UNIVERSIDAD DE ANTIOQUIA

Consolidación de ofertas y reporte basado en aprendizaje automático orientado hacia la toma de decisiones asertivas para el área de preventa

Sebastián Bolívar Vanegas

Informe de práctica presentado para optar al título de Ingeniero de Telecomunicaciones

Modalidad de Práctica

Práctica Empresarial

Asesor

Cristian David Ríos Urrego, Magíster (MSc) en Ingeniería Telecomunicaciones

Universidad de Antioquia

Facultad de Ingeniería

Ingeniería de Telecomunicaciones

Medellín, Antioquia, Colombia

2026

Dedicatoria

A mis padres, Alba Lucía Vanegas y Francisco Bolívar Madrigal.
Por ser el cimiento inquebrantable de mi vida y el impulso constante que me permitió culminar este logro. Gracias por su apoyo incondicional, que no comenzó en las aulas de la universidad, sino mucho antes, en cada lección de vida y en cada sacrificio silencioso.
A ustedes, que sembraron en mí los valores, la disciplina y la fortaleza necesaria para enfrentar los retos académicos y profesionales, les dedico este esfuerzo. Su confianza ha sido mi mayor motivación y su ejemplo, mi mejor guía. Este trabajo y este título son el fruto de su amor y dedicación.

Agradecimientos

Primeramente, elevo mi gratitud a Dios, por guiar mi destino, brindarme la fortaleza necesaria para superar los obstáculos y permitirme culminar con éxito esta etapa fundamental de mi vida.

A mi asesor y maestro, Cristian David Ríos Urrego, por su invaluable acompañamiento técnico y su orientación académica. Sus enseñanzas, recomendaciones y paciencia fueron determinantes para estructurar y llevar a buen término este proyecto de investigación.

Extiendo un agradecimiento especial a Cristian Alfredo Chica Arrieta, por la confianza depositada en mí al abrirme las puertas de Tigo para realizar mis prácticas profesionales. Gracias por el liderazgo, la mentoría y todas las lecciones aprendidas durante este proceso, las cuales han enriquecido significativamente mi perfil profesional.

Finalmente, a mis compañeros y amigos de las diferentes ingenierías, especialmente a aquellos conocidos como "los del 12". Gracias por trascender el ámbito académico y convertirse en un apoyo real. Sus consejos, recomendaciones y compañía hicieron más llevadero y enriquecedor el transcurso de esta carrera.

TABLA DE CONTENIDO

LISTA DE TABLAS

LISTA DE FIGURAS

SIGLAS, ACRÓNIMOS Y ABREVIATURAS

API	*Application programming interfaces* – Interfaces de programación de aplicaciones.
B2B	*Business to business* – servicios de empresa ofrecidos para otra empresa.
EDA	*Exploratory data analysis* – análisis exploratorio de datos.
IA	Inteligencia artificial.
LR	*Long Reach* – Largo alcance.
LLM	*Large language model* – Modelo extenso de lenguaje.
ML.	*Machine learning* – aprendizaje de máquinas.
NER	*Name entity recognition* - reconocimiento de entidades nombradas.
NLP	*Natural language processing* - procesamiento de lenguaje natural.
PCA	*Principal component analysis* - análisis de componentes principales.
PII	*Personal identifiable information* - información de identificación personal.
POE	*Power over Ethernet* – alimentación a través de Ethernet.
RAG.	*Retrieval augmented generation* – generación aumentada por recuperación.
SLA	*Service level agreement* – acuerdo de nivel de servicio.
XAI	*Explainable artificial intelligence* – inteligencia artificial explicable

RESUMEN

En el competitivo entorno de las telecomunicaciones, la agilidad y precisión en la estimación de costos son factores determinantes para el cierre de negocios. Sin embargo, la falta de una visión unificada de los datos históricos genera ineficiencias operativas y sobrecostos, impidiendo capitalizar el conocimiento acumulado para optimizar las estrategias de negociación. Bajo este contexto, la gestión de compras en entornos B2B (*business to business* – negocios o servicios de empresa ofrecidos a empresas) afronta desafíos críticos debido a la dispersión de datos y la heterogeneidad de formatos, lo que dificulta la comparación eficiente de proveedores, tener la información consolidada y lograr tener una visual o recomendación que ayude a realizar compras más sensatas van a ser el factor diferencial para no solo destacar en el entorno competitivo y también tomar mejores decisiones operativas. Este proyecto tiene como objetivo diseñar e implementar un sistema recomendador de compras basado en inteligencia artificial generativa y arquitectura RAG (*Retrieval-Augmented Generation*) para apoyar la toma de decisiones efectivas en el área de preventa de Tigo. La metodología, de enfoque cuantitativo y tecnológico, se desarrolló en tres fases: estructuración y anonimización de datos, creación de un modelo base estadístico (*baseline*) y desarrollo de un sistema semántico avanzado.

Los resultados evidenciaron que, si bien el modelo estadístico es rápido para referencias exactas, carece de precisión en búsquedas complejas. En contraste, la solución de inteligencia artificial con capacidad semántica demostró superioridad al interpretar contextos técnicos y restricciones logísticas, filtrando el ruido y mejorando la pertinencia de los hallazgos. Se concluye que la integración de modelos generativos transforma datos dispersos en conocimiento estratégico, ofreciendo una herramienta transparente que justifica técnicamente la selección de proveedores y supera las limitaciones de los métodos determinísticos tradicionales.

Palabras clave **— sistema recomendador, compras inteligentes, LLM, RAG, búsqueda semántica.**

ABSTRACT

In the competitive telecommunications environment, agility and accuracy in cost estimation are determining factors for closing deals. However, the lack of a unified view of historical data leads to operational inefficiencies and cost overruns, preventing companies from capitalizing on accumulated knowledge to optimize negotiation strategies. In this context, procurement management in B2B (business-to-business) environments faces critical challenges due to data dispersion and format heterogeneity, which makes it difficult to efficiently compare suppliers, Having consolidated information and obtaining a visual or recommendation that helps make more sensible purchases will be the differentiating factor in not only standing out in the competitive environment but also making better operational decisions. This project aims to design and implement a purchasing recommendation system based on generative artificial intelligence and RAG (Retrieval-Augmented Generation) architecture to support effective decision-making in Tigo's pre-sales area. The methodology, which is quantitative and technological in nature, was developed in three phases: data structuring and anonymization, creation of a statistical baseline model, and development of an advanced semantic system.

The results showed that, although the statistical model is fast for exact references, it lacks precision in complex searches. In contrast, the AI solution demonstrated superiority in interpreting technical contexts and logistical constraints, filtering out noise and improving the relevance of findings. It is concluded that the integration of generative models transforms scattered data into strategic knowledge, offering a transparent tool that technically justifies the selection of suppliers and overcomes the limitations of traditional deterministic methods.

Keywords **— recommendation system, smart shopping, LLM, RAG, semantic search.**

I. INTRODUCCIÓN.

En el entorno B2B en el que opera Tigo, las decisiones comerciales del área de preventa dependen de disponer información precisa, oportuna y comparable. En efecto, la capacidad de analizar datos sobre equipamiento de red (*routers*, *switchs*, licencias, módulos y otros componentes), así como las condiciones comerciales ofrecidas por distintos proveedores, es un factor diferencial para la competitividad. Sin embargo, hoy en día el equipo de preventa presenta volúmenes de cotizaciones y hojas comerciales en formatos diversos (CSV, Excel, PDF, sistemas internos). Dado que esta heterogeneidad genera una alta dispersión en la información, se dificulta significativamente su aprovechamiento para el análisis estratégico y la toma de decisiones ágiles. Por consiguiente, la falta de estandarización produce procesos manuales repetitivos, aumenta la probabilidad de errores humanos y extiende los tiempos de respuesta frente a solicitudes comerciales críticas.

Teniendo en cuenta esta situación, se evidencia una limitación en la capacidad de comparar de forma rápida y confiable precios por referencia, evaluar comportamientos históricos de proveedores o identificar tendencias de costos. Una posible herramienta para el manejo de estos datos podría ser modelos tradicionales de análisis de datos, sin embargo, a menudo los modelos tradicionales de análisis de datos a menudo resultan insuficientes cuando se trata de interpretar descripciones técnicas complejas o búsquedas semánticas que no coinciden exactamente con palabras clave predefinidas.

Así pues, el presente trabajo de grado propone la consolidación de los distintos formatos en una única base estructurada, la cual sirve de cimiento para desarrollar y contrastar dos enfoques tecnológicos distintos. Por una parte, se establece un modelo base estadístico (*baseline*), fundamentado en algoritmos determinísticos de filtrado y coincidencia exacta de palabras clave para búsquedas estructuradas. Por otra parte, se implementa una solución avanzada de Inteligencia Artificial, basada en modelos de lenguaje grande (LLM) potenciados por arquitectura RAG, diseñada para interpretar semánticamente descripciones técnicas y contextos complejos que el modelo estadístico no logra capturar."

La implementación de esta arquitectura RAG, resulta fundamental para dotar al sistema de capacidad interpretativa, permitiéndole comprender el contexto de las consultas técnicas, por ejemplo, relacionando conceptos como "transceptor de largo alcance" con referencias específicas como "LR" (*long reach* – largo alcance.) o "10km" y superando así las barreras de los filtros tradicionales. Simultáneamente, este enfoque garantiza la confiabilidad de la información, asegurando que las respuestas generadas por la inteligencia artificial no sean alucinaciones, sino que estén estrictamente ancladas y verificadas contra la base de datos histórica real de la compañía.

En síntesis, el sistema convertirá datos dispersos en conocimiento accionable, entregando respuestas claras, enfocadas y trazables que permitan realizar comparativas y detección de tendencias. De igual manera, cada resultado ofrecerá una perspectiva precisa, vinculada a los registros fuente para garantizar su verificabilidad. En conclusión, con este desarrollo se entregará un prototipo operativo que optimizará la toma de decisiones comerciales en Tigo, pasando de una gestión manual a una asistencia inteligente basada en datos.

II. OBJETIVOS

A. Objetivo general

Diseñar e implementar un sistema de inteligente de análisis de datos que extraiga información relevante de una base de datos de ofertas históricas, con el fin de generar soporte analítico, visualizaciones y reportes interpretativos que faciliten una toma de decisiones ágil y precisa en el área de preventa.

B. Objetivos específicos

1. Estructurar e implementar una base de dato de ofertas históricas mediante procesos integrales de limpieza, estandarización y documentación, garantizando la calidad, integridad y confiabilidad de la información.

2. Diseñar y desarrollar el núcleo de procesamiento analítico que, mediante técnicas de inteligencia artificial, automatice la identificación de información estratégica y la generación de insights accionables a partir de la base de datos consolidada.
3. Evaluar el desempeño y precisión del sistema analítico desarrollado mediante pruebas de validación y comparación de resultados, con el fin de verificar su capacidad para generar información confiable y relevante para el área de preventa.
4. Implementar una capa de presentación que integre los resultados del análisis en informes, reportes y gráficos interactivos, facilitando la comprensión de los datos y la toma de decisiones estratégicas en área de preventa.

III. MARCO TEÓRICO.

La construcción de sistemas inteligentes orientados a la optimización de procesos de abastecimiento en entornos B2B no puede entenderse como una tarea aislada de programación, sino que requiere la convergencia estratégica de diferentes disciplinas: la gestión rigurosa de la calidad de los datos, la capacidad de descubrimiento del aprendizaje automático no supervisado y la potencia interpretativa de las modernas arquitecturas de NLP (*natural language processing* - procesamiento de lenguaje natural.).

En efecto, el ecosistema de compras corporativas se caracteriza por la generación masiva de información heterogénea y no estructurada. Por consiguiente, para transitar de un modelo de gestión documental disperso donde la información reside en hojas de cálculo y archivos PDF a uno de verdadera inteligencia computacional, es necesario establecer los fundamentos teóricos que validan la arquitectura propuesta.

Teniendo en cuenta lo anterior, este marco teórico se estructura sobre tres pilares fundamentales que sustentan la solución tecnológica. En primer lugar, se examina la Gestión de la Calidad del Dato, entendiéndola no como un atributo deseable, sino como un requisito funcional indispensable para la fiabilidad algorítmica. En segundo lugar, se aborda el Aprendizaje Automático No Supervisado, explorando cómo técnicas como el agrupamiento (clustering) permiten revelar patrones latentes de comportamiento en proveedores y costos sin necesidad de etiquetas previas.

Finalmente, se analiza el Procesamiento Semántico Avanzado, fundamentando el uso de modelos generativos y arquitecturas de recuperación RAG como el mecanismo capaz de cerrar la brecha entre el lenguaje técnico complejo y la capacidad de razonamiento de la máquina. Así pues, la integración de estas tres disciplinas constituye la base necesaria para transformar datos históricos inertes en activos estratégicos de decisión.

A. Pilar 1: fundamentos de datos y análisis estadístico

Este primer pilar establece las bases para el tratamiento de la información y la generación del modelo base (baseline), enfocándose en la integridad de la fuente y la identificación de estructuras sin etiquetas previas.

1. Calidad de datos y el enfoque "*data-centric AI*".

El punto de partida para el éxito de cualquier sistema de soporte a la decisión reside, en la integridad y fiabilidad de la fuente de información. En entornos corporativos, las bases de datos históricas acumuladas a partir de múltiples canales de entrada suelen presentar una alta heterogeneidad en sus formatos, así como errores de transcripción, valores faltantes y registros duplicados. Estas deficiencias estructurales comprometen la utilidad de los datos, convirtiéndose en barreras significativas para el análisis automatizado [1], [2].

Sin embargo, el tratamiento sistemático de estos problemas no debe abordarse meramente como una labor de limpieza de datos. Por el contrario, esta fase debe enmarcarse bajo el paradigma de la IA centrada en datos (*Data-Centric AI*). Este enfoque indica un cambio en la estrategia de desarrollo: la calidad de los datos tiene un impacto más significativo en el rendimiento final de los modelos de aprendizaje automático que la sola optimización de los algoritmos o arquitecturas. De hecho, la evidencia respalda esta transición estratégica. Estudios recientes demuestran que invertir esfuerzos en mejorar la consistencia, el etiquetado y la estructura de los datos puede ser de 3 a 5 veces más efectivo para reducir el error del sistema [3]. Por consiguiente, dimensiones clásicas de la calidad como la completitud (ausencia de valores nulos), la consistencia (uniformidad en los

formatos) y la unicidad (ausencia de duplicados) dejan de ser métricas opcionales y se convierten en requisitos obligatorios del sistema.

Ahora bien, para comprender la profundidad teórica de este requisito, es necesario adoptar una perspectiva ontológica. Fundamentada por las investigaciones de Wand y Wang [4], la calidad de los datos se define formalmente como la capacidad del sistema de información para representar de manera no ambigua y completa el estado del mundo real que pretende modelar. Bajo esta perspectiva, un dato erróneo no es solo un fallo técnico, sino una desconexión con la realidad del entorno.

Así pues, las etapas de preprocesamiento de datos —que incluyen la recolección, depuración, y normalización no constituyen una tarea común, sino la fase crítica que garantiza la representación fiel de la realidad comercial del entorno [5]. En conclusión, asegurar esta calidad desde la base es un mecanismo efectivo para mitigar el riesgo de introducir sesgos algorítmicos, alucinaciones o errores de interpretación en las etapas posteriores de inferencia y toma de decisiones estratégicas [6].

2. Arquitectura de análisis y segmentación no supervisada.

Para transformar grandes volúmenes de datos crudos en patrones que respalden la toma de decisiones, no es suficiente con la aplicación aislada de algoritmos; se requiere el diseño de una arquitectura de entorno analítico robusto que soporte el ciclo de vida completo del dato, desde su ingesta hasta su interpretación final. En efecto, Hinojosa et al. [7] describe estas arquitecturas modernas como flujos de trabajo integrados que combinan capas de almacenamiento estructurado con módulos analíticos avanzados, permitiendo el procesamiento eficiente de información heterogénea [7].

Dentro de este módulo analítico, la estrategia para comprender la dinámica de los proveedores y los comportamientos de compra se aborda mediante técnicas de aprendizaje automático no supervisado. A diferencia de los modelos predictivos tradicionales (supervisados), que requieren un conjunto de datos previamente etiquetados, los enfoques no supervisados están diseñados para

operar en escenarios donde no existen categorías predefinidas. Dado que en el EDA (*exploratory data analysis* – análisis exploratorio de datos.) de compras B2B rara vez se cuenta con una clasificación a priori de los proveedores, esta metodología resulta excelente para revelar la estructura de la información. Específicamente, el tipo de comportamiento se implementa mediante algoritmos de agrupamiento o *clustering*, siendo *k-Means* uno de los más utilizados por su eficiencia computacional. Este algoritmo permite descubrir estructuras presentes en los datos, basándose en la similitud matemática de sus características vectoriales. Así, variables cuantitativas como el volumen de compra, la frecuencia de transacciones y el monto monetario se convierten en dimensiones espaciales donde el algoritmo agrupa entidades que comparten comportamientos cercanos [8].

En este contexto, la validación de los resultados genera un desafío crítico, puesto que no existe una "verdad absoluta" contra la cual comparar. Por consiguiente, la evaluación de la calidad de los segmentos generados se realiza mediante métricas de coherencia interna, tales como el coeficiente de silueta la cual cuantifica qué tan similar es un objeto a su propio grupo en comparación con otros grupos (cohesión y separación respectivamente).

En síntesis, la aplicación rigurosa de esta arquitectura y estos métodos permiten identificar grupos estratégicos claros, por ejemplo, diferenciando matemáticamente a proveedores de nicho de proveedores mayoristas sin imponer los sesgos humanos previos que suelen limitar el análisis manual. De igual manera, esto facilita la segmentación dinámica de la cartera de proveedores, adaptándose automáticamente a los cambios en los patrones de consumo de la organización.

B. Pilar 2: representación semántica (NLP).

Este pilar actúa como el puente entre los datos crudos y la inteligencia artificial, permitiendo que el sistema interprete el significado detrás del texto técnico.

1. NLP y Representación Vectorial (Embeddings).

En el contexto del análisis de ofertas comerciales, uno de los desafíos técnicos retadores reside en la naturaleza no estructurada y altamente variable de las descripciones de los productos. En efecto, un mismo ítem tecnológico puede ser descrito de múltiples formas, por ejemplo: "*switch* capa 3 con 24 puertos POE", "Conmutador L3 24p", o simplemente por su referencia técnica. Ante esta realidad, los métodos tradicionales de extracción de información basados en reglas heurísticas o coincidencia exacta de palabras clave (*keyword matching*) terminan siendo insuficientes, ya que carecen de la flexibilidad necesaria para manejar la sinonimia, la polisemia y los errores tipográficos inherentes a la entrada manual de datos [9].

Para superar estas limitaciones, la metodología propuesta adopta el uso de la representación vectorial semántica, conocida técnicamente como *embeddings*. Esta técnica constituye la base funcional del NLP. Básicamente, los *embeddings* transforman unidades lingüísticas discretas (palabras, frases o párrafos completos) en vectores numéricos dentro de un espacio continuo de alta dimensión.

Ahora bien, la propiedad fundamental de este espacio vectorial es que la posición de cada elemento no es arbitraria. Por el contrario, la proximidad geométrica entre dos vectores indica directamente su similitud semántica. Es decir, el sistema tiene la capacidad de comprender matemáticamente que términos como *router* y enrutador, o *Access Point* y antena WiFi, son conceptos cercanos y relacionados (cercanos en ese espacio vectorial), aunque no compartan ninguna raíz léxica o letra en común.

Así pues, esta abstracción termina siendo fundamental para pasar de una búsqueda literal a una búsqueda conceptual. Gracias a ello, se logra la recuperación efectiva de información incluso en escenarios donde la terminología es inconsistente o incompleta, logrando que el sistema identifique y segmente productos funcionalmente equivalentes que, bajo un enfoque tradicional información heterogénea, permanecerían aislados e incomparables.

C. Pilar 3: modelos fundacionales y arquitecturas híbridas.

El tercer pilar integra las tecnologías más avanzadas del proyecto, combinando la capacidad generativa con la precisión de la recuperación de datos para la toma de decisiones.

1. LLM y generación de texto.

La interpretación de la información ha experimentado una evolución que rompe tendencias con la llegada de los LLM. Estos modelos se basan en arquitecturas de redes neuronales profundas conocidas como *transformers*, las cuales originaron mecanismos de atención que permiten procesar secuencias de datos capturando dependencias a largo plazo y patrones complejos de una manera que los modelos anteriores no podían lograr por su propia naturaleza.

Gracias a esta arquitectura, los LLM tienen la capacidad de realizar tareas complejas de NLP con gran fluidez, tales como la generación de resúmenes de textos técnicos extensos, la traducción contextual y la extracción automatizada de entidades nombradas. Estas categorías pueden incluir, pero no se limitan a, nombres de personas, organizaciones, ubicaciones, expresiones de tiempos, cantidades, códigos médicos, valores monetarios y porcentajes, entre otros. Así, estas técnicas permiten transformar datos no estructurados en sentidos coherentes, facilitando la comprensión humana de grandes volúmenes de información técnica.

Ahora bien, la implementación de estas tecnologías en un entorno empresarial crítico, como el área de preventa, no está exenta de riesgos significativos. Dado que los LLM son, en esencia, motores probabilísticos diseñados para predecir la siguiente palabra más probable en una secuencia, presentan una limitación intrínseca conocida como "alucinación". Este fenómeno se manifiesta cuando el modelo genera información que es sintácticamente correcta, coherente y altamente persuasiva, pero presuntamente falsa o inventada, careciendo de un anclaje en la realidad de los datos. Por tanto, no resulta viable utilizar un LLM genérico de manera aislada o "de caja negra" para tareas que requieren precisión absoluta, como la consulta de precios históricos, la verificación de inventarios o la validación de referencias técnicas específicas. En conclusión, en escenarios donde la exactitud del dato numérico es innegociable, el uso del modelo generativo debe estar estrictamente controlado y complementado por mecanismos de verificación externa.

2. Arquitectura RAG.

Para mitigar los riesgos de las alucinaciones y garantizar la precisión técnica requerida en el análisis de costos y proveedores, este proyecto implementa la arquitectura conocida como RAG. En efecto, esta técnica híbrida representa un cambio de enfoque, puesto que combina la fluidez lingüística y la capacidad de razonamiento de los LLM con la exactitud de los sistemas de recuperación de información basados en bases de datos vectoriales. [11]

El funcionamiento teórico de esta arquitectura se articula en un flujo secuencial que transforma la interacción con la IA. A continuación, se describen sus fases:

1. **Fase de recuperación (*retrieval*):** ante una consulta del usuario, el sistema no interroga directamente al modelo de lenguaje. En cambio, convierte la pregunta en un vector matemático y realiza una búsqueda de similitud semántica en la base de datos vectorial del proyecto. Específicamente, el algoritmo identifica y extrae los fragmentos de información histórica (cotizaciones, descripciones técnicas, precios) que son más relevantes para el contexto de la pregunta, actuando como un mecanismo de memoria externa precisa.
2. **Fase de generación (*generation*):** acto seguido, el sistema construye un "prompt" o una instrucción enriquecida, enviando al LLM tanto la consulta original del usuario como los datos reales recuperados en la fase anterior. De este modo, el modelo actúa como un procesador de razonamiento y síntesis, obligado a generar su respuesta basándose única y exclusivamente en la evidencia suministrada, ignorando cualquier información externa o inventada.

En síntesis, la implementación de RAG permite que el sistema razone sobre datos corporativos privados, actualizados y específicos sin la necesidad de costosos procesos de reentrenamiento del modelo (*fine-tuning*). Por consiguiente, se asegura que cada recomendación de precio o selección de proveedor no sea una inferencia probabilística abstracta, sino una conclusión fundamentada y verificable en un registro histórico real.

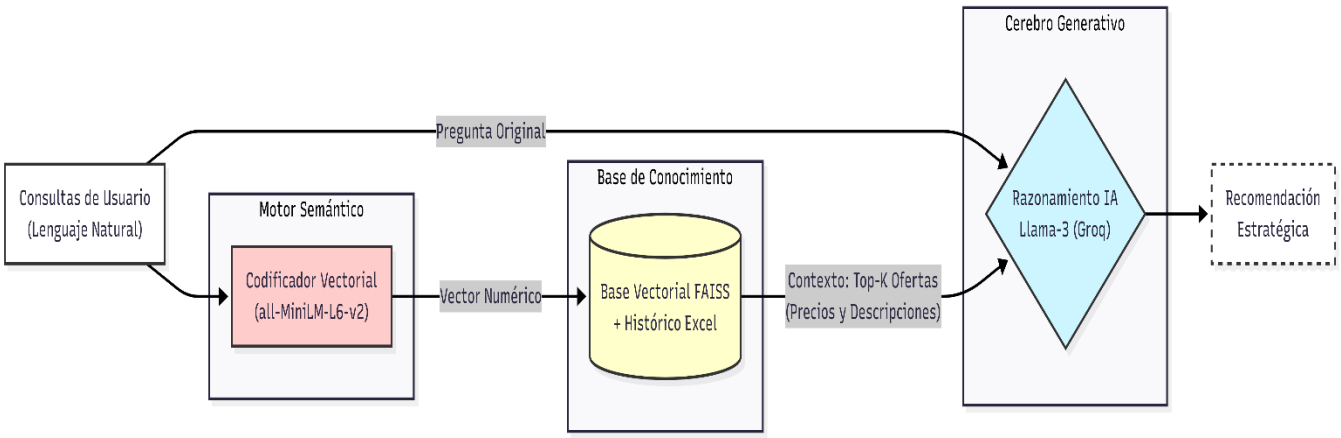


Fig 1 Esquema de la arquitectura LLM + RAG.

3. Confianza en la decisión (*score*).

La integración de sistemas de IA en procesos críticos de toma de decisiones, como lo es el abastecimiento corporativo, exige superar la barrera del desconocimiento algorítmico. En efecto, los modelos avanzados de aprendizaje profundo (*deep learning*) suelen comportarse como "cajas negras", donde la complejidad de sus operaciones internas dificulta comprender cómo se llegó a un resultado específico. Para abordar este desafío, se adoptan los principios de puntuación o score el cual va dando puntos según la búsqueda realizada por el usuario se parezca en gran medida con los datos recuperados, un campo de estudio que busca desarrollar algoritmos y técnicas que permitan hacer que los resultados de los sistemas automatizados sean transparentes, interpretables y comprensibles para los seres humanos [12].

De igual manera, en el contexto del sistema propuesto, la explicabilidad trasciende su condición de característica técnica para convertirse en un requisito de negocio inevitable. Dado que las recomendaciones del sistema implican asignaciones de presupuesto y relaciones comerciales, los usuarios expertos no pueden depender de una "fe ciega" en el algoritmo.

Específicamente, al proporcionar una recomendación de compra, el sistema no se limita a entregar un nombre o un precio. En cambio, debe exponer la evidencia que sustenta dicha conclusión: visualizar los registros históricos exactos que se utilizaron para la comparación,

explicitar la lógica de selección aplicada (ponderación entre costo y tiempo de entrega) y presentar métricas de confianza, como puntuaciones de similitud semántica.

En conclusión, esta transparencia transforma la interacción usuario-máquina. Así pues, el analista de preventa no es reemplazado, sino dotado con herramientas que le permiten validar la sugerencia de la IA contrastándola con datos reales, fomentando una adopción tecnológica segura, fundamentada y alineada con los estándares de auditoría corporativa.

IV. METODOLOGÍA

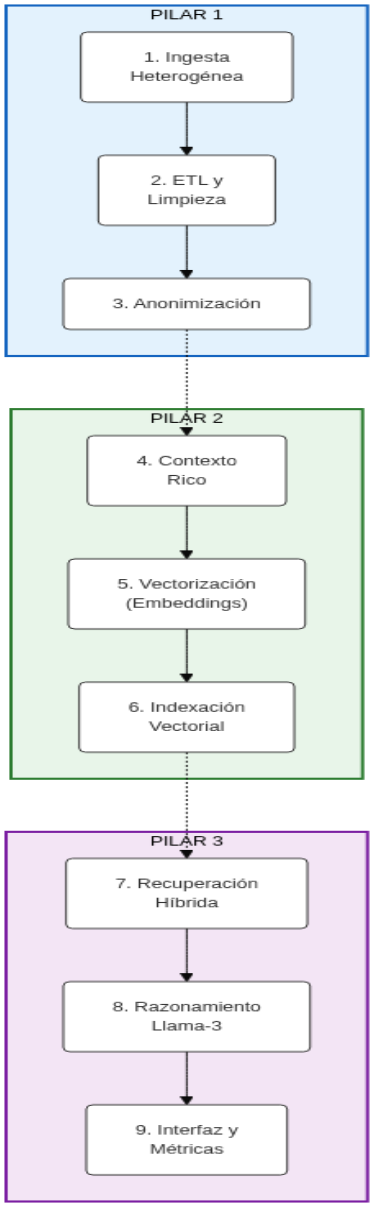


Fig 2. Metodología del recomedador.

El proceso metodológico del proyecto de grado realizado durante el semestre de industria se centra en un enfoque cuantitativo, dentro de la modalidad aplicada y tecnológica. En efecto, este diseño metodológico no busca únicamente la generación de conocimiento teórico abstracto, sino la resolución efectiva de un problema práctico de ingeniería: la optimización de los procesos de toma de decisiones de compra en un entorno corporativo B2B. Para tal fin, se procede a la construcción y despliegue de desarrollos de software avanzados, específicamente algoritmos de recuperación de información y modelos generativos.

Por consiguiente, el proceso metodológico mejora la simple observación pasiva de los fenómenos. En cambio, implica una intervención directa y controlada sobre los conjuntos de datos históricos para validar una hipótesis central: que los LLM, al operar sobre contextos semánticos complejos, ofrecen una superioridad técnica significativa frente a los métodos estadísticos tradicionales y determinísticos en tareas de recomendación y recuperación de información.

Así pues, para garantizar el rigor científico, la trazabilidad de los experimentos y el cumplimiento de los objetivos planteados, el desarrollo se estructuró en tres fases secuenciales fuertemente relacionadas. Este diseño abarca desde la gestión del ciclo de vida del dato hasta la implementación de inteligencia artificial de vanguardia: recopilación y tratamiento de datos, establecimiento de un modelo base (*baseline*) y desarrollo del sistema avanzado de IA (LLM + RAG).

FASE 1: recopilación, preprocesamiento y anonimización de datos.

Esta fase constituyó las bases del proyecto, enfocándose en transformar un conjunto de datos crudos, ofertas en formatos PDF, word, excel en un activo analítico confiable y seguro. Dado que la información original provenía de fuentes heterogéneas, se consolidó la información de valor en un archivo de hoja de cálculo ("CONSOLIDADO-OFERTAS") con información sensible, se aplicaron protocolos estrictos de ingeniería de datos.

Caracterización del Conjunto de Datos:

El activo de información base está conformado por un total de 374 registros transaccionales históricos, organizados en un esquema tabular de 15 variables. La estructura de datos abarca dimensiones de identificación y trazabilidad (oferta_id), entidades comerciales (cliente, Proveedor) y aspectos temporales (fecha_oferta, tiempo_entrega). La especificación técnica del requerimiento se detalla a través de campos categóricos y descriptivos como categoria, sku y descripcion. Por último, la dimensión económica se presenta con un desglose financiero detallado que incluye la cantidad, el tipo de divisa (moneda_orig), los costos base (valor_unitario_orig, valor_total_orig), el componente impositivo (iva, valor_lote_iva) y el monto final consolidado de la transacción (valor_total_con_iva).

1. **Ingesta y limpieza de datos:**

   El proceso comenzó con la carga de los registros históricos utilizando técnicas de ciencia de datos de *python*, apoyándose principalmente en la librería pandas. Acto seguido, se ejecutaron procesos de limpieza profunda para sanear la información. Específicamente, se abordó la normalización de tipos de datos, corrigiendo formatos de fecha para permitir análisis temporales y depurando columnas numéricas críticas (como precios unitarios y tiempos de entrega) que presentaban caracteres no procesables o símbolos de moneda que impedían el cálculo matemático. De igual manera, se gestionaron los valores nulos mediante estrategias de imputación o exclusión según la relevancia del campo, asegurando una base libre de ruido computacional.

2. **Anonimización y seguridad de la información:**

   Teniendo en cuenta los compromisos de confidencialidad comercial y la sensibilidad de los datos corporativos, se implementó un mecanismo de supresión de datos sensibles antes de cualquier procesamiento avanzado. Las columnas identificativas, tales como cliente y oferta_id, fueron eliminadas programáticamente del *dataframe* en memoria. Esta medida garantizó que ninguna PII (*personal identifiable information* - información de identificación personal.) o datos estratégicos de clientes fueran expuestos o transmitidos durante las interacciones posteriores con modelos externos o API (*application programming interfaces* – interfaces de programación de aplicaciones.) de terceros. Por

consiguiente, el análisis se centró exclusivamente en las características técnicas del producto y el comportamiento del proveedor, eliminando sesgos asociados a la identidad del cliente.

3. **Enriquecimiento de datos y construcción de contexto:**
   Para disponer del análisis semántico, fue necesario mejorar la estructura de la información de la base de datos. Así pues, se desarrolló un proceso de ingeniería de características orientado a la generación de "Texto Rico". Se creó una nueva variable sintética denominada texto_rag, la cual concatena en un solo párrafo las características clave de cada transacción: categoría, descripción técnica detallada, referencia SKU y precio. Este paso fue crucial, pues preparó los datos para que el LLM pudiera interpretar el contexto completo de cada oferta de una manera semántica coherente, en lugar de como celdas aisladas, facilitando posteriormente la vectorización y la recuperación de información.

   Adicionalmente, se ejecutó un proceso de transformación numérica para derivar métricas de comportamiento comercial. A partir de los datos transaccionales crudos, se calcularon y normalizaron tres indicadores clave por entidad (cliente/proveedor): el volumen total de unidades adquiridas, la frecuencia transaccional (conteo de operaciones) y el monto total facturado. Estas nuevas variables cuantitativas fueron estandarizadas para eliminar sesgos de escala, constituyendo los insumos vectoriales necesarios para los modelos de segmentación posteriores.

FASE 2: Desarrollo del modelo base estadístico (*baseline*).

Una vez consolidados y estructurados los datos, se procedió al desarrollo de un modelo de referencia o *baseline*. El objetivo primordial de esta fase fue establecer una métrica de rendimiento inicial sólida, utilizando técnicas estadísticas tradicionales y algoritmos de ML (*machine learning* – aprendizaje de máquinas). De este modo, se construyó un estándar comparativo cuantitativo que permitiría comparar, en etapas posteriores, el rendimiento la solución de inteligencia artificial generativa.

1. **Segmentación no supervisada:**

   Para comprender la estructura de la información de clientes y proveedores, se aplicó el algoritmo de agrupamiento *k-means*. Específicamente, el modelo se entrenó utilizando variables numéricas clave normalizadas: volumen de compra, frecuencia de transacciones y monto total facturado. Simultáneamente, dado que la visualización de datos multidimensionales presenta complejidades gráficas, se utilizaron técnicas de reducción de dimensionalidad, concretamente PCA (*principal component analysis* - análisis de componentes principales.). Esto permitió proyectar los grupos detectados en un plano bidimensional para su interpretación visual, la validez matemática de los clústeres generados se verificó mediante el cálculo del coeficiente de silueta, una métrica que evalúa la cohesión interna y la separación entre grupos, garantizando que la segmentación no fuera producto del azar.

2. **Implementación de método de búsqueda al *baseline*:**

   Por otra parte, se implementó un sistema de búsqueda y recomendación basado en reglas determinísticas, operando mediante coincidencias exactas de texto (*keyword matching*) y el cálculo de promedios históricos de precios. Sin embargo, durante la evaluación de este modelo, se notó una limitación crítica inherente a los enfoques estadísticos: la "incapacidad de entendimiento semántico".

   Los resultados mostraron un desempeño dispar según la naturaleza del producto. En categorías homogéneas, caracterizadas por una baja variabilidad técnica (como hardware estandarizado), el modelo alcanzó niveles de consistencia superiores al 35%, logrando identificar proveedores dominantes. En contraste, en categorías heterogéneas como "Servicios", la métrica de consistencia (Win Rate — probabilidad de que un mismo proveedor sea la mejor opción recurrente) cayó drásticamente al 0.5%. Este valor cercano a cero indica una dispersión total, demostrando que el enfoque estadístico es incapaz de encontrar patrones estables cuando se mezclan conceptos dispares (ej. "licencias anuales" vs. "horas de soporte"), lo cual justificó la transición hacia modelos semánticos.

FASE 3: desarrollo del sistema inteligente (LLM + RAG).

En respuesta a las limitaciones de consistencia y comprensión semántica detectadas durante la evaluación del *baseline*, la fase final del proyecto se centró en la implementación de una arquitectura RAG. El propósito fundamental fue integrar modelos de lenguaje de última generación para dotar al sistema de capacidades de razonamiento y comprensión contextual.

1. **Implementación del *embedding*:**
   Para trascender la búsqueda literal, se incorporó un componente de vectorización utilizando el modelo $all - MiniLM - L6 - v2$[1]. Este algoritmo tiene la función de transformar las descripciones técnicas no estructuradas de los productos en vectores numéricos de alta dimensión (*embeddings*). De ese modo, el sistema adquirió la capacidad matemática de entender la similitud semántica y conceptual entre términos técnicos dispares por ejemplo, relacionar correctamente "KEM" con "botonera" o "*access point*" con "antena", eliminando la dependencia de la coincidencia exacta de palabras que restringía al modelo anterior.
2. **Despliegue de base de datos vectorial:**
   Para la gestión eficiente de los vectores generados, se implementó un índice FAISS (*facebook AI similarity search*). Esta tecnología permitió almacenar y recuperar la información mediante búsquedas de similitud en espacios densos, operando en tiempos de respuesta del orden de milisegundos. Por consiguiente, el sistema logró recuperar los registros históricos más relevantes para cada consulta basándose en el significado de la necesidad del usuario y no solo en su sintaxis.
3. **Generación y razonamiento:**
   Como cerebro del sistema, se integró el modelo de lenguaje masivo $llama - 3$[2], ejecutado a través de la infraestructura de inferencia de baja latencia de Groq. A diferencia del modelo estadístico que operaba por promedios, este componente recibe los datos recuperados y aplica lógica deductiva para seleccionar el mejor proveedor. Específicamente, el modelo analiza las especificaciones técnicas, descarta opciones incompatibles y justifica su decisión final mediante lenguaje natural, proporcionando un análisis cualitativo que complementa los datos cuantitativos.

4. **Búsqueda híbrida y preprocesamiento Inteligente:**
   Para robustecer la interacción humano-máquina y mitigar errores de entrada, se desarrolló una capa de preprocesamiento inteligente. Esta funcionalidad utiliza el LLM para realizar una corrección ortográfica y técnica de la consulta del usuario (por ejemplo, interpretando y corrigiendo "Roter" a "Router") antes de ejecutar la búsqueda vectorial. En conclusión, este mecanismo de autocorrección contextual aseguró una alta tasa de recuperación de información pertinente, incluso ante entradas de texto imperfectas o con ausencia de precisión a la hora de realizar la consulta.

V. ANÁLISIS DE RESULTADOS

Segmentación de clientes mediante reducción de dimensionalidad (PCA):

Para visualizar la estructura subyacente de la cartera de clientes y validar la separación de los grupos identificados por el algoritmo *k-means*, se aplicó PCA. Esta técnica permitió proyectar las múltiples variables originales (gasto total, frecuencia de compra y volumen) en un espacio bidimensional que captura la mayor varianza de los datos.

Como se observa en la Fig 3. La distribución espacial de los puntos revela tres comportamientos claramente diferenciados:

1. Concentración y homogeneidad (grupo 2 - amarillo): Se evidencia una alta densidad de observaciones agrupadas en la región negativa del componente principal 1. La poca dispersión entre estos puntos indica que, en este segmento, que representa la mayoría de la muestra, presenta un comportamiento transaccional muy homogéneo. En términos de negocio, esto corresponde a la "base de clientes" o clientes transaccionales, cuyos volúmenes de compra y frecuencia son bajos y similares entre sí.
2. Dispersión y magnitud (grupo 1 – azul claro): En el extremo opuesto (valores positivos altos del Componente 1 y 2), se identifican puntos aislados y altamente dispersos. Esta separación visual confirma matemáticamente la existencia de clientes "atípicos" o VIPs. Su ubicación lejana del centroide sugiere que sus métricas de consumo (gasto y volumen) son

magnitudes superiores al resto de la población, validando la hipótesis de que una pequeña fracción de clientes aporta una variabilidad significativa al negocio.

3. Transición (Grupo 0 - morado): Se observa un grupo intermedio que actúa como puente entre la base y los valores extremos. Aunque presentan mayor variabilidad que el grupo 2, no alcanzan las magnitudes del Grupo 1, sugiriendo un segmento de clientes en desarrollo o de consumo medio.

El gráfico de PCA demuestra que la segmentación propuesta no es arbitraria, sino que responde a una estructura latente en los datos donde el Componente Principal 1 parece estar fuertemente correlacionado con la magnitud de la compra; separando a los clientes grandes de los pequeños, validando así la capacidad del modelo para discriminar perfiles de valor.

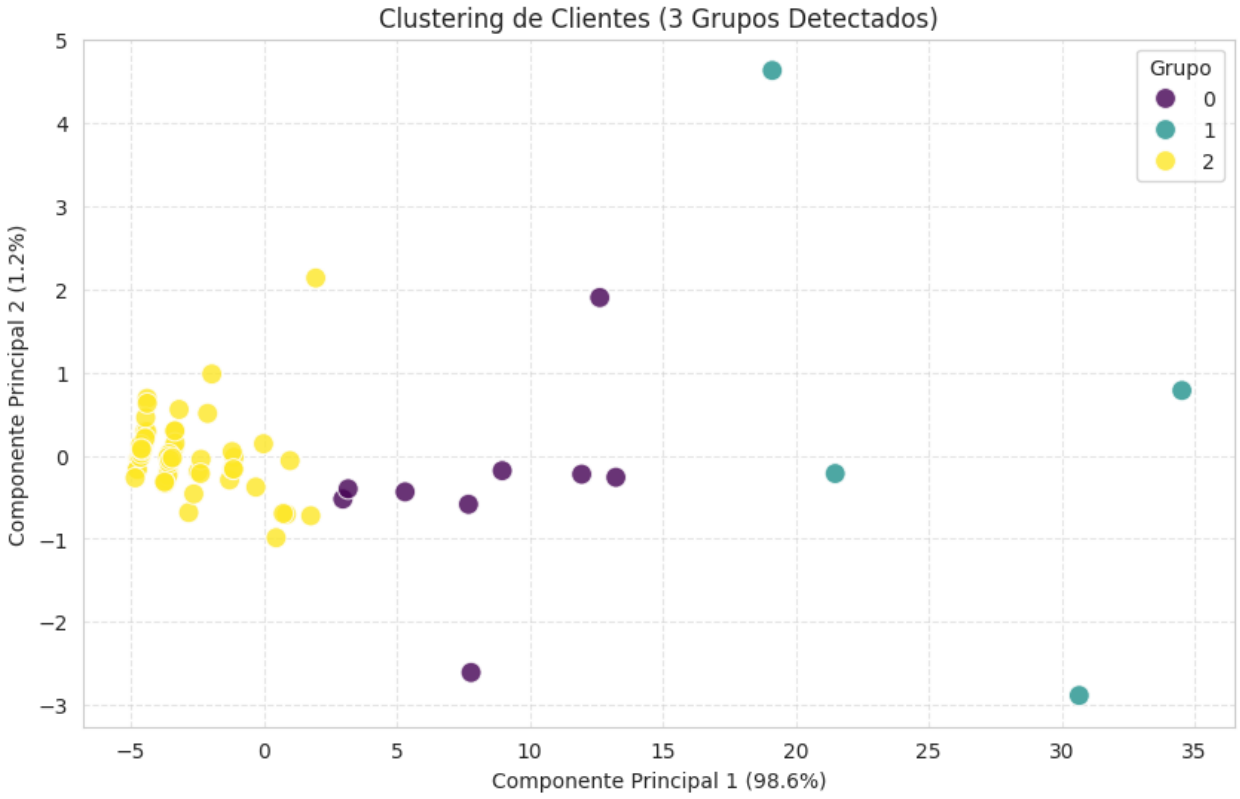


Fig. 3 clustering clientes.

Perfilamiento estratégico de proveedores:

De manera análoga al análisis de clientes, se aplicó la técnica de reducción de dimensionalidad para examinar la estructura de la base de proveedores. El modelo se alimentó con tres variables cuantitativas clave derivadas del historial transaccional: monto total de ventas (venta_total), volumen de operaciones (transacciones) y ticket promedio por compra (ticket_promedio). **La Fig 4** ilustra la proyección de estos datos multidimensionales en el espacio de componentes principales, revelando una segmentación tripartita que sugiere roles diferenciados en la cadena de suministro.

El análisis conjunto de la distribución espacial y las estadísticas descriptivas permite caracterizar los grupos de la siguiente manera:

1. Proveedores críticos de volumen (Grupo 1 – azul claro): Se identifica un punto extremo altamente diferenciado en el eje positivo de la componente 1. Este *cluster*, aunque reducido en número de integrantes, representa la base del abastecimiento operativo. Con un promedio de 73 transacciones y una venta total logarítmica de 37.0 (la más alta del conjunto), este grupo corresponde a los mayoristas o distribuidores principales con los que se mantiene una relación comercial continua y de alto flujo.
2. Socios estratégicos (grupo 2 - amarillo): En una posición intermedia se ubican proveedores que, si bien no alcanzan el volumen masivo del grupo 1, mantienen una relevancia significativa con un promedio de 38 transacciones. Su dispersión en el gráfico sugiere una variabilidad en los tipos de contrato, consolidándose como alternativas sólidas para el abastecimiento recurrente o especializado.
3. Proveedores de nicho u ocasionales (grupo 0 - morado): La gran mayoría de los proveedores se agrupan en la zona izquierda del gráfico, caracterizándose por una baja frecuencia transaccional (promedio de 3.4 operaciones). Sin embargo, es muy importante destacar un hallazgo contraintuitivo revelado en la tabla de promedios: este grupo presenta el *ticket* promedio más alto (0.54) de todos los segmentos. Esto indica que, aunque la relación comercial es esporádica, se utiliza a estos proveedores para adquisiciones puntuales de alto valor unitario, posiblemente correspondientes a infraestructura especializada, licencias complejas o proyectos de consultoría que no son recurrentes, pero sí costosos.

TABLA I

CLASIFICACION DE PROVEEDORES

Cluster	venta_total	transacciones	ticket_promedio
0	1.73	3.44	0.54
1	37.01	73	0.50
2	18.96	38	0.49

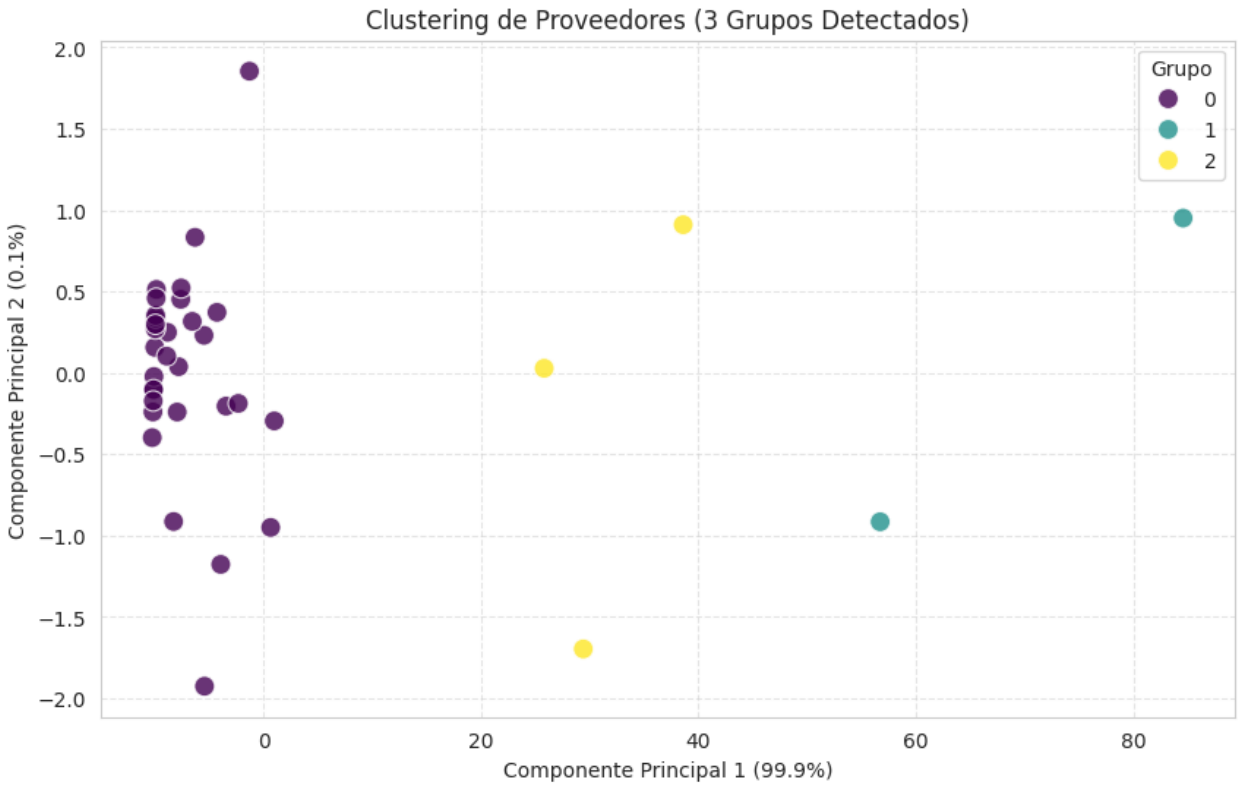


Fig. 4 clustering proveedores.

De manera análoga al análisis de clientes, se aplicó la técnica de reducción de dimensionalidad para visualizar la estructura de la base de proveedores en un gráfico de dispersión. Complementariamente, la **TABLA I** presenta la caracterización cuantitativa de los tres segmentos identificados, detallando el comportamiento promedio de los proveedores que conforman cada grupo (centroides).

Nota Aclaratoria: Es importante resaltar que los valores expresados en la Tabla I para las variables monetarias (venta_total y ticket_promedio) se encuentran en escala logarítmica, resultado de la normalización aplicada en la etapa de preprocesamiento para mitigar el sesgo de valores extremos. Por consiguiente, un valor de 37.0 en la escala logarítmica representa una magnitud financiera exponencialmente superior a un valor de 1.7, evidenciando la brecha masiva entre proveedores mayoristas y de nicho.

Descripción de la Interfaz: recomendador histórico (*baseline*).

Fig. 5 interfaz del recomendador baseline.

Como se aprecia en la **Fig. 5**, esta herramienta constituye el modelo de referencia (baseline) del proyecto. Su diseño visual, estructurado mediante campos de selección secuencial, está orientado a la búsqueda paramétrica, un enfoque donde el usuario debe conocer de antemano la clasificación o referencia técnica exacta del producto que desea consultar para obtener resultados.

Componentes de la Interfaz:

A. Categoría (filtro jerárquico superior):

- **Función:** es el primer nivel de segmentación. Despliega una lista cerrada de las categorias de productos existentes en la base de datos (ej. Accesorios VoIP, Switches, Routers).
- **Propósito:** evitar la comparación de productos de diferentes categorías. Al seleccionar una categoría, el sistema reduce la búsqueda exclusivamente a ese subconjunto de datos, disminuyendo el ruido.

B. SKU (filtro de precisión - cascada):

- **Función:** es un menú desplegable dinámico que se activa y actualiza después de seleccionar una Categoría. Muestra solo los códigos o referencias técnicas disponibles para esa categoría específica**.**
- **Propósito:** Permitir una comparación exacta de productos con un alto nivel de similitud. Si el usuario selecciona un SKU específico (ej. CP-8841-K9), el sistema ignora cualquier descripción y busca exclusivamente el historial de precios de esa referencia exacta.

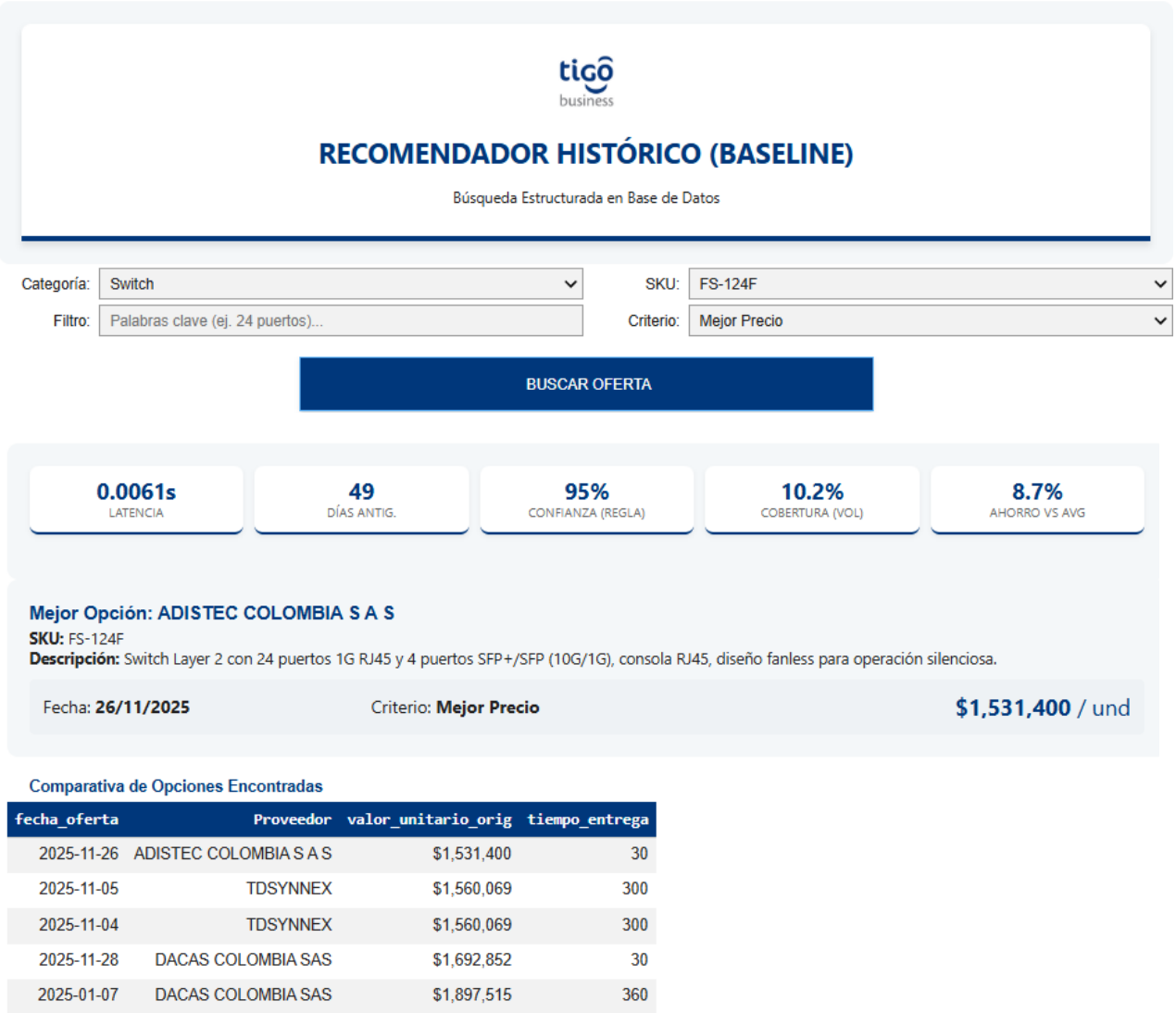
C. Filtro (búsqueda por palabras clave):

- **Función:** un campo de texto libre para búsquedas aproximadas.
- **Uso:** se utiliza cuando el usuario no conoce el SKU exacto. Funciona mediante coincidencia literal de cadenas de texto (ej. escribir "24 puertos" buscará registros que contengan exactamente esa frase).
- **Limitación:** a diferencia de la IA, este campo no entiende sinónimos; si el usuario escribe "veinticuatro ptos", el sistema no encontrará nada.

D. Criterio (lógica de ordenamiento):

- **Función:** define la prioridad matemática para seleccionar al "ganador".
  Mejor precio: ordena los resultados de menor a mayor valor unitario.
  Rapidez: ordena de menor a mayor tiempo de entrega (días).
  Confianza: ordena de mayor a menor frecuencia de compra (proveedor más utilizado).

Esta interfaz demuestra la eficiencia para compras repetitivas y estandarizadas (donde el SKU es conocido), pero evidencia la rigidez que el modelo de IA (RAG) busca solucionar en la siguiente fase del proyecto.



fecha_oferta	Proveedor	valor_unitario_orig	tiempo_entrega
2025-11-26	ADISTEC COLOMBIA S A S	$1,531,400	30
2025-11-05	TDSYNNEX	$1,560,069	300
2025-11-04	TDSYNNEX	$1,560,069	300
2025-11-28	DACAS COLOMBIA SAS	$1,692,852	30
2025-01-07	DACAS COLOMBIA SAS	$1,897,515	360

Fig. 6 recomendador baseline (búsqueda por SKU).

Para evaluar la efectividad del modelo de referencia (baseline), se realizó una prueba controlada utilizando un caso de uso de alta especificidad técnica: la búsqueda de un switch mediante su referencia exacta de fabricante (SKU: FS-124F), bajo el criterio de priorización de “mejor precio”. Como se observa en la **Fig 6**, los resultados visibles en la interfaz de usuario permiten caracterizar el comportamiento del sistema en escenarios de compras estandarizadas.

1. **Eficiencia computacional y latencia:**
   El sistema demostró una velocidad de respuesta inmediata, registrando una latencia de 0.0061 segundos. Este tiempo, prácticamente imperceptible para el usuario humano, valida la eficiencia de los algoritmos de filtrado directo sobre estructuras de datos indexadas. Al no requerir inferencia vectorial ni procesamiento de lenguaje natural, el *baseline* se consolida como la herramienta ideal para consultas transaccionales rápidas donde los parámetros de búsqueda son conocidos y fijos.

2. **Interpretación de métricas de negocio:** El modelo retornó indicadores importantes que contextualizan la recomendación:

   - **Confianza ~ 95%:** Aunque la coincidencia por SKU exacto es de naturaleza binaria (el producto corresponde o no a la referencia solicitada), el nivel de confianza reportado por el sistema incorpora factores adicionales como la calidad del registro, la consistencia de los atributos asociados y la evaluación del criterio de priorización por “mejor precio”. Por esta razón, la confianza se expresa como un valor alto pero no absoluto, reflejando una validación casi completa del cumplimiento de la intención de búsqueda.
   - **Antigüedad del Dato (49 días):** La recomendación se basa en una oferta de hace aproximadamente un mes y medio. Esto alerta al usuario sobre la vigencia del precio, sugiriendo que, aunque el dato es válido, podría requerir una revalidación si el mercado es muy volátil en lo que respecta a los precios de los productos.
   - **Cobertura ~ 10.2%:** Esta métrica revela la limitación estructural más importante del modelo. Indica que el SKU consultado (FS-124F) representa apenas el 10.2% del volumen total de transacciones dentro de la categoría "Switch". Esto significa que el *baseline* es "ciego" ante el 90% restante de la categoría si el usuario no conoce los códigos exactos de los otros modelos. Este bajo porcentaje justifica la necesidad de implementar técnicas como LLM + RAG que puedan explorar el resto del inventario mediante similitud semántica, y no solo por coincidencia exacta de código.

- **Ahorro vs. promedio ~ 8.7%:** Esta métrica es fundamental para la toma de decisiones financieras. El sistema detectó que la opción ganadora (Adistec a $1.5M) es un 8.7% más económica que el promedio del mercado para ese mismo equipo. Esto demuestra la capacidad del *baseline* para generar valor tangible inmediato al identificar dispersión de precios.

3. **Análisis comparativo de proveedores:** Es de importancia destacar cómo el modelo discriminó no solo por precio sino implícitamente por condiciones logísticas: la opción ganadora ofrece un tiempo de entrega de 30 días, contrastando drásticamente con las ofertas de TDSYNNEX que, aunque cercanas en precio, presentan tiempos de entrega de 300 días. Aunque el criterio seleccionado fue "precio", la visualización tabular permite al usuario detectar estos costos ocultos de tiempo, demostrando que el *baseline* actúa como de manera eficaz de soporte a la decisión en escenarios de reposición de hardware específico. En resumidas cuentas, para búsquedas basadas en referencias técnicas exactas (SKU), el modelo *baseline* ofrece un rendimiento técnico impecable (baja latencia) y una utilidad financiera clara (detección de ahorro), constituyendo el punto de referencia contra el cual se medirá la inteligencia artificial en consultas complejas o ambiguas.

fecha_oferta	Proveedor	valor_unitario_orig	tiempo_entrega
2025-09-22	TICLINE S.A.S.	$62,254	300
2025-08-28	TICLINE S.A.S.	$1,025,143	30
2025-11-26	ADISTEC COLOMBIA S A S	$1,531,400	30

Fig. 7 recomendador baseline (búsqueda por similitud de palabras).

Análisis crítico de la búsqueda por texto libre:

Para contrastar con la precisión del SKU, se ejecutó una consulta abierta utilizando el término "24 puertos" dentro de la categoría switch. Como se observa en la **Fig 7**, los resultados obtenidos demuestran una alta variación ligada a los métodos de búsqueda literal.

1. Falsa sensación de ahorro (ahorro Vs avg 98%): El sistema reporta un Ahorro vs. Promedio del 98.1%. A simple vista, parece un éxito rotundo. Sin embargo, un análisis cualitativo revela que este indicador está sesgado por la heterogeneidad de los productos encontrados. El sistema está comparando un "*switch D-link"* básico de 62.254 contra equipos empresariales de alta gama como los de los proveedores Aidstec o Fortinet que cuestan 1.531.400. Matemáticamente el ahorro es

real, pero funcionalmente es una comparación inválida provocada por la falta de entendimiento técnico del algoritmo.

2. Degradación de la confianza: La métrica de Confianza cae al 70% (frente al 95% del caso SKU). Esto indica que, aunque los registros cumplen con tener la palabra "24 puertos", existe una alta varianza en las especificaciones técnicas y precios dentro del grupo recuperado. El usuario se ve obligado a filtrar manualmente para entender por qué hay una diferencia de precio de 20 veces entre la primera y la tercera opción.

3. Latencia y cobertura:

- **Latencia (0.0047s):** El modelo mantiene su velocidad extrema, confirmando que la búsqueda de texto simple es computacionalmente eficiente.
- **Cobertura (46.9%):** Al usar una palabra clave genérica ("24 puertos"), el sistema logró "ver" casi la mitad de la base de datos de *switchs*. Esto es positivo para la exploración, pero peligroso sin un filtro de calidad, ya que trae demasiados resultados dispares.

La búsqueda por texto en el baseline es útil para barridos rápidos, pero peligrosa para la toma de decisiones directas. El sistema es incapaz de distinguir gamas de productos, entregando al usuario la responsabilidad total de validar si el "ahorro del 98%" corresponde a un producto equivalente o a una tecnología inferior. Esta limitación justifica la implementación de modelos de lenguaje capaces de entender que un *D-link* y un *fortiswitch* pertenecen a segmentos de mercado distintos.

Descripción de la interfaz: asistente de compras inteligente (LLM + RAG)

Fig 8 Interfaz del recomendador con interpretación semántica (LLM + RAG).

Esta interfaz representa la mejora hacia un modelo de interacción conversacional, diseñado para soportar consultas complejas, ambiguas o que requieren razonamiento técnico. Tal como se evidencia en la **Fig 8**, la arquitectura visual prioriza la simplicidad para el usuario final, ocultando la complejidad de los procesos de inteligencia artificial subyacentes.

1. Componentes de la Interfaz:

A. Campo de entrada de lenguaje natural:

- **Función:** Una caja de texto abierta que invita al usuario a escribir su necesidad tal como la pensaría o la hablaría.
- **Propósito:** Eliminar la necesidad de tener que categorizar o conocer códigos técnicos. Permite consultas multi-variables (ej. "*switch* de 24 puertos que llegue rápido" o "reemplazo para el equipo Cisco"), aprovechando la capacidad del modelo para interpretar la intención del usuario.

B. **Botón consultar:**

- Función: disparador del proceso. Al hacer clic, activa la cadena de procesamiento: corrección ortográfica inteligente, búsqueda híbrida (palabras clave + vectores) y generación de respuesta con el LLM (Llama-3).

C. Área de resultados:

- **Función:** espacio donde se despliega la respuesta generada.
- **Estructura:** Presenta primero una respuesta de inteligencia artificial (análisis cualitativo y justificación) seguido de un panel de métricas y una tabla de evidencia histórica para garantizar la transparencia, es decir la explicabilidad de la decisión.

2. Flujo de uso: El diseño propone un cambio de enfoque: de "buscar datos" a "consultar a un experto" (recomendación).

- **Formulación:** El usuario ingresa su requerimiento en lenguaje natural, pudiendo combinar especificaciones técnicas, restricciones logísticas (tiempo) y preferencias de marca en una sola frase.
- **Procesamiento:** El sistema interpreta la semántica de la frase, corrigiendo errores técnicos (ej. "Roter" -> "Router") y expandiendo la búsqueda a conceptos relacionados.
- **Respuesta Aumentada:** El usuario recibe no solo una lista de precios, sino una recomendación argumentada que explica por qué esa opción es la mejor técnica y económicamente, acompañada de métricas que le indican qué tan confiable es la recomendación.

Evaluación del modelo inteligente (LLM + RAG) en escenarios cualitativos.

	Fecha	Proveedor	sku	Precio Unitario	descripcion
167	2025-03-20	DACAS COLOMBIA SAS	FSW-124F	$1,771,014	Switch FortiSwitch 124F administrado, con 24 puertos GE y 4 puertos SFP, integración con FortiGate para seguridad y gestión centralizada.
178	2025-05-26	DACAS COLOMBIA SAS	FS-124F-FPOE	$3,859,556	Switch FortiSwitch 124F con soporte Full PoE, 24 puertos GE y 4 puertos SFP, integración con FortiGate para gestión segura y centralizada.
197	2025-09-22	TICLINE S.A.S.	DGS-1210-26	$62,254	Switch D-Link 24 puertos Gigabit PoE, gestión inteligente, ideal para redes empresariales con soporte VLAN y QoS.

Fig 9 recomendador con interpretación semántica (LLM + RAG).

Para contrastar la capacidad interpretativa del modelo basado en inteligencia artificial generativa frente al baseline determinístico, se sometió al sistema a una consulta de complejidad semántica: “Necesito un switch de 24 puertos para una red empresarial robusta”. Los resultados presentados en la **Fig 9** permiten evaluar si el modelo podía diferenciar atributos de calidad más allá de la simple coincidencia de palabras clave.

Análisis de la respuesta generada: a diferencia del listado del *baseline*, el modelo RAG generó una narrativa de justificación técnica. Identificó la opción de TICLINE S.A.S. ($62,254) no solo por ser la más económica, sino argumentando explícitamente las características encontradas en la descripción no estructurada: ″gestión inteligente, soporte VLAN y QoS″.

Este resultado evidencia dos hallazgos críticos:

1. **Comprensión de atributos**: el sistema demostró capacidad para leer y extraer especificaciones técnicas ("VLAN", "QoS") que validan el requisito de "red empresarial", algo que un filtro tradicional no puede hacer sin columnas específicas pre-etiquetadas.
2. **Sensibilidad al Dato Fuente:** El modelo recomendó un equipo con un precio atípicamente bajo ($62k) porque la descripción original del proveedor lo catalogaba textualmente como ″ideal para redes empresariales″. Esto confirma que la IA actúa como un amplificador de la información contenida en los datos; si el dato histórico promete robustez a bajo costo, la IA lo priorizará basándose en su lógica de eficiencia.

Métricas de desempeño comparativo:

- **Confianza semántica (99.9%):** El modelo reportó una certeza casi absoluta en su selección. Esto se debe a que el motor vectorial encontró una similitud casi perfecta entre la intención del usuario (red empresarial) y la semántica de la descripción del producto.

- **Eficiencia unitaria (100%):** Al seleccionar la opción de menor costo absoluto entre los candidatos técnicamente viables, el modelo maximizó la métrica de ahorro teórico.

- **Cobertura (13.2%):** al comparar esta métrica con el 46.9% obtenido en la búsqueda por texto libre del *baseline*, se evidencia la capacidad de focalización del modelo LLM + RAG. Mientras que el filtro de palabras clave trajo casi la mitad de la categoría (mezclando gamas altas y bajas), la IA restringió su análisis al 13.2% más relevante semánticamente. Esta reducción significativa del espacio de búsqueda (de 46.9% a 13.2%) indica que el algoritmo descartó eficazmente el "ruido" o los equipos que, aunque contenían las palabras "switch" y "24", no se alineaban con el contexto de "red empresarial robusta", mejorando así la calidad de la recomendación final.

  .

Mientras el *baseline* entregó el dato por pura coincidencia de caracteres ("24 puertos"), el LLM + RAG validó la pertinencia técnica del hallazgo. Aunque ambos llegaron al mismo producto debido a la agresiva diferencia de precios en la base de datos, el RAG aportó una capa de explicabilidad (XAI) que permite al usuario entender por qué ese equipo cumple con los requisitos empresariales, transformando un dato frío en una recomendación argumentada.

Evaluación de razonamiento logístico y restricciones temporales.

	Fecha	Proveedor	sku	Precio Unitario	descripcion
167	2025-03-20	DACAS COLOMBIA SAS	FSW-124F	$1,771,014	Switch FortiSwitch 124F administrado, con 24 puertos GE y 4 puertos SFP, integración con FortiGate para seguridad y gestión centralizada.
178	2025-05-26	DACAS COLOMBIA SAS	FS-124F-FPOE	$3,859,556	Switch FortiSwitch 124F con soporte Full PoE, 24 puertos GE y 4 puertos SFP, integración con FortiGate para gestión segura y centralizada.
332	2025-08-28	TICLINE S.A.S.	SW-POE-24P	$2,848,462	Switch administrable con 24 puertos Gigabit PoE+, soporte VLAN, QoS y alimentación para dispositivos como APs y teléfonos IP.

Fig 10 Caso de uso del recomendador con interpretación semántica (LLM + RAG) y limite de tiempo.

En el segundo experimento, se introdujo un contexto crítico para la operatividad del negocio: la urgencia. La consulta “Necesito un switch de 24 puertos, pero me urge, la entrega debe ser en menos de 45 días” buscaba desafiar al modelo para que descartara opciones económicas pero rápidas en lo que respecta a tiempos de entrega. Los resultados presentados en la **Fig 10** permiten analizar el comportamiento del modelo bajo esta restricción temporal.

1. Decisión estratégica del modelo: El sistema recomendó como mejor opción a ADISTEC COLOMBIA S.A.S. con el equipo FS-124F a un precio de $1,531,400. Esta elección es una victoria algorítmica significativa si se contrasta con los datos de fondo:

- Existían opciones mucho más baratas (TICLINE a $62k), pero con tiempos de entrega de 300 días.
- El modelo ignoró el precio más bajo para cumplir con la restricción de tiempo, priorizando a un proveedor que, según la base de datos histórica, tiene un tiempo de entrega registrado de 30 días (inferior al límite de 45 días solicitado).

Aunque el modelo indicó en su respuesta textual cierta cautela ("no se menciona explícitamente"), su selección final fue matemáticamente correcta, alineándose con el proveedor que cumplía la restricción de tiempos de entrega.

2. Métricas de desempeño.

El sistema proporcionó un conjunto de cinco indicadores cuantitativos que permiten auditar la calidad de la recomendación:

- **Eficiencia (100%):** se confirma que, dentro del subconjunto de proveedores que cumplían con la restricción de tiempo (< 45 días), la opción seleccionada (ADISTEC) fue la más económica. El modelo no sacrificó precio innecesariamente; eligió la opción más barata posible dentro de las condiciones logísticas dadas.
- **Confianza Semántica (99.9%):** el modelo reportó una certeza técnica casi absoluta. Esto indica que la descripción del producto encontrado ("*switch layer* 2

con 24 puertos...") se alineó perfectamente con la solicitud del usuario ("switch de 24 puertos"). No hubo ambigüedad lingüística; el sistema entendió que lo encontrado era exactamente lo pedido.

- **Cobertura (13.2%):** Esta métrica muestra que el filtro de "urgencia" fue altamente selectivo. De todo el conjunto de switches disponibles en la base de datos, solo el 13.2% calificó como relevante para esta consulta específica. Esto indica la capacidad del LLM + RAG para descartar la gran mayoría de opciones (el 86.8% restante) que, aunque eran switches, no cumplían con los criterios de entrega rápida o robustez, enfocándose la atención del comprador solo en lo viable.
- **Antigüedad del Dato (154 días):** El recomendador advierte que la información utilizada tiene aproximadamente 5 meses de antigüedad. Esta métrica es importante para el manejo de riesgos: aunque la recomendación es técnicamente sólida, le indica al comprador que debe revalidar la vigencia de la oferta con el proveedor, ya que el stock o el precio podrían haber variado en ese lapso.
- **Latencia (0.0730s):** El tiempo de procesamiento fue muy bajo, aumentando marginalmente respecto a consultas más simples. Esto valida que la arquitectura vectorial (FAISS) mantiene su rendimiento incluso cuando se le exige aplicar lógica condicional compleja ("menos de X días").

El modelo LLM + RAG demostró capacidad para gestionar *trade-offs* (intercambios de valor). Entendió que la urgencia ("45 días") era un criterio de veto superior al precio. Mientras un buscador tradicional hubiera ordenado por precio y mostrado opciones de 300 días de espera, la IA filtró eficazmente para entregar una solución viable logísticamente, protegiendo al usuario de tomar una decisión de compra que hubiera resultado en un quiebre de stock operativo.

TABLA II

COMPARATIVA DE LOS RECOMENDADORES.

Criterio de Evaluación	Baseline (Búsqueda por SKU)	Baseline (Búsqueda por Texto)	Sistema inteligente (LLM + RAG)
Tipo de Consulta	Exacta / Paramétrica	Literal / Abierta	Semántica / Compleja
Tecnología Base	Filtrado SQL / Pandas	Coincidencia de características	Embeddings vectoriales + generación de texto
Latencia Promedio	0.0061s	0.0047s	0.0730s
Nivel de Confianza	95.0% (Alta certeza binaria)	70.0% (Degradada por ruido)	99.9% (Alta alineación semántica)
Cobertura de Búsqueda	10.2% (Muy limitada)	10.2% (Muy limitada)	13.2% (Focalizada / Precisa)
Gestión de Precios	Identificación precisa del menor precio para el ítem exacto.	"Falso Ahorro" 98% Mezcla gamas incompatibles	Eficiencia (100%): Encuentra el menor precio dentro del contexto técnico válido.
Capacidad de Razonamiento	Nula. Solo ordena datos.	Nula. No distingue categorías de producto.	Alta. Interpreta atributos

VI. CONCLUSIONES Y RECOMENDACIONES

A. Conclusiones.

La finalización de este proyecto de investigación permite afirmar que el uso de Inteligencia artificial generativa, específicamente bajo la arquitectura RAG, representa un salto cualitativo hacia la modernización de los procesos de compras B2B. Al contrastar los resultados obtenidos con los objetivos trazados inicialmente, se derivan las siguientes conclusiones que sintetizan el valor de la solución:

1. Transformación de datos en activos seguros.

Más allá de la simple limpieza de archivos, el proyecto logró convertir un conjunto de datos históricos dispersos y heterogéneos en una base de conocimiento estructurada y segura. En efecto, la implementación de protocolos de anonimización fue decisiva: permitió utilizar la

potencia los modelos de lenguaje sin comprometer la confidencialidad de los clientes estratégicos de Tigo. Se demostró que la creación de campos de *embeddings* es el puente necesario para que una máquina pueda interpretar el contexto de una oferta comercial, superando las barreras que imponen las bases de datos tradicionales donde la información suele estar fragmentada en celdas aisladas.

2. **La IA como intérprete del lenguaje técnico.**

   El desarrollo del sistema semántico rompió con la rigidez de los buscadores convencionales. Dado que en el mundo de la tecnología un mismo producto puede nombrarse de muchas formas (por ejemplo, "KEM", "Módulo de expansión" o "Botonera"), los sistemas tradicionales fallaban al no encontrar coincidencias exactas. Por el contrario, el modelo implementado demostró una capacidad casi humana para entender la intención detrás de la consulta, identificando oportunidades de compra que antes permanecían ocultas. Esto confirma que la verdadera inteligencia en este contexto no reside en calcular números, sino en entender el lenguaje técnico de los productos.

3. **Calidad sobre cantidad en los resultados.**

   Las pruebas comparativas revelaron una lección crítica sobre el riesgo de los métodos estadísticos simples. El baseline tendía a entregar una falsa sensación de ahorro (reportando economías del 98%) al comparar erróneamente equipos de uso doméstico con infraestructura empresarial. En cambio, la solución basada en IA demostró superioridad al filtrar el "ruido": redujo la cantidad de resultados mostrados (del 46% al 13% del catálogo), pero aseguró que esos pocos resultados fueran técnicamente pertinentes y logísticamente viables. Así pues, se concluye que la IA actúa como un filtro de calidad que protege al usuario de tomar decisiones basadas en comparaciones incompatibles.

4. **Confianza del usuario mediante transparencia.**

   Finalmente, la interfaz desarrollada no se limitó a mostrar únicamente una cifra, sino que transformó la experiencia de consulta en un diálogo asistido. Al incluir explicaciones narrativas generadas por el modelo (IA Explicable), el sistema ofrece al equipo de preventa el "porqué" detrás de cada recomendación. Esto es fundamental, pues permite que los expertos humanos validen el criterio de las respuestas retornadas por los modelos generativos que están en auge actualmente, influyendo en la ética de uso de las inteligencias artificiales, por ejemplo, entendiendo que se eligió un proveedor más costoso porque era el

único que cumplía con la urgencia de entrega, fomentando así la confianza y la adopción de la herramienta en el día a día.

B. Recomendaciones.

El prototipo desarrollado planta las bases para un sistema robusto de inteligencia de negocios. Sin embargo, para maximizar su impacto y garantizar su sostenibilidad en el tiempo, se plantean las siguientes recomendaciones para trabajos futuros:

- **Evolución hacia el tiempo real:** actualmente, el sistema se basa con los datos presentes en el *dataset* basándose en lo que sucedió en el pasado. Teniendo en cuenta que los precios de tecnología fluctúan con el dólar y el stock cambia a diario, se recomienda conectar el sistema directamente a las fuentes vivas de información (ERPs o APIs de proveedores). Esto transformaría la herramienta de un consultor histórico a un asistente operativo en tiempo real, capaz de confirmar disponibilidad inmediata.
- **Aprendizaje colaborativo:** ningún modelo o desarrollo es perfecto. Se sugiere implementar un mecanismo simple donde los usuarios puedan calificar las respuestas del recomendador (con un "me sirvió" o "no es preciso"). De esta manera, el sistema podría aprender de la experiencia acumulada de los expertos de Tigo, afinando sus criterios de búsqueda y adaptándose a las políticas de compra de la compañía de manera orgánica y continua.
- **Visión financiera:** La decisión de compra no siempre depende solo del precio de lista. Sería valioso enriquecer el modelo de datos con variables ocultas como costos de flete, aranceles de importación o descuentos por volumen. Así, la IA podría calcular y recomendar basándose en el costo total de propiedad (TCO), ofreciendo una visión mucho más realista de la rentabilidad de cada operación.
- **Capacidad de "Leer" documentos de ofertas:** Gran parte del conocimiento técnico reside hoy en archivos PDF, fichas técnicas o imágenes de catálogos que no están digitalizados en texto. Se recomienda explorar el uso de modelos multimodales que puedan extraer y leer estos documentos directamente. Esto reduciría la carga manual de ingresar descripciones y permitiría que el sistema

responda preguntas sobre especificaciones muy detalladas que solo aparecen en los manuales de los fabricantes.

REFERENCIAS

[1] I. Barrera, "Los 12 problemas de calidad de datos más comunes y su origen," Data Ladder, 2022. [En línea]. Disponible en: https://bit.ly/3LxSNrS.

[2] PowerData, "El problema de la duplicidad de datos y cómo corregirlo," *El valor de la gestión de datos*, 2016. [En línea]. Disponible en: https://bit.ly/3LGozCV.

[3] J. Jakubik, M. Vössing and J. Walk, "Data-centric artificial intelligence," *arXiv*, 2024. [En línea]. Disponible en: https://bit.ly/3NP02MA.

[4] Y. Wand and R. Y. Wang, "Anchoring data quality dimensions on ontological foundations," *Communications of the ACM*, MIT, 1996. [En línea]. Disponible en: https://bit.ly/3YIR31S.

[5] K. Yasar and G. Lawron, "What is data preprocessing? Key steps and techniques," TechTarget, 2025. [En línea]. Disponible en: https://bit.ly/4b3GnlQ.

[6] S. Mohammed, A. Nathansen and F. Naumann, "The effects of data quality on machine learning performance on tabular data," *arXiv*, 2025. [En línea]. Disponible en: https://smurl.es/1469aV.

[7] E. A. Hinojosa Palafox, O. M. Rodríguez Elías, J. A. Hoyo Montaño, J. H. Pacheco-Ramírez and J. M. Nieto Jalil, "An analytics environment architecture for industrial cyber-physical systems big data solutions," Sensors (MDPI), vol. 21, no. 13, 2021. [En línea]. Disponible en: https://smurl.es/147pYq.

[8] M. Alves Gomes and T. Meisen, "A review on customer segmentation methods for personalized customer targeting in e-commerce use cases," Springer Nature, 2023. [En línea]. Disponible en: https://smurl.es/148Ldb.

[9] C. Niklaus, M. Cetto, A. Freitas and S. Handschuh, "A survey on open information extraction," ACL Anthology, University of Manchester, 2018. [En línea]. Disponible en: https://smurl.es/149T7a.

[10] Supriyon, A. P. Wibawa and F. Kurniawan, "A survey of text summarization: Techniques, evaluation and challenges," ScienceDirect, 2024. [En línea]. Disponible en: https://smurl.es/14bEBf.

[12] J. García Castillón, "IA explicable: algoritmos y métodos para interpretar modelos de caja negra," DEV Community, 2024. [En línea]. Disponible en: https://smurl.es/14cEEf.

[11] A. Hernandez Suarez, "Sistema de búsqueda inteligente aplicando retrieval-augmented generation (RAG)" Escuela técnica superior de ingenieros informáticos, En línea]. Disponible en: https://smurl.es/17MARk.

ANEXOS

Anexo A. Autoarchivo en Repositorio y documentos de interés

Consulte el instructivo de autoarchivo en el este enlace.

Anexo B. Póster Autoarchivo en Repositorio y documentos de interés