# UNIVERSIDAD EUROPEA DE MADRID

## ESCUELA DE ARQUITECTURA, INGENIERÍA Y DISEÑO

### MASTER DE FORMACIÓN PERMANENTE EN INTELIGENCIA ARTIFICIAL

TRABAJO FIN DE MÁSTER

# CockRoach

**XAVIER RAMBLA CENTELLAS**

**Dirigido por**

**Dr. Jorge Moratalla Collado**

**CURSO 2024-2024**

**TÍTULO**: CockRoach

**AUTOR**: XAVIER RAMBLA CENTELLAS

**TITULACIÓN**: MASTER DE FORMACIÓN PERMANENTE EN INTELIGENCIA ARTIFICIAL

**DIRECTOR/ES DEL PROYECTO**: Dr. Jorge Moratalla Collado

**FECHA**: septiembre de 2024

# RESUMEN

La industria del retail ha experimentado una evolución en los últimos años, transformándose de tienda física en una entidad compleja, altamente automatizada y con altos volúmenes de negocio. En este contexto, la tecnología ha sido el principal catalizador de avances disruptivos en todas las áreas operativas de las empresas.

En el entorno competitivo actual, la implementación de técnicas avanzadas de fijación dinámica de precios, apoyadas en herramientas y algoritmos, se ha convertido en una necesidad estratégica. Estas soluciones permiten no solo maximizar beneficios, sino también aumentar volúmenes de venta, o incluso lograr ambos objetivos.

Conocer los precios de la competencia proporciona una ventaja competitiva crucial, ya que permite comprender mejor el mercado, ajustar las estrategias comerciales, mejorar los márgenes de ganancia y tomar decisiones empresariales basadas en datos. El análisis también permite identificar patrones y estrategias de comportamiento de los competidores, lo que facilita la anticipación de movimientos futuros y la formulación de respuestas proactivas.

CockRoach tiene como objetivo recuperar y analizar los precios de la competencia para extraer información estratégica valiosa para la empresa.

CockRoach satisface las necesidades siguientes:

- Ecommerces que buscan una ventaja competitiva.
- Fabricantes para supervisar los rangos de precios definidos en los canales de distribución externos.
- Departamentos de Compras que buscan obtener precios más competitivos.

CockRoach ha sido desarrollado en colaboración con Beral Projects S.L., una empresa dedicada al retail y al ecommerce.

Actualmente, el proyecto se encuentra operativo en Producción y se está diseñando nuevas funcionalidades para una nueva versión.

**Palabras clave:**

- Precio Dinámico
- Seguimiento de Precios de la Competencia
- Optimización de Precios Basada en Reglas
- Monitoreo de Precios Competitivos
- Scraping de Precios
- Investigación de Mercado

# ABSTRACT

The Retail industry has undergone significant evolution in recent years, transforming from physical stores into a complex, highly automated entity with substantial business volumes. In this context, technology has been the main catalyst for disruptive advancements across all operational areas of companies.

In today's competitive environment, the implementation of advanced dynamic pricing techniques, supported by tools and algorithms, has become a strategic necessity. These solutions not only allow for profit maximization but also increase sales volumes, or even achieve both objectives simultaneously.

Understanding competitor pricing provides a crucial competitive advantage, as it allows for a better understanding of the market, adjusting commercial strategies, improving profit margins, and making data-driven business decisions. Analysis also helps identify patterns and competitor behavior strategies, enabling companies to anticipate future moves and formulate proactive responses.

CockRoach aims to retrieve and analyze competitor prices to extract valuable strategic information for the company.

CockRoach responds to the following business needs:

E-commerce businesses seeking a competitive edge.
Manufacturers looking to monitor price ranges set in external distribution channels.
Purchasing departments aiming to secure more competitive prices.
CockRoach has been developed in collaboration with Beral Projects S.L., a company dedicated to retail and e-commerce.

Currently, the project is fully operational in Production, and new features are being designed for the next version.

**Keywords:**

- Dynamic Pricing
- Competitor Price Tracking
- Rule-based pricing optimization
- Competitive Price Monitoring
- Price Scraping
- Market Research

## AGRADECIMIENTOS

Quiero expresar mi más sincero agradecimiento a todas las personas que me han acompañado a lo largo de este camino, tanto en lo personal como en lo académico.

En primer lugar, a mi mujer y mis hijos, por su paciencia, comprensión y apoyo incondicional durante todo este proceso. Gracias por estar siempre a mi lado, por brindarme fuerzas en los momentos difíciles y por recordarme cada día lo verdaderamente importante. Sin vuestro amor y sacrificio, nada de esto hubiera sido posible.

A mis padres, por inculcarme desde pequeño el valor del esfuerzo y la perseverancia, y por darme siempre su confianza y aliento. A mi hermana, por estar siempre dispuesta a escucharme y darme su apoyo incondicional. A toda mi familia, que ha sido un pilar fundamental a lo largo de este viaje académico.

Un agradecimiento especial a Jorge Moratalla, mi tutor, por su orientación experta, sus valiosos consejos y por guiarme de manera firme y generosa durante todo el desarrollo del proyecto. Ha sido un honor y un placer aprender de su vasta experiencia y conocimiento. A Miguel Ángel Rodríguez López por sus clases y por despertar mi interés en el campo de la IA.

Agradecimiento a Jonathan y Óscar por su apoyo incondicional, por sus conocimientos, consejos y ayuda. A todo el equipo de Beral Projects que ha hecho posible la realización de este proyecto y ha aportado su granito de arena.

Finalmente, a mis compañeros del máster, José Camacho y Jesús Garcia, gracias por su compañerismo, por las largas horas de trabajo compartido y por el constante intercambio de ideas y apoyo mutuo. El viaje ha sido más fácil, divertido y enriquecedor gracias a vosotros.

A todos, gracias de corazón. Este logro es tanto mío como vuestro.

**Cita - frase célebre / Dedicatoria**

La inteligencia artificial es la puerta hacia un futuro donde la creatividad y la tecnología se entrelazan para resolver los desafíos más complejos de la humanidad

# TABLA RESUMEN

| | **DATOS** |
|---|---|
| **Nombre y apellidos:** | Xavi Rambla Centellas |
| **Título del proyecto:** | CockRoach |
| **Directores del proyecto:** | Jorge Moratalla |
| **El proyecto se ha realizado en colaboración de una empresa o a petición de una empresa:** | SI |
| **El proyecto ha implementado un producto:** (esta entrada se puede marcar junto a la siguiente) | SI |
| **El proyecto ha consistido en el desarrollo de una investigación o innovación:** (esta entrada se puede marcar junto a la anterior) | NO |
| **Objetivo general del proyecto:** | Realización de estudios basados en el precio del producto |

# Índice

# Índice de Figuras

# Índice de Tablas

# Capítulo 1. RESUMEN DEL PROYECTO

El presente documento constituye el documento del Trabajo Final de Máster (TFM) que se desarrolla en el marco del programa de Máster de Formación Permanente en Inteligencia Artificial en la Universidad Europea. El propósito de este trabajo es documentar la solución tecnológica para la búsqueda de una ventaja competitiva empresarial basada en la obtención, tratamiento y análisis de la información relacionada con los precios de productos de la competencia, con el fin de facilitar/automatizar la toma de decisiones estratégicas en los contextos empresariales.

Con el uso de esta solución, las empresas dispondrán de una actualización continua de los precios de los productos de la competencia y dispondrán de soluciones gráficas avanzadas que permitirán disponer de información visual para la comprensión y visualización de patrones, tendencias y cambios para la toma de decisiones.

## 1.1 Contexto y justificación

El proyecto se sitúa en un contexto empresarial cada vez más competitivo y dinámico, donde la capacidad para adaptarse rápidamente a los cambios del mercado y tomar decisiones informadas se ha convertido en un factor determinante para el éxito. En este sentido, la inteligencia competitiva juega un papel crucial, ya que permite a las empresas comprender mejor el posicionamiento del mercado, ajustar sus estrategias de precios y diferenciarse de la competencia. Por ejemplo,[i] hallazgos cómo que la publicidad de precios tiene un efecto positivo en la sensibilidad al precio (Kaul and Wittink 1995[ii]).

Sin embargo, la recopilación manual de datos de precios de la competencia puede ser una tarea laboriosa y propensa a errores, especialmente en entornos donde el mercado está en constante evolución y la información es abundante y dispersa. Es aquí donde la aplicación de técnicas de inteligencia artificial puede ofrecer una solución eficiente y automatizada para la recopilación, procesamiento y análisis de datos de precios, permitiendo a las empresas acceder a información actualizada y precisa de manera rápida y efectiva.

En este contexto, el presente proyecto se justifica como una iniciativa destinada a desarrollar y aplicar técnicas avanzadas de inteligencia artificial para la obtención y análisis de precios de la

competencia. La implementación de estas técnicas no solo permitirá a las empresas optimizar su estrategia de precios, sino también identificar oportunidades de mercado, anticipar tendencias y mejorar su posición competitiva en el mercado.

Basado el trabajo en una versión del framework de Kopalle[iii] se buscará dar respuesta a las preguntas relacionadas con el precio.

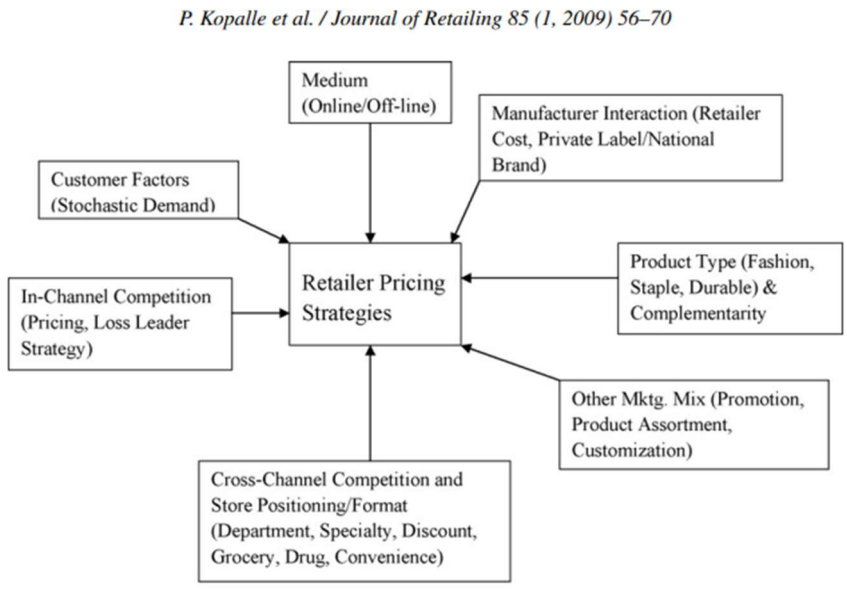



*Figura 1 P. Kopalle et al. / Journal of Retailing 85 (1, 2009) 56–70*

Asimismo, el desarrollo de este proyecto contribuirá al avance del conocimiento en el campo de la inteligencia artificial aplicada a la gestión empresarial, al tiempo que proporcionará al estudiante la oportunidad de aplicar los conocimientos y habilidades adquiridos durante el programa de máster en un contexto práctico y relevante.

## 1.2 Planteamiento del problema

En el entorno empresarial altamente competitivo de hoy, las empresas necesitan tener una comprensión clara y actualizada de los precios de los productos de la competencia para tomar decisiones informadas. Un software de estudio de mercado que monitoree y analice estos

precios puede proporcionar una ventaja estratégica significativa. Este planteamiento del problema describe la necesidad, los objetivos, los requisitos y los beneficios esperados de desarrollar un software para realizar estudios de mercado sobre los precios de los productos de la competencia.

## 1.3 Objetivos del proyecto

El objetivo principal de este estudio es desarrollar un software de inteligencia competitiva usando inteligencia artificial capaz de recopilar, analizar y procesar información relevante sobre los precios de productos o servicios ofrecidos por competidores en el mercado. Este sistema se centrará en la identificación de patrones, tendencias y comportamientos de precios, con el fin de proporcionar a las empresas una visión integral y actualizada del entorno competitivo para poder comprender las oportunidades y desafíos del entorno.

Específicamente, los objetivos del estudio son los siguientes:

- Desarrollar un sistema de recopilación de datos.
- Aplicar técnicas avanzadas de análisis de datos.
- Cobertura amplia
- Generar informes, insights, alertas y recomendaciones.
- Evaluar la efectividad del sistema.


Al alcanzar estos objetivos, se espera que este estudio contribuya significativamente al campo de la inteligencia competitiva innovando con las tecnologías de inteligencia artificial aplicada al análisis competitivo y la toma de decisiones empresariales.

Aunque el proyecto está destinado a resolver un problema genérico de cualquier sector aplicando el framework de Kopalle con la premisa de obtener los precios públicos y accesibles a través de internet ( se excluyen los precios de tienda física), se va a acotar el alcance específicamente al nicho de productos relacionados con las piscinas y Wellness. Esta decisión es consecuencia de la financiación de este máster por una empresa del sector y adaptaremos algunas de las tareas del diagrama de Kopalle para ajustarlas a las características del mercado concreto.

## 1.4 Resultados obtenidos

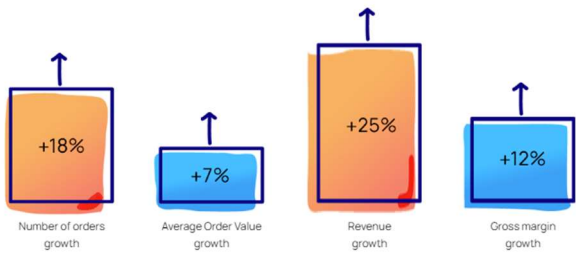
En este apartado se resume en grandes rasgos los resultados obtenidos gracias a la solución implementada. Comentar que únicamente tiene un mes de vida en producción con lo que todavía es pronto para rellenar esta sección, aunque ya hay algunas afirmaciones que pueden realizarse.

### 1.4.1 Resultados obtenidos:

- Solución tecnológica basada en un portal web administrable y configurable.
- Arquitectura diseñada para ser totalmente escalable (a excepción de la Base de datos que es centralizada).
- Obtención de información en 24h, y preparado para obtenerla en tiempo real.
- Sistema de Scrapping avanzado y configurable por el usuario en el portal (desde cambiar el navegador, cambiar el buscador, …. ).
- Independencia de LLM, configurable y facilidad en la integración con los futuros LLMs que saldrán.
- Sistema de notificaciones configurable.

### 1.4.2 Resultados obtenidos por negocio en su corto recorrido de vida.

En su corto periodo de vida del proyecto en producción ya se dispone de algunos resultados para negocio que permiten obtener una perspectiva más precisa de la competencia y del entorno.




- Posicionamiento del precio de nuestros productos respecto la competencia
- Posicionamiento de nuestra empresa respecto al resto en cuestión de precios.
- Avances tecnológicos de la competencia.
- Estrategias irregulares cómo bajar los precios en fin de semana por debajo del límite marcado por el proveedor.
- Estrategias irregulares cómo unir varios productos en una promoción para vender por debajo del límite marcado por el proveedor.
- Detección de las rebajas de precio de la competencia

- Detección de la manipulación del precio de un producto concreto en un corto periodo de tiempo (10 minutos).

## 1.5 Estructura de la memoria

La presente memoria se compone de nueve capítulos donde detalla exhaustivamente el desarrollo y los resultados del proyecto. A continuación, se describe brevemente el contenido de cada capítulo para facilitar al lector una visión global de la estructura del documento:

**Resumen y Abstract**

Se ofrecen un resumen en español y un abstract en inglés que sintetizan los objetivos, metodología y conclusiones principales del proyecto, permitiendo una comprensión rápida de su alcance y resultados.

**Tabla Resumen**

Proporciona una visión general de los aspectos clave del proyecto, incluyendo datos relevantes y conclusiones destacadas, facilitando así una rápida referencia.

**Capítulo 1: Resumen del Proyecto**

Este capítulo introduce el contexto y justificación del proyecto, presenta el planteamiento del problema, define los objetivos generales y específicos, expone los resultados obtenidos y finaliza con la estructura de la memoria, orientando al lector sobre el contenido del documento.

**Capítulo 2: Antecedentes / Estado del Arte**

Se realiza un análisis detallado del estado del arte, examinando las soluciones existentes y las investigaciones previas relacionadas con el tema. Se profundiza en el contexto y justificación adicionales, se reitera el planteamiento del problema y se lleva a cabo un análisis crítico que identifica las brechas que el proyecto pretende abordar.

**Capítulo 3: Objetivos**

Se detallan los objetivos generales y específicos del proyecto, estableciendo claramente las metas a alcanzar. Además, se destacan los beneficios del proyecto, tanto desde una perspectiva académica como práctica, y se presentan unas conclusiones preliminares que sirven de guía para el desarrollo posterior.

**Capítulo 4: Desarrollo del Proyecto**

Este capítulo constituye el núcleo del trabajo, abarcando la planificación del proyecto, la descripción de la solución propuesta, las metodologías y herramientas empleadas, y los recursos requeridos. Se incluye un análisis del presupuesto y la viabilidad del proyecto. Se presentan los resultados obtenidos, los beneficios para el negocio, y se discuten las adaptaciones realizadas para satisfacer las necesidades específicas. Finalmente, se ofrece una conclusión final sobre el desarrollo realizado.

**Capítulo 5: Discusión**

Se exploran en profundidad los desafíos y problemas encontrados durante el proyecto, como los relacionados con los buscadores y el diseño de la arquitectura del módulo de IA. Se analiza cómo se abordaron estos problemas y las soluciones implementadas, proporcionando una reflexión crítica sobre el proceso.

**Capítulo 6: Conclusiones**

Se presentan las conclusiones del trabajo, evaluando el grado de cumplimiento de los objetivos y el impacto de los resultados. Se incluyen también conclusiones personales, donde se reflexiona sobre el aprendizaje obtenido y el desarrollo profesional a lo largo del proyecto.

**Capítulo 7: Futuras Líneas de Trabajo**

Se proponen mejoras y extensiones para el proyecto, identificando áreas donde se puede profundizar, como en el frontend web, las mejoras en IA y posibles innovaciones que podrían implementarse en trabajos futuros.

**Capítulo 8: Anexos**

Se incluyen materiales complementarios que apoyan y enriquecen el contenido de la memoria, como códigos fuente, diagramas detallados, resultados adicionales de pruebas y cualquier otra información relevante que pueda ser de interés para el lector.

**Capítulo 9: Referencias**

Se recopilan todas las fuentes bibliográficas y recursos consultados, siguiendo las normas académicas de citación, lo que proporciona respaldo teórico y contextual al proyecto.

# Capítulo 2. ANTECEDENTES / ESTADO DEL ARTE

En el dinámico y competitivo mundo del ecommerce, el conocimiento preciso y actualizado de los precios de la competencia es crucial para mantener una ventaja competitiva. Las empresas que pueden ajustar sus precios rápidamente y de manera informada tienen mayores probabilidades de atraer y retener clientes. Este estudio explora el estado del arte en las tecnologías y métodos utilizados para monitorear y analizar los precios de la competencia en el ámbito del comercio electrónico. Se abordarán tres secciones principales: contexto y justificación, planteamiento del problema y un análisis exhaustivo de las técnicas y herramientas actuales.

## 2.1 Estado del arte

El secreto del éxito está en saber algo que nadie más sabe. Aristóteles Onassis

## 2.2 Contexto y justificación

El comercio electrónico ha experimentado un crecimiento exponencial en la última década, impulsado por la globalización, la digitalización y el cambio en los hábitos de consumo. Este crecimiento ha intensificado la competencia, obligando a los minoristas online a buscar constantemente nuevas formas de diferenciarse y mejorar su propuesta de valor. Uno de los factores críticos en esta competencia es la estrategia de precios.


- **Necesidad de Monitoreo de Precios**: Los precios son uno de los principales factores que influyen en las decisiones [iv]de compra de los consumidores. La capacidad de ofrecer precios competitivos sin sacrificar márgenes de beneficio es esencial para el éxito en el ecommerce.

- **Avances Tecnológicos**: Las tecnologías avanzadas, como el scraping de datos, la inteligencia artificial y el machine learning, han abierto nuevas posibilidades para el

monitoreo y análisis de precios. Estas tecnologías permiten a las empresas recopilar y procesar grandes volúmenes de datos en tiempo real, proporcionando insights valiosos para la toma de decisiones.

- **Competencia Dinámica**: Los precios en el ecommerce pueden cambiar con gran frecuencia debido a promociones, descuentos, cambios en los costos de producción y otros factores. La capacidad de rastrear y reaccionar rápidamente a estos cambios es crucial para mantenerse competitivo.

Otro aspecto importante es que el proyecto es financiado por una empresa del ramo del ocio de las piscinas. Este sector tiene una serie de particularidades muy diferentes al resto de sectores tradicionales que requieren de soluciones específicas.

Algunas de estas particularidades son:

- Mayoría de clientes son esporádicos
- Los clientes que repiten suelen tardar meses o años en volver.
- Ventas se concentran en 4 meses, resto del año es residual.

## 2.3 Planteamiento del problema

A pesar de los avances tecnológicos, muchas empresas de ecommerce enfrentan desafíos significativos al intentar monitorear los precios de la competencia de manera efectiva y eficiente.

Algunos de los problemas que se encuentran son:

- **Falta de Datos Actualizados y Precisos**: Obtener datos precisos y actualizados sobre los precios de la competencia puede ser complicado debido a la gran cantidad de competidores y productos disponibles en el mercado.

- **Procesamiento y Análisis de Grandes Volúmenes de Datos**: El análisis de grandes volúmenes de datos requiere recursos significativos y capacidades avanzadas de procesamiento, lo que puede ser un desafío para muchas empresas.

- **Identificación de Estrategias de Precios Efectivas**: No basta con conocer los precios de la competencia; las empresas deben poder interpretar estos datos para desarrollar estrategias de precios efectivas que consideren tanto la competitividad como la rentabilidad.

- **Integración de Datos de Diferentes Fuentes**: Los precios pueden variar significativamente entre diferentes canales de venta y regiones geográficas. Integrar y analizar datos de diversas fuentes es un desafío técnico importante.

- **Limitaciones Legales y Éticas**: La recopilación de datos de precios puede estar sujeta a regulaciones y consideraciones éticas que las empresas deben cumplir.

Todos esos problemas son genéricos y afecta a la mayoría de ecommerce, en nuestra solución se deben añadir los problemas específicos derivados del sector.

- Ventas concentradas en un periodo de tiempo

- Imposibilidad de conocer las estrategias de precios de antemano, ya que se ejecutan cuando empieza la temporada.

- Los efectos del tiempo cómo el frío, la lluvia o el calor intenso afectan directamente al volumen de la demanda.

- Los precios suelen tener una ponderación menor en situaciones determinadas. Por ejemplo, en general en el mes de julio los clientes dan prioridad a la recepción del producto que al precio.

- Leyes que afectan a la demanda. Ejemplo: Limitaciones del gobierno para la construcción de piscinas, limitación/imposibilidad de llenado de piscinas por la Ley de Sequedad, …

## 2.4 Análisis del Estado del Arte

En esta sección se detalla el análisis del Estado del arte para los principales componentes que componen la solución tecnológica:

### 2.4.1 Scrapping de datos

El scraping de datos es una técnica comúnmente utilizada para extraer información de sitios web de manera automatizada. Herramientas como BeautifulSoup, Scrapy y Selenium son populares en esta área. Estas herramientas permiten a las empresas recopilar datos de precios de los sitios web de la competencia en tiempo real.

**Ventajas:**

- Proporciona acceso directo a los datos visibles en los sitios web.
- Permite la recopilación de datos de múltiples competidores y productos simultáneamente.

**Desafíos:**

- Puede ser bloqueado por los sitios web que implementan medidas de seguridad contra el scraping.
- Requiere mantenimiento continuo para adaptarse a los cambios en la estructura de los sitios web.

### 2.4.2 APIs de Marketplaces

Muchas plataformas de comercio electrónico y marketplaces ofrecen APIs que permiten el acceso a datos de precios y otros detalles de productos. Amazon, eBay y otros grandes marketplaces tienen APIs que los desarrolladores pueden utilizar para obtener datos estructurados.

**Ventajas:**

- Acceso a datos precisos y actualizados directamente de la fuente.
- Menos riesgo de ser bloqueado en comparación con el scraping.

**Desafíos**:

- Puede estar sujeto a limitaciones de uso y tarifas.
- No todos los competidores pueden tener APIs accesibles.

### 2.4.3 FrontEnd

Toda solución requiere de un frontend fácil, amigable y sencillo para el usuario que ofrezca toda la información, potencia y control requeridos para una solución como esta.

Además, debe poder estar disponible para trabajar desde cualquier dispositivo, desde cualquier lugar y estar disponible las 24h del día. También debe ser multiusuario, permitir una restricción de acceso por permisos, asignación de grupos, …

**Ventajas:**

- Acceso a un frontend amigable, sencillo y conocido por el usuario.
- Configurable para que el usuario se adapte su solución a sus necesidades y/o gustos.
- Disponibilidad las 24h del día
- Disponibilidad en cualquier aparato con acceso a internet.
- Sin necesidad de instalaciones en el Sistema Operativo.

**Desventajas:**

- Limitaciones del navegador y/o tecnología.
- Rendimiento para el tratamiento de grandes volúmenes de información.
- Requiere de servidor/es para albergar la solución web.

# Capítulo 3. OBJETIVOS

El objetivo principal de este estudio es desarrollar un software para la visualización de precios basada en la inteligencia competitiva usando inteligencia artificial capaz de recopilar, analizar y procesar información relevante sobre los precios de productos o servicios ofrecidos por competidores en el mercado. Este sistema se centrará en la identificación de patrones, tendencias y comportamientos de precios, con el fin de proporcionar a las empresas una visión integral y actualizada del entorno competitivo para poder comprender las oportunidades y desafíos del entorno.

## 3.1 Objetivos generales

El objetivo general del presente trabajo consiste en ampliar el conocimiento actual de precios de la competencia por parte de los directivos con la finalidad de poder tomar decisiones más precisas y completas,

El objetivo final es utilizar todo el conocimiento generado por la solución para tomar decisiones que comporten un aumento de las ventas, y, en consecuencia, de los beneficios.

## 3.2 Objetivos específicos

Es difícil determinar los objetivos específicos de la solución ya que van aumentando según los usuarios visualizan nuevas funcionalidades de la solución. En esta sección voy a especificar los objetivos del estudio iniciales definidos al arrancar el proyecto:

- **Desarrollar un sistema de recopilación de datos**: Implementar una infraestructura tecnológica que permita la extracción de información de diversas fuentes de datos en línea, incluidos sitios web de competidores, Marketplaces y/o portales cómo Amazon o Google.

- **Aplicar técnicas de análisis de datos**: Utilizar técnicas avanzadas de análisis de datos y minería de texto para identificar y extraer información relevante sobre precios, como

variaciones de precios, descuentos, promociones y estrategias de fijación de precios de la competencia.

- **Análisis de Precios**: Proporcionar herramientas analíticas que permitan a los usuarios comparar precios, identificar tendencias y realizar pronósticos.

- **Cobertura Amplia**: Monitorear una amplia gama de competidores y productos en diferentes mercados y canales.

- **Generar insights y recomendaciones**: A partir del análisis de los datos recopilados, generar insights accionables y recomendaciones estratégicas para ayudar a las empresas a tomar decisiones basadas sobre las estrategias de precios y posicionamiento en el mercado.

- **Gráficos avanzados:** Estudiar y generar gráficos e informes avanzados que permitan visualizar la información de manera precisa, clara y simple.


- **Alertas y Notificaciones**: Generar alertas cuando se detecten cambios significativos en los precios de los competidores.

- **Evaluar la efectividad del sistema**: Evaluar el rendimiento y la precisión del sistema desarrollado en términos de su capacidad para recopilar, analizar y proporcionar información útil sobre los precios de la competencia.

Al alcanzar estos objetivos, se espera que este proyecto contribuya significativamente al campo de la inteligencia competitiva innovando con las tecnologías de inteligencia artificial aplicada al análisis competitivo y la toma de decisiones empresariales.

Cómo ya se ha comentado, estos objetivos son los iniciales y se irán ampliando según la solución se vaya utilizando y los usuarios requieran de nuevas funcionalidades para cubrir sus necesidades.

Aunque el proyecto está destinado a resolver un problema genérico de cualquier sector aplicando el framework de Kopalle con la premisa de obtener los precios públicos y accesibles a través de internet (se excluyen los precios de tienda física), se va a acotar el alcance específicamente al nicho de productos relacionados con las piscinas y Wellness. Esta decisión es consecuencia de la financiación de este máster por una empresa del sector y adaptaremos algunas de las tareas del diagrama de Kopalle para ajustarlas a las características del mercado concreto.

## 3.3 Beneficios del proyecto

Este proyecto tiene varios beneficios significativos que pueden contribuir tanto a la academia como a la industria. A continuación, se detallan los beneficios esperados al alcanzar los objetivos establecidos:


- **Decisiones Estratégicas Informadas**: El sistema proporcionará a los directivos de empresas información detallada y actualizada sobre los precios de la competencia, lo que les permitirá tomar decisiones más informadas y estratégicas. Esto puede resultar en una optimización de las estrategias de precios y un mejor posicionamiento en el mercado.

- **Ventaja Competitiva**: Al contar con un conocimiento profundo y actualizado del entorno competitivo, las empresas pueden reaccionar rápidamente a los cambios del mercado y ajustar sus estrategias de precios en consecuencia, obteniendo así una ventaja competitiva significativa.

- **Aumento de Ventas y Beneficios**: Al implementar estrategias de precios optimizadas basadas en datos precisos y actualizados, se espera que las empresas experimenten un aumento en las ventas y los beneficios. La capacidad de ajustar precios en tiempo real en respuesta a las acciones de la competencia puede mejorar significativamente el rendimiento financiero.

- **Eficiencia Operativa**: La automatización de la recopilación y análisis de datos reducirá el tiempo y los recursos necesarios para llevar a cabo estas tareas manualmente. Esto permitirá a las empresas centrar sus esfuerzos en la toma de decisiones estratégicas y en otras áreas críticas de negocio.

- **Innovación en Inteligencia Competitiva**: Este proyecto contribuirá al campo de la inteligencia competitiva al integrar tecnologías avanzadas de inteligencia artificial y análisis de datos. La innovación en estas áreas puede abrir nuevas posibilidades para el análisis competitivo y la toma de decisiones empresariales.

- **Adaptación al Mercado Específico**: Aunque el proyecto está diseñado para ser aplicable a cualquier sector, se centrará específicamente en el nicho de productos

relacionados con las piscinas y el wellness. Esta adaptación permitirá un enfoque más detallado y preciso, proporcionando insights específicos y relevantes para este mercado.

- **Contribución Académica**: El desarrollo y la evaluación del sistema proporcionarán un valioso aporte académico al campo de la inteligencia competitiva y el análisis de precios. Los resultados y las metodologías utilizadas pueden servir como base para futuras investigaciones y desarrollos en este campo.

## 3.4 Conclusiones

En conclusión, este proyecto tiene el potencial de transformar la forma en que las empresas de ecommerce comprenden y responden a los precios de la competencia. Al alcanzar los objetivos establecidos, se espera que el sistema desarrollado proporcione a las empresas las herramientas necesarias para mejorar su competitividad, optimizar sus estrategias de precios y, en última instancia, aumentar sus ventas y beneficios. La combinación de técnicas avanzadas de inteligencia artificial y análisis de datos con un enfoque específico en el mercado de piscinas y wellness asegura que este proyecto no solo sea innovador, sino también altamente relevante y aplicable.

# Capítulo 4. DESARROLLO DEL PROYECTO

En este capítulo se detalla las diferentes tareas requeridas para el desarrollo del proyecto, así cómo sus tiempos, fechas estimadas de arranque, de finalización, costes, …

Esta información es una estimación que puede variar según se vaya desarrollando las distintas fases del proyecto, pero debe ser una guía para controlar y corregir las desviaciones que pueden producirse.

## 4.1 Planificación del proyecto

El siguiente cronograma muestra la previsión del desarrollo del proyecto dividido en las distintas tareas detectadas junto a la fecha prevista y su estado actual.

Este mes

| Tarea | Fecha | Estado | Cronograma |
|---|---|---|---|
| Requerimientos | | Listo | 1 - 11 may. |
| Anteproyecto | | Listo | 20 may. |
| Fase Funcional | | Listo | 8 - 31 may. |
| Base solución Web | | Listo | 1 - 16 jun. |
| Frontend Scrapping | | Listo | 11 - 28 jun. |
| Scrapping Buscadores | | Listo | 28 jun. - 5 jul. |
| Anteproyecto Versión Inicial | | Listo | 15 jul. |
| Scrapping Portales Web | | Listo | 5 - 24 jul. |
| Integración Scrapping - Web | | Listo | 24 - 31 jul. |
| Modelos - Tratamiento de datos | | Listo | 1 - 8 ago. |
| Integración Modelos -Web | | Listo | 8 - 14 ago. |
| Gestión Datos a Información | | Listo | 14 - 19 ago. |
| Dashboards | | En curso | 17 - 31 ago. |
| Configuración y Parametrización | | Listo | 1 - 8 sep. |
| Testing | | Listo | 15 jul. - 31 ago. |
| Documentación | | Listo | 1 ago. - 15 sep. |
| Entrega Final -Primera Versión | | En curso | 12 sep. |
| Entrega Final | | No iniciado | 19 sep. |

*Figura 2 Cronograma del proyecto*

En el siguiente diagrama se Gantt se pueden ver las distintas fases en el espacio temporal y los checkpoints previstos para controlar el desarrollo de la solución.

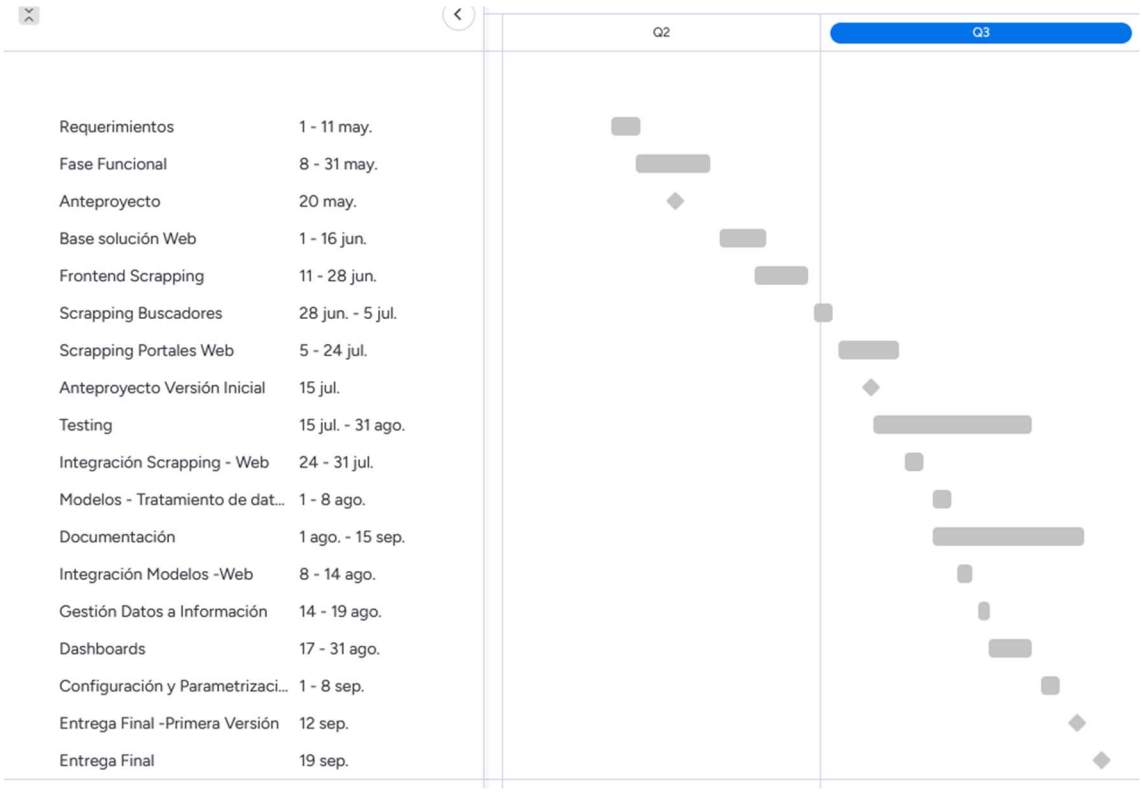



*Figura 3 Diagrama de Gantt del Proyecto*

A continuación, se define con un pequeño texto explicativo cada una de las fases del proyecto así como una predicción del coste en días que requiere para su realización.

Comentar que los datos de esfuerzo están basados en día/Persona. Es decir, si pone 2 días significa que destino un recurso durante 8h cada día.

**Requerimientos**

Esfuerzo: 10 días

En esta etapa se llevará a cabo la investigación exhaustiva de la problemática específica abordada en el TFM, así como la recopilación de los datos relevantes necesarios para su análisis y posterior desarrollo. También se realizarán reuniones con los departamentos implicados para entender sus necesidades y cubrir sus expectativas.

**Anteproyecto**

En este checkpoint se realiza y presenta un documento que describe el proyecto y algunas de sus fases. Se define todos los componentes que tendrá el proyecto y se detalla con mayor precisión aquellos dónde no hay incertidumbres o los riesgos/problemas son bajos.

Este documento también sirve cómo indicador de si se debe seguir trabajando en el proyecto o debe descartarse por la potencia de algunos riesgos.

**Fase funcional**

Esfuerzo: 23 días

En esta etapa se detalla en un grado mucho mayor los puntos detectados en la fase de requerimientos y, en base a los datos relevantes recopilados se detalla las funcionalidades que se desarrollan y cómo se implementarán en las fases posteriores del proyecto.

**Base solución Web**

Esfuerzo: 15 días

En esta etapa se busca la arquitectura web que mejor se adapta a las necesidades del proyecto y del equipo. Se decide seleccionar tres arquitecturas distintas para posteriormente escoger la mejor solución.

**Frontend Scrapping**

Esfuerzo: 17 días

Ante la cantidad de dificultades y riesgos que tiene las técnicas de Scrapping, definiremos un frontend configurable para paliar, reducir o eliminar los problemas detectados al usar técnicas de Scrapping. Decisiones cómo el navegador a utilizar, el motor de búsqueda, o el motor de Scrapping serán tomadas por el usuario mediante la configuración del sistema y se intentará realizar adaptaciones cuando precise.

**Scrapping Buscadores**

Esfuerzo: 7 días

Una vez el usuario informe del producto a estudiar, la solución debe buscar los ecommerce en los buscadores donde poder adquirir el producto. Aunque el proceso será automático debe existir la posibilidad de introducir manualmente ecommerce por parte del usuario (petición expresa de negocio).

**Anteproyecto versión inicial**

Presentación del documento de anteproyecto donde ya se tiene un conocimiento mayor sobre todas las fases del proyecto y se detallan tiempos, costes y duración de cada una de las fases.

Este documento ya presenta una visión de cuando se tendrá la primera versión implementada, su coste económico y el esfuerzo que requiere.

**Scrapping Portales Web**

Esfuerzo: 19 días

En esta etapa se buscarán soluciones mediante inteligencia artificial para obtener la información y el precio de los productos en los distintos ecommerce que se pueden indicar en la fase 'Scrapping Buscadores'.

Los portales web pueden ser variados y presentar todo tipo de dificultades, desde detectores de Scrapping, tecnologías complejas basadas en javascript, htmls enormes, o página web con varios precios.

**Testing**

Esfuerzo: 46 días

A partir de esta etapa las integraciones, complejidad y dificultades van en aumento y es preciso tener herramientas automatizadas capaces de probar y comprobar que las funcionalidades ya programadas funcionan correctamente ante los nuevos desarrollos e integraciones que se irán produciendo.

Esta fase se va a estar realizando continuamente durante el resto de vida del proyecto. Es decir, se ejecutará durante toda la fase hasta su entrega y posteriormente se seguirá ejecutando cuando se vayan implementando nuevas novedades, modificaciones o adaptaciones.

**Integración Scrapping Buscadores -Web**

Esfuerzo: 7 días

Una vez finalizadas las etapas de Scrapping de buscadores, debe recogerse esta información para facilitarla al módulo de Scrapping de portales web. Ya que será este último el encargado de recuperar la información del precio.

**Modelos, tratamiento de datos**

Esfuerzo: 7 días

En esta fase se estudian diferentes modelos en la búsqueda de los mejores modelos que nos ayuden a encontrar la información. La solución debe permitir seleccionar el modelo deseado y facilitar la integración con nuevos modelos. Actualmente, el sistema permite trabajar en Llama 3 y ChatGPT. Otros Modelos cómo Gemini, Rasa o Mistral han sido descartados por los malos resultados alcanzados, aunque no se descarta que puedan ser recuperados en el futuro ante una mejora de su rendimiento.

Está pendiente realizar estudios con LLMs muy prometedores cómo pueden ser Claude, Anthropic o DeepSeek.

**Documentación**

Esfuerzo: 14 días

Aunque esta fase se repite durante todo el proyecto desde sus inicios hasta su fin, es en este punto cuando se controla, se detalla, se facilita a los usuarios y se comprueba que todos los desarrollos se encuentran debidamente documentos y claramente explicados para su comprensión.

Esta fase, cómo la de testing, acompañará el proyecto durante el resto de su vida hasta la finalización de su ciclo de vida.

**Integración Modelos Web**

Esfuerzo: 6 días

Cómo ya se ha comentado en una fase anterior, los modelos LLM deben poder seleccionarse, configurarse y adaptarse a las necesidades. Razón por la que se requiere una integración, configuración y ajuste que permita obtener los mejores resultados.

**Gestión de datos a información**

Esfuerzo: 5 días

Después de obtener los precios de los distintos productos e ecommerce, es necesario realizar un correcto estudio de las necesidades e datos que aporte información fácil, entendible y de utilidad que permita la correcta toma de decisiones.

Junto con negocio se programa una batería de reuniones donde se define las necesidades, la presentación y la exportación de datos. También debe 'definirse' las apis que tendrá la solución para que otras soluciones puedan interactuar y/o recoger información.

**Dashboards**

Esfuerzo: 14 días.

Ante las necesidades detectadas en la fase anterior, se define los distintos gráficos, informes y datos que se requieren. También es preciso definir los periodos a visualizar, fechas de generación, cómo se recibirán, …

Importante en este apartado indicar también cuando se producen las notificaciones y los filtros que deben pasarse para generarse dichas notificaciones ( Ej: Envía un email cuando el precio del producto se reduzca un 15% ).

**Configuración y parametrización**

Esfuerzo: 7 días.

La solución ya se encuentra desarrollada en su fase inicial y debe configurarse y parametrizarse para que funcione y se adapte a las necesidades de los usuarios.

También está previsto algún tipo de acción de importación inicial para recibir los productos a los que se debe controlar.

**Entrega final – Primera versión**:

Entrega del documento del proyecto para su lectura y revisión por parte de todos los implicados.

**Entrega final:**

Entrega del documento, presentación y visualización de la solución realizada.

## 4.2 Descripción de la solución, metodologías y herramientas empleadas

En este apartado se describe la solución implementada, sus componentes que la componen, como se relacionan entre ellos y conviven de manera conjunta para generar el resultado final.

El proyecto se compone de varios módulos que trabajan de manera independiente donde existe un módulo encargado de orquestrar y realizar la toma de decisiones, eventos y acciones que activan/desactivan las funcionalidades del resto de módulos.

Los módulos implicados hasta ahora son:

- Módulo de Scrapping: módulo responsable de realizar todas las tareas de Scrapping , recibe un input con la tarea solicitada, su parametrización y se encarga de visitar internet para recoger la información solicitada para finalmente devolverla mediante una llamada a una API.

- Módulo de IA: módulo encargado de realizar las acciones de IA para obtener información. Este módulo debe permitir la inserción y actualización de diferentes LLMs tanto presentes cómo futuros.

- Portal Web encargada de orquestar, configurar e integrar la solución, además presenta un frontend visual amigable, sencillo y fácil para que los usuarios puedan realizar toda la operativa.

### 4.2.1 Módulo de Scrapping

El módulo de Scrapping es la pieza de la solución encargada de realizar todas las operaciones de búsqueda en internet y recoger la información de distintos webs, portales y marketplaces para poder facilitarse al resto de módulos.

Se pretende que este módulo sea capaz de realizar operaciones cómo:

- Obtener los resultados de las búsquedas sobre un buscador de internet, comúnmente Google pero debería ser capaz de conectarse a otros cómo Bing, Yahoo o Duck Duck Go con el fin de no concentrar la obtención de datos en una única fuente.

- Obtener la información de un Marketplace cómo Amazon para recoger tanto los datos de Amazon como de terceros que venden en Amazon.

- Obtener la información de una página concreta de un ecommerce.

#### *4.2.1.1 Tareas*

Cada 'tarea' de Scrapping se define desde la opción de menús 'Tareas'.

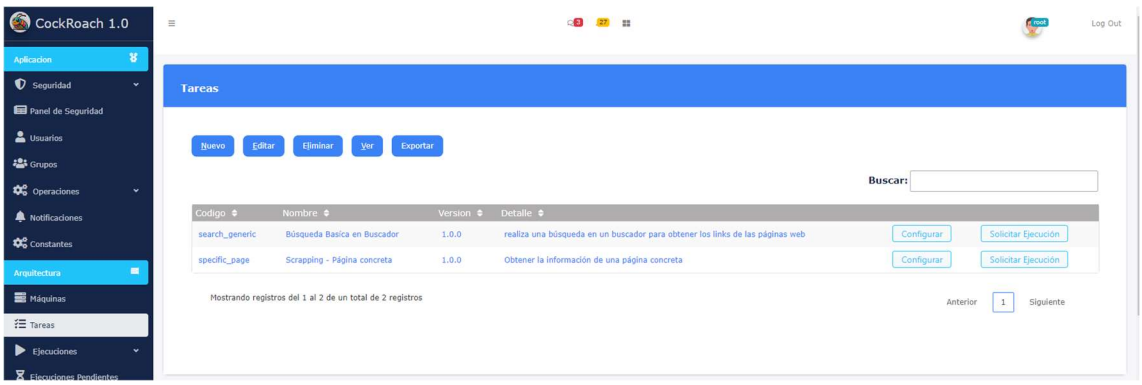



*Figura 4 Gestión de Tareas*

Este módulo está diseñado para ser independiente de cualquier tecnología específica, utilizando diversos motores de búsqueda, navegadores y bibliotecas para procesar datos. El objetivo de esta diversidad es minimizar los riesgos y la dependencia de una tecnología particular.

Desde la opción 'Editar', se pueden configurar varios elementos, como la descripción, la versión y, lo más importante, el código que se ejecutará al finalizar la tarea.

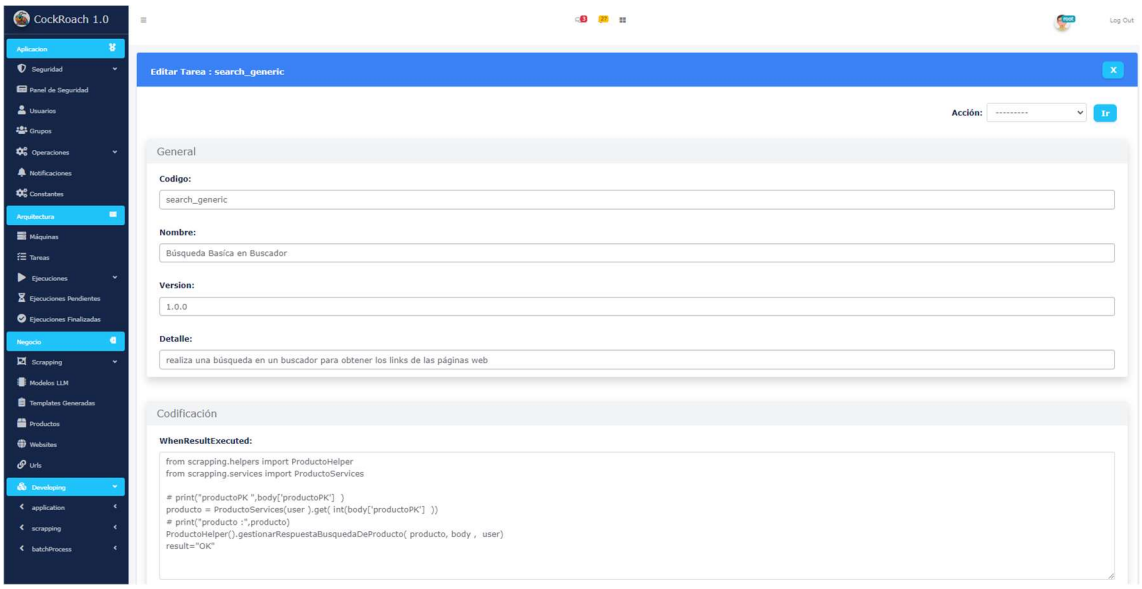



*Figura 5 Edición de Tarea*

Actualmente, el proyecto contempla dos tareas principales:

- Search_generic: Esta tarea consiste en consultar diversos motores de búsqueda para identificar las páginas web que se someterán a scraping.

- Specific_page: accede a la página web de cada comercio electrónico para recopilar información, como los precios.

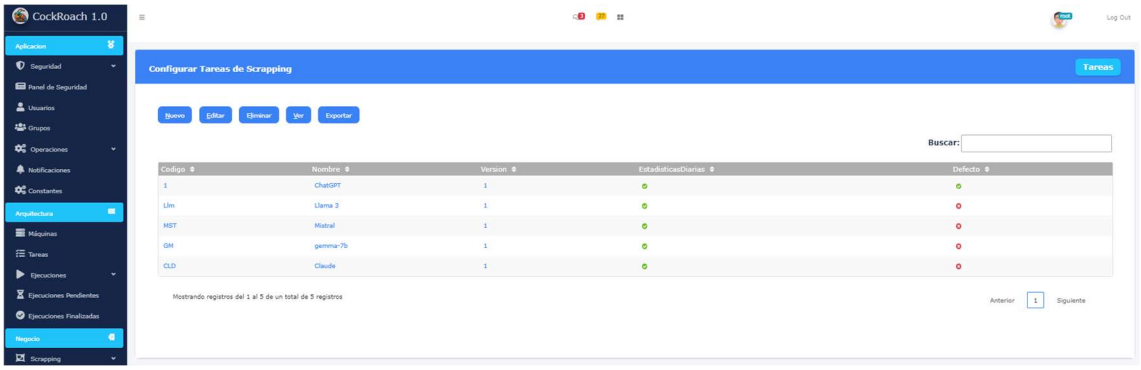



*Figura 6 Visualización de Tareas*

Cada tarea tiene su configuración particular:

#### 4.2.1.1.1 Search_generic

La tarea de búsqueda genérica se encarga de buscar, en base al nombre de un producto específico, todos los enlaces de los principales e-commerce que permiten su adquisición en línea.

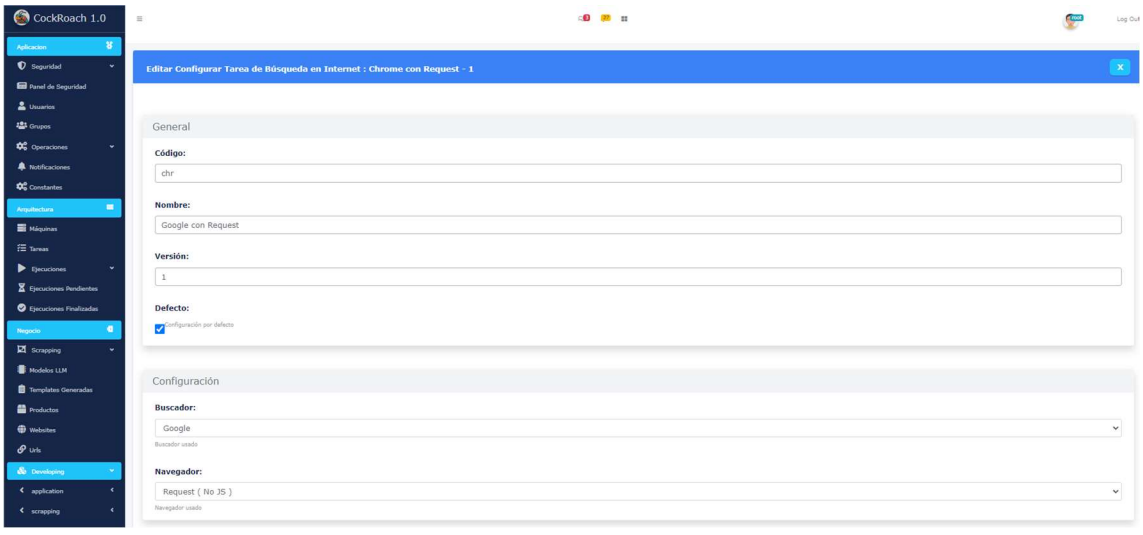



*Figura 7 Configuración particular de la tarea de Búsquedas en Internet*

Entre los parámetros configurables, se destacan dos por su importancia:

**Buscador:** El software permite hacer scraping en los cuatro motores de búsqueda más utilizados y conocidos: Google, Bing, DuckDuckGo y Yahoo. Es posible integrar nuevos buscadores, aunque esto requiere un pequeño desarrollo adicional para su completa integración con la solución.

**Navegador**: Debido a los diversos niveles de seguridad que cada buscador implementa, es posible que se requiera cambiar el navegador y la forma en que se procesa la información. *

Para soluciones que no requieran JavaScript, se dispone de:

- Requests
- Mechanize

Para soluciones que requieran JavaScript, se dispone de:

- Chrome
- Firefox

* El navegador por defecto es el informado en la tarea pero es posible especificar en la configuración del Website usar otro navegador en particular. En este caso, el website prevalece sobre la tarea.

Para el tema de control de entidades, la solución va integrada con un navegador TOR para intentar anonimizar la IP y disponer de un número 'infinito' de IPs cambiable de manera rápida y fácil de manera automática.

#### 4.2.1.1.2 Anonimización de entidades

Este caso particular aborda la resolución de un problema específico y requiere una sección especial para detallar tanto el problema como la solución implementada.

Los motores de búsqueda y las páginas web están equipados con tecnologías y sistemas diseñados para detectar técnicas de scraping y rechazar todas las solicitudes cuando se identifica el uso de dichas técnicas.

Debido al volumen de solicitudes de productos generadas por la solución, estas pueden ser identificadas como scraping, resultando en el bloqueo de todas las peticiones subsecuentes.

Existen varias soluciones para abordar esta problemática, que van desde el uso de múltiples ordenadores/servidores para distribuir las solicitudes hasta la implementación de tecnologías más avanzadas.

Tor (The Onion Router) es un navegador web que permite a los usuarios a acceder a una red que anonimiza el tráfico web para proporcionar una navegación en línea totalmente privada.

En este proyecto, utilizaremos un Gateway (puerta de enlace) entre el navegador y el sitio web. Esta puerta de enlace emplea la tecnología TOR, que permite anonimizar nuestra identidad y cambiarla en cualquier momento. Con esta técnica efectiva, las solicitudes detectadas como scrapping se reenviarán con una identidad diferente.

#### *4.2.1.2 Máquinas*

Las tareas definidas anteriormente consumen una gran cantidad de recursos tanto de memoria cómo de cpu y espacio en disco. La solución se ha montado para que dichas tareas deban ejecutarse en una máquina independiente para que el rendimiento del portal Web no se vea impactado, además se puede escalar y añadir un número indefinido de máquinas.

En el siguiente gráfico se visualiza conceptualmente la arquitectura diseñada.

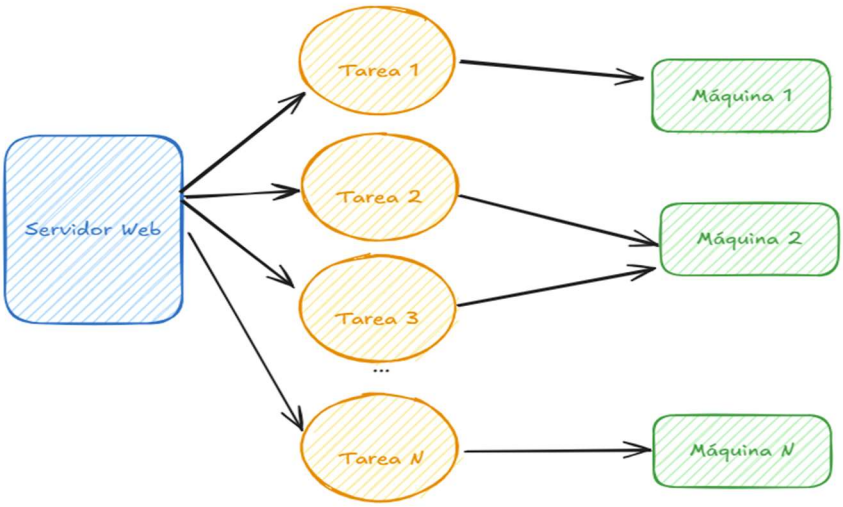



*Figura 8 Diagrama de la relación entre Tareas y Máquinas*

Las tareas pueden ser ejecutadas de forma manual o automática. En la actualidad en el entorno de producción, se programan para su ejecución diaria durante la madrugada y se distribuyen a las diferentes máquinas mediante un algoritmo especializado configurable desde la 'Política de Asignación de Máquinas' disponible en la opción de 'Máquinas'.

La gestión y configuración de las máquinas debe realizarse a través del panel de control.

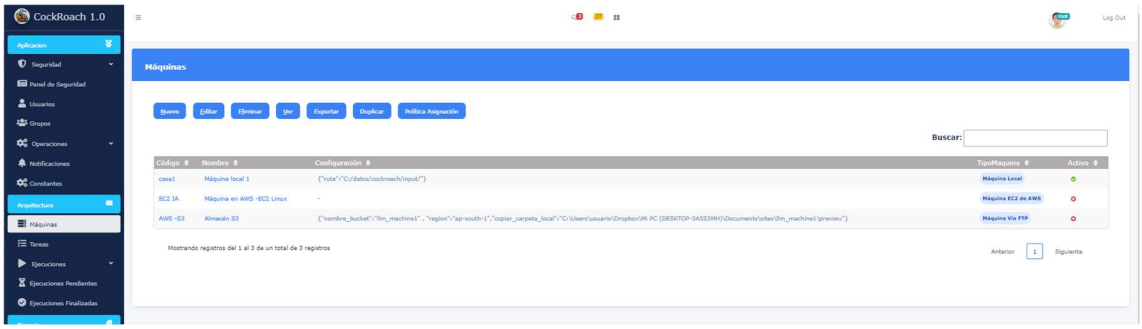



*Figura 9 Grid de Gestión de máquinas*

En el panel se pueden realizar todas las acciones necesarias, tales cómo crear, editar, activar/desactivar máquinas.

Las máquinas disponen de varios campos configurables. Entre los cuales destacan los siguientes:

- **TipoMáquina**: Define el tipo de máquina representado por el registro. Actualmente, existen 4 tipos distintos de máquinas:
  - Local
  - FTP
  - EC2 de Amazon
  - S3 de Amazon

- **CodeToStart**: Script que se ejecuta cuando arranca la máquina.

- **CodeToRun**: Script que se ejecuta cuando se le asigna la ejecución de una tarea.

- **CodeToCheck:** Script que se ejecuta para comprobar el estado de la máquina.

- **CodeToStop**: Script que se ejecuta cuando se para la máquina.

- Es importante tener en cuenta que el script cuando finaliza la tarea se encuentra guardado en la Tarea y no en la Máquina, razón por la que no existe un parámetro en este lugar.

**Configuración**: ese campo se utiliza para parametrizar los scripts y reducir las modificaciones realizadas en ellos.

Los scripts de un tipo de máquina concreto, como por ejemplo S3, deberían ser todos iguales ya que deben realizarse las mismas acciones y únicamente cambia los datos de acceso, las rutas, el usuario y password, … La opción Configuración facilita la parametrización de esta información, evitando la necesidad de scripts distintos para máquinas del mismo tipo.

Independientemente de lo anterior, si el administrador considera que una máquina concreta requiere un script customizado puede modificar el código.

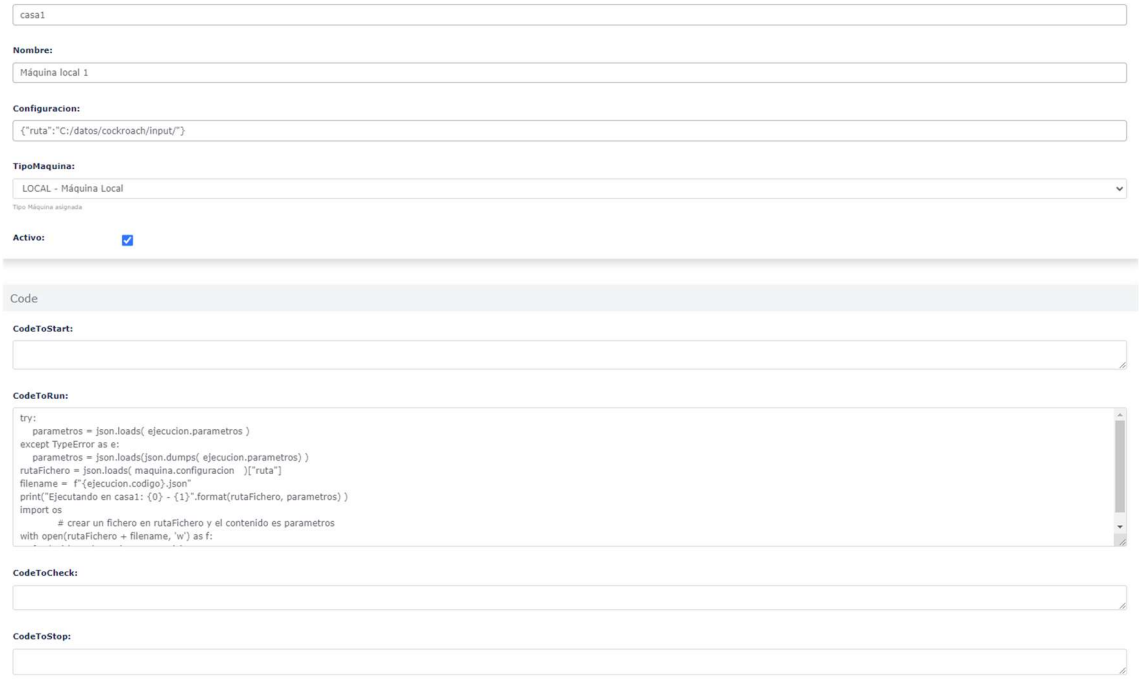



*Figura 10 Edición de la configuración de una máquina*

#### *4.2.1.3 Ejecuciones y Ejecuciones Finalizadas*

Cada vez que se ejecuta una tarea, se genera una nueva ejecución que contiene todos los datos necesarios para el éxito de la tarea. Una vez esta ejecución ha finalizado, pasa a ejecución finalizada.

Las ejecuciones contienen varios tipos de información:

- Estado de la ejecución: este campo permite saber dónde se encuentra la ejecución trabajando o paralizada.
- Detalle: para que el usuario tenga una referencia de la ejecución.
- Parámetros: lista de parámetros enviados al Módulo de IA y al Módulo de LLM. Estos datos se informan automáticamente de Tareas, LLMs, websites, productos y Máquinas
- WhenResultExecuted: Una vez la Ejecución obtiene el Precio, se ejecuta el script guardado en este campo proveniente de Tareas.
- Tarea: Tarea que se ha solicitado ejecutar.
- Máquina: Máquina asignada y donde ese ejecutará esta tarea.

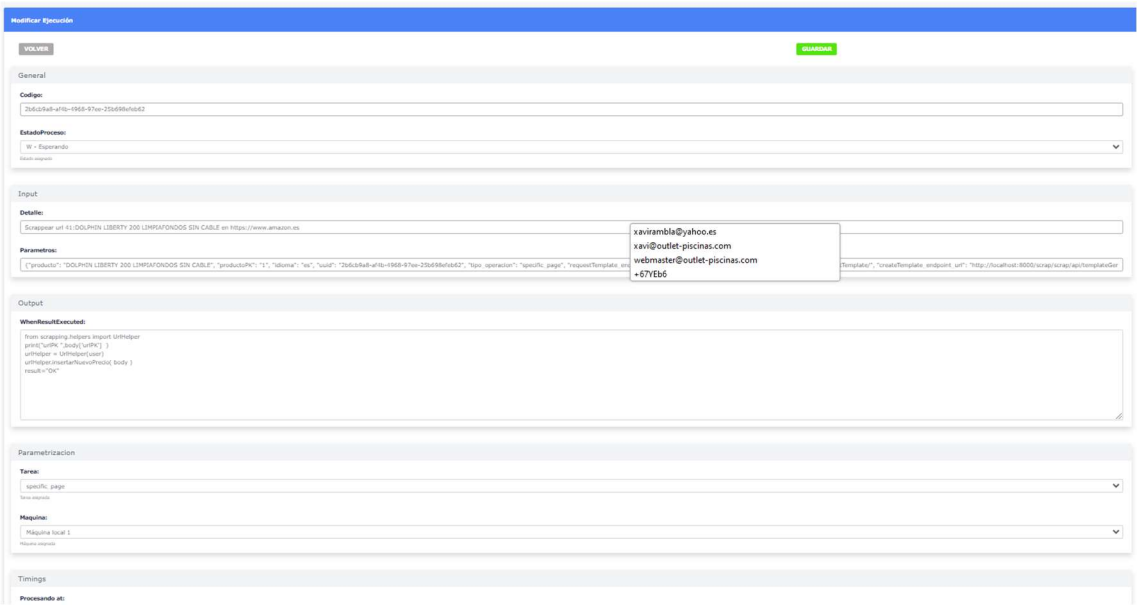



*Figura 11 Edición de una configuración de una Máquina*

Debido a la cantidad de ejecuciones que se ejecutan cada día, se ha dividido las ejecuciones en dos tipos, las ejecuciones pendientes y las ejecuciones terminadas.

Las ejecuciones pendientes son todas aquellas ejecuciones que, por alguna razón, todavía no han finalizado.

Las ejecuciones terminadas son todas aquellas ejecuciones que han finalizado, independientemente del resultado de la ejecución.

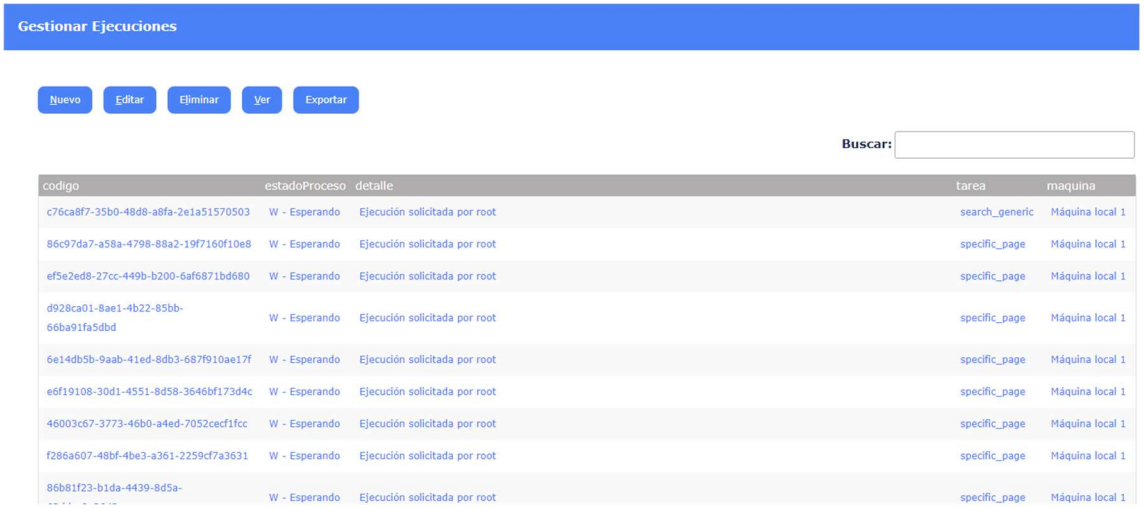



*Figura 12 Listado de ejecuciones*

### 4.2.2 Módulo de IA

Aunque se ha comentado que el módulo de IA cómo una opción independiente del Módulo de Scrapping, finalmente se han integrado con el módulo de Scrapping ya que van muy ligados y así se reducen tanto la complejidad cómo las interacciones entre ellos.

Igualmente, por la complejidad del módulo y la cantidad de componentes que lo componen es preferible explicarlo en un apartado distinto.

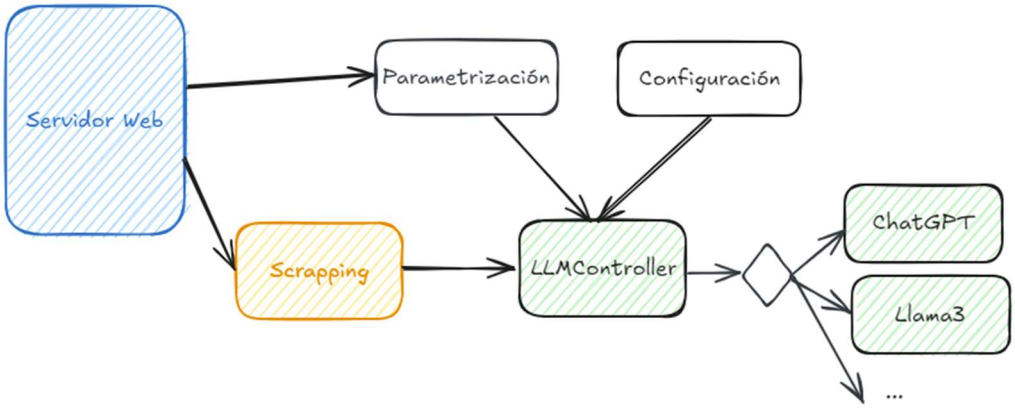



*Figura 13 Arquitectura del módulo de IA*

A grandes rasgos, todas las máquinas parten que tienen los mismos LLMs y con el mismo software y capacidades. Es importante para reducir la complejidad del diagrama y partir con la premisa que cualquier tarea puede ser ejecutada en cualquier máquina.

En el servidor Web se detecta el lanzamiento de una tarea, que desencadena en la creación, la parametrización y la petición de una ejecución.

El proceso de Scrapping recibe todos los datos de la ejecución y junto con la configuración local de cada máquina, se realiza el proceso de Scrapping y llama al controller de LLM cuando se requiera.

#### *4.2.2.1 Funcionamiento del Módulo de IA*

El módulo de IA se ha reconstruido varias veces para encontrar una fórmula válida que cumpliera con las necesidades del proyecto. La misión del módulo es bastante simple pero su implementación es algo más compleja.

En base a la página web del producto scrappeado se consigue el html que es enviado cómo parámetro al módulo de IA, éste debe utilizar su tecnología para encontrar y devolver el precio del producto.

Se han implementado varios desarrollos para realizar la operación con la finalidad de cumplir una serie de requisitos para la validez del módulo.

Dichos requisitos son los siguientes:

- Fiabilidad del resultado
- Obtención de únicamente la información solicitada (precio).
- Velocidad aceptable.
- Configurable y modulable.

En este documento únicamente se explica el desarrollo final construido.

#### 4.2.2.1.1 Construcción de Templates.

Inicialmente la idea era mediante el código html y el Modelo LLM conseguir el precio del producto. Esta solución es factible con GPT-3, y algo menos fiable en Llama3. Pero tiene un inconveniente muy grande que es el tiempo, es una buena solución para pedir el precio de un producto concreto pero no para realizar 100 solicitudes.


*Figura 14 Versión inicial de funcionamiento de la Generación de Template*

Trabajando el contexto y añadiendo condiciones y reglas se consigue unos decentes resultados [1]aunque la fiabilidad es aleatoria

[1] Se han realizado un número bajo de pruebas, debería realizarse un mayor número.

| Modelo LLM | Sin modificar resultado | Modificando levemente | Modificando |
|---|---|---|---|
| GPT3 | 84% | 13% | 3% |
| Llama 3 | 15% | 80% | 5% |

*Tabla 1 Tabla orientativa del funcionamiento de la solución*

Lo que resulta preocupante acerca de la solución actual son los tiempos de respuesta. Con GPT-3, los tiempos son aceptables debido a que el modelo se ejecuta en los servidores de OpenAI. Sin embargo, con Llama3, los tiempos de respuesta pueden extenderse hasta un minuto por ejecución, principalmente debido al uso de una máquina con hardware no optimizado específicamente para la inteligencia artificial y los procesos de arranque.

Se busca una solución más elaborada que consiste en solicitar al modelo LLM que genere el código en Python que recupere el precio de la página HTML para la creación de un template.

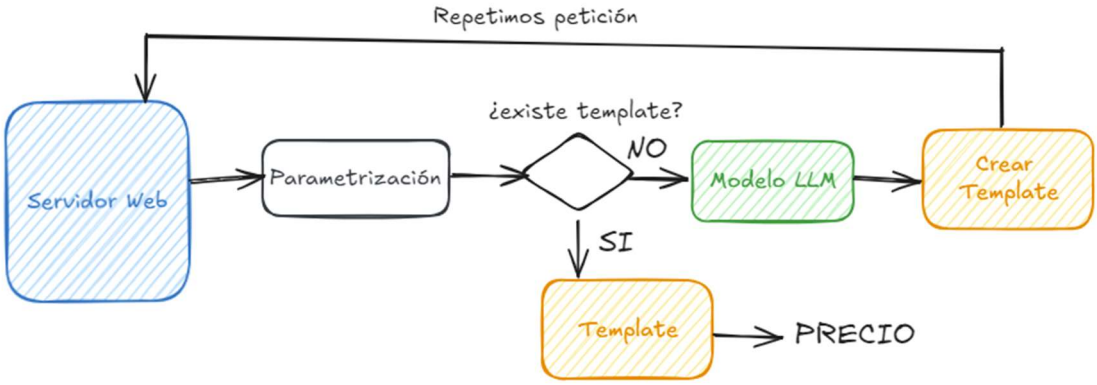



*Figura 15 Diagrama de funcionalidad del Módulo de IA*

El proceso operativo se basa en un conmutador que verifica la existencia de una plantilla de código en Python para la recuperación de precios. Si se encuentra una plantilla funcional, se procede a ejecutar el código para obtener el precio correspondiente.

En caso de no disponer de una plantilla preexistente, se utiliza el HTML y los parámetros disponibles para solicitar a un modelo de lenguaje grande (LLM) la generación del código Python necesario para extraer el precio del HTML proporcionado por el módulo de Scrapping. Una vez generado el código Python, se envía al servidor para su almacenamiento. Posteriormente, se repite la ejecución de la tarea utilizando la plantilla ahora disponible para recuperar el precio.

#### 4.2.2.1.2 Arquitectura del Módulo de IA

En la arquitectura de Módulo de IA se compone de varios elementos necesarios para el correcto y buen funcionamiento.

- Modelo LLM
- Templates Generadas

##### *4.2.2.1.2.1 Modelo LLM*

Esta entidad almacena todos los modelos disponibles de LLMs en el módulo de IA.

Indicar que esta entidad es únicamente informativa y la inclusión de LLMs requiere la instalación, integración y algún desarrollo para su funcionamiento.

Se ha realizado un estudio sobre varios LLMs donde se puede ver un breve resumen en el anexo CockRoach-Xavi Rambla_Anexo_Resultados.docx. En este documento se pueden visualizar los diferentes resultados obtenidos por los LLMs en base al mismo contexto.

En conclusión, el estudio anterior da cómo mejor LLM para la tarea chatGPT y posteriormente Llama. Comentar que el estudio se hizo en abril, y los datos ya se encuentran desfasados con lo que el ranking de resultados podría variar con los nuevos LLM ya disponibles.

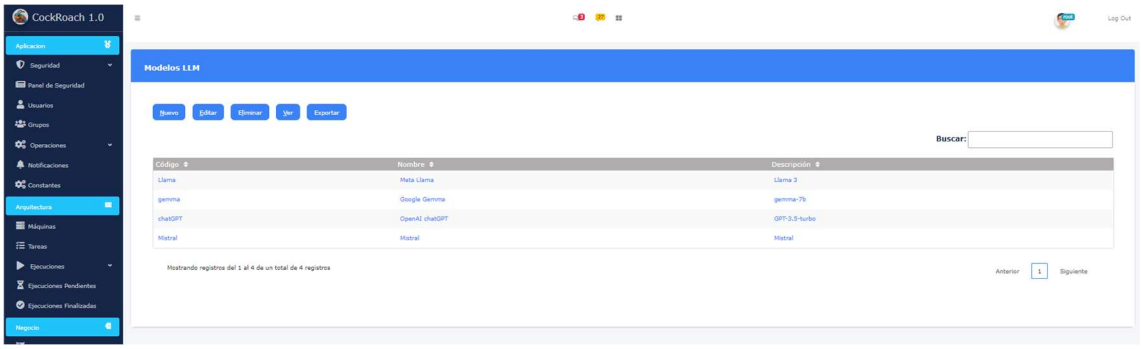



*Figura 16 Grid de la gestión de Modelos LLM disponibles*

#### *4.2.2.1.2.2 Templates generadas*

Las templates generadas por el modelo de IA se almacenan en esta opción de menú y permite al Administrador visualizar el código generado por el modelo LLM.

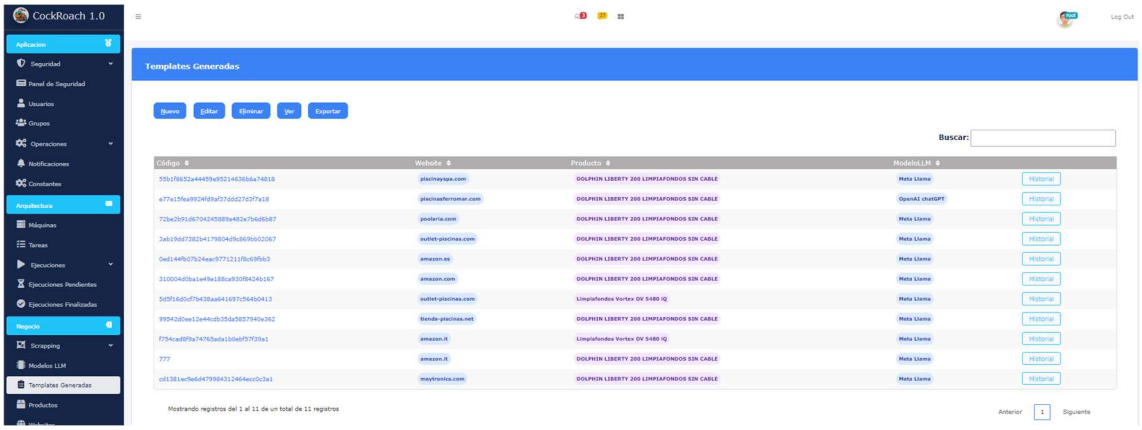



*Figura 17 Listado de templates generadas*

Cada template generada tiene la siguiente información:

- **Origen**: Url de la Página web utilizada para generar el template.
- **Website**: Ecommerce al que se puede utilizar este template.
- **Producto**: Producto utilizado para generar la template inicialmente.
- **Modelo LLM**: Modelo usado para generar el template
- **Code**: Código generado por el modelo LLM. Es posible modificarlo para adecuarlo correctamente.

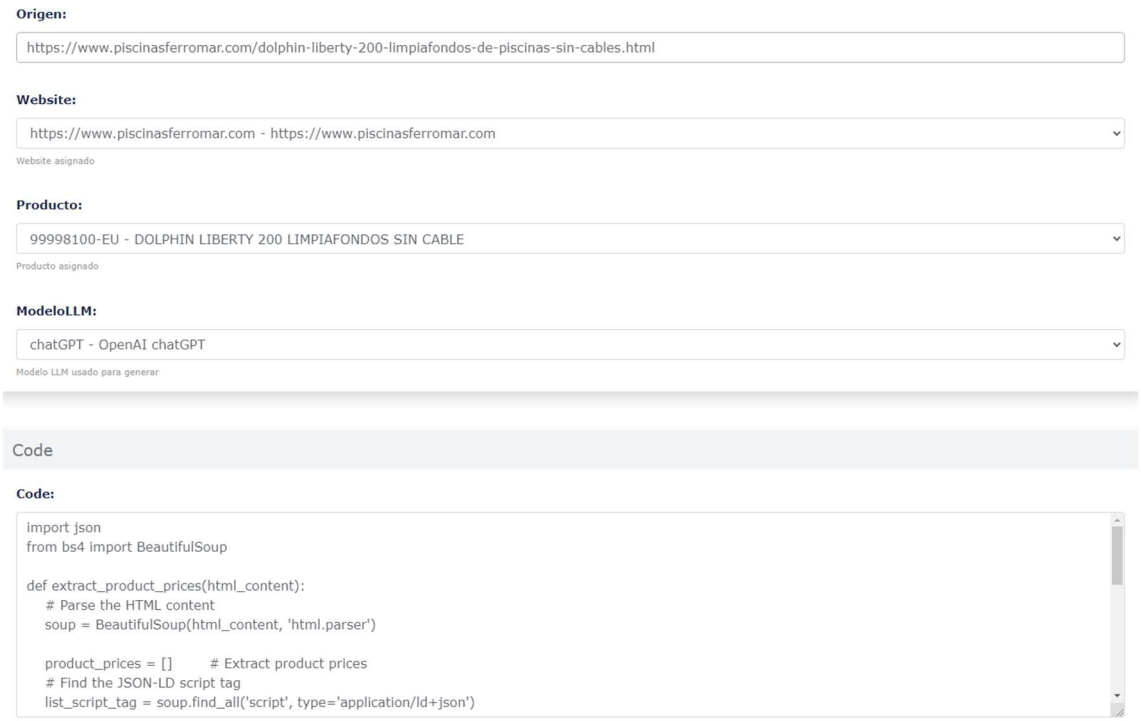



*Figura 18 Configuración de un template generado*

Es fundamental que el 'code' tenga integrado las siguientes reglas:

- Se recibe en la variable *html_content* el contenido de la página web
- Debe crearse la variable *result* donde se almacenará el precio recuperado de la página web.

### 4.2.3 Módulos de FrontEnd

La plataforma web que estamos desarrollando está estructurada para optimizar la gestión y la independencia de sus componentes, con el objetivo de mejorar la modularidad y la mantenibilidad del sistema. Actualmente, la solución se organiza en tres aplicaciones principales que segmentan responsabilidades y establecen estructuras independientes para minimizar el impacto de los cambios y reducir la posibilidad de errores.

#### *4.2.3.1 Aplicación application*

La aplicación "application" constituye el núcleo esencial del proyecto, proporcionando la infraestructura base sobre la cual se construyen todas las demás aplicaciones del sistema. Desarrollada utilizando el robusto framework Django, esta aplicación actúa como una capa superior que añade funcionalidades, facilidades y mejoras al diseño original de Django. Su arquitectura está diseñada para incluir componentes y funcionalidades esenciales que son reutilizables en otras partes del sistema, asegurando así una mayor cohesión y eficiencia en el desarrollo.

##### 4.2.3.1.1 Componentes y Funcionalidades Clave

A continuación, se detallan los principales módulos que componen esta aplicación:

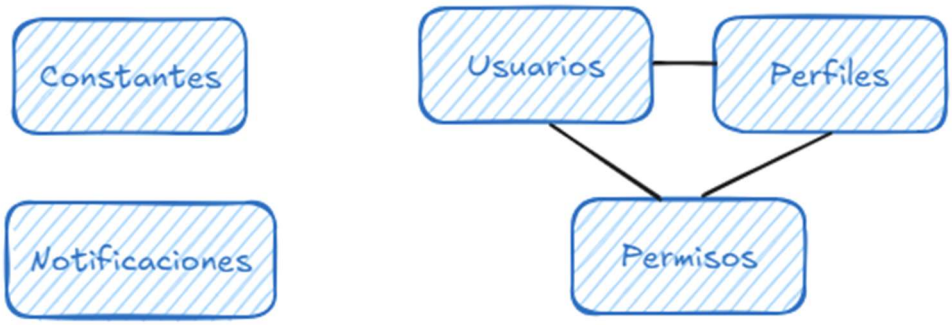



*Figura 19 Diagrama de la Aplicación 'Application'*

*Módulo de Constantes*:

Este módulo es fundamental para la gestión de constantes, valores fijos y diccionarios utilizados a lo largo de la solución web. Su objetivo es centralizar estos elementos para facilitar su mantenimiento y actualización, mejorando así la consistencia, reduciendo errores en el sistema y reduciendo código al tener esta funcionalidad centralizada.

Los valores indicados en esta opción son muy delicados y modifican el comportamiento de la aplicación con lo que es conveniente que sean gestionados únicamente por el administrador.

*Módulo de Seguridad*:

En este módulo se gestionan todos los aspectos relacionados con la seguridad de la solución web. Esto incluye la administración de usuarios, roles y permisos, garantizando que solo las personas autorizadas tengan acceso a las distintas funcionalidades del sistema. La implementación de este módulo sigue las mejores prácticas en seguridad informática para proteger los datos y la integridad del sistema.

Técnicamente, se mantiene la estructura y funcionalidad ya implantada por el equipo de Django pero se añaden funcionalidades, adaptaciones y mejoras para dar mayores y mejores servicios.

*Módulo de Notificaciones*:

Este módulo es crucial para la gestión de notificaciones tanto de la aplicación como del negocio. Permite enviar alertas y mensajes a los usuarios, informándolos sobre eventos importantes, actualizaciones y otras actividades relevantes. La flexibilidad de este módulo permite adaptarse a diferentes tipos de notificaciones, asegurando una comunicación eficiente y efectiva con los usuarios.

#### 4.2.3.1.2 Beneficios de la aplicación ‘application’.

La implementación de la aplicación "application" sobre el framework Django no solo facilita el desarrollo y la integración de nuevas funcionalidades, sino que también mejora la escalabilidad y la mantenibilidad del sistema. Al ofrecer una base sólida y bien estructurada, permite a los desarrolladores centrarse en la creación de aplicaciones complementarias que se integran perfectamente con el núcleo principal.

Otro valor añadido es que esta aplicación es reutilizable en otros proyectos ya que todas las conexiones con las otras aplicaciones del proyecto son unidireccionales y son éstas últimas las que tienen dependencias con ella.

#### *4.2.3.2 Aplicación Scrapping*

La aplicación 'Scrapping' constituye la base del proyecto, proporciona las tablas, funciones y operaciones para gestionar toda la operativa de Scrapping requerida por el proyecto.

##### 4.2.3.2.1 Componentes y Funcionalidades Clave

A continuación, se detallan los principales módulos que componen esta aplicación:

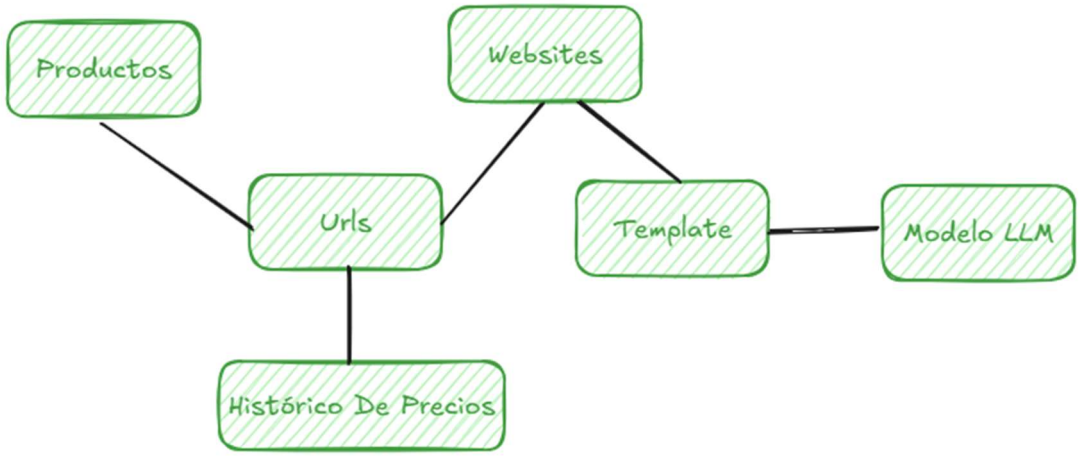



*Figura 20 Diagrama de la Aplicación 'Scrapping'*

*Módulo de Productos*:

El módulo de productos almacena toda la información relacionada con los productos que se quiere obtener el precio y van a pasar el proceso de scrappear.

*Módulo de Websites*:

El módulo de websites gestiona toda la información relacionada con las páginas web y ecommerces que serán objeto de scraping de productos.

**Automatización y Registro Manual**

La incorporación de registros de websites se realiza de manera automatizada, basándose en los resultados obtenidos de motores de búsqueda como Google, Yahoo, Bing, entre otros. No obstante, también se ofrece la opción de introducir registros manualmente por parte del usuario.

**Limitaciones de los Resultados Automatizados**

Los resultados proporcionados por los buscadores suelen estar relacionados con las páginas que tienen mayor volumen de visitas o que invierten en publicidad. Sin embargo, es posible que pequeños nichos o websites muy especializados no aparezcan en los resultados de búsqueda o lo hagan en posiciones muy bajas en el ranking de SEO. Esta es la razón por la cual nuestra solución permite la introducción manual de registros, asegurando así una cobertura más amplia y precisa de los websites relevantes para el scraping.

*Módulo de Urls*:

Los productos y los sitios web están interrelacionados a través de las URLs. Estas URLs especifican qué productos están disponibles en el sitio web y proporcionan el enlace directo a la página del producto específico en ese comercio electrónico.

Las URLs son el núcleo del módulo de Scraping, ya que contienen tantos registros como páginas web se visitarán para extraer información.

El módulo de URLs incluye todas las funcionalidades y acciones necesarias para gestionar las URLs de manera eficiente y proporciona métodos y funciones para agilizar/facilitar las tareas al resto de módulos y aplicaciones que requerirán de sus servicios.

Las URLs se agregan automáticamente en base a los resultados obtenidos de los motores de búsqueda, aunque también se permite su manipulación manual. Esto permite a los usuarios corregir, añadir o eliminar URLs según sea necesario.

### *Módulo de Histórico de Precios*:

El proceso de extracción de datos desde una URL permite obtener el precio de un producto específico en un ecommerce determinado. Estos datos se almacenan en el módulo denominado "Histórico de Precios", facilitando su posterior tratamiento, manipulación y presentación al usuario final.

El sistema está diseñado para almacenar los precios en intervalos temporales diversos, que van desde segundos y minutos, hasta horas y días. Actualmente, el sistema se encuentra configurado para realizar capturas de precios diarias, es decir, se registra un precio por cada día del año.

El sistema está preparado para almacenar los precios en cualquier rango temporal: segundos, minutos, horas, … Debido al gran volumen de datos recibidos, se está trabajando a nivel diario, es decir, un precio para cada día del año.

### *Módulo de Templates*:

El raspado de páginas web se fundamenta en plantillas generadas por un Modelo de Lenguaje Grande (LLM). Este módulo se encarga de gestionar dichas plantillas, implementando todas las funcionalidades y acciones necesarias para su manejo eficiente.

Además, este módulo simula un MOE (Mixture Of Experts) para seleccionar la plantilla adecuada según los datos proporcionados, escoger el LLM a utilizar en caso de no tener plantillas preconstruidas para el caso concreto. También se encarga de almacenar los resultados de su

ejecución, permitiendo al administrador tomar decisiones informadas y ajustar el funcionamiento del sistema con base en los datos recopilados.

*Módulo de LLMS*:

Módulo que almacena los datos de los LLMs, su funcionalidad es totalmente informativa para ser visualizada en el frontend.

La parte técnica, configuraciones, … debido a la gran variedad de customizaciones se ha delegado a los scripts y/o configuración de la máquina.

#### *4.2.3.3 Aplicación BatchProcess*

Se ha creado una aplicación específica llamada 'batchProcess' para gestionar todos los procesos batch que requiere la solución tecnológica. Tener en el frontend esta aplicación permite a los administradores configurar, ejecutar y saber el estado de todos los procesos batch que hay en la solución.

La gestión de la escalabilidad usando el frontend es muy fácil y sencilla permitiendo crear nuevas tareas, nuevas máquinas, gestionar el reparto de las tareas entre las máquinas, saber el estado de las ejecuciones, …


*Figura 21 Diagrama del Módulo de BatchProcess*

*Módulo de Máquinas*:

El módulo de máquinas está diseñado para administrar las máquinas disponibles y facilitar la realización de procesos BatchProcess. Su objetivo principal es permitir la gestión de las máquinas desde el frontend, ofreciendo la posibilidad de modificar su disponibilidad según la demanda y las necesidades específicas.

Cada máquina dispone de scripts específicos para configurar la ejecución de tareas cómo respuesta a ciertos eventos proporcionando al administrador un mayor nivel de control.

Los eventos administrables son:

- Arranque de la máquina.
- Ejecución de una tarea en la máquina.
- Parada de la máquina.
- Comprobación del estado de la máquina.

Se añade funcionalidades de disponibilidad para que las máquinas pueden activarse/desactivarse para no tener que registrar de nuevo la máquina en caso de parada o cierre temporal.

Las máquinas se han categorizado por tipología para permitir la compartición de scripts. Se ha añadido un campo de Configuración para personalizar parámetros específicos de cada máquina.

La configuración es posible modificarla directamente desde el listado para ayudar al administrador a realizar sus modificaciones de manera más rápida.

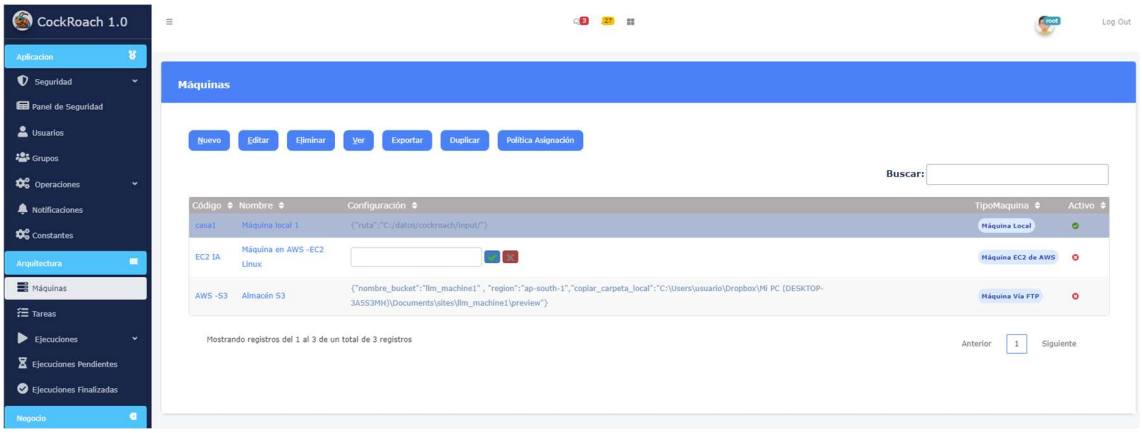



*Figura 22 Formulario en Grid para rápidas modificaciones*

Desde el Grid de Máquinas se configura la Política de Asignación de Máquinas.

#### *4.2.3.3.1.1.1 Política de Asignación de Máquinas*

Las tareas pendientes de ejecución son repartidas en las máquinas siguiendo la política marcada en la aplicación.

Desde esta opción, el usuario puede modificar dicha política para adaptarse a las distintas necesidades de negocio.

Actualmente, se han definido 3 tipos de políticas distintas para repartir las tareas:

- **Aleatoria**: Asignación aleatoria de máquinas.
- **Secuencial**: Asignación secuencial a cada máquina, las ejecuciones de las tareas se reparten secuencialmente asignando el mismo volumen de carga en todas las máquinas.
- **X Volumen**: Se asigna la ejecución de la tarea a la máquina que tenga menos ejecuciones pendientes, es decir, la que menos volumen de trabajo tiene asignada.

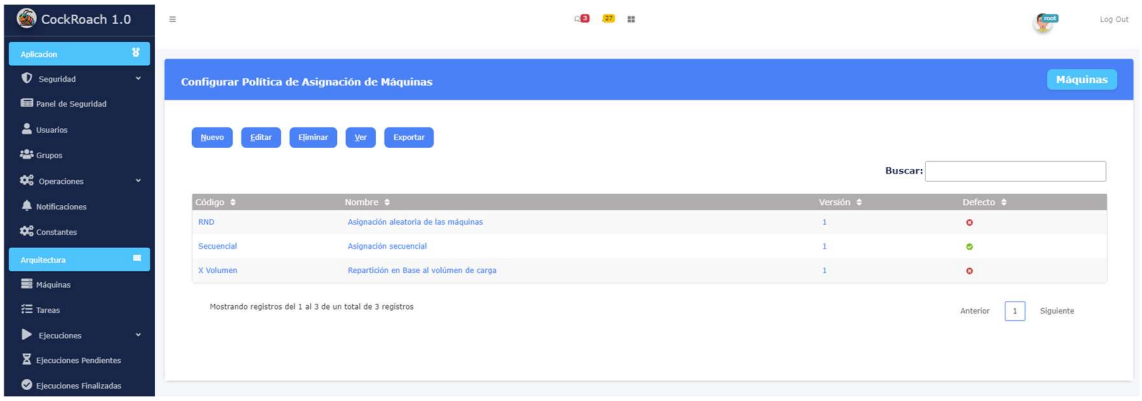



*Figura 23 Política de Asignación de Máquinas*

Por defecto, únicamente es posible disponer de una política de Asignación de Máquinas activa.

Aunque estas son las políticas actualmente definidas, no se descarta añadir nuevas políticas para adaptarse a las necesidades de negocio.

### *Módulo de Tareas*:

El módulo de tareas gestiona todas las tareas que requieren ser ejecutadas en un proceso Batch.

Cuando se recibe la respuesta de la ejecución de la tarea en la máquina correspondiente, se puede configurar un script para su ejecución y procesamiento del resultado recibido.

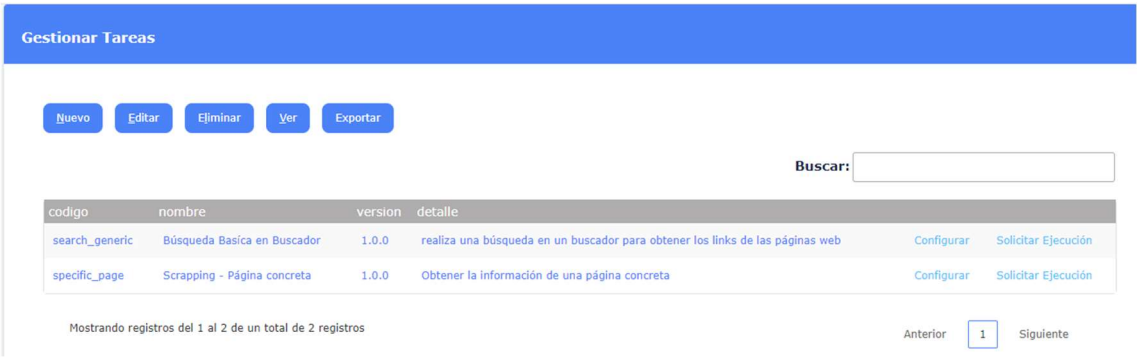



*Figura 24 Listado de tareas disponibles*

Las tareas pueden ser ejecutadas manualmente para el administrador. Esta opción no está pensada para usarse en producción, sino para temas de debug, testing, ...

Cada tarea además, tiene su propia configuración que se puede acceder desde la opción de 'Configurar' del listado de tareas.

En esta opción se visualizan todas las configuraciones disponibles para esta tarea y la configuración por defecto que se usará para su ejecución.

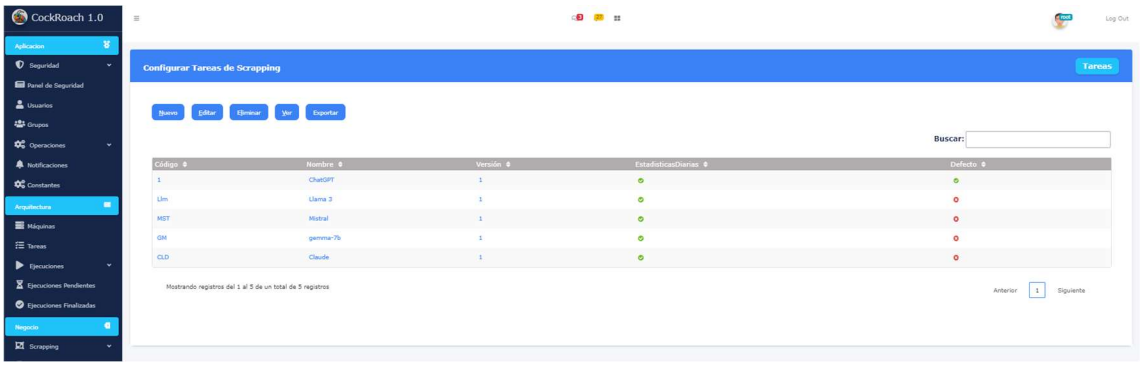



*Figura 25 Gestión de las configuraciones disponibles para configurar Tarea de Scrapping*

### *Módulo de Ejecuciones y Ejecuciones Finalizadas*:

Toda ejecución de una tarea en una máquina tiene reflejada la acción en un registro en la tabla de Ejecuciones en caso de que no haya finalizado o en la tabla de Ejecuciones Finalizadas si ya se dispone de un resultado.

Este módulo ofrece todas las funcionalidades para poder gestionar las ejecuciones, saber el estado de las mismas, la configuración usada, los tiempos de respuesta, el script ejecutado cuando se finaliza su ejecución, …

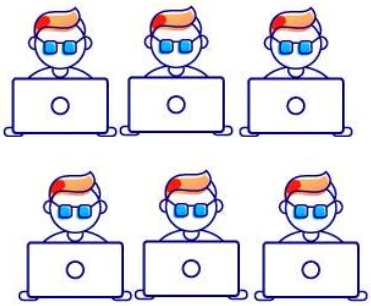
La tabla de Ejecuciones es muy útil para el administrador para detectar el correcto funcionamiento de la arquitectura, encontrar ejecuciones paralizadas y encontrar el stopper para remediar la solución.

La tabla de Ejecuciones Finalizadas almacena el resultado de todas las ejecuciones y sus datos para la realización de estudios y estadísticas.

Esta tabla se va vaciando cada cierto tiempo cómo consecuencia del gran volumen de información que recibe diariamente.

## 4.3 Recursos requeridos

A continuación, se enumeran los recursos tecnológicos usados en el proyecto:



- Mysql
- Apache
- Django
- Xampp
- Windows
- Linux
- Llama3
- ChatGPT
- Gemma-7b
- Mistral
- GPT-3
- GPT-4
- GPT-4 o
- Chrome
- Ollama
- Firefox
- Tor Browser

Librerías utilizadas:

- Requests
- Mechanize
- Selenium
- Transformers
- Ollama
- Openai
- Colorama
- Bs4
- Stem

- Django 5.0.6
- Pymysql
- Django-mysql-pymysql
- Djangorestframework
- Django-cors-headers
- Django-simple-history
- Mysqlclient
- Django-import-export
- Plotly
- Kaleido
- Pandas
- Dash
- Datatables.net
- Bootstrap
- AdminIte
- Toastr
- Torch
- Fake-useragent
- JQuery
- Jszip
- Pdfmake
- Plotly
- Htmx

Aplicaciones utilizadas:

- TaskCoach
- LowCode2
- WinMerge
- Baregrep
- Excalidraw
- VSCode
- Notepad++
- PyCharm
- AnyDesk

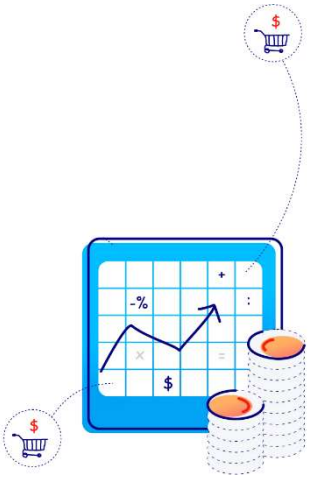
En la documentación se añaden los anexos para la configuración:

Frontend: CockRoach-Xavi Rambla_Anexo_Instalación FrontEnd

Aplicación: CockRoach-Xavi Rambla_Anexo_Instalación Aplicación y LLMs

## 4.4 Presupuesto

Para desarrollar esta solución, es fundamental llevar a cabo un estudio económico exhaustivo que permita evaluar diversos aspectos financieros y operativos del proyecto. Este estudio proporcionará una visión clara sobre los recursos necesarios, los costos asociados y los posibles beneficios, garantizando una planificación adecuada y sostenible en el tiempo. A continuación, se presenta un análisis estructurado y orientativo, que ofrece una visión resumida de los principales factores a considerar:



| | | Rango Estimado (Euros) | |
|---|---|---|---|
| **Categoría** | **Descripción** | **Mínimo** | **Máximo** |
| **Desarrollo del Software** | | | |
| Costos de Desarrollo Inicial | | | |
| Personal de Desarrollo | Coste 6 meses por personal | 35.000 | 45.000 |
| Jefe de Proyecto | 20% de 6 meses | 15.000 | 20.000 |
| Herramientas y Licencias | | | |
| Herramientas de Desarrollo | | 0 | 1.000 |
| Licencias de Software y Bibliotecas | Costo de licencias | 100 | 1.000 |
| Servicios de Almacenamiento en la Nube | Costo anual | 300 | 3.000 |
| **Costes operativos** | | | |
| Mantenimiento y Actualizaciones | | | |
| Personal de Mantenimiento | Costo anual por personal | 40.000 | 75.000 |
| Infraestructura | | 300 | 3.000 |
| Servidores y Almacenamiento en la Nube | Costo anual | 500 | 2.000 |
| Herramientas de Monitoreo y Seguridad | Costo anual | 1.000 | 2.000 |
| **Cumplimiento Legal** | | | |
| Asesoría Legal | Costo anual en asesoría | 0 | 10.000 |
| **Costos Totales Estimados** | | 92.200 | 162.000 |
| **Costos Anuales Estimados** | | 42.200 | 87.000 |

*Tabla 2 Tabla de costes resumido*

Se debe tener en cuenta que el desarrollo del software tiene una previsión de realizarse en 6 meses mientras que los costes operativos por Mantenimiento y Actualizaciones deberán realizarse anualmente mientras perdure el proyecto en activo.

La parte del Cumplimiento Legal, se prevé una partida para realizar un estudio por una asesoría legal para cumplir la ley.

### 4.4.1 Presupuesto Mínimo Estimado

En este aparto se visualiza el reparto de los costes de la solución en caso de ejecutar el Presupuesto Mínimo estimado indicado anteriormente.

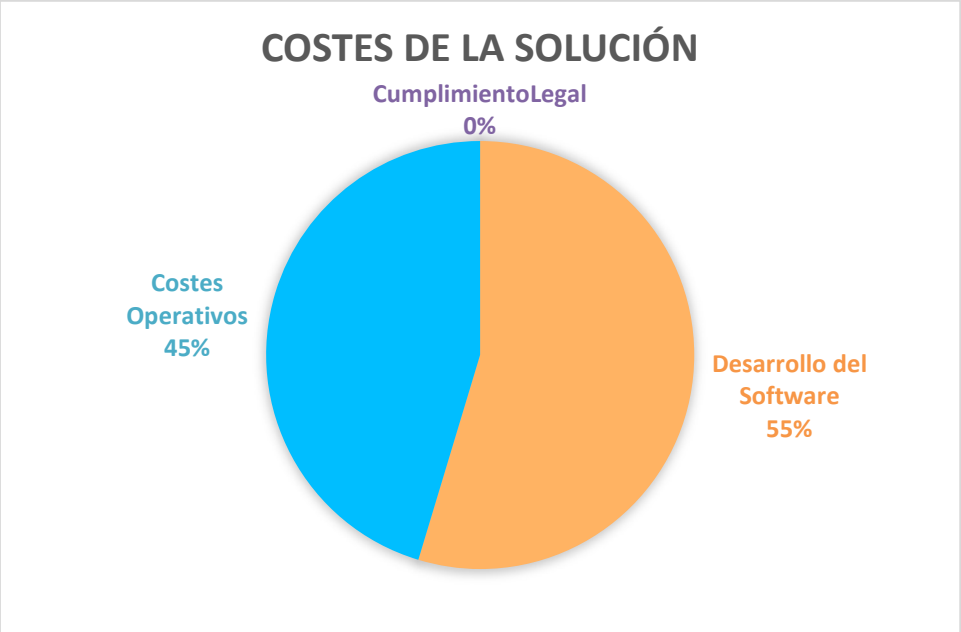



*Figura 26 Costes de la solución para el Presupuesto Mínimo Estimado*

Se puede observar cómo el desarrollo del Software se lleva el 55% del presupuesto estimado mientras que los costes Operativas representan el 45%. Los temas legales no tienen ningún valor y por eso tienen un 0%.

### 4.4.2 Presupuesto Máximo Estimado

En este aparto se visualiza el reparto de los costes de la solución en caso de ejecutar el Presupuesto Máximo estimado indicado anteriormente.

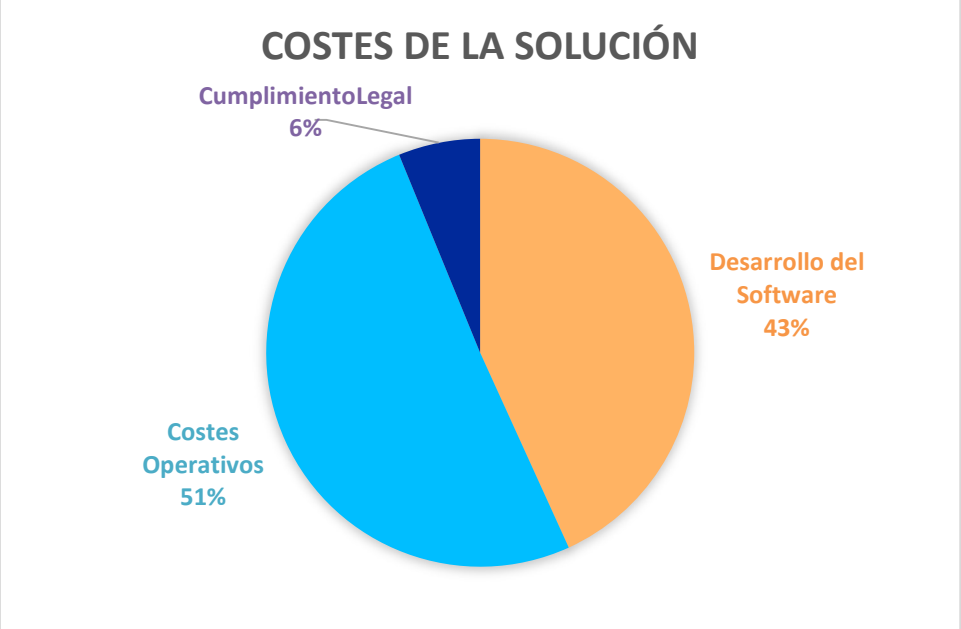



*Figura 27 Costes de la solución para el Presupuesto Máximo Estimado*

Se puede observar cómo el desarrollo del Software se lleva el 43% del presupuesto estimado mientras que los costes Operativas representan el 51%. Los temas legales no tienen ningún valor y por eso tienen un 6%.

### 4.4.3 Conclusiones del Presupuesto

Las conclusiones sobre los dos presupuestos son los siguientes:

Mientras que los Costes de Desarrollo del Software varían ligeramente, el resto de las partidas se ven incrementadas tanto en importe cómo en % respecto al importe presupuestado.

La partida de 'Cumplimiento Legal' es lógico su ascenso debido a que la partida no tenía asignado ningún coste al realizar el Presupuesto Mínimo Estimado.

Los Costes operativos se han doblado entre el presupuesto mínimo y máximo, esta diferencia se debe a qué todas las subpartidas de este apartado han crecido exponencialmente consecuencia de la necesidad de multiplicar los recursos requeridos. Esta situación está pensada en caso de que el volumen de clientes sea importante y se deba aumentar los recursos para hacer frente a la demanda.

Los Costes del Desarrollo del Software se han mantenido bastante estables ya que un aumento de la demanda tiene un mínimo efecto sobre el Desarrollo en esta primera versión del proyecto. En posteriores versiones, se tiene previsto la creación e integración con APIs dónde los costes De Desarrollo se verán más impactados con el volumen de Clientes.

### 4.4.4 Análisis Detallado del Presupuesto

A continuación, se añade el análisis estructurado orientativo completo:

| | | Rango Estimado (Euros) | |
|---|---|---|---|
| **Categoría** | **Descripción** | **Mínimo** | **Máximo** |
| **Desarrollo del Software** | | | |
| Costos de Desarrollo Inicial | | | |
| Personal de Desarrollo | Coste 6 meses por personal | 35.000 | 45.000 |
| Jefe de Proyecto | 20% de 6 meses | 15.000 | 20.000 |
| Herramientas y Licencias | | | |
| Herramientas de Desarrollo | | 0 * | 1000 * |
| VSCode | | | |
| Notepad++ | | | |
| XAMPP | | | |
| Gimp | | | |
| Monday.com | | | |
| TaskCoach | | | |
| Bitbucket | | | |
| ChatGPT | | | 200 |
| Microsoft Copilot | | | 70 |
| Licencias de Software y Bibliotecas | Costo de licencias | 100 * | 1000 * |
| Mysql | | | |
| Apache | | | |

| | | | |
|---|---|---|---|
| Django | | | |
| Win | | | |
| Apache | | | |
| Llama3 | | | |
| Ollama | | | |
| ChatGPT | | | |
| ChatGPT | | | |
| Gemma-7B | | | |
| Mistral | | | |
| GPT-3 | | | |
| GPT-4 | | | |
| GPT-4o | | | |
| Firefox | | | |
| Chrome | | | |
| Tor Browser | | | |
| Requests | | | |
| Colorama | | | |
| Mechanize | | | |
| Selenium | | | |
| Transformers | | | |
| OpenAI | | | 200 |
| BS4 | | | |
| Stem | | | |
| PyMysql | | | |
| Django-mysql-pymysql | | | |
| DjangoRestFramework | | | |
| Django-cors-headers | | | |
| Django-simple-history | | | |
| Mysqlclient | | | |
| Django-import-export | | | |
| Plotly | | | |
| Kaleido | | | |
| Pandas | | | |
| Dash | | | |
| Datatables.net | | | |
| bootstrap | | | |
| adminlte | | | |
| Toastr | | | |
| Jquery | | | |
| jszip | | | |
| pdfmake | | | |

| | | | |
|---|---|---|---|
| Plotly | | | |
| htmx | | | |
| Servicios de Almacenamiento en la Nube | Costo anual | 300 | 3.000 |
| **Costes operativos** | | | |
| Mantenimiento y Actualizaciones | | | |
| Windows | | | |
| Linux | | | |
| Antivirus | | | |
| Personal de Mantenimiento | Costo anual por personal | 40.000 | 75.000 |
| Infraestructura | | 300 | 3.000 |
| Servidores y Almacenamiento en la Nube | Costo anual | 500 | 2.000 |
| Herramientas de Monitoreo y Seguridad | Costo anual | 1.000 | 2.000 |
| **Cumplimiento Legal** | | | |
| Asesoría Legal | Costo anual en asesoría | 0 | 10.000 |
| **Costos Totales Estimados** | | 92.100 | 160.470 |
| **Costos Anuales Estimados** | | 42.100 | 85.470 |

** Importe orientativo*
*Tabla 3 Tabla de costes completo*

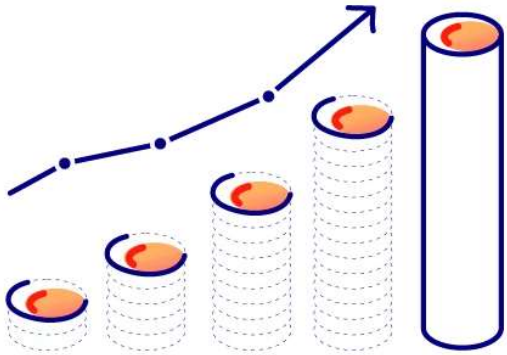
## 4.5 Viabilidad

La viabilidad económica es un aspecto crítico en la evaluación de cualquier proyecto o iniciativa empresarial. En este capítulo, se abordarán los elementos clave que determinan si un proyecto es financieramente viable y sostenible a largo plazo. La viabilidad económica no solo se refiere a la capacidad de un proyecto para generar beneficios, sino también a su capacidad para manejar riesgos financieros, atraer inversiones y mantenerse competitivo en el mercado.

### 4.5.1 Análisis de Retorno de Inversión (ROI)

Este proyecto no tiene previsión de ser vendido cómo servicio, aunque se plantea dicha opción como una oportunidad a explotar a largo plazo.

El SaaS es un software que se ofrece a los clientes para su uso por una tarifa de 2000€ por cliente, inicialmente se planifica realizar un descuento del 50% para atraer clientes con la finalidad de obtener su feedback y tener una solución más robusta y completa.



*Ingresos Anuales.*

El ingreso anual se calcula multiplicando el número de clientes por la tarifa anual.

*Costes Anuales.*

Los costes anuales se estiman redondeando entre 25.000 y 50.000€. Cómo ya se ha detallado en el capítulo anterior, estos costes incluyen el desarrollo. mantenimiento del software, servidores, entre otros.

Quedan de este punto excluidos los costes colaterales cómo serían costes de marketing, gastos de asesoramiento financiero, gastos legales, y otros costes importantes para el negocio, pero independientes a la solución.

#### *4.5.1.1 Cálculo del Punto de Equilibrio (Break-even Point):*

El punto de equilibrio es el número de clientes necesario para que los ingresos sean iguales a los costes. Se calcula dividiendo los costes anuales por la tarifa anual por cliente.

ROI = Ganancias netas / Coste de la inversión * 100

##### 4.5.1.1.1 Cálculo del Punto de equilibrio

Comentar que los cálculos se han realizado sobre un importe redondeado al alza de la tabla de costes para facilitar la comprensión

Costes Mínimos: 50.000 € (cálculo real 42.200 €)
Costes máximos: 100.000 € (cálculo real 87.000 €)

Ingreso anual por cliente (primer año): 1000€

**Punto de equilibrio mínimo = 50.000/1000 = 50 clientes**
**Punto de equilibrio máximo = 100.000/1000 = 100 clientes**

Es decir, el software requiere entre 50 y 100 clientes mínimo para poder cubrir los costes mínimos necesarios para operar la solución.

#### 4.5.1.1.2 ROI

Supongamos que se adquieren N clientes.

Ingresos: 1000€ * N
Costes: Entre 50.000€ y 100.000€

Entonces, las ganancias netas serían:

Para costes mínimos: 1000€ * N - 50.000€
Para costes máximos: 1000€ * N - 100.000€

A continuación, se plantean varios escenarios hipotéticos para el primer año del software.

##### *4.5.1.1.2.1 Ejemplo del cálculo del ROI para 100 clientes*

| | Cuota | Núm. Clientes | Total |
|---|---|---|---|
| **Ingresos** | 1.000 € | 100 | **100.000 €** |

| | Ingresos | Costes mínimos | Total |
|---|---|---|---|
| **Ganancias Netas** | 100.000 € | 50.000 € | **50.000 €** |

| | Ingresos | Costes máximos | Total |
|---|---|---|---|
| **Ganancias Netas** | 100.000 € | 100.000 € | **0 €** |

| | Ingresos | Costes | |
|---|---|---|---|
| **ROI Costes máximos** | 100.000 € | 50.000 € | **200%** |
| **ROI Costes mínimos** | 100.000 € | 100.000 € | **100%** |

*Tabla 4 Cálculo del ROI para 100 clientes*

**Resumen**
Dependerá del número de clientes y de los costes exactos. Para 100 clientes, el ROI variará entre 0% y 200% dependiendo de los costes.

#### *4.5.1.1.2.2 Ejemplo del cálculo del ROI para 150 clientes*

| | Cuota | Num. Clientes | Total |
|---|---|---|---|
| **Ingresos** | 1.000 € | 150 | **150.000 €** |

| | Ingresos | Costes mínimos | Total |
|---|---|---|---|
| **Ganancias Netas** | 150.000 € | 50.000 € | **100.000 €** |

| | Ingresos | Costes máximos | Total |
|---|---|---|---|
| **Ganancias Netas** | 150.000 € | 100.000 € | **50.000 €** |

| | Ingresos | Costes | |
|---|---|---|---|
| **ROI Costes máximos** | 150.000 € | 50.000 € | **300%** |
| **ROI Costes mínimos** | 150.000 € | 100.000 € | **150%** |

*Tabla 5 Cálculo del ROI para 150 clientes*

**Resumen**

Dependerá del número de clientes y de los costes exactos. Para 150 clientes, el ROI variará entre 100% y 300% dependiendo de los costes.

#### *4.5.1.1.2.3 Ejemplo del cálculo del ROI para 200 clientes*

| | Cuota | Num. Clientes | Total |
|---|---|---|---|
| **Ingresos** | 1.000 € | 200 | **200.000 €** |

| | Ingresos | Costes mínimos | Total |
|---|---|---|---|
| **Ganancias Netas** | 200.000 € | 50.000 € | **150.000 €** |

| | Ingresos | Costes máximos | Total |
|---|---|---|---|
| **Ganancias Netas** | 200.000 € | 100.000 € | **100.000 €** |

| | Ingresos | Costes | |
|---|---|---|---|
| **ROI Costes máximos** | 200.000 € | 50.000 € | **400%** |
| **ROI Costes mínimos** | 200.000 € | 100.000 € | **200%** |

*Tabla 6 Cálculo del ROI para 200 clientes*

**Resumen**

Dependerá del número de clientes y de los costes exactos. Para 200 clientes, el ROI variará entre 200% y 400% dependiendo de los costes.

En caso de cumplirse este escenario, se debe replantear el aumento de los costes para hacer una adecuación de las estructuras de la empresa y reducir y/o eliminar la política de descuentos sobre el precio para los nuevos clientes.

*Tabla 7 Tabla de Costes Mínimos*

| Tabla de Costes Mínimos | | | | |
|---|---|---|---|---|
| **Núm. Clientes** | **Ingresos** | **Costes Mínimos** | **Resultado** | **% sobre Coste** |
| 100 | 100000 | 50000 | 50000 | 100% |
| 150 | 150000 | 50000 | 100000 | 200% |
| 200 | 200000 | 50000 | 150000 | 300% |
| 250 | 250000 | 50000 | 200000 | 400% |
| 500 | 500000 | 50000 | 450000 | 900% |
| 1000 | 1000000 | 50000 | 950000 | 1900% |

*Tabla 8 Tabla de Costes Máximos*

| Tabla de Costes Máximos | | | | |
|---|---|---|---|---|
| **Núm. Clientes** | **Ingresos** | **Costes Máximos** | **Resultado** | **% sobre Coste** |
| 100 | 100000 | 100000 | 0 | 0% |
| 150 | 150000 | 100000 | 50000 | 50% |
| 200 | 200000 | 100000 | 100000 | 100% |
| 250 | 250000 | 100000 | 150000 | 150% |
| 500 | 500000 | 100000 | 400000 | 400% |
| 1000 | 1000000 | 100000 | 900000 | 900% |

## 4.6 Resultados del proyecto

Los resultados obtenidos en este proyecto son altamente satisfactorios, dado que se ha logrado desarrollar una solución integral para la captura de precios de la competencia. Esta solución no solo es capaz de adaptarse a diversas situaciones y escenarios del entorno competitivo, sino que también permite la visualización tanto a nivel local como global de la posición de los productos y su comparación directa con los de otras empresas competidoras.

Esta capacidad de análisis situacional es clave para la toma de decisiones informadas y estratégicas en entornos dinámicos de mercado.

### 4.6.1 Arquitectura Tecnológica

Desde un punto de vista arquitectónico, el sistema presenta varias características que lo hacen tecnológicamente avanzado y altamente adaptable a las necesidades actuales.

La solución es distribuible y fácilmente instalable, permitiendo una instalación completa en menos de 8 horas.

Además, su capacidad de ser escalable es otro de sus puntos fuertes: los módulos de la aplicación son replicables de manera ágil para ajustarse a aumentos en la demanda, mientras que la única restricción es la necesidad de centralizar la base de datos. Aun así, esta limitación podría ser superada con un cambio en el motor de base de datos, lo cual permitiría mayor flexibilidad en su arquitectura.

Desde el frontend, la herramienta ofrece un amplio rango de opciones configurables que permiten personalizar el comportamiento de la solución según los requerimientos del negocio.

Estas incluyen, por ejemplo, la configuración de políticas para la asignación de recursos computacionales, la selección del modelo de lenguaje LLM a utilizar para diversas tareas, la elección del navegador para realizar operaciones de scraping, …. Esto asegura que la solución no solo sea potente, sino también flexible y adaptable a distintos contextos operacionales.

### 4.6.2 Visión Global y Gestión Centralizada

En términos de software, se ha diseñado la herramienta con un enfoque hacia la observabilidad y el control centralizado. Todo evento, acción o interacción dentro del sistema es registrado y puede ser gestionado desde el frontend, permitiendo al usuario una visión global del funcionamiento del software en tiempo real. Esta trazabilidad incluye información clave, como

el uso de templates generados, el número de errores, el porcentaje de éxito en la obtención de datos, y detalles precisos sobre las URLs visitadas para la recolección de datos.

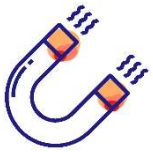
Asimismo, el sistema informa continuamente del estado de las ejecuciones, lo que proporciona una visión clara y detallada del progreso de las operaciones, así como del estado real de cada proceso. Esta capacidad de monitoreo mejora significativamente la eficiencia operativa, facilitando la identificación y resolución de posibles puntos de falla en tiempo real.

#### *4.6.2.1 Capacidad de Notificación y Gestión de Anomalías*

Un aspecto adicional que potencia el uso de esta herramienta es la capa de notificaciones, la cual está diseñada para cubrir una posible falta informativa y sirve para alertar a los usuarios sobre eventos o situaciones críticas. Esta capa también incluye una serie de disparadores que notifican en caso de detectar situaciones anómalas o fuera de lo esperado. Esta funcionalidad mejora la capacidad de respuesta ante contingencias y permite una gestión proactiva del sistema.

### 4.6.3 Optimización de los Modelos de Lenguaje (LLM)

Para la optimización de los Modelos de Lenguaje LLM se trata cómo un punto a parte debido a su importancia y relevancia en este proyecto.



En el ámbito de los modelos de lenguaje (LLMs), se realizaron importantes modificaciones en las políticas de uso debido a los resultados iniciales subóptimos, tanto en términos de rendimiento como de precisión en los resultados obtenidos.

Inicialmente, se implementó una política simple basada en recuperar el HTML de las páginas web, asignar un contexto y recibir el precio cómo respuesta del LLM.

Este enfoque, aunque simple, demostró ser insuficiente en cuanto a rendimiento y precisión de los resultados. Conseguir el precio era complejo ya que se consigue después de realizar un volumen de llamadas LLM (no especifico números ya que depende del LLM pero entre 1 y 30 llamadas), prácticamente ninguna vez se conseguía a la primera aunque, cuando devolvían los precios raramente se equivocaban.

La mejora del enfoque se basó en las siguientes tres palancas:

- **Solicitar el código Python**: En las primeras versiones se busca el precio usando el LLM, en las versiones actuales se solicita al LLM que genere el código Python para recuperar el precio. Este cambio mantiene o empeora el rendimiento de la solución ante nuevas y desconocidas solicitudes pero supone un salto sustancial cuando la situación ya es conocida y se dispone del código Python.

- **Mejora del contexto**: Se mejora y detalla de forma precisa un nuevo contexto proporcionado al LLM, guiándolo con mayor claridad y enriqueciendo la información con mayor precisión. Ejemplo, un gran avance en este campo fue la traducción al inglés del contexto.

- **Automatización del bucle de generación**: Implementación de un bucle iterativo para solicitar diferentes versiones de código al LLM hasta obtener el resultado esperado o alcanzar un límite de solicitudes predeterminado desde la Configuración de la Tarea en el Frontend.

Toda esta batería de cambios, logran mejorar los resultados obtenidos tanto en rendimiento cómo en precisión, y donde se mantiene el mal rendimiento es únicamente durante la fase de exploración de templates ante situaciones desconocidas.

A continuación, se muestra una tabla de los distintos LLMs y sus resultados

*Tabla 9 Resultados LLM con 10 webs*

| Proveedor | Modelo | Código OK | Media de llamadas al LLM | Media de llamadas cuando acierta el LLM |
|---|---|---|---|---|
| Facebook | Llama 3.1 | 4 | 14,4 | 4,25 |
| Facebook | Llama 3 | 5 | 13 | 3,8 |
| Mistral AI | Mistral | 3 | 16,3 | 7,667 |
| Google | Gemma2 | 3 | 16,2 | 7,333 |
| **OpenAI** | **ChatGPT** | **7** | **7,9** | **2,714** |
| Alibaba | Qwen2 | 4 | 14,7 | 6,75 |

En septiembre, los distintos LLMs fueron evaluados a través de 10 pruebas[2], donde ChatGPT demostró ser el modelo con los mejores resultados. Sin embargo, es importante señalar que se utilizó una versión de pago de ChatGPT, mientras que los demás modelos evaluados eran gratuitos, lo que podría haber influido en los resultados. Cabe destacar que los LLMs evolucionan constantemente, y es previsible que los modelos gratuitos mejoren con el tiempo, por lo que este estudio podría volverse obsoleto rápidamente.

La tabla únicamente ofrece resultados del LLM, los datos de rendimiento no se representan ya que son engañosos y se llega a falsas conclusiones. ChatGPT tiene unos resultados excelentes ya que utiliza un servicio en la nube mientras que el resto de LLMs se ejecutan en local en una máquina no optimizada.

[2] Información únicamente ilustrativa ya que el volumen de pruebas es insuficiente para la toma de decisiones.

## 4.7 Beneficios para Negocio

El software actual lleva únicamente unas semanas en producción con lo que no es posible precisar conclusiones acerca de los beneficios que aporta la solución a negocio.

Actualmente, los beneficios obtenidos son:

- Detección de un proveedor bajando precios en fin de semana por debajo del valor marcado por el proveedor
- Definición de patrones de las Políticas de precios de algunos proveedores.
- Detección de cambios de precio en momentos puntuales del día.

Estos beneficios deben ampliarse y para negocio se esperan estos tres pilares:

- Beneficio económico
- Beneficio operativo
- Pricing avanzado

## 4.8 Adaptaciones a las Necesidades del Negocio

En cuanto a la alineación con las necesidades del negocio, la solución ha experimentado varios cambios significativos. Un ejemplo radical es la creación de los dashboards en la plataforma empresarial existente, a solicitud del equipo de negocio, para centralizar toda la visualización de resultados y nuevas funcionalidades en una única herramienta para el usuario. Esta integración requiere la creación de APIs y la comunicación con otros sistemas, permitiendo así un intercambio fluido de información entre plataformas.

## 4.9 Conclusión

Como conclusión general, este proyecto ha demostrado el potencial que los LLMs pueden ofrecer en la personalización de soluciones adaptadas a entornos competitivos ya no únicamente en el mundo del Pricing sino también en otras áreas y mercados. La capacidad de automatizar procesos complejos y obtener datos precisos de forma ágil, rápida y automática señala un camino prometedor para su uso en la toma de decisiones empresariales.

Además, el bajo nivel de error cuando consigue un resultado es relevante y sorprende ante la cantidad de información que procesa.

Grandes empresas como TradeInn han desarrollado soluciones similares, invirtiendo varios años y presupuestos superiores a 1M €, lo que destaca la relevancia y el valor de la solución presentada en este proyecto, ya que ha alcanzado niveles de eficiencia comparables en un tiempo aproximadamente de 4 meses y unos costes significativamente menores gracias a los LLMs.

# Capítulo 5. DISCUSIÓN

En este capítulo se van a tratar las principales discusiones que se han generado al realizar este proyecto y las conclusiones o acuerdos alcanzados ante tal problema.

## 5.1 Problemas con los buscadores

Los Buscadores de internet tratan de eliminar todas las consultas que se realizan mediante Scrapping con lo que se han realizado varios estudios para superar dichos problemas.

Algunas de las soluciones propuestas son:

- Utilizar varios navegadores
- Utilizar varios ordenadores distintos.
- Utilizar VPNs y tecnologías parecidas.

### 5.1.1 Conclusión

El objetivo es implementar un sistema que utilice diferentes direcciones IP y que modifique la IP automáticamente cuando el buscador o el ecommerce detecte un posible ataque de scraping.

La red Tor proporciona una solución eficaz, permitiendo cambiar de identidad en cualquier momento, lo que facilita la continuación del scraping utilizando diferentes identidades.

Tor funciona enviando el tráfico a través de tres servidores aleatorios (también conocidos como repetidores) en la red Tor. El último repetidor en el circuito (el "repetidor de salida") envía el tráfico hacia el Buscador/Ecommerce solicitado.

Esta solución es la seleccionada por ser gratuita, fácil de implementar y sencilla de entender. Aunque tiene algunas deficiencias que se solucionan cambiando la identidad nuevamente hasta encontrar una identidad ‘legal’.

## 5.2 Diseño arquitectura del Módulo de IA

Cómo se ha comentado en el documento, se han realizado varios diseños de arquitectura para procesar la información buscando principalmente tres características:

- Fiabilidad
- Velocidad
- Coste económico

### 5.2.1 Acceso directo al Modelo LLM

Para obtener el precio a partir del HTML de una página web, se utiliza un modelo de lenguaje grande (LLM) con un contexto predefinido. En nuestras pruebas, esta opción ha funcionado correctamente únicamente con GPT-3.5 y, de manera parcial, con Llama3. Otros modelos, como Mistral y Gemini, han sido descartados debido a resultados significativamente imprecisos.

Desde una perspectiva de rendimiento, Llama3 se descarta debido a su tiempo de respuesta de aproximadamente un minuto por solicitud, mientras que GPT-3.5, al ser un servicio externo, ofrece tiempos de respuesta mucho más rápidos.

Sin embargo, en términos de costo, GPT-3.5 resulta ser considerablemente más caro, mientras que Llama3 no incurre en costos.

Dada esta situación, se explora nuevas soluciones alternativas que puedan alcanzar el mismo objetivo con mayor eficiencia y menor costo.

### 5.2.2 MultiDemanda al Modelo LLM

Una estrategia propuesta para mejorar el rendimiento y acelerar el proceso consiste en enviar múltiples páginas al modelo de lenguaje (LLM) y solicitar los precios correspondientes.

Sin embargo, esta solución se descarta por completo debido a las limitaciones en el tamaño del contexto que la mayoría de los LLMs pueden manejar, lo cual impide procesar tal cantidad de información de manera efectiva.

### 5.2.3 Acceso a Internet del Modelo LLM

Para resolver el problema del tamaño de contexto se plantea pedir al modelo LLM que coja los datos de internet para recuperar el Precio.

Solución descartada ya que no existe actualmente una tecnología con esa funcionalidad. Lo más parecido es LangChain pero es una tecnología muy nueva, poco madura y tampoco tengo la seguridad que se pueda conseguir con los desarrollos actuales.

### 5.2.4 Dashboards o APIs

En un principio se abordó la solución para que desde el frontend de la web los usuarios realicen todas las tareas y estudios. Según se ha desarrollado el proyecto y se ha madurado la idea ha surgido la idea de utilizar APIs para que los usuarios puedan tener la información directamente en sus aplicaciones internas.


Las comunicaciones entre CockRoach y dichas aplicaciones serían multidireccionales y servirían para traspasar información, actualizar datos y realizar las oportunas acciones.

La idea de las APIs nace con la finalidad de facilitar las tareas al usuario y no añadir un nuevo portal, y darle la información en sus aplicaciones actuales.

### 5.2.5 Conclusión

Cómo ya se ha comentado en este documento, se modifica la filosofía de uso del LLM para solicitar que genere el código Python que permita recoger los datos de la página Html.

Esta solución ofrece unas mejoras de rendimiento muy buenos ya que únicamente se llama al modelo LLM en la creación de Templates, después se utilizan dichas plantillas para recuperar la información obteniendo unos tiempos de respuesta excelentes, una gran precisión y sin alucinaciones.

Ante la solución anterior, se define una arquitectura montada por una gran variedad de templates generados por el LLM.

### 5.2.6 Facilitando la labor al LLM

La solución previa resultó en un código Python considerablemente complejo y propenso a errores debido a la gran cantidad de variaciones en el HTML. Por lo tanto, se están explorando alternativas que permitan obtener el mismo resultado de manera más eficiente y robusta.

#### *5.2.6.1 Datos Estructurados*

Los 'Datos estructurados' son un estándar promovido por Google que permite a los motores de búsqueda acceder y comprender la información de una página web de manera eficiente, utilizando un formato JSON identificable y accesible. Este JSON de datos estructurados incluye todos los detalles que se desean comunicar a los motores de búsqueda para que se incorporen a los resultados de búsqueda, facilitando así una indexación rápida y precisa, y mejorando el SEO (Search Engine Optimization).

Una de las secciones clave de los datos estructurados es la de “Product”, donde se encuentra la sección "Offer" que incluye el campo "Price". Esta metodología es utilizada por Google para poblar la opción 'Shopping' en sus resultados de búsqueda.

Ajustamos el contexto y el modelo de scraping para enviar al LLM (Language Model) los datos estructurados cuando estén disponibles, optimizando así la precisión y relevancia de los datos extraídos.

#### *5.2.6.2 Conclusión*

Los resultados de realizar este cambio son extraordinarios obteniendo una mejora sustantiva en el código Python generado.

El punto negativo de esta tecnología es que no todos los ecommerce tratados contienen datos estructurados.

# Capítulo 6. CONCLUSIONES

En este capítulo se presentan las conclusiones derivadas del análisis y desarrollo realizado a lo largo de este trabajo. A partir de los resultados obtenidos, se reflexionará sobre el cumplimiento de los objetivos propuestos y se discutirán los principales hallazgos y aportaciones.

## 6.1 Conclusiones del trabajo

El desarrollo del software CockRoach ha permitido cumplir con el objetivo principal de este proyecto: crear una herramienta eficiente y flexible para el Scrapping de precios en diversas plataformas online. A lo largo del desarrollo, se abordaron varios desafíos técnicos y conceptuales, desde la obtención confiable de datos hasta la gestión de estos de manera estructurada y funcional.

A continuación, detallo algunos de los hitos más destacables:

- **Eficiencia en la recolección de datos**: CockRoach ha demostrado ser capaz de extraer grandes volúmenes de datos con precisión y en tiempos reducidos. Gracias a la integración de bibliotecas avanzadas de scraping, como Selenium o BeautifulSoup, el software logra acceder a precios de productos de manera rápida y eficaz, incluso en sitios que requieren interacción dinámica. Además, la automatización de tareas y la planificación de ejecuciones programadas han optimizado el proceso de extracción.

- **Adaptabilidad a diferentes plataformas**: Uno de los retos más importantes del scrapping es la variabilidad en la estructura de las páginas web. CockRoach se diseñó con un enfoque modular, generando la configuración específica para cada plataforma de comercio electrónico usando LLMs. Esta flexibilidad ha resultado en una herramienta capaz de adaptarse fácilmente a cambios en la estructura HTML de las páginas, lo que extiende su ciclo de vida sin necesidad de constantes ajustes manuales.

- **Tratamiento de datos y almacenamiento**: La estructuración de los datos extraídos y su almacenamiento ha sido otro de los pilares clave de este proyecto. CockRoach no solo extrae información, sino que la organiza de manera coherente en bases de datos optimizadas, permitiendo un acceso rápido para el análisis. Los precios capturados se almacenan con metadatos relevantes, como fechas, nombres de productos y fuentes, facilitando el análisis de tendencias de precios a lo largo del tiempo.

- **Gestión de bloqueos y restricciones**: Muchos sitios web implementan medidas de seguridad para evitar el scrapping, como CAPTCHAs, tasas de acceso limitadas o restricciones basadas en IP. CockRoach ha incorporado soluciones para mitigar estos problemas, como el uso de la red Tor y la simulación de comportamiento humano, lo que ha permitido minimizar los bloqueos durante la recolección de datos.

- **Impacto en la toma de decisiones comerciales**: La implementación de CockRoach ha mostrado un claro potencial en el análisis competitivo y la estrategia de precios. La capacidad de recopilar y analizar precios en tiempo real otorga a las empresas una ventaja competitiva al proporcionar datos precisos sobre el mercado, lo que permite ajustes de precios rápidos y eficientes. El software se posiciona como una herramienta valiosa para cualquier empresa que busque optimizar su modelo de precios basado en la observación del entorno competitivo.

En resumen, CockRoach se ha consolidado como una herramienta eficiente y versátil para el scraping de precios, con un impacto directo en la estrategia de precios y la competitividad en el mercado.

## 6.2 Conclusiones personales

A lo largo del desarrollo del proyecto CockRoach, he tenido la oportunidad de profundizar en las técnicas y herramientas necesarias para la creación de un software de scrapping, específicamente enfocado en la recolección de precios. Este trabajo me ha permitido no solo mejorar mis habilidades técnicas, sino también comprender de manera más completa las implicaciones y el potencial de este tipo de herramientas en diversos sectores.

Una de las principales conclusiones es el poder y la versatilidad del scrapping como herramienta para extraer datos valiosos en tiempo real. En un entorno donde los precios cambian constantemente, especialmente en el comercio electrónico, tener acceso a esta información de manera automatizada se convierte en una ventaja competitiva fundamental para las empresas. El desarrollo de CockRoach me permitió confirmar la importancia de los datos en la toma de decisiones comerciales y cómo la precisión y actualización oportuna de la información son clave para un proceso eficiente.

Durante el proceso, me enfrenté a varios desafíos técnicos, como la gestión de restricciones impuestas por algunas páginas web, la detección y manejo de bloqueos de IP, así como la necesidad de garantizar que el software pudiera mantenerse escalable y eficiente al procesar

grandes volúmenes de datos. Sin embargo, estos retos también me ofrecieron una valiosa experiencia para explorar soluciones alternativas como el uso de la red Tor, el manejo eficiente del contenido mediante los 'Datos Estructurados', y la optimización de los algoritmos para que el proceso de extracción fuera más rápido y preciso.

Utilizar las herramientas, técnicas y conocimientos adquiridos en el Máster han sido fundamentales para conectar CockRoach con los LLMs, trabajar con ellos y procesar su respuesta. En un mundo tan cambiante como es el de la Intelligencia Artificial crear una solución preparada para conectarse, trabajar e interactual con cualquier LLM del mercado permitirá larga vida a la solución y adaptación a las mejoras que se vayan produciendo en este sector.

Otro gran salto cualitativo, ha sido el uso de LowCode2, solución desarrollada personalmente durante años en mi tiempo libre que me ha permitido crear un frontend en Django en un tiempo récord con una riqueza, calidad y personalización que nunca imaginé cuando empecé a pensar en ella.

Por otro lado, una reflexión importante que emerge de este trabajo es la ética del scraping. Aunque es una técnica poderosa, su uso responsable es crucial para respetar las políticas de las páginas web y las normativas de protección de datos. Como desarrollador, he comprendido la necesidad de encontrar un equilibrio entre la explotación de estos datos y el respeto a la legalidad, lo que llevó a implementar medidas que respeten los términos de uso de las fuentes de información.

En cuanto al futuro del proyecto, creo que CockRoach tiene el potencial de evolucionar para ser utilizado en múltiples contextos, no solo para la recolección de precios, sino también para cualquier tipo de información crítica que esté disponible públicamente en internet como por ejemplo el SEO. La modularidad del software abre las puertas a nuevas funcionalidades, como la predicción de tendencias de precios mediante la incorporación de inteligencia artificial, o la personalización de los resultados según las preferencias del usuario.

Todos los conocimientos aportados tanto en el máster cómo en la realización de este proyecto me han elevado mis habilidades y mi experiencia para afrontar retos mucho más complejos y observar el potencial profesional todavía pendiente de explotar que reside en mí. Además, me ha ayudado a tejer una red de amistad más grande con mis compañeros de máster que también han aportado su granito de arena y unas relaciones profesionales con mis compañeros de trabajo mucho más empáticas y cercanas gracias a las dificultades y retos hallados que hemos superado juntos con ganas, pasión y sobre todo con buen ambiente y alegría.

En resumen, este trabajo me ha permitido consolidar mis conocimientos en scrapping, desarrollo de grandes plataformas de software, lenguaje Python, el uso de LLMs, … al mismo tiempo que me ha enseñado a valorar la importancia de los datos en el entorno empresarial moderno. El desarrollo de CockRoach no solo ha sido una experiencia técnica enriquecedora, sino también una reflexión sobre la responsabilidad y el impacto que este tipo de herramientas puede tener en los distintos sectores.

# Capítulo 7. FUTURAS LÍNEAS DE TRABAJO

El proyecto tiene infinitas líneas de mejora, de aportaciones y ampliaciones. El abanico es infinito y aquí únicamente se indican algunas de ellas. Pero ante un mundo tan cambiante y competitivo, las necesidades cambian continuamente y se está en constante movimiento.

Las líneas de trabajo se dividen en distintas categorías en base a su naturaleza:

## 7.1 Mejoras en Web - FrontEnd.

- Actualización del frontend por uno nuevo visualmente más actual con colores más friendly, librerías Javascript más modernas.
- Rediseño de algunos formularios para customizarlos a los datos solicitados.
- Multiidioma.
- Módulo para compartir datos entre varias soluciones.
- Dashboards más complejos y completos.
- Posibilidad de customizar Dashboards a las necesidades del usuario
- Inclusión de las llamadas síncronas al servidor para la realización de gráficos más potentes, actualización de notificaciones en tiempo real, …
- Sustituir motor de Base de Datos Mysql por un motor más potente y que permita grandes volúmenes de procesamiento de información.
- Conectar la solución a un servidor documental para almacenar todos los estudios y datos generados para su consulta.
- Conexión y creación de APIs.

## 7.2 Mejoras en IA.

- Inclusión de nuevos modelos de LLMs.
- Facilidad/Simplificación para implantar nuevos LLMs.
- Comunicación mucho más amplia entre el módulo Scrapping/IA con el FrontEnd
- Mejora del contexto enviado al LLM para solicitar la generación de código de Python.
- Innovaciones de explicabilidad del resultado para poder precisar mejor el contexto enviado al LLM.

- Creación de un módulo de LangChain para poder realizar consultas SQL, gráficos e informes.
- Ampliación del proyecto para recuperar más información acerca de los productos cómo puede ser la demanda, número de visitas / interés, stocks, …
- Utilizar los modelos LLMs para recuperar datos macroeconómicos para detectar sus efectos en la demanda, poder comprenderlos, analizarlos y realizar proyecciones.

## 7.3 Innovaciones.

- Tecnologías de DataLake y tratamiento de grandes volúmenes de información.
- Más Uso de tecnologías de alto rendimiento para mejorar en velocidad de respuesta.
- Scripts para la creación automática de máquinas que permitan aumentar/reducir el parquet y así adaptarse a la demanda.
- Almacenamiento y tratamiento del SEO recuperado del Scrapping en los buscadores de Internet.

# Capítulo 8. ANEXOS

CockRoach-Xavi Rambla_Anexo_Resultados.docx Estudio de los resultados de distintos LLM ante el mismo contexto.

CockRoach-Xavi Rambla_Anexo_Instalación FrontEnd Instalación del frontend en una máquina con Windows.

CockRoach-Xavi Rambla_Anexo_Instalación Aplicación y LLMs Instalación del frontend en una máquina con Windows.

# Capítulo 9. REFERENCIAS

En este apartado figurará el conjunto de libros, revistas u otros textos que el autor considere de interés para justificar las soluciones adoptadas en el Proyecto.

**Mamta, Padam Saini, Vaibhab Bansal, Manish Kr. Mishra. 2021**. House Price Prediction using Machine Learning and Web Scrapping

https://www.academia.edu/50061366/House_Price_Prediction_using_Machine_Learning_and_Web_Scrapping?sm=b

**Nur Indah Riwajanti 2023.** Shares Price Forecasting Using Simple Moving Average Method and Web Scrapping

https://www.academia.edu/115197989/Shares_Price_Forecasting_Using_Simple_Moving_Average_Method_and_Web_Scrapping?sm=b

**Pradeep Chintagunta** 2014, Price Transparency and Retail Prices

https://www.academia.edu/83258928/Price_Transparency_and_Retail_Prices?sm=b

**Carlos Fenollosa**, Optimización de Precios en Ecommerce on Inteligencia artificial
https://www.youtube.com/watch?v=fwkvwPjAS3Y

**Itnig** CasoTradeInn: e-commerce de a 430M
https://www.youtube.com/watch?v=8BVW9ie354M

**ITnig** Deporvillage https://www.youtube.com/watch?v=ZGYtbHcVneI

Análisis de precios con netRivals https://www.youtube.com/watch?v=oZ8kx6ONiw0

---

i www.theguardian.com/money/2013/may/14/motorway-service-stations-fuel-prices

ii **Anil Kaul , Dick R. Wittink 1995** . Empirical Generalizations About the Impact of Advertising on Price Sensitivity and Price

https://ideas.repec.org/a/inm/ormksc/v14y1995i3_supplementpg151-g160.html

iii **Praveen Kopalle a, Dipayan Biswas b 1, Pradeep K. Chintagunta c 2, Jia Fan d 3, Koen Pauwels a 4, Brian T. Ratchford e 5, James A. Sills f 6 2009**.Retailer Pricing and Competitive Effects

DOI:10.1016/j.jretai.2008.11.005

https://www.sciencedirect.com/science/article/abs/pii/S0022435908000870?via%3Dihub

iv

For 44 to 66% of consumers, price is a very or extremely significant factor in determining where to shop, depending on the retail segment.
retail-week.com/tech/strategic-report-five-winning-pricing-strategies-for-2021/7036479.article?authent=1