colaboración sostenida y complementariedad en capacidades científicas, tecnológicas y estratégicas.

Además, el grafo pone en evidencia que estos Super Aliados no solo actúan como hubs individuales, sino que forman una comunidad cohesiva, maximizando el flujo de información, la eficiencia de proyectos compartidos y la posibilidad de escalar soluciones a nivel regional o global.

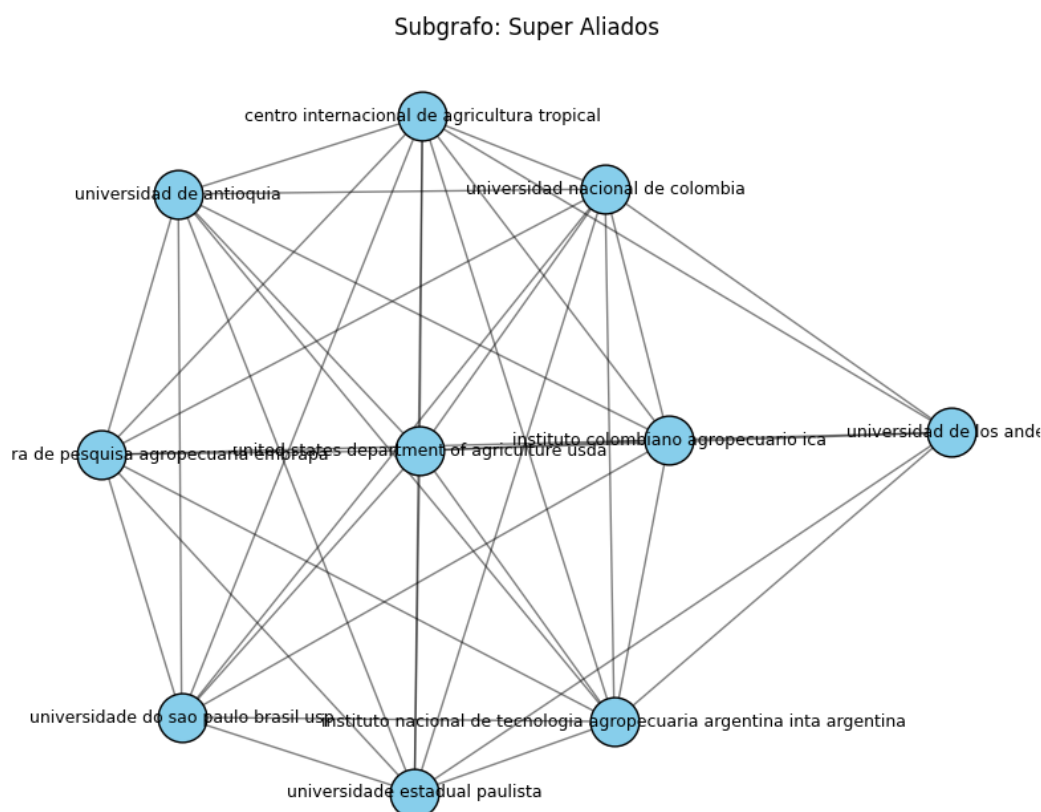


Figura 18: Subgrafo Super Aliados

Además de los Super Aliados, se identificó un conjunto de instituciones catalogadas como aliados clave, definidos como aquellos actores que, además de mantener un alto grado de conectividad y centralidad en la red, participan activamente en al menos dos de los tres tipos de alianza estratégica establecidos en este estudio.

Nº	Nombre de la Institución	Betweenness	Grado	Tipos de alianza
1	Instituto Interamericano de Cooperación para la Agricultura (IICA)	17	195	Vínculo negocial, Espacio para la construcción de alianzas
2	Wageningen University, Países Bajos	795	107	Gestión del conocimiento, Espacio para la construcción de alianzas
3	University of California, EE. UU.	788	101	Gestión del conocimiento, Espacio para la construcción de alianzas
4	Universidad de Buenos Aires, Argentina	2734	80	Gestión del conocimiento, Espacio para la construcción de alianzas
5	Swedish University of Agricultural Sciences, Suecia	5264	74	Gestión del conocimiento, Espacio para la construcción de alianzas
6	Asociación Hortifrutícola de Colombia (ASOHOFRUCOL)	108	73	Vínculo negocial, Espacio para la construcción de alianzas
7	Ohio State University, EE. UU.	4568	64	Gestión del conocimiento, Espacio para la construcción de alianzas

8	Pennsylvania State University, EE.UU.	780	60	Gestión del conocimiento, Espacio para la construcción de alianzas
9	Universidade Federal de Viçosa, Brasil	1127	57	Gestión del conocimiento, Espacio para la construcción de alianzas
10	Commonwealth Scientific and Industrial Research Organisation (CSIRO), AU	771	49	Gestión del conocimiento, Espacio para la construcción de alianzas

Tabla 10: Top 10 aliados clave

El análisis revela que estos aliados clave desempeñan un rol intermedio entre los Super Aliados y los actores periféricos, conectando comunidades diversas y facilitando el flujo de conocimiento, recursos y oportunidades colaborativas en contextos estratégicos. Instituciones como el Instituto Interamericano de Cooperación para la Agricultura (IICA), Wageningen University (Países Bajos) y la University of California se destacan por su capacidad de conexión transnacional y su participación tanto en entornos formales (como convenios) como en espacios académicos y científicos.

También figuran organizaciones del sector agroempresarial y gremial como ASOHOFRUCOL (Colombia), que aparece como una de las pocas organizaciones productivas dentro del listado, destacando su papel en el fortalecimiento de alianzas de tipo negocial y técnico. De igual forma, universidades e instituciones líderes en investigación como la Swedish University of Agricultural Sciences, la Universidad de Buenos Aires, la Ohio State University, y la Pennsylvania State University reflejan una fuerte vocación colaborativa que trasciende fronteras y temáticas.

Subgrafo: Aliados Clave

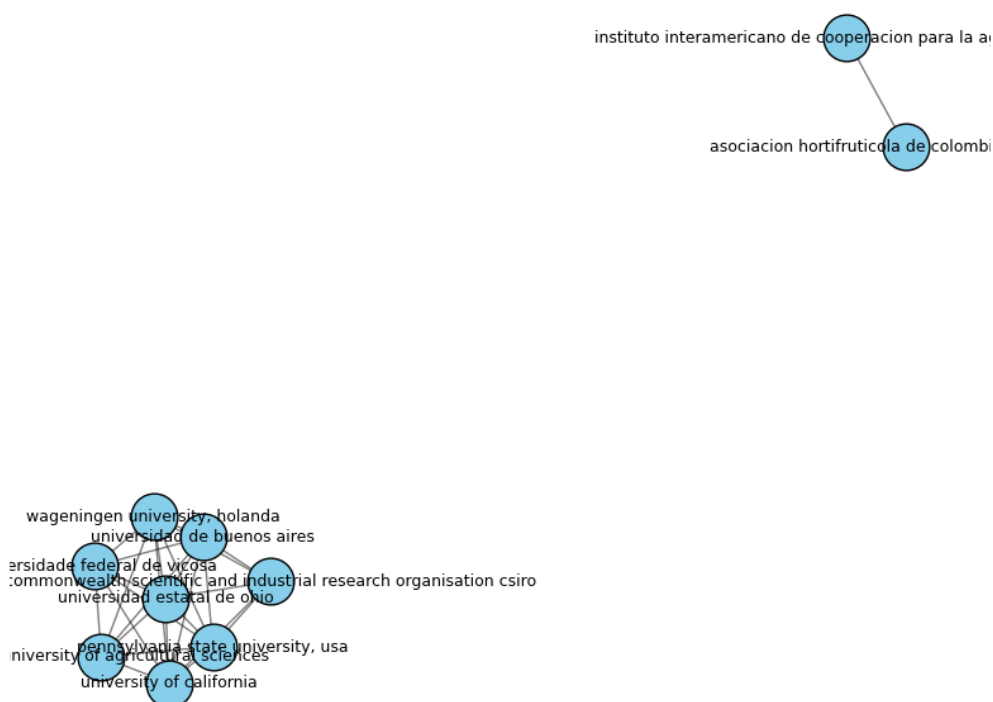


Figura 19: Subgrafo aliados clave



El subgrafo de los aliados clave, si bien no presentan el mismo nivel de centralidad e interacción que los Super Aliados, muestra que se divide principalmente en dos componentes:

- Un componente central denso, conformado por universidades y centros de investigación, estas instituciones mantienen entre sí un alto grado de conectividad, generando una subred colaborativa transnacional robusta, enfocada principalmente en investigación científica, publicaciones conjuntas y desarrollo tecnológico agropecuario.
- Un componente aislado, que incluye al Instituto Interamericano de Cooperación para la Agricultura (IICA) y ASOHOFRUCOL (Asociación Hortifrutícola de Colombia), aunque este par presenta una conexión directa entre sí, no está vinculado al resto del subgrafo, lo cual sugiere un nodo estratégico en el contexto latinoamericano, pero con baja integración a la red internacional principal de colaboración. Esta situación podría representar una oportunidad de mejora en la articulación, especialmente en actividades que vinculen a instituciones productivas con actores académicos y científicos.

#### 4.2.4. Tablero de Análisis Interactivo para la Visualización de Relacionamientos

Como parte del proceso de análisis exploratorio y visualización de redes de colaboración, se desarrolló un tablero interactivo en Power BI que permite examinar de forma detallada los relacionamientos establecidos por AGROSAVIA. Este tablero integra múltiples visualizaciones dinámicas que facilitan la interpretación de datos, apoyando la toma de decisiones estratégicas en torno a la gestión de alianzas.

Entre los elementos más relevantes se encuentra un contador total de alianzas, que revela la magnitud del universo colaborativo analizado (5.972 relaciones en total). Además, se incorporaron filtros por año y estado del relacionamiento, lo que permite realizar análisis evolutivos o situacionales según los intereses del usuario.

El tablero incluye visualizaciones clave como:

- Distribución de los tipos de relacionamientos, destacando que el vínculo comercial y la gestión del conocimiento son los formatos predominantes.
- Evolución temporal de las alianzas, la cual permite evidenciar picos de colaboración institucional en años específicos.
- Un mapa geográfico interactivo que muestra la localización global de los aliados, permitiendo identificar zonas de concentración regional.
- Rankings del Top 10 de aliados por tipo de relación, mostrando instituciones como el CIAT y la Universidad Nacional de Colombia con alta frecuencia de participación.

- Un análisis del Top 10 de instituciones con mayor capacidad de generar relacionamientos y su contribución a la red general.
- Gráfico de alianzas por tipo de aliado, útil para entender la diversidad institucional que participa en las redes de AGROSAVIA.

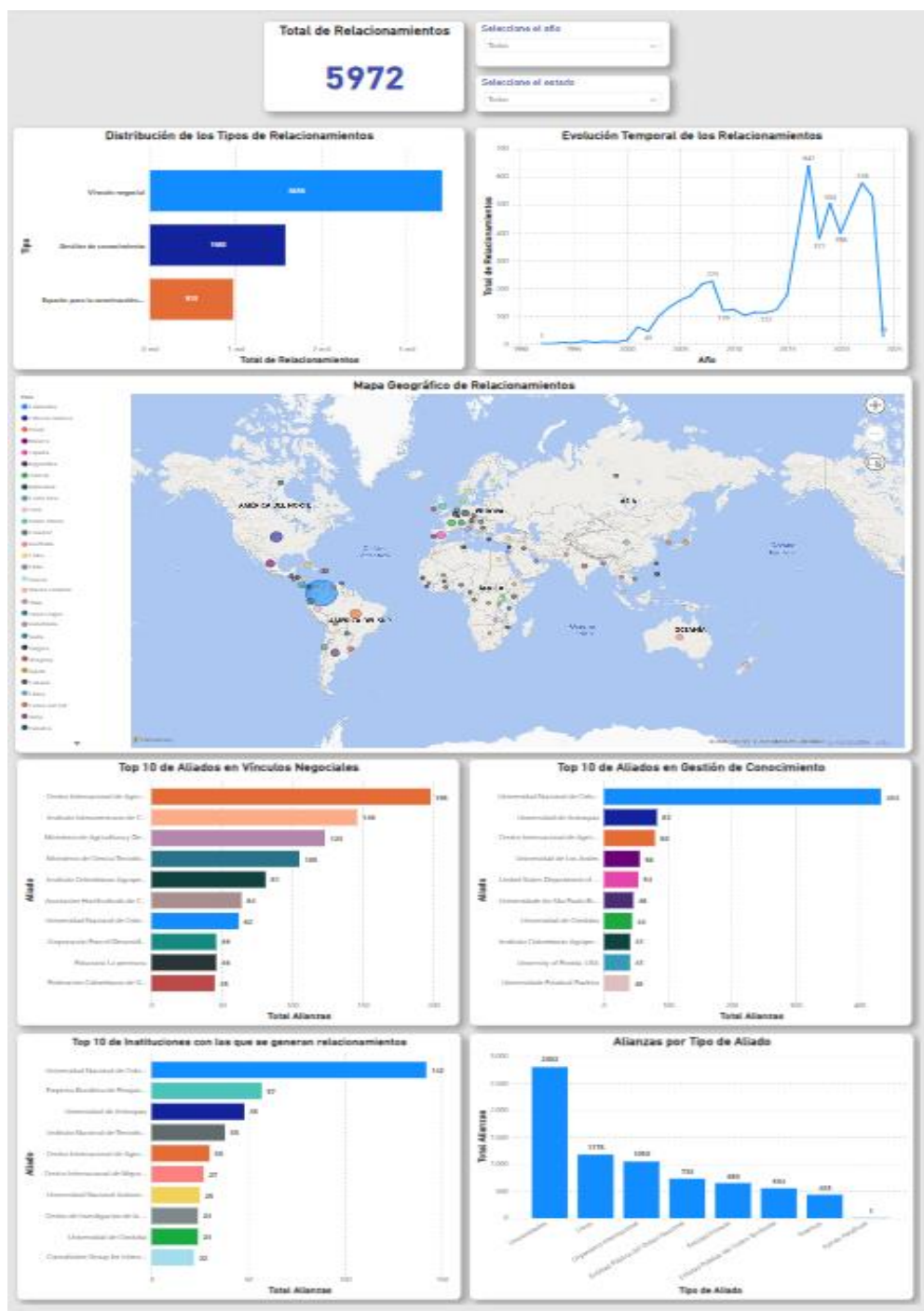


Figura 20: Visualización de Relacionamientos

Este tablero no solo complementa los análisis computacionales y de predicción realizados, sino que también funciona como un instrumento de inteligencia estratégica institucional, ofreciendo una forma amigable, interactiva y útil de comunicar los hallazgos a diversos públicos interesados.

N.º	Producto	Enlace
1	Notebook TFM – Código para la generación de grafos y demás funcionalidades	<a href="#">Enlace Notebook</a>
2	Red de colaboración institucional	<a href="#">Enlace Red</a>
3	Tablero de control relacionamientos	<a href="#">Enlace Tablero</a>
4	Nodos red de colaboración	<a href="#">Enlace nodos</a>
5	Aristas red de colaboración	<a href="#">Enlace Aristas</a>

Tabla 11: Productos Análisis exploratorio y visualización de redes de colaboración

## 4.3. Desarrollo de Modelos Predictivos

### 4.3.1. Análisis de redes bipartitas, técnicas de Link Prediction y métodos de filtrado colaborativo adaptado

Para realizar la validación del modelo de recomendación institucional basado en análisis de redes, se implementó una prueba seleccionando como unidad organizacional base a la Universidad Nacional de Colombia, reconocida por su papel central en la red de colaboración de AGROSAVIA como Super Aliado.

#### 1. Frecuencia histórica de co-participación en proyectos

El modelo, calcula el número de veces que otras instituciones han coincidido en proyectos donde también ha participado la institución base, y con base en esto genera un score de recomendación. Este score no refleja una métrica de similitud matemática, sino la intensidad del relacionamiento previo, proporcionando una visión del historial colaborativo.

Nº	Nombre de la Institución	ID Institución	Score
1	Centro Internacional de Agricultura Tropical (CIAT)	15	39
2	Universidad de Antioquia	62	24
3	Universidad de Caldas	432	13
4	Pontificia Universidad Javeriana	620	12
5	Universidad Pedagógica y Tecnológica de Colombia (UPTC)	617	12
6	Instituto Colombiano Agropecuario (ICA)	6	11
7	Universidad de La Salle	196	11

8	Universidad de Nariño	318	10
9	Universidad del Valle	583	10
10	United States Department of Agriculture (USDA)	169	8

Tabla 12: Co-partición en alianzas - Institución base Universidad Nacional de Colombia

Los resultados muestran un conjunto de instituciones con altos niveles de interacción histórica con la Universidad Nacional. En primer lugar, el Centro Internacional de Agricultura Tropical (CIAT) aparece con un score de 39 proyectos compartidos, seguido por la Universidad de Antioquia (24), la Universidad de Caldas (13), y otras instituciones relevantes como la Pontificia Universidad Javeriana, la Universidad Pedagógica y Tecnológica de Colombia (UPTC), y el Instituto Colombiano Agropecuario (ICA). También destaca la inclusión del USDA (United States Department of Agriculture) con 8 proyectos compartidos, lo cual evidencia vínculos previos de carácter internacional.

Además del puntaje de co-participación, el sistema presenta los identificadores de los proyectos compartidos, lo cual permite contextualizar la relación entre las instituciones y orientar futuras decisiones colaborativas con base en antecedentes concretos.

## 2. Similitud Coseno

Se implementó un segundo modelo de recomendación utilizando similitud coseno aplicada sobre una matriz institucional de co-participación en alianzas históricas. A diferencia del enfoque anterior basado en frecuencia absoluta de co-participación, esta metodología calcula el grado de semejanza entre los vectores institucionales, es decir, cuán similares son los patrones de colaboración de cada institución con respecto a la entidad base, en este caso la Universidad Nacional de Colombia.

Nº	Nombre de la Institución	ID Institución	Score
1	Centro Internacional de Agricultura Tropical (CIAT)	15	0.087979
2	Universidad de Caldas	432	0.084612
3	Universidad de Zaragoza (España)	354	0.079181
4	Humboldt University (Alemania)	5088	0.079181
5	Universidad de Antioquia	62	0.078103
6	Pontificia Universidad Javeriana	620	0.067187
7	Universidad de Costa Rica	2700	0.065984
8	Agroscope Liebefeld-Posieux (Suiza)	4948	0.064651
9	Universidad Pedagógica y Tecnológica de Colombia (UPTC)	617	0.064060
10	Universidad de Queensland (Australia)	4196	0.061333

Tabla 13: Predicción de aliados por similitud de coseno - Institución base Universidad Nacional de Colombia

Los resultados obtenidos evidencian que este enfoque identifica no solo instituciones con alta frecuencia de colaboración directa, sino también aquellas que, aunque no necesariamente han participado en un gran número de proyectos con la entidad base, comparten estructuras de relacionamiento similares dentro de la red global de alianzas. En los primeros lugares del ranking aparecen actores como el Centro Internacional de Agricultura Tropical (CIAT), con una alta coincidencia tanto en similitud de coseno como en proyectos compartidos (0.087), seguido por

instituciones con patrones afines como la Universidad de Caldas, la Universidad de Zaragoza (España), Humboldt University (Alemania), y Agroscope (Suiza).

Una observación clave es que este modelo destaca aliados internacionales con menor frecuencia absoluta de co-participación, pero con perfiles colaborativos similares. Por ejemplo, entidades como la Universidad de Queensland (Australia), la Universidad de Costa Rica, y la Universidad de Zaragoza aparecen en posiciones prioritarias, lo cual sugiere una ampliación del radar estratégico más allá de los vínculos históricos inmediatos.

### 3. Índice de Adamic-Adar

En la tercera estrategia evaluada para la recomendación de aliados estratégicos, se implementó un modelo de predicción de enlaces utilizando el índice de Adamic-Adar, una métrica de teoría de grafos que asigna puntuaciones más altas a pares de nodos que comparten muchos vecinos, penalizando aquellos vecinos comunes con alta conectividad. Esta aproximación es particularmente útil para identificar conexiones potenciales no observadas aún, pero con fuerte evidencia estructural en la red de colaboración.

En esta prueba, aplicada desde la perspectiva de la Universidad Nacional de Colombia, los resultados muestran un conjunto distinto de instituciones sugeridas, en comparación con los métodos anteriores. Encabeza la lista la Corporación Observatorio del Caribe Colombiano, seguida de la Universidad Tecnológica de Bolívar, la Universidad de Bonn (Alemania), y diversas asociaciones, empresas privadas y entidades sin ánimo de lucro como Pizano S.A., Diligent Energy Systems B.V., y ASOINTO (Asociación Integral de Tomateros).

Nº	Nombre de la Institución	ID Institución	Score
1	Corporación Observatorio del Caribe Colombiano	5028	40.582565
2	Universidad Tecnológica de Bolívar	2935	38.744925
3	Universidad de Bonn	806	38.485786
4	Acuerdo con Productor	3845	38.130330
5	Instituto de Ciencia y Tecnología Alimentaria - ICTA	720	38.094326
6	Pizano S.A	316	37.858431
7	Diligent Energy Systems B.V.	312	37.674543
8	Asociación Integral de Tomateros - ASOINTO	500	37.500124
9	Asociación Municipal de Usuarios Campesinos de Colombia	499	37.500124
10	Asociación Colombiana de Pensionados de INRAVID	590	37.343677

Tabla 14: Predicción de aliados por Índice de Adamic-Adar - Institución base Universidad Nacional de Colombia

Una diferencia clave con respecto al modelo de frecuencia de co-participación y al de similitud coseno, es que el modelo de Adamic-Adar no requiere un historial directo de proyectos compartidos. En su lugar, la recomendación se basa en la existencia de vecinos comunes con la institución base dentro del grafo general. Es decir, se privilegia la posibilidad de colaboración futura sobre la base de contextos relacionales compartidos, más que de antecedentes colaborativos concretos.

Este enfoque tiene la capacidad de detectar aliados emergentes o infrautilizados, que operan en redes similares o comparten socios estratégicos, pero con los que aún no se ha establecido una

relación directa. En este sentido, el modelo de Adamic-Adar complementa a los anteriores al ofrecer una visión más predictiva y exploratoria de las alianzas institucionales.

## 4.3.2. Integración de Embeddings

### 1. Factorización SVD

En esta etapa se aplicó un enfoque basado en la técnica de factorización matricial mediante SVD (Singular Value Decomposition), a partir de la matriz bipartita Alianzas x Instituciones, con el fin de generar una representación latente de las instituciones colaboradoras en un espacio vectorial reducido. Este método permite capturar relaciones implícitas de co-participación entre entidades, incluso si no han colaborado directamente en el pasado, identificando instituciones con perfiles colaborativos similares.

Al seleccionar como institución base a la Universidad Nacional de Colombia, el modelo arrojó un ranking de instituciones con mayor similitud estructural en el espacio latente.

Los valores altos de similitud coseno (todos mayores a 0.99) evidencian una fuerte coincidencia en los patrones de colaboración de estas instituciones con respecto a la institución base, lo que sugiere que comparten proyectos, áreas temáticas o estructuras de cooperación similares.

Nº	Nombre de la Institución	ID Institución	Similitud Coseno
1	Universidad del Pacífico	552	0.994774
2	Red Interinstitucional de Cambio Climático y Seguridad Alimentaria	679	0.994450
3	Universitat de Lleida	2678	0.994431
4	Asociación AGROPEBEL	4778	0.994162
5	Universidad del Quindío	2043	0.993903
6	Asociación de Productores de Frutas y Hortalizas del Norte de Nariño	3882	0.993803
7	Instituto Biológico Campinas	5105	0.993790
8	Gobernación de Caldas	165	0.993595
9	Universidad de Costa Rica	2700	0.993146
10	Cancer Research Center	4988	0.990000

Tabla 15: Predicción de aliados por factorización SVD – Institución base Universidad Nacional de Colombia

A diferencia de otros enfoques, como el basado en la frecuencia de co-participación o el índice de Adamic-Adar, que se enfocan en vínculos directos o vecinos comunes, la factorización SVD permite identificar instituciones con similitud estructural latente en sus patrones de colaboración, incluso si no han trabajado juntas previamente. Este método ofrece una visión más global y robusta ante datos dispersos, y al reducir el ruido y la esparsidad de la matriz de relaciones, optimiza la precisión de las recomendaciones, superando también las limitaciones del filtrado colaborativo clásico.



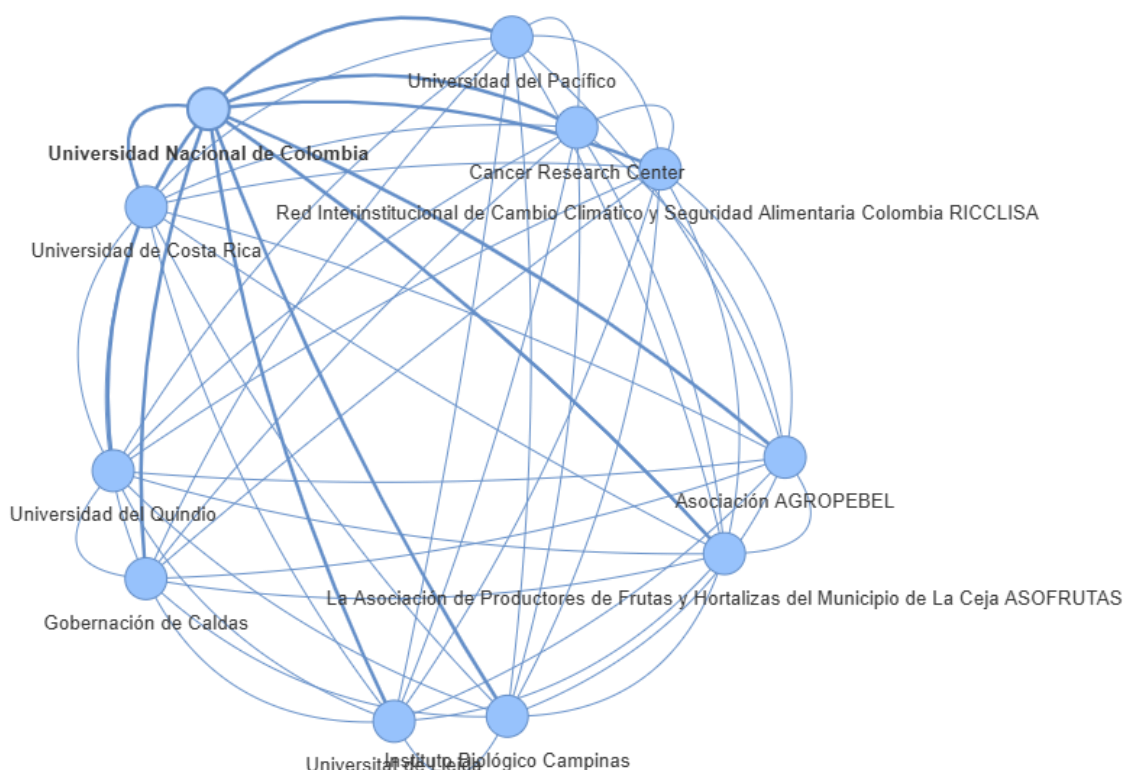


Figura 21: Subgrafo Predicción de aliados por factorización SVD – Institución base Universidad Nacional de Colombia

## 2. Co-ocurrencia Word2Vec

Se implementó el modelo de Word2Vec para capturar relaciones latentes entre instituciones a partir de su co-ocurrencia en alianzas previas. A diferencia de métodos como la factorización SVD que modelan relaciones globales en una matriz densa de colaboración, Word2Vec se basa en secuencias locales de participación, tratando a las instituciones como "palabras" en el contexto de un "documento" (alianza o proyecto). Esto permite identificar entidades con patrones similares de colaboración aunque no hayan trabajado juntas directamente.

Nº	Nombre de la Institución	ID Institución	Similitud Coseno
0	Universidade Federal do Rio Grande Do Sul	2673	0.966783
1	Centro Nacional de Investigaciones de Café CEN...	631	0.929404
2	Instituto Nacional de Tecnología Agropecuaria ...	1066	0.924454
3	United States Department of Agriculture USDA	169	0.923796
4	Universidad Pública de Navarra	419	0.915442
5	Swedish University of Agricultural Sciences	5264	0.905676
6	Centro Internacional de Agricultura Tropical	15	0.900364



7	Universidad de Pamplona	353	0.898621
8	Asociación Hortifrutícola de Colombia ASOHOFRUCOL	108	0.897870
9	Empresa Brasileira de Pesquisa Agropecuaria EM...	8	0.895867

Tabla 16: Predicción de aliados por co-ocurrencia Word2Vec - Institución base Universidad Nacional de Colombia

Entre las instituciones más cercanas a la Universidad Nacional de Colombia se destacan la Universidade Federal do Rio Grande do Sul (similitud 0.967), el Centro Nacional de Investigaciones de Café - CENICAFÉ (0.929) y el Instituto Nacional de Tecnología Agropecuaria de Argentina (0.924). Este tipo de modelo resulta particularmente útil para predecir colaboraciones emergentes con base en patrones de comportamiento colaborativo similares, incluso en contextos internacionales.

A diferencia del enfoque de score por co-participación histórica, Word2Vec no requiere vínculos directos previos, lo que permite generar recomendaciones más exploratorias. Frente al método Adamic-Adar, que valora la especificidad de los vínculos indirectos, Word2Vec captura patrones generales sin necesidad de alta conectividad. Comparado con SVD, Word2Vec ofrece una alternativa más sensible a relaciones locales y específicas de contexto.

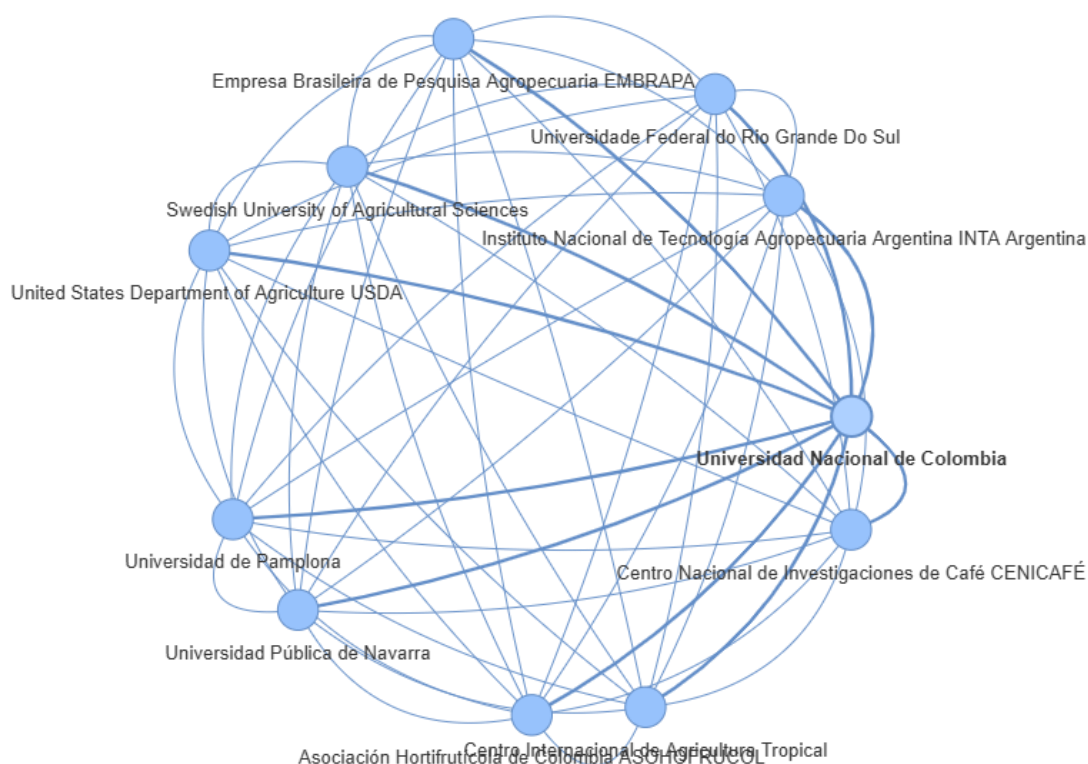


Figura 22: Subgrafo predicción de aliados por co-ocurrencia Word2Vec - Institución base Universidad Nacional de Colombia

### 3. Similitud Semántica de Texto

Se implementó un sistema de recomendación de aliados fundamentado en similitud semántica de texto. A diferencia de otros métodos basados en la estructura de la red o en la co-participación histórica, este enfoque permite ingresar de forma libre una idea de proyecto, y obtener recomendaciones de instituciones aliadas cuyas descripciones de colaboración previas presenten alta similitud semántica con dicho texto.

Para lograrlo, se integraron modelos de lenguaje como SentenceTransformer con la arquitectura BGE-M3, capaces de generar representaciones vectoriales del campo `ali_nombre` del dataset original. Posteriormente, se aplicó la similitud de coseno para calcular el grado de cercanía semántica entre la idea ingresada y cada uno de los títulos de alianzas históricas.

En la prueba realizada con la consulta “sostenibilidad agropecuaria”, el sistema recomendó aliados estratégicos altamente pertinentes. Entre los primeros lugares se destacan instituciones como la Fundación Social, el Ministerio de Agricultura y Desarrollo Rural, la Universidad de La Amazonía, y la Corporación para el Desarrollo Sostenible del Sur de la Amazonía. Estas entidades han participado en proyectos cuyos títulos reflejan objetivos alineados con sostenibilidad, productividad, conservación de ecosistemas y cooperación técnica ambiental.

Nº	Nombre de la Institución	ID Institución	Similitud Coseno
0	Fundación Social	185	0.616
1	Departamento Técnico Administrativo del Medio Ambiente del Valle del Cauca	61	0.615
2	Universidad de La Amazonía	562	0.606
3	Corporación para el Desarrollo Sostenible y Mitigación del Cambio Climático	1135	0.606
4	Ministerio de Agricultura y Desarrollo Rural (MADR)	2	0.602
5	Ministerio de Ambiente y Desarrollo Sostenible	10	0.591
6	Sociedad Agraria de Transformación Cosechar SAT	4328	0.590
7	Fundación Centro para la Investigación en Sistemas Sostenibles de Producción Agropecuaria	773	0.589
8	Organización de la Naciones Unidas para la Agricultura y la Alimentación (FAO)	16	0.586
9	Ministerio de Ciencia Tecnología e Innovación (MinCiencias)	4	0.577

Tabla 17: Recomendación de aliados basada en similitud semántica – Consulta: “sostenibilidad agropecuaria”

Una ventaja clave de este enfoque es su capacidad de descubrimiento semántico, lo que permite sugerir alianzas incluso con instituciones que no han tenido relaciones previas con AGROSAVIA pero presentan un perfil temático compatible con la idea propuesta. Esto lo diferencia sustancialmente de métodos como: Score por co-participación, que se basa exclusivamente en vínculos históricos; Similitud estructural (coseno o Adamic-Adar), que explora la red de colaboraciones existentes; y modelos predictivos supervisados, que dependen de datos etiquetados de relaciones previas.

En cambio, la recomendación semántica no requiere una estructura de red previa y se basa en el contenido y propósito de las alianzas, abriendo nuevas posibilidades para la exploración prospectiva y el diseño de convocatorias institucionales.

Este tipo de sistema es particularmente útil en contextos de ideación de proyectos, apertura de nuevas líneas estratégicas o formulación de alianzas internacionales, ya que posibilita encontrar actores no evidentes pero con alto potencial de sinergia temática.

### 4.3.3. Modelo Predictivo Basado en Características Estructurales de Redes

En esta etapa, se desarrolló un modelo de predicción de enlaces (Link Prediction) a partir de características estructurales de la red de colaboración construida con datos de nodos (instituciones) y aristas (colaboraciones) entre coactores. El objetivo fue predecir la probabilidad de formación de nuevas colaboraciones basándose en la estructura actual de la red.

Para ello, se utilizaron las siguientes métricas topológicas como variables predictoras: número de vecinos comunes, coeficiente de Jaccard, preferential attachment y índice Adamic-Adar, las cuales son ampliamente reconocidas por capturar propiedades relevantes de conectividad entre pares de nodos.

El flujo de trabajo se estructuró en los siguientes pasos:

- Construcción del grafo real a partir de los archivos de nodos y aristas.
- Generación de enlaces positivos (existentes en la red) y negativos (pares de nodos no conectados).
- Cálculo de las características estructurales para cada par de nodos.
- Entrenamiento del modelo de regresión logística, usando un conjunto balanceado de enlaces positivos y negativos.
- Evaluación del desempeño del modelo, empleando métricas estándar.

El modelo fue evaluado con una división del 70% de los datos para entrenamiento y 30% para prueba. Se obtuvo un AUC-ROC de 0.984, lo que refleja un excelente poder discriminativo del modelo para diferenciar entre enlaces existentes y no existentes. El reporte de clasificación indicó una precisión del 97% para los enlaces positivos (clase 1) y del 87% para los negativos (clase 0). La exactitud global del modelo fue del 91%, con un balance sólido entre precisión y sensibilidad, lo cual es clave en contextos de predicción de redes donde desequilibrios pueden afectar la confiabilidad.

Las probabilidades generadas por el modelo (columna pred\_proba) reflejan un elevado grado de similitud estructural entre los nodos analizados, con valores que oscilan entre 0.9999 y 1.0. Este resultado sugiere que muchas de las relaciones predichas se basan en patrones sólidos de interacción previos, tales como la existencia de vecinos comunes, similitud de conectividad o relaciones indirectas fuertes.

Nº	Institución 1	Institución 2	Probabilidad colaboración
1	Centro Internacional de Agricultura Tropical	Universidad Nacional de Colombia	1.0000000000000000
2	Universidad de Antioquia	Universidad Nacional de Colombia	0.9999999999999999
3	Universidad Agraria La Molina - UNALM	University of Yucatán - UADY	0.9999999999999999
4	Pennsylvania State University, USA	University of Yucatán - UADY	0.9999999999999999
5	Universidad Federal de Paraná UFPR	University of Yucatán - UADY	0.9999999999999999
6	Pennsylvania State University, USA	Universidad Agraria La Molina - UNALM	0.9999999999999999
7	Universidad Agraria La Molina - UNALM	Universidad Federal de Paraná UFPR	0.9999999999999999
8	Pennsylvania State University, USA	Universidad Federal de Paraná UFPR	0.9999999999999999
9	Universidad Estatal de Ohio	University of Yucatán - UADY	0.9999999999999999
10	Centro Internacional de Agricultura Tropical	Instituto Interamericano de Cooperación para la Agricultura	0.9999999999999999

Tabla 18: Top 10 de probabilidad - Modelo Predictivo Basado en Características Estructurales de Redes

Entre las predicciones destacadas figuran instituciones reconocidas como el Centro Internacional de Agricultura Tropical, la Universidad Nacional de Colombia, la Universidad de Antioquia, y entidades internacionales como la Pennsylvania State University o la Universidad Federal de Paraná, cuyas posiciones en la red y el número de interacciones previas hacen altamente probable el fortalecimiento de sus colaboraciones. El hecho de que muchas de estas instituciones ya hayan establecido colaboraciones entre sí valida la capacidad del modelo para reconocer patrones consistentes y sugiere un alto potencial para identificar nuevos vínculos estratégicos basados en la estructura global de la red.

## 4.4. Interpretación y Recomendaciones

Una vez realizados los análisis exploratorios, estructurales y predictivos de las redes de colaboración científica de AGROSAVIA, fue posible generar un conjunto de recomendaciones estratégicas orientadas al fortalecimiento de las alianzas actuales y la identificación de nuevas oportunidades de colaboración con alto potencial de impacto. Estas estrategias se fundamentan en

la interpretación de métricas de red (grado, centralidad de intermediación, cercanía, densidad, modularidad), en los patrones históricos de co-participación institucional, y en las predicciones generadas mediante diferentes enfoques, incluyendo modelos de similitud semántica, embeddings y análisis de redes bipartitas.

Nº	Tipo de Acción	Estrategia	Justificación
1	Mantenimiento de alianzas clave	Priorizar el fortalecimiento de relaciones con Super Aliados identificados (ej. Universidad Nacional, CIAT)	Alta centralidad, participación transversal en todos los tipos de alianzas
2	Expansión de redes existentes	Incentivar nuevas alianzas entre instituciones que ya tienen vecinos comunes o colaboraciones cercanas	Modelos de predicción por adamic-jaccard, jaccard y vecinos comunes
3	Interconexión entre comunidades	Fomentar alianzas entre instituciones de diferentes clústeres geográficos o temáticos	Análisis de comunidades (Louvain) y subgrafos por región
4	Visibilización de actores periféricos	Incluir en convocatorias o programas de fortalecimiento a aliados con baja frecuencia de colaboración	Identificación de nodos periféricos con alto potencial de conexión semántica o estructural
5	Recomendación de aliados por temática	Usar el motor de recomendación semántica para sugerir aliados según una idea textual de alianza propuesta	Implementación del modelo BGE-M3 con similitud coseno sobre ali_nombre
6	Planificación estratégica	Utilizar el dashboard de Power BI y los filtros interactivos en Python para tomar decisiones informadas	Visualización de redes filtradas por tipo de alianza, año, país, etc.
7	Ampliación internacional	Fortalecer vínculos con redes internacionales (UE, Asia, Oceanía) que muestran conexiones débiles pero crecientes	Subgrafos por región y análisis de grado promedio de instituciones extranjeras
8	Identificación de vacíos temáticos	Detectar áreas de I+D+i con baja densidad de colaboración y proponer alianzas orientadas a cerrar esas brechas	Cruce de keywords/títulos con presencia institucional en los proyectos

Tabla 19: Estrategias para el fortalecimiento de redes de colaboración en AGROSAVIA

N.º	Producto	Enlace
1	Notebook TFM – Código para la generación de grafos y demás funcionalidades	<a href="#">Enlace Notebook</a>
2	Notebook TFM2 – Código para el desarrollo de modelos predictivos	<a href="#">Enlace Notebook</a>
3	Notebook TFM2 – 3. Similitud Semántica de Texto	<a href="#">Enlace Notebook</a>

Tabla 20: Productos Desarrollo de Modelos Predictivos

## 5. Conclusiones y trabajos futuros

Para concluir este Trabajo Final de Máster, se presentan las conclusiones que responden a los objetivos inicialmente planteados. Además, se proponen posibles funcionalidades adicionales y líneas de trabajo futuro que podrían enriquecer y ampliar el alcance del proyecto.

### 5.1. Conclusiones

El desarrollo del presente proyecto permitió demostrar la viabilidad y el valor de implementar un modelo de análisis predictivo basado en ciencia de redes para fortalecer y anticipar colaboraciones estratégicas en el contexto científico-académico de AGROSAVIA. Los resultados obtenidos alcanzaron las expectativas iniciales en cuanto a la calidad de las visualizaciones, la capacidad de predicción de los modelos, y la identificación de actores clave. La combinación de técnicas de análisis de redes, modelado predictivo y herramientas de visualización interactiva permitió establecer un enfoque integral para la comprensión profunda del ecosistema de alianzas de la Corporación.

Los resultados no solo confirmaron hipótesis preexistentes, como el rol central de ciertas instituciones nacionales e internacionales, sino que también revelaron patrones inesperados de colaboración y la existencia de actores periféricos con alto potencial estratégico. Esta dualidad entre confirmación y descubrimiento fortaleció el valor de la metodología empleada.

Se desarrolló la línea base del estado actual de los relacionamientos y alianzas estratégicas de AGROSAVIA, mediante la consolidación de un dataset maestro estructurado en PostgreSQL, resultado de procesos de extracción, transformación y carga (ETL) aplicados a múltiples fuentes de información de alianzas. Este proceso implicó una limpieza y normalización rigurosa de los datos y la integración coherente de tablas maestras, lo que permitió contar con una base de datos relacional sólida y reutilizable para distintos fines analíticos y estratégicos.

El análisis exploratorio y visualización de redes permitió identificar clústeres de colaboración entre instituciones, calcular métricas de centralidad (grado, intermediación, cercanía) y detectar comunidades mediante el algoritmo de Louvain. Estos análisis revelaron instituciones con altos niveles de conectividad, “Super Aliados” y “Aliados Clave”, así como brechas de relacionamiento entre regiones o sectores. La herramienta interactiva desarrollada permitió analizar subgrafos según filtros como tipo de alianza, año, comunidad y grado, fortaleciendo el entendimiento de las dinámicas internas y externas.

Con base en los resultados anteriores, se propusieron recomendaciones específicas para mejorar el posicionamiento estratégico de AGROSAVIA mediante alianzas con actores emergentes o infrautilizados. Se destacaron los beneficios de monitorear las colaboraciones a través de dashboards interactivos y de fortalecer los vínculos con instituciones clave ya existentes. Además, se propusieron funcionalidades adicionales basadas en inteligencia artificial para apoyar la toma de



decisiones estratégicas sobre nuevas alianzas, como el motor de recomendación textual basado en similitud semántica de objetivos de proyectos.

La planificación del proyecto se cumplió en términos generales, con la adecuada implementación de todas las etapas previstas: recolección de datos, procesamiento, análisis exploratorio, modelado predictivo y visualización. Se aplicaron metodologías iterativas y ágiles que permitieron realizar ajustes sin perder de vista los entregables principales. La metodología fue adecuada para abordar la complejidad del problema, aunque en fases como la integración de embeddings y el análisis semántico textual se requirió extender los tiempos inicialmente previstos.

El proyecto incorporó la Competencia de Compromiso Ético y Global (CCEG), y los resultados reflejan un cumplimiento de los principios asociados a esta competencia. Se garantizó en todo momento un uso ético, seguro y responsable de los datos, respetando los criterios de privacidad y confidencialidad institucional, sin exponer información sensible o de carácter personal. En términos ético-sociales, el proyecto logró fortalecer la transparencia en la gestión de alianzas y fomentar decisiones informadas basadas en evidencia, promoviendo a su vez la inclusión de actores con baja visibilidad en las redes tradicionales.

Desde la perspectiva de sostenibilidad, el modelo propuesto permite optimizar los esfuerzos institucionales, facilitando la identificación de alianzas de alto valor estratégico y reduciendo la duplicidad de esfuerzos. Además, la herramienta desarrollada contribuye al diseño de políticas institucionales más alineadas con los retos de sostenibilidad agroalimentaria y climática. En cuanto a diversidad, el análisis estructural y predictivo de las redes favoreció la identificación de instituciones de diferentes escalas, geografías y tipos organizacionales, promoviendo una participación más equitativa y representativa en la construcción de conocimiento agropecuario.

En cuanto a impactos no previstos, se identificó el potencial del sistema para ser replicado en otras áreas de la organización o en entidades similares del ámbito público, lo cual representa un aporte positivo no anticipado, que abre nuevas líneas de trabajo institucional en inteligencia estratégica. Asimismo, el nivel de apropiación por parte de los usuarios internos superó las expectativas, fortaleciendo la cultura de análisis de datos y promoviendo nuevas oportunidades de formación y colaboración interdisciplinaria dentro de AGROSAVIA.

## 5.2. Trabajos Futuros

A pesar de los avances logrados en este trabajo, existen diversas líneas de trabajo que se proponen como oportunidades de desarrollo futuro, con el fin de ampliar el alcance, la aplicabilidad y el impacto estratégico del modelo propuesto:

- Integración de variables temáticas específicas asociadas a los proyectos y alianzas, tales como las áreas de investigación, sistemas productivos, resumen, descripción, etc. La incorporación de estas variables permitiría un análisis temático más robusto, segmentar las



redes por líneas estratégicas de I+D+i, y generar recomendaciones más contextualizadas a necesidades institucionales o territoriales.

- Desarrollo de un módulo avanzado de análisis temporal de las redes de colaboración, que permita visualizar y analizar la evolución de las alianzas en función del tiempo. Esto facilitaría la identificación de dinámicas de consolidación, crecimiento o debilitamiento de relaciones institucionales, así como el análisis de impacto de eventos específicos sobre la estructura de la red.
- Automatización del sistema de actualización de datos, mediante la conexión directa con fuentes institucionales internas (por ejemplo, sistemas de información de proyectos, convenios y eventos) o incluso externas (repositorios abiertos, bases bibliográficas, redes internacionales), lo cual garantizaría la sostenibilidad del modelo en el tiempo y su capacidad de mantenerse vigente sin depender exclusivamente de procesos manuales.
- Desarrollo de una plataforma Web institucional donde el sistema pueda ser desplegado con una interfaz amigable para los diferentes usuarios, integrando filtros, visualizaciones y módulos de recomendación de manera accesible. Esta plataforma permitiría ampliar la apropiación institucional de la herramienta y favorecer su incorporación en procesos de planificación, monitoreo y evaluación de alianzas.
- Incorporación de técnicas de aprendizaje profundo (deep learning) y redes neuronales gráficas (Graph Neural Networks – GNN) para refinar las predicciones de vínculos futuros y caracterizar patrones ocultos en las interacciones complejas. Estos enfoques podrían aumentar la precisión del modelo y abrir nuevas posibilidades para analizar colaboraciones no evidentes.
- Incluir mecanismos de evaluación participativa de los resultados del modelo, involucrando a investigadores y personal de apoyo, con el fin de validar, retroalimentar y ajustar los hallazgos del sistema con el conocimiento experto del entorno institucional. Esto permitiría fortalecer el proceso de toma de decisiones y generar una cultura institucional basada en inteligencia estratégica y uso ético de los datos.

## 6. Glosario

**AGROSAVIA:** Corporación Colombiana de Investigación Agropecuaria. Entidad nacional encargada de liderar la investigación científica, el desarrollo tecnológico y la innovación en el sector agropecuario colombiano.

**Aliado Estratégico:** Institución, organización o entidad con la que AGROSAVIA establece relaciones formales o informales de colaboración, con el propósito de ejecutar proyectos, compartir conocimiento o fortalecer capacidades técnicas y científicas.

**Análisis de Redes Sociales (ARS):** Metodología que permite estudiar las relaciones entre actores (nodos) y las interacciones entre ellos (aristas o enlaces), con el fin de entender estructuras, dinámicas y patrones de colaboración.

**AUC-ROC (Area Under Curve - Receiver Operating Characteristic):** Métrica estadística que evalúa el desempeño de un modelo de clasificación, indicando su capacidad para distinguir entre clases positivas y negativas.

**Clúster:** Agrupación de nodos dentro de una red que presentan una alta densidad de conexiones entre sí, en comparación con el resto de la red.

**Cosine Similarity (Similitud Coseno):** Medida utilizada para evaluar la similitud entre dos vectores de texto, útil para comparar campos de texto mediante representaciones semánticas.

**CSV (Comma-Separated Values):** Formato de archivo plano ampliamente utilizado para almacenar y compartir datos tabulares, separados por comas.

**Embeddings:** Representaciones vectoriales densas de entidades (como textos, instituciones o nodos) utilizadas en modelos de lenguaje o de redes para capturar similitudes latentes y relaciones contextuales.

**ETL (Extract, Transform, Load):** Proceso de ingeniería de datos que implica extraer información desde múltiples fuentes, transformarla para integrarla y cargarla en un sistema de almacenamiento estructurado como una base de datos.

**Grafo:** Representación matemática y visual de una red, compuesta por nodos (actores) y aristas (relaciones).

**GNN (Graph Neural Network):** Tipo de red neuronal diseñada para operar directamente sobre grafos, capturando relaciones complejas entre nodos y prediciendo dinámicas en redes estructuradas.

**IpyWidgets:** Librería de Python que permite crear interfaces gráficas interactivas en notebooks, facilitando la exploración dinámica de datos y resultados.

**Link Prediction (Predicción de Enlaces):** Técnica utilizada para anticipar relaciones futuras en una red con base en su estructura actual.

**Louvain:** Algoritmo para la detección de comunidades en grafos, basado en la optimización de la modularidad. Es eficiente y ampliamente utilizado en el análisis de redes sociales y científicas.

**Modelo Predictivo:** Algoritmo o sistema basado en datos históricos y técnicas estadísticas o de aprendizaje automático que permite anticipar eventos futuros.

**ODS (Objetivos de Desarrollo Sostenible):** Conjunto de 17 objetivos globales establecidos por las Naciones Unidas para erradicar la pobreza, proteger el planeta y asegurar la prosperidad para todos.

**PCA (Análisis de Componentes Principales):** Técnica estadística de reducción de dimensionalidad que permite visualizar datos complejos en dos o tres dimensiones, facilitando la interpretación de los embeddings.

**Power BI:** Herramienta de visualización de datos desarrollada por Microsoft.

**Python:** Lenguaje de programación ampliamente utilizado en ciencia de datos, inteligencia artificial y análisis de redes.

**Red Bipartita:** Tipo de grafo en el que los nodos se dividen en dos conjuntos disjuntos y las conexiones solo ocurren entre conjuntos diferentes.

**SVD (Singular Value Decomposition):** Técnica de factorización matricial utilizada para obtener embeddings a partir de matrices de relaciones, útil en sistemas de recomendación y reducción de dimensiones.

**t-SNE (t-distributed Stochastic Neighbor Embedding):** Algoritmo de reducción de dimensionalidad no lineal utilizado para visualizar estructuras complejas en datos de alta dimensión.

## 7. Bibliografia

- Abbasi, A., Altmann, J. y Hossain, L., 2011. Identifying the effects of co-authorship networks on the performance of scholars: A correlation and regression analysis of performance measures and social network analysis measures. *Journal of Informetrics*, 5(4), pp.594–607.
- Azrai, M., Bahrún, A.H., Efendi, R., Andayani, N.N., Jihad, M., Bahtiar, B., Zainuddin, B. y Aqil, M., 2024. Global drought tolerant maize research and development: Analysis and visualization of cutting-edge scientific technologies, *Journal of Agriculture and Food Research*, 16, p. 100514
- Barabási, A.-L. & Pósfai, M., 2016. *Network Science*. Cambridge: Cambridge University Press
- Barabási, A.-L., Jeong, H., Néda, Z., Ravasz, E., Schubert, A. y Vicsek, T., 2002. Evolution of the social network of scientific collaborations. *Physica A: Statistical Mechanics and its Applications*, 311(3-4), pp.590-614.
- Bastian, M., Heymann, S. y Jacomy, M., 2009. Gephi: An Open-Source Software for Exploring and Manipulating Networks. *Proceedings of the International AAAI Conference on Weblogs and Social Media*, 3(1), pp.361-362.
- Börner, K., 2015. *Atlas of Knowledge: Anyone Can Map*. MIT Press.
- Bródka, P., Chmiel, A., Magnani, M. y Ragozini, G., 2018. Quantifying layer similarity in multiplex networks: a systematic study. *Royal Society Open Science*, 5(7), 171747.
- Fuglie, K.O., Heisey, P.W., King, J.L., Pray, C.E., Day-Rubenstein, K., Schimmelpfennig, D., Wang, S.L. y Karmarkar-Deshmukh, R., 2011. *Research Investments and Market Structure in the Food Processing, Agricultural Input, and Biofuel Industries Worldwide*. Economic Research Report No. 130. Washington, D.C.: U.S. Department of Agriculture, Economic Research Service.
- Hagberg, A.A., Schult, D.A. y Swart, P.J., 2008. Exploring Network Structure, Dynamics, and Function using NetworkX. *Proceedings of the 7th Python in Science Conference (SciPy2008)*, pp.11-15.
- Lande, D., Fu, M., Guo, W., Balagura, I., Gorbov, I. y Yang, H., 2020. Link prediction of scientific collaboration networks based on information retrieval. *World Wide Web*, 23(4), pp.2239–2257.

- Liu, C., Han, Y., Xu, H., Yang, S., Wang, K. y Su, Y., 2024. A Community Detection and Graph-Neural-Network-Based Link Prediction Approach for Scientific Literature. *Mathematics*, 12(3), p.369.
- Liben-Nowell, D. y Kleinberg, J., 2007. The link-prediction problem for social networks. *Journal of the American Society for Information Science and Technology*, 58(7), pp.1019-1031.
- Newman, M. E. J., 2001. The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences*, 98(2), 404-409.
- Pardey, P.G., Chan-Kang, C., Dehmer, S.P. y Beddow, J.M., 2016. Agricultural R&D is on the move. *Nature*, 537(7620), pp.301–303
- Plex, A., De Col, V., Etherton, B., Xing, Y., Agarwal, A., Ramic, L., Bonaiuti, E., Friedmann, M., Proietti, C., Thiele, G.P.C. y Garrett, K., 2024. What traits of collaboration networks are associated with project success? The case of two CGIAR agricultural research programs for development. *Agricultural Systems*, 219, p.104013.
- Resce, G., Zinilli, A. & Cerulli, G., 2022. Machine learning prediction of academic collaboration networks. *Scientific Reports*, 12, 22100.
- ResearchGate (2022) About us. Disponible en: <https://www.researchgate.net>
- Sepúlveda-Casadiego, Y., 2020. Machine learning en la agricultura, revisión y aplicaciones. Documento de Trabajo, Universidad Nacional Abierta y a Distancia, Centro de Investigación en Agricultura y Biotecnología (CIAB).
- SER - Radio Palma (2024) [El Observatorio Citrícola dota de herramientas digitales inteligentes al sector para tomar decisiones.](#)
- SIDALC (2022) Alianza de Servicios de Información Agropecuaria de las Américas. Disponible en: <https://www.sidalc.net>
- Sun, Y. y Han, J., 2012. Mining Heterogeneous Information Networks: Principles and Methodologies. *Synthesis Lectures on Data Mining and Knowledge Discovery*. Morgan & Claypool Publishers.
- Tuninetti, M., Aleta, A., Paolotti, D., Moreno, Y. y Starnini, M., 2021. Prediction of new scientific collaborations through multiplex networks. *EPJ Data Science*, 10(1), p.44.

- Weraikat, D., Šorič, K., Žagar, M. y Sokač, M., 2024. Data Analytics in Agriculture: Enhancing Decision-Making for Crop Yield Optimization and Sustainable Practices. Sustainability, 16(17), p.7331.
- Zhang, Z., Cui, P., & Zhu, W., 2020. Graph Neural Networks: A Review of Methods and Applications. AI Open, 1, pp.57–81.
- Zhu, Y., Kim, D., Jiang, T., Zhao, Y., He, J., Chen, X. y Lou, W., 2024. Dependency, reciprocity, and informal mentorship in predicting long-term research collaboration: A co-authorship matrix-based multivariate time series analysis. Journal of Informetrics, 18(1), p.101486.