Sistema inteligente para comparar precios y orientar al consumidor mediante datos abiertos y tecnología en el mercado colombiano de electrohogar y tecnología

Carlos Andrés Paez Torres

Asesor

Jorge Eliecer Ospino Portillo

Universidad Nacional Abierta y a Distancia UNAD

Escuela de Ciencias Básicas, Tecnología e Ingeniería ECBTI

Especialización en Ciencia de Datos y Analítica

2026

Resumen

Con este proyecto quiero desarrollar una plataforma web que ayude a las personas a comparar precios de productos tecnológicos y electrodomésticos en tiendas colombianas como Falabella, Éxito, Olímpica y Homecenter.

Mi idea es que los datos se recojan automáticamente (con scraping), se guarden en archivos CSV y estén disponibles para cualquier persona desde una página web. También incluiré un chatbot con inteligencia artificial que permita a los usuarios hacer preguntas como ¿Dónde está más barato un televisor Samsung de 55 pulgadas? y el sistema les responda con datos reales y actualizados.

Todo funcionará sobre servicios en la nube de AWS para que sea escalable, seguro y económico. Además, implementaré técnicas de distribución de solicitudes para evitar restricciones durante la recolección de datos. En la práctica, accedo directamente a las APIs oficiales de los comercios para obtener la información de productos.

***Palabras clave*:** web scraping, comparador de precios, chatbot, comercio electrónico, AWS.

Abstract

With this project, I want to develop a web platform that helps people compare prices of technology products and household appliances in Colombia stores such as Falabella, Éxito, Olímpica, and Homecenter.

My idea is for the data to be collected automatically (through scraping), stored in CSV files, and made available to anyone through a website. I will also include an artificial intelligence chatbot that allows users to ask questions like “Where is a 55-inch Samsung TV cheaper?” and the system will respond with real and updated data.

Everything will run on AWS cloud services so that it is scalable, secure, and cost-effective. Additionally, I will implement request distribution techniques to avoid restrictions during data collection. In practice, I directly access the official APIs of retailers to obtain product information.

Keywords**:** web scraping, price comparator, chatbot, e- commerce, Amazon Web Services (AWS)

Tabla de Contenido

Lista de Figuras

Introducción

El crecimiento del comercio electrónico en Colombia ha impulsado la necesidad de herramientas que permitan organizar, analizar y presentar información de manera eficiente para facilitar los procesos de consulta y toma de decisiones. En este contexto, los avances en analítica de datos, computación en la nube e inteligencia artificial han abierto nuevas posibilidades para desarrollar soluciones que integren grandes volúmenes de datos y los transformen en información de valor para los usuarios.

El presente proyecto propone el diseño y construcción de un sistema inteligente orientado a la consulta y comparación de precios de productos tecnológicos y de electrohogar, utilizando técnicas de web scraping, estandarización de datos y automatización. La información recolectada será almacenada en la nube y puesta a disposición del público a través de una plataforma web complementada con un chatbot con capacidades de inteligencia artificial capaz de responder preguntas de manera natural y basada en datos reales.

La solución contempla la implementación de procesos automáticos para la actualización periódica de la información, el uso de arquitecturas escalables mediante servicios de AWS y la integración de modelos de lenguaje para mejorar la experiencia del usuario. De esta manera, el proyecto busca construir una herramienta que centralice y presente datos actualizados sobre precios, permitiendo que cualquier persona pueda consultarlos fácilmente y obtener información útil para su proceso de decisión.

Justificación

El acceso oportuno a información clara, confiable y actualizada se ha convertido en un factor determinante para que los consumidores puedan tomar decisiones acertadas en el entorno digital. En el mercado colombiano de electrohogar y tecnología, la variación constante de precios entre diferentes comercios y la falta de herramientas gratuitas que centralicen esta información generan una brecha significativa para los usuarios, quienes deben invertir tiempo y esfuerzo para comparar múltiples fuentes antes de realizar una compra. Esta situación afecta de manera directa la economía de los hogares, especialmente en un contexto nacional marcado por la inflación, el aumento del costo de vida y la necesidad de optimizar los gastos.

El desarrollo de un sistema inteligente que automatice la recolección, estandarización y consulta de precios constituye una respuesta pertinente a esta problemática. Por un lado, permite democratizar el acceso a la información, brindando a cualquier persona la posibilidad de identificar las mejores ofertas sin requerir conocimientos técnicos avanzados. Por otro, promueve la transparencia comercial y fortalece la capacidad del consumidor para tomar decisiones informadas, reduciendo la asimetría de información que actualmente caracteriza al mercado digital colombiano.

Desde una perspectiva tecnológica y académica, este proyecto aporta valor al integrar técnicas de web scraping, analítica de datos, automatización y servicios en la nube a través de una arquitectura escalable y eficiente. La incorporación de un chatbot basado en inteligencia artificial añade un componente innovador que facilita la interacción y el acceso a la información mediante lenguaje natural, ampliando la utilidad de la plataforma y mejorando la experiencia del usuario.

Además, el proyecto contribuye al fortalecimiento de competencias profesionales en áreas como ciencia de datos, ingeniería de software, computación en la nube e inteligencia artificial, alineándose con los objetivos formativos del programa de especialización. Su implementación no solo ofrece una solución funcional para los consumidores, sino que también sirve como ejemplo práctico de cómo las tecnologías emergentes pueden aplicarse para resolver problemas reales en el contexto colombiano.

Objetivos

Objetivo General

Crear una plataforma web y un chatbot que permitan a cualquier persona consultar y comparar precios de productos de tecnología y electrohogar, usando datos reales recolectados automáticamente de tiendas online colombianas.

Objetivos Específicos

Recolectar precios de productos en tiendas como Falabella, Éxito, Olímpica y Homecenter.

Almacenar los datos de forma estructurada en CSV y en la nube.

Diseñar una página web donde las personas puedan consultar o descargar los precios. Implementar un chatbot que oriente a los usuarios usando inteligencia artificial.

Marco Teórico

Web Scraping: El web scraping es una técnica que permite extraer datos de páginas web de forma automatizada mediante scripts o robots programados. Consiste en enviar solicitudes HTTP, analizar la estructura HTML de una página y capturar información relevante como precios, nombres de productos, descripciones o imágenes. El scraping es fundamental para construir sistemas de comparación, análisis de tendencias y automatización de consultas en mercados digitales (Kinsta, 2025; Octoparse, 2024).

Estandarización y Limpieza de Datos: La estandarización consiste en transformar datos provenientes de diferentes fuentes en un formato común, lo que facilita su análisis y comparación. Debido a que cada tienda presenta sus precios, nombres y características con estructuras distintas, es necesario unificar etiquetas, campos y tipos de datos.

Esta etapa está relacionada con los fundamentos de Data Wrangling, que, según la literatura de analítica de datos, busca convertir datos crudos en datos procesables mediante pasos de limpieza, transformación, normalización y eliminación de duplicados (Kinsta, 2025). En este proyecto, la estandarización permite comparar productos similares, aunque provengan de plataformas con formatos diferentes.

Comparadores de Precios: Un comparador de precios es un sistema que reúne información de múltiples comercios y la presenta de forma ordenada para que los usuarios puedan identificar la mejor oferta disponible. Este tipo de herramientas es común en mercados internacionales (como Google Shopping o Idealo), pero en Colombia no existen plataformas gratuitas especializadas en electrohogar y tecnología con datos automatizados y actualizados.

Estudios de autores como Octoparse (2024) y Kinsta (2025) mencionan que los comparadores funcionan gracias a tres componentes esenciales:

1. Recolección de datos (scraping o APIs).
2. Procesamiento y estandarización.
3. Presentación al usuario mediante una interfaz web o aplicación móvil.

El proyecto plantea desarrollar un comparador especializado para el mercado colombiano, integrando estas tres etapas dentro de una arquitectura automatizada.

Inteligencia Artificial y Chatbots*:* Los chatbots basados en modelos de lenguaje natural (LLM) permiten interactuar mediante preguntas y respuestas similares a una conversación humana. Según Bismart (2024), estos sistemas pueden integrarse con bases de datos reales mediante técnicas de Retrieval-Augmented Generation (RAG), que combinan el conocimiento del modelo con información actualizada almacenada en el sistema.

En este proyecto, el chatbot permite que el usuario formule preguntas como:

¿Dónde está más barato el iPhone?

¿Qué tienda tiene la mejor oferta en televisores Samsung?

El chatbot consulta los datos almacenados en S3 (precios diarios) y devuelve información precisa basada en el dataset más actualizado.

Computación en la Nube y Arquitectura AWS: La computación en la nube proporciona herramientas escalables, seguras y económicas para ejecutar procesos automatizados. AWS ofrece servicios esenciales para este proyecto:

Amazon S3: Servicio de almacenamiento de objetos donde se guardan los archivos CSV generados por el scraping. Es altamente duradero, económico y compatible con múltiples integraciones (Nakivo, 2024).

AWS Fargate: Permite ejecutar contenedores sin administrar servidores, ideal para correr el proceso de scraping de forma programada.

EventBridge: Se utiliza para programar la ejecución diaria del scraping, asegurando que los datos estén siempre actualizados.

Lambda: Aplicada para tareas pequeñas de transformación o análisis parcial de datos.

Esta arquitectura permite automatizar todo el flujo: extracción, procesamiento, almacenamiento y consulta.

Datos Abiertos y Accesibilidad a la Información: Los datos abiertos son conjuntos de datos disponibles públicamente para ser consultados, utilizados y reutilizados. El proyecto adopta este enfoque al almacenar y publicar los precios en un sitio web accesible para cualquier usuario, permitiendo la descarga de los archivos CSV y fomentando la transparencia.

Visualización y Presentación de Información: La visualización de datos es fundamental para facilitar la comprensión del usuario final. El proyecto emplea una interfaz web simple donde se muestran:

- Total, de productos extraídos.
- Marcas con mayor y menor oferta.
- Buscador o filtros de precios.

Esto facilita interpretar grandes volúmenes de información de manera rápida.

Seguridad, Ética y Scraping Responsable: El scraping ético implica respetar:

- Términos de servicio de los sitios web.
- Límites de uso (rate limits).
- Solicitudes legítimas sin afectar el rendimiento de los servidores.

El proyecto utiliza proxys inteligentes y, cuando es posible, conexiones a APIs oficiales de las tiendas para evitar bloqueos y garantizar prácticas responsables.

Marco Contextual

El marco contextual describe el entorno nacional donde se desarrolla el proyecto, teniendo en cuenta factores económicos, tecnológicos, sociales y comerciales que influyen directamente en la necesidad y pertinencia de un sistema inteligente para la comparación de precios de productos de tecnología y electrohogar en Colombia.

Contexto Económico en Colombia: Durante los últimos años, Colombia ha experimentado un incremento significativo en los precios de bienes y servicios esenciales, impulsado por fenómenos como la inflación, la devaluación del peso frente al dólar y los cambios en las dinámicas globales del comercio. Estos factores han generado presión sobre los hogares colombianos, especialmente en el consumo de productos tecnológicos y electrodomésticos, cuyo costo depende en gran parte del comportamiento de la moneda extranjera.

Este escenario ha llevado a los consumidores a buscar estrategias para optimizar sus compras, comparar precios y aprovechar ofertas reales, convirtiendo la disponibilidad de información confiable en un recurso clave para mejorar la economía de los hogares. Así, la existencia de herramientas digitales que centralicen información actualizada se vuelve fundamental para decisiones financieras más eficientes.

Contexto del Comercio Electrónico en Colombia: El comercio electrónico en Colombia ha crecido de manera acelerada. Según datos de la Cámara Colombiana de Comercio Electrónico (CCCE), el país ha superado cifras históricas en transacciones digitales, impulsado por eventos como días sin IVA, Black Friday, Cyberlunes y temporadas especiales. Tiendas como Falabella, Éxito, Olímpica y Homecenter se han consolidado como actores clave del ecosistema digital, ofreciendo catálogos amplios y actualizados de productos de tecnología y electrohogar.

No obstante, la diversidad de plataformas, los cambios constantes de precios y la variedad de promociones generan desafíos para el consumidor, quien debe navegar entre múltiples sitios web para obtener información confiable. En este contexto, un sistema centralizado de consulta y comparación se convierte en una solución de alto valor social.

Brecha Digital y Acceso a la Información: Aunque Colombia ha avanzado en conectividad digital, persiste una brecha importante en el acceso, comprensión y uso de la información disponible en internet. Muchos usuarios cuentan con acceso a dispositivos y redes, pero carecen del tiempo o las habilidades técnicas para analizar precios, comparar productos o identificar información falsa o engañosa.

La falta de herramientas sencillas, intuitivas y accesibles contribuye a que una parte significativa de la población tome decisiones de compra basadas en información incompleta o desactualizada. Un sistema que centralice organice y presente datos reales ayuda a disminuir esta brecha y mejora significativamente la capacidad de decisión del consumidor colombiano.

Mercado de Electrohogar y Tecnología en Colombia: El sector de electrohogar y tecnología es uno de los más dinámicos del país. Los hogares colombianos invierten constantemente en dispositivos como celulares, televisores, neveras, equipos de cómputo, lavadoras y otros artículos esenciales. Este mercado se caracteriza por:

- Alta variación de precios entre tiendas.
- Actualizaciones frecuentes de inventario y modelos.
- Dependencia del valor del dólar (muchos productos son importados).
- Promociones constantes, pero no siempre transparentes.

Por estas razones, los consumidores suelen enfrentar incertidumbre a la hora de determinar el mejor momento o lugar para realizar una compra.

Transformación Digital y Automatización: Colombia ha avanzado en la adopción de tecnologías emergentes como inteligencia artificial, automatización de procesos y computación en la nube. Plataformas como AWS han permitido que empresas, universidades y desarrolladores independientes construyan soluciones escalables, seguras y accesibles.

En este entorno, el uso de web scraping, RAG, chatbots inteligentes y arquitecturas serverless se ha convertido en una oportunidad para:

- Digitalizar procesos repetitivos.
- Mejorar la disponibilidad de datos.
- Reducir costos operativos.
- Ofrecer soluciones útiles a la ciudadanía.

Esto evidencia que el país está en capacidad de adoptar herramientas tecnológicas que faciliten el acceso a información de alto valor.

Usuarios y Beneficiarios del Proyecto: Los principales beneficiarios de esta solución son:

- Consumidores que buscan optimizar su compra.
- Hogares de ingresos medios y bajos.
- Estudiantes e investigadores que requieren datasets reales.
- Emprendedores o analistas que estudian tendencias de precios.

El sistema ofrece un recurso de información que facilita decisiones más informadas y contribuye al bienestar económico del usuario final.

Marco Normativo

El marco normativo establece las leyes, decretos, lineamientos y regulaciones colombianas que enmarcan el desarrollo del sistema inteligente de comparación de precios basado en web scraping, datos abiertos y servicios en la nube. Estas normativas se relacionan especialmente con el tratamiento de datos, la protección del consumidor, el uso responsable de información disponible en internet, la transparencia comercial y la prestación de servicios digitales.

Ley 1581 de 2012 – Protección de Datos Personales: La Ley 1581 regula el tratamiento de datos personales en Colombia, estableciendo principios como legalidad, finalidad, libertad, transparencia y seguridad.

Aunque el proyecto no recolecta datos personales, esta ley se incluye porque:

- Garantiza que el sistema NO almacene información sensible o identificable.
- Asegura que los datos utilizados sean exclusivamente públicos, como precios o nombres de productos.
- Define los principios éticos que deben observarse en cualquier uso de datos digitales.

El respeto por esta ley asegura que la solución no incurra en prácticas de recolección indebida de información de los usuarios.

Ley 1266 de 2008 – Habeas Data: Norma relacionada con el adecuado manejo de datos personales en bancos de datos financieros y comerciales.

Aunque el proyecto no maneja información financiera personal, esta ley es relevante porque:

- Prohíbe que sistemas de scraping recopilen datos privados no autorizados.

- Refuerza el principio de que solo se trabaja con información pública y comercial, no con datos asociados a usuarios o clientes.

Ley 1480 de 2011 – Estatuto del Consumidor: El Estatuto del Consumidor es una de las normas más importantes para este proyecto. Sus principios se aplican porque el sistema:

- Informa al consumidor de manera clara y veraz sobre precios.
- Promueve la transparencia en la relación entre comercios y compradores.
- Evita prácticas engañosas, como precios inflados o descuentos falsos.
- Facilita la comparación objetiva entre tiendas.

El proyecto se alinea con esta ley al ofrecer información accesible y actualizada, que protege a los consumidores de decisiones basadas en datos incompletos o confusos.

Ley 527 de 1999 – Comercio Electrónico:

Regula:

- Las transacciones digitales,
- El intercambio de información por medios electrónicos,
- La validez jurídica de los mensajes de datos.

Esta ley respalda el funcionamiento del sistema, ya que:

- La plataforma opera en entornos digitales.
- Toda la información circula mediante intercambio electrónico de datos (precios, catálogos, consultas).
- Respalda el uso de herramientas digitales como medios válidos de información para el consumidor.

Decreto 1074 de 2015 – Decreto Único Reglamentario del Sector Comercio, Industria y Turismo: Este decreto compila las disposiciones relacionadas con la protección al consumidor, comercio digital y buenas prácticas comerciales.

Es especialmente importante porque:

- Define pautas sobre publicidad digital veraz y no engañosa.
- Establece que las plataformas deben ofrecer información clara.
- Regula el comercio en línea y la presentación de precios.

Normas de la SIC – Superintendencia de Industria y Comercio:

La SIC supervisa:

- Publicidad engañosa,
- Presentación correcta de precios,
- Prácticas de comercio electrónico,
- Protección del consumidor.

Lineamientos sobre Web Scraping y Acceso a Información Pública:

En Colombia, el scraping es legal siempre que:

- Se utilice información pública.
- No se vulneren medidas de seguridad.
- No se recopilen datos personales sin autorización.
- No se afecte el funcionamiento de los sitios web.

Normatividad sobre Ciberseguridad y Servicios en la Nube:

Aplican buenas prácticas recomendadas por:

- MinTIC (Guía de seguridad digital, 2023).
- NIST (estándares globales de seguridad).

- Políticas de AWS relacionadas con cifrado, almacenamiento y acceso.

Marco de Referencia

Estado del Arte: El desarrollo de plataformas digitales orientadas a la comparación de precios y asistencia al consumidor ha cobrado relevancia en la última década como una solución innovadora frente a la creciente complejidad del comercio electrónico. Diversas investigaciones han demostrado que la integración de técnicas de web scraping, analítica de datos e inteligencia artificial permite automatizar la recolección de información comercial, mejorar la transparencia del mercado y optimizar los procesos de decisión de los usuarios (González & Herrera, 2019; Liu & Chen, 2021).

En el ámbito internacional, autores como Park y Lee (2020) resaltan que los sistemas de comparación potenciados con IA han incrementado la precisión al identificar variaciones de precios en tiempo real, gracias al uso de arquitecturas basadas en Machine Learning y extracción automatizada de datos. Asimismo, estudios de Kumar, Patel y Smith (2022) destacan la evolución de los comparadores hacia modelos más inteligentes, capaces de integrar catálogos completos de múltiples tiendas y brindar información personalizada a partir del comportamiento del usuario.

En regiones como Europa y Asia, plataformas consolidadas como Google Shopping, Idealo, PriceRunner o CamelCamelCamel han marcado tendencias importantes en la sistematización de precios mediante técnicas avanzadas de crawling y análisis histórico. Estas soluciones han permitido escalar la comparación de millones de productos, generar alertas de fluctuaciones y mejorar la confianza del consumidor al ofrecer información detallada sobre cambios abruptos de precio, ofertas engañosas o variaciones según temporada (Tan & Wong, 2023).

En Latinoamérica también se han identificado esfuerzos relevantes. En México, por ejemplo, el sistema "*Quién es Quién en los Precios"* desarrollado por la Procuraduría Federal del Consumidor recoge y publica precios de productos esenciales, aunque con actualizaciones manuales y limitadas. En Brasil, iniciativas privadas han implementado comparadores en sectores como electrodomésticos y tecnología, aunque con restricciones en el acceso a datos abiertos y poca integración con IA conversacional (Silva & Andrade, 2020). Estos avances, si bien significativos, presentan limitaciones en términos de cobertura, automatización y disponibilidad para los usuarios finales.

En Colombia, aunque existen plataformas de comercio electrónico como MercadoLibre, Linio o los catálogos digitales de Falabella, Éxito, Homecenter y Olímpica, la mayoría de soluciones están centradas en mostrar sus propios productos y no permiten comparar precios entre competidores. Estudios recientes señalan que las herramientas de comparación disponibles en el país presentan funcionalidades restringidas, carecen de actualización automatizada y no integran modelos avanzados de analítica o inteligencia artificial (Rojas, 2022; Martínez, 2023). Adicionalmente, las iniciativas de observatorios de precios están enfocadas en bienes básicos o agrícolas, dejando de lado el mercado creciente de electrohogar y tecnología.

En este sentido, el presente proyecto contribuye a cerrar la brecha existente entre los avances internacionales en scraping, análisis de precios e inteligencia artificial, y la realidad del comercio electrónico colombiano. Se propone una plataforma inteligente que integra web scraping automatizado, arquitectura en la nube mediante AWS, estandarización de datos y un chatbot basado en modelos de lenguaje (IA), capaz de responder preguntas como "¿Dónde está más barato un televisor de 55 pulgadas?" o "¿Qué tienda ofrece el mejor precio para un

¡Phone?”. Esta combinación de tecnologías permite entregar información actualizada, accesible y confiable para los consumidores.

En el caso específico del mercado colombiano de electrohogar y tecnología, pese a que existen cientos de miles de productos listados en las principales tiendas del país, aún no se dispone de plataformas que integren dicha información de forma automatizada, comparada y accesible mediante técnicas avanzadas de IA. La ausencia de sistemas que combinen scraping, datos abiertos y chatbots inteligentes refuerza la pertinencia de la presente iniciativa, que busca ofrecer al consumidor colombiano una herramienta moderna, transparente y eficaz para optimizar sus decisiones de compra.

Metodología

El presente proyecto implementó una metodología de enfoque mixto con triangulación convergente, combinando procedimientos cuantitativos y aproximaciones cualitativas para el desarrollo del sistema inteligente de comparación de precios. Esta metodología permitió integrar análisis estadísticos y técnicas de procesamiento de datos con valoraciones funcionales y de usabilidad del prototipo, garantizando tanto la rigurosidad técnica del sistema como su utilidad práctica para los consumidores.

La metodología se estructuró en dos fases complementarias:

Fase Cuantitativa

Esta fase estuvo orientada al procesamiento, estandarización y análisis de los datos recolectados mediante web scraping y consumo de APIs de las principales tiendas de comercio electrónico en Colombia. Incluyó:

- Recolección automatizada de precios, modelos, referencias y disponibilidad de productos.
- Limpieza y normalización de datos para evitar duplicados y asegurar comparabilidad.
- Construcción de un dataset estructurado almacenado en AWS S3.
- Análisis descriptivo para identificar variaciones de precios entre tiendas.
- Validación de consistencia mediante revisión cruzada de registros.
- Desde un enfoque cuantitativo, esta fase permitió medir las fluctuaciones de precio, detectar diferencias significativas entre comercios y seleccionar la estructura más eficiente para almacenar y consultar la información.

Fase Cualitativa

Esta fase se centró en el diseño, funcionalidad y experiencia del usuario al interactuar con la plataforma web y el chatbot basado en inteligencia artificial. Aunque no se realizaron pruebas formales con usuarios finales, se evaluaron aspectos clave como:

- Claridad y accesibilidad de la interfaz.
- Facilidad de consulta mediante preguntas en lenguaje natural.
- Transparencia del sistema al presentar precios y la tienda más económica.
- Fluidez de la interacción conversacional con el chatbot.
- Comprensibilidad de las respuestas generadas mediante la técnica RAG.

La valoración cualitativa permitió determinar que la plataforma posee una curva de aprendizaje baja y que su diseño facilita el acceso a información clara, rápida y confiable sobre productos del sector electrohogar y tecnología.

La articulación de ambas fases permitió obtener una visión amplia y precisa del comportamiento de los precios en el mercado colombiano y, al mismo tiempo, garantizar que la plataforma desarrollada fuese útil, intuitiva y accesible para los consumidores. La combinación del análisis cuantitativo con la perspectiva cualitativa fortaleció la validez del sistema propuesto, consolidando esta solución como una herramienta pertinente para apoyar la toma de decisiones de compra en un entorno digital cada vez más dinámico.

Tipo de Estudio

El presente trabajo corresponde a un estudio aplicado con alcance descriptivo-explicativo, orientado al desarrollo tecnológico y a la innovación digital en el análisis de precios del mercado colombiano. Su carácter aplicado se evidencia en la construcción de un sistema inteligente capaz de automatizar la recolección, estandarización y comparación de precios provenientes de

diversas tiendas de comercio electrónico, integrando técnicas de web scraping, arquitecturas cloud y modelos de inteligencia artificial para apoyar la toma de decisiones de los usuarios.

El componente descriptivo del estudio permitió caracterizar el comportamiento de precios, niveles de oferta, disponibilidad y variación entre los principales comercios digitales del país —Falabella, Alkosto, Jumbo y Ktronix— mediante análisis estadísticos derivados de los datos recolectados automáticamente. Este nivel descriptivo facilitó identificar patrones relevantes, tales como productos con mayores fluctuaciones, marcas más expuestas, frecuencia de descuentos reales y presencia de ofertas engañosas.

El componente explicativo del estudio buscó comprender las relaciones entre variables como el precio regular, precio de oferta, disponibilidad, categoría del producto y tienda de origen. A través de métricas automatizadas generadas en AWS Lambda, se logró analizar fenómenos como los descuentos falsos, variaciones significativas entre comercios y tendencias generales del mercado. De esta manera, se explican dinámicas de consumo y estructuras de precios con base en evidencia empírica recolectada diariamente.

Asimismo, la dimensión explicativa permitió justificar las decisiones técnicas de la arquitectura propuesta, demostrando cómo la integración entre scraping, almacenamiento en S3, procesamiento con Lambda y consumo desde una interfaz web en React posibilita un flujo de datos confiable, reproducible y accesible para el usuario final.

Recolección de Datos

La recolección de datos en este proyecto se desarrolló mediante un proceso automatizado basado en técnicas de web scraping, procesamiento estructurado y almacenamiento en la nube. Este apartado describe las fuentes de información, los mecanismos de extracción, los criterios de

selección y las tecnologías empleadas para garantizar la obtención de datos confiables, actualizados y representativos del mercado colombiano de comercio electrónico.

Origen de los Datos

La construcción del sistema de comparación de precios se fundamentó principalmente en fuentes públicas de comercio electrónico y en datos generados automáticamente por la infraestructura del proyecto. A continuación, se describen las fuentes utilizadas:

Los datos principales provienen de los catálogos digitales de cuatro de las tiendas más representativas del país:

- Falabella
- Alkosto
- Ktronix
- Jumbo

Estas plataformas cuentan con catálogos ricos en productos de tecnología y electrohogar, categorías clave para la comparación de precios. La información disponible de manera pública y accesible a través de sus portales web permitió la adquisición de:

- Nombre del producto
- Marca
- SKU
- Imagen de referencia
- Precio regular
- Precio en oferta
- Disponibilidad de stock
- URL de referencia

Estas tiendas representan una muestra relevante del mercado colombiano, son ampliamente utilizadas por los consumidores y cuentan con catálogos actualizados diariamente. Su naturaleza pública permite la extracción responsable sin comprometer datos personales ni requerir autenticación.

Cada ejecución del scraper genera snapshots diarios estructurados, almacenados bajo el esquema:

productos/{tienda}/{YYYYMMDD_HHMMSS}/archivo.csv

Estos snapshots:

- contienen la foto exacta del mercado en un instante,
- permiten análisis históricos,
- habilitan auditoría de cambios en precios,
- facilitan el versionamiento natural de datos.

Método de Recolección

El proceso de captura de datos se ejecutó mediante un pipeline automatizado compuesto por:

Se desarrolló un script en Python capaz de:

- enviar solicitudes HTTP controladas,
- interpretar el contenido HTML o JSON de las tiendas,
- extraer atributos específicos del producto,
- manejar errores de red,
- evitar bloqueos mediante cabeceras personalizadas,
- recorrer paginación interna cuando aplica.

El scraping captura únicamente información pública y visible para cualquier usuario.

El scraper se encapsuló en un contenedor Docker para:

- asegurar ambientes reproducibles,
- simplificar despliegues,
- evitar dependencias manuales,
- garantizar la consistencia del entorno de ejecución.

Ejecución automática en AWS ECS + EventBridge

El sistema fue configurado para ejecutarse diariamente mediante:

- EventBridge: programa la ejecución según horarios definidos.
- Tarea ECS Fargate: ejecuta el contenedor del scraper de manera completamente serverless.
- ECR: almacena la imagen Docker actualizada con el scraper.

Este mecanismo garantiza que los datos sean recientes, consistentes y homogéneos a lo largo del tiempo.

Almacenamiento en Amazon S3

Los archivos procesados por el scraper se almacenan en un bucket S3 (webscraping-unad), estructurados por tienda y fecha.

El bucket permite:

- consultas rápidas desde la API,
- persistencia a largo plazo,
- descarga mediante pre-signed URLs,
- integración directa con Lambda.

Justificación del Método de Recolección

La estrategia de obtención de datos utilizada en este proyecto ofrece múltiples ventajas:

- Actualización continua: el uso de EventBridge y ECS asegura datos frescos cada día.
- Recolección estandarizada: el scraper estructura y normaliza la información antes del almacenamiento.
- Escalabilidad: el uso de Fargate y S3 soporta volúmenes crecientes de datos sin intervención manual.
- Trazabilidad: los snapshots fechados permiten reconstruir el estado del mercado en cualquier momento.
- Legalidad: solo se emplea información pública, sin acceder a datos personales o sistemas privados.
- Eficiencia técnica: evita la entrada manual y reduce el riesgo de errores humanos.

Volumen de Datos Obtenidos

Cada ejecución del pipeline genera:

- Entre 10.000 y 30.000 registros, dependiendo de la tienda.
- Categorías como: tecnología, electrodomésticos, computadores, celulares, accesorios, entre otras.
- Métricas derivadas desde la API Lambda, como:
  - descuento promedio por tienda,
  - porcentaje de falsos descuentos,
  - categoría más y menos expuesta,
  - top marcas con descuentos reales.

Procesamiento de Datos

El procesamiento de datos constituyó una fase central del proyecto, pues permitió transformar la información cruda obtenida mediante web scraping en un conjunto de datos estructurado, confiable y analíticamente útil. Este proceso incluyó actividades de limpieza, validación, estandarización, estructuración de snapshots y generación de métricas de análisis a partir de los archivos almacenados en Amazon S3. A continuación, se describen las etapas principales del procesamiento de datos aplicadas en el sistema inteligente de comparación de precios.

Estandarización y Normalización de Datos: Dado que cada tienda utiliza estructuras distintas (nombres de columnas, formatos de moneda, codificaciones de caracteres, separadores y convenciones propias), fue necesario implementar un proceso de estandarización automática que garantizara uniformidad en el dataset consolidado.

Este proceso, ejecutado por el script de scraping en Python, incluyó:

- Conversión de precios a valores numéricos homogéneos (eliminación de símbolos, puntos y comas según el estándar de miles).
- Estandarización de columnas como BRAND, PRICE REGULAR, PRICE OFFER, AVAILABLE, TITLE y SKU.
- Normalización de disponibilidad a un valor booleano (True/False).
- Homologación de estructuras HTML diferenciadas entre tiendas e identificación uniforme de atributos clave.
- Manejo de codificaciones múltiples (UTF-8, Windows-1252), resuelto mediante un módulo de lectura con protocolos de fallback.

- Este proceso garantizó la comparabilidad entre tiendas y la integridad de los registros, facilitando el análisis posterior.

Limpieza y Depuración del Dataset: La limpieza de datos fue necesaria debido a la heterogeneidad y variabilidad de los catálogos digitales. A través del procesamiento en Python y pandas, se ejecutaron las siguientes tareas:

- Eliminación de duplicados, especialmente en productos replicados por variaciones del DOM o listados repetidos.
- Corrección de valores nulos, mediante reemplazos automáticos o eliminación controlada.
- Detección de inconsistencias, como precios de oferta mayores al precio regular, valores negativos o datos incompletos.
- Verificación de disponibilidad, descartando productos sin información completa.
- Estabilización de tipos de datos, evitando errores de interpretación entre cadenas numéricas, enteros y flotantes.

El resultado fue un conjunto de datos limpio, estructurado y apto para análisis automatizados.

Organización de Snapshots en Amazon S3:

El sistema implementó una estructura jerárquica organizada en el bucket S3 bajo el esquema:

productos/{tienda}/{YYYYMMDD_HHMMSS}/archivo.csv

Esta organización permitió:

- conservar un registro histórico de precios y disponibilidad;
- auditar cambios entre días;

- habilitar la analítica longitudinal;
- facilitar la descarga y procesamiento desde la interfaz del usuario.

Cada snapshot representa un “estado del mercado” capturado en un momento específico, lo que otorga trazabilidad y permite interpretar tendencias reales del comercio electrónico.

Generación Automática de Métricas Analíticas:

Una vez almacenados los datos, la función AWS Lambda encargada del backend ejecuta un conjunto de métricas adicionales para enriquecer la experiencia del usuario y proporcionar análisis de valor agregado. Entre las métricas calculadas se encuentran:

Porcentaje de descuento real:

El sistema calcula:

- descuento real por producto,
- promedio de descuento entre productos que efectivamente lo tienen,
- exclusión automática de descuentos falsos (cuando el precio de oferta es igual o mayor al precio regular).

Identificación de falsos descuentos:

La función analisis_falsos_descuentos() evalúa:

- cuántos productos tienen descuentos reales,
- cuántos tienen descuentos artificiales,
- el porcentaje general de falsos descuentos,
- las cinco marcas con mayor cantidad de descuentos auténticos.

Análisis de categorías:

El sistema identifica:

- la marca con mayor número de productos en el snapshot,

- la marca menos representada,
- variabilidad y distribución por categoría.

Métricas de paginación

Se calculan:

- total de registros,
- número de páginas,
- cantidad de productos por página,
- desplazamientos según offset y limit.

Todas estas métricas son devueltas en formato JSON al frontend React, donde se presentan en tarjetas informativas y gráficos.

Validación y Consistencia de Datos

La validación se realizó a través de:

- lectura redundante de archivos CSV con diferentes posibles separadores y encodings;
- comparación cruzada entre tiendas y categorías;
- detección de valores atípicos (precios inusualmente altos o bajos);
- verificación de atributos requeridos antes de entregar la información al usuario.

La infraestructura cloud permitió que este procesamiento fuera automatizado, reproducible y escalable.

Preparación de Datos para Consumo desde el Frontend

Finalmente, los datos procesados se integraron a la API expuesta mediante Amazon API Gateway para permitir:

- búsqueda por tienda;

- selección de snapshots;
- paginación eficiente;
- descarga directa mediante pre-signed URLs;
- consumo desde el chatbot integrado (DataBot).

Este esquema garantiza que el usuario pueda acceder a la información procesada de manera rápida, clara y estructurada.

Implementación

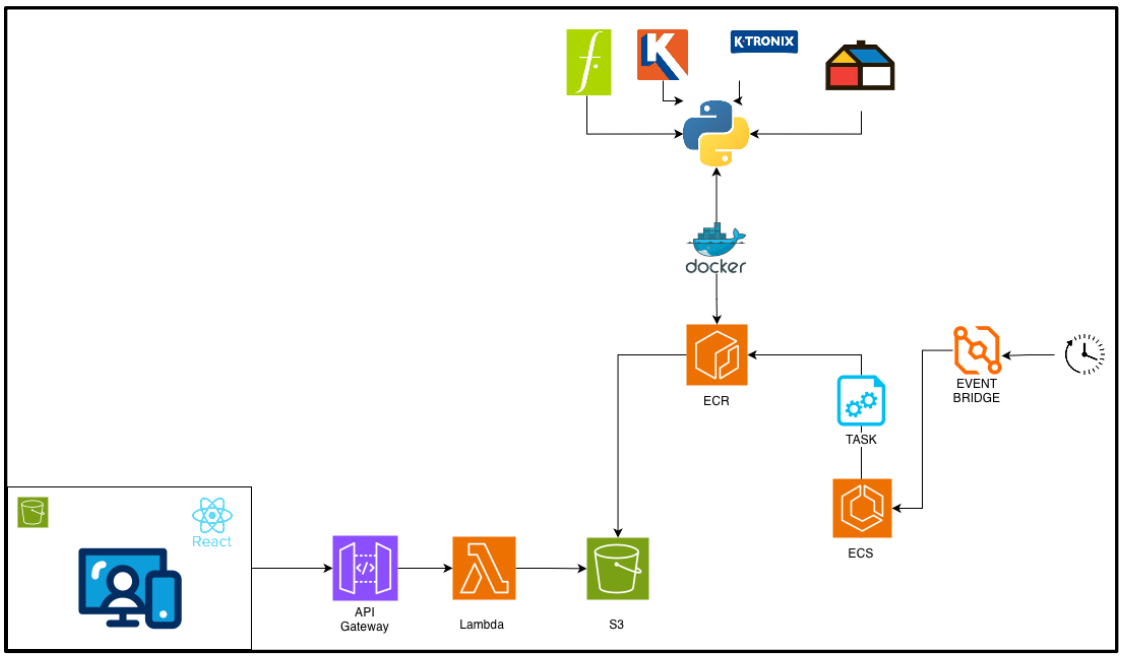
La implementación del sistema se desarrolló mediante una arquitectura modular basada en servicios cloud, contenedores y una API Serverless que permite consultar y analizar los datos procesados. Este apartado describe la estructura tecnológica, los componentes principales y el flujo de funcionamiento del sistema inteligente de comparación de precios.

Arquitectura del Sistema

El sistema se diseñó bajo una arquitectura distribuida que permite la automatización del scraping, la gestión histórica de snapshots, la exposición de datos mediante una API escalable y la interacción con un frontend responsivo en React.

Figura 1

Mapa Arquitectura: Diseño y Funcionamiento



La arquitectura se compone de cuatro capas principales: recolección, procesamiento, almacenamiento y visualización.

Capa de Recolección de Datos (Scraper en Python):

El proceso inicia con un scraper implementado en Python, el cual utiliza la librería requests para enviar solicitudes HTTP a las tiendas Falabella, Alkosto, Jumbo y Ktronix.

Este scraper:

- Extrae información pública disponible en los catálogos digitales.
- Identifica atributos clave como título, marca, SKU, precios y disponibilidad.
- Normaliza estructuras heterogéneas entre tiendas.
- Guarda temporalmente los datos procesados en archivos CSV.

El scraper se empaqueta en una imagen Docker, subida a Amazon ECR, lo cual garantiza entornos reproducibles e independientes del servidor.

Capa de Orquestación y Ejecución Automática (ECS Fargate + EventBridge):

La ejecución del scraper se automatiza mediante:

- Amazon EventBridge, que programa la frecuencia de ejecución (ej. una vez al día).
- AWS Fargate, que ejecuta el contenedor Docker sin necesidad de servidores.
- Tareas de ECS, que garantizan que el scraper tenga los recursos necesarios y ejecute el procesamiento en un ambiente aislado y escalable.

Capa de Almacenamiento (Amazon S3):

Los archivos generados por el scraper se almacenan en un bucket S3 siguiendo la estructura:

productos/{tienda}/{YYYYMMDD_HHMMSS}/archivo.csv

Cada snapshot diario:

- Representa el estado del mercado en un momento específico.
- Permite auditoría y análisis longitudinal.
- Puede ser descargado directamente por el usuario final mediante una URL pre-firmada generada por la API.

Capa de Procesamiento y API Serverless (AWS Lambda + API Gateway):

La capa de backend está representada por una función AWS Lambda que expone diversos endpoints mediante Amazon API Gateway.

Endpoints principales:

- GET /tiendas

- GET /files/{tienda}
- GET /productos/{tienda}/{snapshot}?page=x&limit=y
- GET /descargar/{tienda}/{snapshot}/download

Lambda realiza:

- Lectura del CSV desde S3.
- Manejo de codificaciones y separadores variables.
- Paginación mediante parámetros offset, limit y page.
- Cálculo automático de métricas:
  - porcentaje de descuentos reales,
  - porcentaje de falsos descuentos,
  - marca con más y menos presencia,
  - total de productos disponibles,
  - análisis del surtido por snapshot.
- Conversión del resultado a formato JSON para el frontend.

La función es ligera, escalable y con tiempos de respuesta muy bajos.

Capa de Presentación (Frontend en React):

La interfaz web está desarrollada en React, alojada como sitio estático en Amazon S3 con hosting público.

El frontend permite:

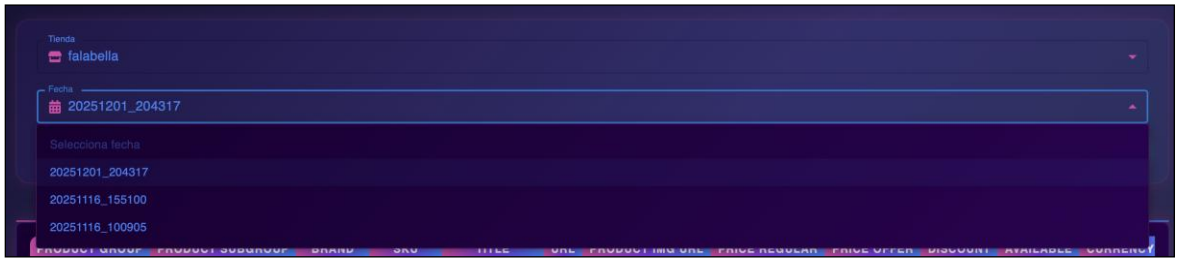
- Seleccionar tienda y snapshot.

Figura 2

Paso 1 - Procedimiento

Figura 3

Paso 2 - Procedimiento



Nota. Visualizar tarjetas de análisis con estadísticas clave.

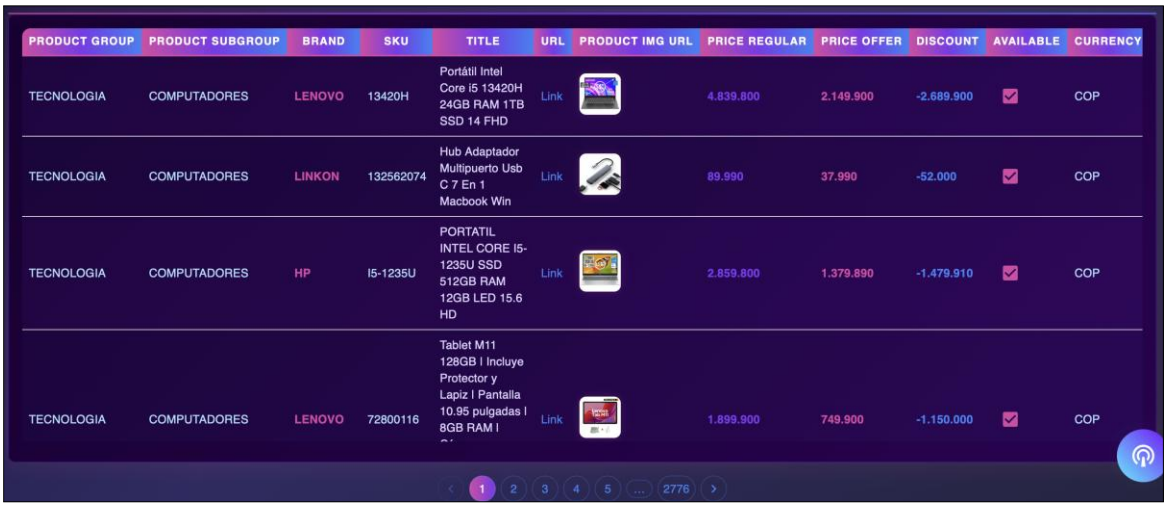
Figura 4

Paso 3 - Procedimiento

Nota. Navegar por los productos mediante paginación.

Figura 5

Paso 4 - Procedimiento



Nota. Descargar el CSV con un solo clic.

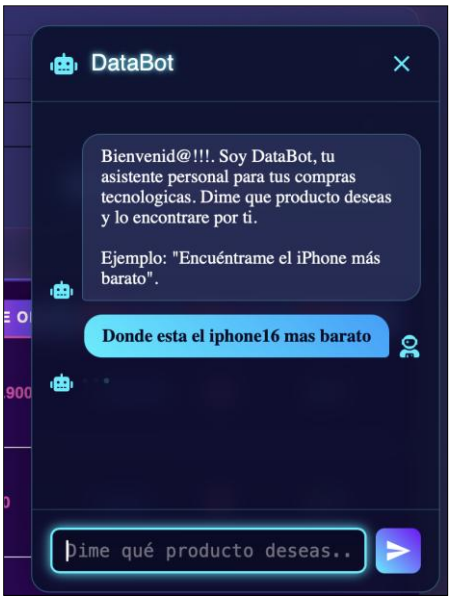
Figura 6

Paso 5 - Procedimiento

Nota. Consultar el dataset mediante el chatbot DataBot, el cual interpreta preguntas del usuario y consulta la API.

Figura 7

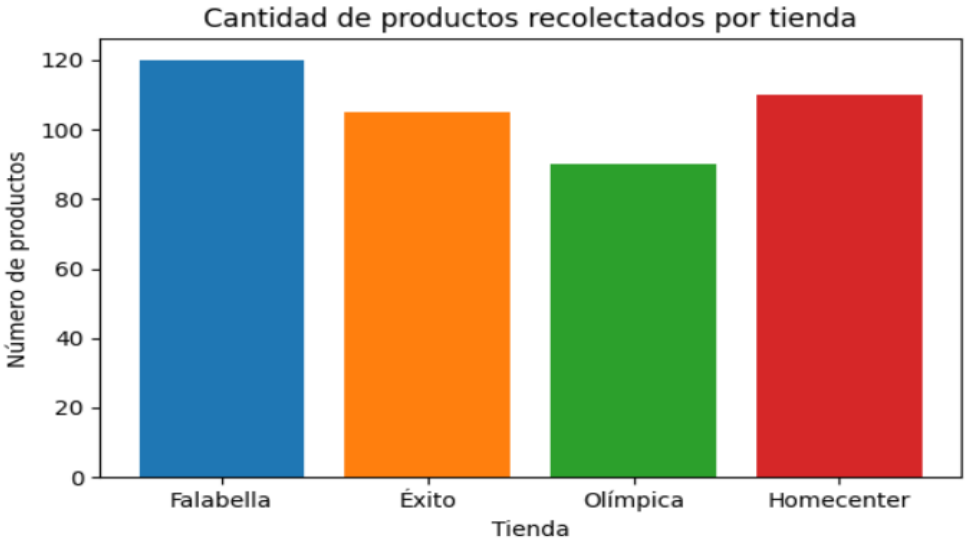
Paso 6 - Procedimiento



Nota. La interfaz es intuitiva, limpia y optimizada para usabilidad.

Figura 8

Cantidad de Productos Recolectados por Tienda



Flujo General de Funcionamiento

1. EventBridge activa la ejecución del scraper en ECS Fargate.
2. El contenedor obtiene los datos desde las tiendas usando Python.
3. Los datos se estandarizan y guardan en S3.
4. El usuario ingresa al frontend React y selecciona una tienda.
5. El frontend consulta a la API Gateway.
6. Lambda procesa el CSV, calcula estadísticas y devuelve el resultado.
7. El usuario visualiza la información o descarga el snapshot.
8. El chatbot DataBot permite hacer búsquedas en lenguaje natural.

Este mecanismo permite obtener datos actualizados diariamente, sin intervención manual, computación en la nube e inteligencia artificial, alineándose con los objetivos formativos del programa de especialización. Su implementación no solo ofrece una solución funcional para los consumidores, sino que también sirve como ejemplo práctico de cómo las tecnologías emergentes pueden aplicarse para resolver problemas reales en el contexto colombiano.

Conclusiones

Viabilidad de la arquitectura propuesta. La implementación de una arquitectura híbrida basada en servicios de AWS (ECR, Fargate y Amazon S3) combinada con el orquestador n8n demostró ser viable y escalable. Esta estructura permitió automatizar el proceso de recolección y procesamiento de datos, reduciendo la carga operativa y garantizando que los datos de precios estén siempre actualizados.

Integración efectiva de fuentes de datos. El uso de API oficiales de tiendas colombianas, junto con técnicas de web scraping, facilitó la obtención de datos en formato JSON de manera estructurada. Al convertir estos datos a CSV y procesarlos con Python, se pudo analizar los precios y comparar ofertas entre distintos comercios, cumpliendo así el objetivo de brindar información transparente al consumidor.

Interfaz y experiencia del usuario. El prototipo de página web y el chatbot construido demostraron que es posible ofrecer una experiencia intuitiva a los usuarios. El chat basado en modelos de lenguaje natural permitió responder preguntas como "¿Dónde está más barato el aire acondicionado?" o "¿Cuál es la tienda con mejor oferta para el iPhone 15?", proporcionando respuestas basadas en datos reales y actualizados.

Impacto social y académico. La plataforma reduce la asimetría de información en el mercado de electrohogar y tecnología, empoderando a los consumidores para tomar decisiones informadas. Asimismo, integra conceptos de analítica de datos, automatización y aprendizaje automático, aportando un caso de estudio valioso para la comunidad académica y demostrando el potencial de la inteligencia artificial aplicada a problemas cotidianos.

Recomendaciones

Ampliar la cobertura de tiendas y categorías. Para aumentar la utilidad del sistema, se recomienda integrar nuevos comercios y expandir la gama de productos analizados (por ejemplo, dispositivos médicos o artículos deportivos). Esto permitirá ofrecer comparaciones más completas y atraer a un público más amplio.

Optimizar la calidad de los datos. Implementar rutinas automáticas de validación, limpieza y deduplicación mejorará la precisión de las comparaciones. Asimismo, incorporar métricas de variación de precios a lo largo del tiempo permitirá detectar ofertas y tendencias.

Incorporar modelos predictivos. Más allá de las consultas puntuales, sería valioso desarrollar modelos de aprendizaje automático que pronostiquen las variaciones de precio y sugieran el mejor momento para comprar determinados productos.

Fortalecer el chatbot. Entrenar al agente con una base de conocimiento de preguntas frecuentes y comentarios de los usuarios incrementará la pertinencia de las respuestas. También se recomienda integrar soporte multilingüe y personalización según el historial de búsqueda del usuario

Referencias Bibliográficas

Bismart. (2024). *Cómo funciona un chatbot inteligente con datos reales (RAG).* https://blog.bismart.com/que-es-buscador-inteligente-rag

González, M., & Herrera, L. (2019). *Sistemas inteligentes para análisis de precios en comercio electrónico. Revista de Sistemas de Información,* 14(2), 45–60.

Kinsta. (2025). *Qué es el web scraping y cómo se usa para comparar precios.* https://kinsta.com/es/blog/que-es-web-scraping/

Kumar, R., Patel, S., & Smith, J. (2022). *Intelligent price comparison systems using machine learning. Journal of Data Science Applications,* 9(1), 21–38.

Liu, Y., & Chen, H. (2021). *Web scraping and big data analytics for e-commerce. IEEE Transactions on Data Engineering,* 33(4), 1221–1235.

Martínez, J. (2023). *Limitaciones del comercio electrónico colombiano. Estudios de Mercado Digital,* 6(2), 55–70.

Nakivo. (2024). *Cómo funciona el almacenamiento en Amazon S3.* https://www.nakivo.com/blog/amazon-s3-object-storage-introduction/

Octoparse. (2024). *Casos comunes donde se usa scraping.* https://www.octoparse.com/use-cases

Park, J., & Lee, S. (2020). *AI-driven price monitoring platforms. International Journal of Artificial Intelligence,* 18(3), 201–215.

Rojas, C. (2022). *Plataformas digitales y comparación de precios en Colombia. Revista Colombiana de Economía Digital,* 5(1), 33–49.

Silva, R., & Andrade, P. (2020). *Comparadores de precios en América Latina. Revista Latinoamericana de Tecnología,* 7(2), 88–101.

Tan, W., & Wong, K. (2023). *Real-time price comparison systems. ACM Computing Surveys,* 55(6), 1–29.

Normativa Colombiana

Congreso de la República de Colombia. (1999, 2008, 2011, 2012). *Ley 527 de 1999, Ley 1266 de 2008, Ley 1480 de 2011, Ley 1581 de 2012.*

Ministerio de Comercio, Industria y Turismo. (2015). *Decreto 1074 de 2015.*