



Universidad Abierta Interamericana

Universidad Abierta Interamericana

Facultad de Tecnología Informática

Inteligencia Artificial Explicable y Procesamiento de Lenguaje Natural en sistemas de recomendación de contenidos para streaming en Argentina

Trabajo final de carrera presentado obtener el título de Lic. en Gestión de Tecnología
Informática

Alumno: Luis Ariel Pollio

Profesora: Dra. Marcela Samela

Julio 2025

Resumen

Las empresas que operan plataformas de distribución de contenidos en Argentina enfrentan el desafío de anticiparse a las preferencias de sus usuarios. Para lograrlo deben ofrecer recomendaciones personalizadas y relevantes. Aunque existen soluciones basadas en motores de recomendación, su implementación suele ser costosa, poco efectiva o limitada en su capacidad de adaptación en tiempo real. Este trabajo propone el diseño de un sistema de recomendación basado en algoritmos de inteligencia artificial que pueda procesar dinámicamente patrones de consumo individuales y colectivos, generando recomendaciones ajustadas al perfil del usuario. Como diferencial, se incorpora el uso de técnicas de Procesamiento de Lenguaje Natural (PLN) para interpretar la metadata existente con el fin de optimizar las sugerencias en función de su significado y de Inteligencia Artificial Explicable (XAI) para ofrecer recomendaciones comprensibles y justificadas. La arquitectura propuesta busca mejorar la experiencia del usuario, reducir los tiempos de búsqueda, simplificar la navegación y aumentar la confianza en el sistema, contribuyendo así a una menor tasa de abandono, mayor consumo de contenidos y mejor satisfacción general.

Palabras clave: inteligencia artificial explicable, patrones de consumo, perfilado de usuario, procesamiento de lenguaje natural, recomendaciones personalizadas, sistemas de recomendación

Abstract

Companies operating content distribution platforms in Argentina face the challenge of anticipating user preferences. To achieve this, they must provide personalized and relevant recommendations. While recommendation engines already exist, their implementation is often costly, ineffective, or limited in their ability to adapt in real time. This work proposes the design of a recommendation system based on artificial intelligence algorithms capable of dynamically processing both individual and collective consumption patterns, generating suggestions tailored to each user profile. As a key differentiator, Natural Language Processing (NLP) techniques are employed to interpret existing content metadata, enhancing recommendations based on semantic meaning. In addition, Explainable Artificial Intelligence (XAI) is integrated to deliver transparent and justifiable recommendations. The proposed architecture aims to improve user experience, reduce search time, simplify navigation, and increase trust in the system, ultimately leading to lower churn, greater content consumption, and higher overall user satisfaction.

Keywords: consumption patterns, explainable artificial intelligence, natural language processing, personalized recommendations, recommender systems, user profiling

Dedicatoria

A mi esposa Maira, mi gran compañera de vida que sin su acompañamiento y apoyo todo hubiese sido mucho más difícil. A mi hija Felicitas, quien esperó cada resultado de mis parciales y finales con todo su amor. Para mis padres, quienes supieron ver desde mi infancia lo que me apasionaba. A mis compañeros y docentes de los cuales aprendo cada día. Gracias a todos, son parte importante de este trabajo.

Índice general

Resumen.....	1
Abstract.....	2
Dedicatoria.....	3
Índice general.....	4
Índice de tablas	8
Índice de figuras.....	9
1. Introducción.....	11
1.1 Planteamiento del problema.....	11
1.1 Contexto y justificación	12
1.2 Marco institucional.....	12
1.3 Objetivo general y específicos	14
2. Estado del arte	15
2.1 Evolución, tendencias y contexto argentino.....	15
2.2 Desafíos del descubrimiento de contenido y la importancia de la personalización.....	19
2.3 Avances tecnológicos en la recomendación. Personalización y explicabilidad.....	22
2.3.1 Integración con modelos de lenguaje	22
2.3.2 Mejora de la precisión aplicando IA Explicada.....	23
2.3.3 Explicabilidad aplicada en Spotify	25
2.3.4 Netflix y el uso de PLN e IA generativa en sus sistemas de recomendación.....	26
2.3 Conclusiones	27
3. Marco teórico.....	28

3.1	Fundamentos de la Inteligencia Artificial en sistemas de recomendación	28
3.2	Contexto argentino	31
3.3	Prueba de Turing	32
3.4	Machine learning.....	33
3.4.1	Conceptos fundamentales	33
3.4.1	Deep Learning	34
3.5	Motor de recomendaciones	38
3.5.1	Contexto y fundamentos conceptuales	38
3.5.2	Fases de un motor de recomendaciones	39
3.5.3	Modelos colaborativos, basados en contenido e híbridos.....	39
3.6	Ética y seguridad de la información.....	42
3.6.1	Principios éticos en la inteligencia artificial	42
3.6.2	Buenas prácticas	43
3.7	Inteligencia Artificial Explicada	45
3.7.1	Conceptos de IA explicable	45
3.7.2	Comparación entre IA y la XAI	46
3.7.3	Métodos de explicabilidad.....	46
3.8	Procesamiento de lenguaje natural.....	47
3.8.1	Fundamentos del procesamiento de lenguaje natural	47
3.8.2	Aplicaciones específicas del procesamiento de lenguaje natural	49
3.8.3	Aplicación del PLN e IA en sistemas de recomendación.....	51

4.	Propuesta a implementar	52
4.1	Recursos tecnológicos.....	53
4.1.1	Lenguaje de programación	53
4.1.2	Infraestructura en la nube	53
4.1.3	Almacenamiento y bases de datos	55
4.1.4	Procesamiento en tiempo real.....	56
4.1.5	Monitoreo y visualización	58
4.1.6	Arquitectura de microservicios e integración mediante APIs	59
4.2	Recursos metodológicos.....	61
4.2.1	Inteligencia Artificial Explicable XAI	61
4.2.2	Perfilado de usuario	64
4.2.3	Integración metodológica con la arquitectura del sistema.....	66
4.3	Validación de resultados	67
5.	Propuesta de cronograma y recursos para la implementación.....	72
5.1	Consideraciones generales	72
5.2	Recursos requeridos	72
5.2.1.	Recursos humanos	72
5.2.2.	Recursos técnicos	74
5.3	Cronograma de implementación	77
5.4	Propuesta económica de la solución	79
5.4.1.	Costos de capital humano	80

5.4.2. Costos técnicos	81
5.4.3. Evaluación integral de costos y modelo comercial	83
5.5 Analisis de riesgos	84
5.5.1. Factores externos condicionantes	84
5.5.2. Factores internos condicionantes	85
6. Conclusiones.....	87
7. Trabajos futuros.....	89
7.1 Interacción conversacional basada en lenguaje natural	89
7.2 Privacidad diferencial y preservación de datos personales en sistemas de recomendación	90
8. Acrónimos	91
9. Referencias	93

Índice de tablas

Tabla 1 <i>Detalle de tipos de datos a almacenar</i>	55
Tabla 2 <i>Pipeline de procesamiento en tiempo real para recomendaciones personalizadas</i> ..	58
Tabla 3 <i>Indicadores para auditoría de recomendaciones</i>	64
Tabla 4 <i>Ejemplo de perfilado de usuario</i>	65
Tabla 5 <i>Capacidades técnicas estimadas según recurso</i>	75
Tabla 6 <i>Costos del personal del proyecto</i>	81
Tabla 7 <i>Costos técnicos estimados de infraestructura</i>	83
Tabla 8 <i>Resumen total del costo del proyecto</i>	84

Índice de figuras

Figura 1 <i>Uso de dispositivos en Argentina</i>	16
Figura 2 <i>Distribución velocidades internet en hogares argentinos.</i>	17
Figura 3 <i>Tecnologías de acceso a internet fijo</i>	18
Figura 4 <i>Promedio de tiempo para encontrar contenido</i>	20
Figura 5 <i>Estructura de InteRecAgent</i>	23
Figura 6 <i>Índice de Preparación para la Inteligencia Artificial 2024</i>	31
Figura 7 <i>Esquema conceptual del test de Turing</i>	32
Figura 8 <i>Áreas de la Inteligencia artificial</i>	34
Figura 9 <i>Aprendizaje en base a la experiencia</i>	35
Figura 10 <i>Redes Neuronales Convolucionales (CNN)</i>	37
Figura 11 <i>Redes Neuronales Recurrentes (RNN)</i>	37
Figura 12 <i>Fases de un motor de recomendaciones</i>	39
Figura 13 <i>Sistema de Recomendación basado en Filtrado de Contenido</i>	41
Figura 14 <i>Países adherentes a los Principios de IA de la OECD</i>	44
Figura 15 <i>Concepto IA Explicada</i>	45
Figura 16 <i>Disciplinas dentro de la inteligencia artificial</i>	48
Figura 17 <i>Comunicación de Microservicios y Plataformas mediante APIs RESTful</i>	60
Figura 18 <i>Recomendaciones con IA Explicada</i>	63
Figura 19 <i>Integración metodológica con la arquitectura del sistema</i>	67

Figura 20 <i>Interfaz de usuario para encuestas</i>	70
Figura 21 <i>Cronograma tentativo</i>	79
Figura 22 <i>Simulación de una interfaz conversacional</i>	89

1. Introducción

En este capítulo se presentan los fundamentos que motivaron el desarrollo del presente trabajo. Se introducen conceptos clave relacionados con la problemática actual en torno a los sistemas de recomendación de contenidos digitales en Argentina y se explicitan los objetivos que guían la investigación y propuesta de solución.

1.1 Planteamiento del problema

En la era digital actual, las plataformas de streaming han transformado el acceso y consumo de contenido audiovisual, consolidándose como una de las principales formas de entretenimiento. En Argentina, este fenómeno es particularmente notable, 60% de los argentinos mira streaming de video, según el informe *Inside Video 2024 Argentina* de Kantar IBOPE (2024). Sin embargo, estas plataformas enfrentan el desafío de mantener a los usuarios comprometidos y satisfechos en un entorno cada vez más competitivo. Un elemento central para lograrlo es la capacidad de ofrecer recomendaciones precisas, personalizadas y adaptadas dinámicamente a los gustos y comportamientos individuales.

Aunque existen motores de recomendación, muchos de ellos presentan dificultades en su implementación, son costosos y no siempre cumplen con las expectativas en términos de precisión y eficiencia. Estos sistemas a menudo fallan en adaptarse a los patrones de consumo específicos y cambiantes de los usuarios argentinos, lo que puede llevar a una experiencia de usuario no personalizada, marcada por largos tiempos de búsqueda y una navegación ineficaz por el catálogo de contenido. En promedio, una persona tarda diez minutos y medio buscando contenido (Nielsen, 2023). La falta de recomendaciones relevantes no solo frustra a los usuarios, sino que también puede resultar en una menor retención de suscriptores y una disminución en el uso de la plataforma.

En este contexto, la ausencia de recomendaciones relevantes impacta negativamente en la retención de usuarios y en el aprovechamiento del catálogo disponible. Por ello, se vuelve necesario explorar nuevas estrategias que combinen tecnologías avanzadas de inteligencia

artificial con enfoques centrados en el usuario, para construir sistemas más inteligentes, explicables y sensibles a los hábitos de consumo actuales.

1.1 Contexto y justificación

El consumo de contenido digital ha crecido de manera sostenida en todo el mundo, y Argentina no ha sido la excepción. Las plataformas de streaming se consolidaron como una de las principales fuentes de entretenimiento, ofreciendo a los usuarios acceso inmediato a un catálogo cada vez más amplio de series, películas y transmisiones en vivo. Según el informe de Simon-Kucher (2024), el mercado global continúa expandiéndose, aunque a un ritmo más moderado que en años anteriores.

En este escenario, uno de los desafíos centrales es lograr que cada usuario acceda a contenido que realmente se adapte a sus intereses y hábitos. Según Accenture (2022) el 54% de los usuarios considera que el uso de servicios de búsqueda alternativos podría mejorar significativamente su experiencia, lo que refleja una insatisfacción con las recomendaciones actuales.

Los motores de recomendación cumplen un rol clave, ya que influyen directamente en la forma en que el usuario interactúa con el catálogo y percibe la oferta disponible. Frente a esto, se considera fundamental analizar nuevas estrategias que permitan mejorar la precisión, la relevancia y la comprensión de las recomendaciones ofrecidas.

1.2 Marco institucional

La solución propuesta está orientada a plataformas de streaming de video que operan en el mercado argentino y que buscan integrar un sistema inteligente de recomendaciones personalizadas en sus interfaces. El objetivo es que las sugerencias generadas por el motor puedan ser consumidas directamente por los usuarios finales a través de los distintos dispositivos compatibles. Se considera como usuario a cada perfil individual dentro de una cuenta, entendiendo que el modelo de uso más común en los hogares contempla una única

cuenta con múltiples perfiles asociados, lo que permite personalizar la experiencia según cada integrante del hogar. Para garantizar una experiencia completa, se prevé que el sistema pueda integrarse en dispositivos compatibles con aplicaciones de streaming, como teléfonos móviles, tablets, notebooks y TV boxes con sistema Android TV.

Cabe destacar que este trabajo no contempla el desarrollo de software funcional, sino que se enfoca en el diseño conceptual y técnico de un motor de recomendación basado en algoritmos de inteligencia artificial, que integra técnicas de PLN para analizar tanto la información de consumo como la metadata de los contenidos, y mecanismos de Inteligencia Artificial Explicable XAI para ofrecer recomendaciones comprensibles y justificadas, con potencial de implementación en entornos reales.

1.3 Objetivo general y específicos

Objetivo general

Diseñar un sistema inteligente de recomendación de contenidos basado en técnicas de XAI y PLN para integrar a una plataforma de streaming con el fin de maximizar la experiencia del usuario.

Objetivos específicos

- Analizar los patrones de consumo de los usuarios de una plataforma de streaming
- Aplicar técnicas de procesamiento de lenguaje natural para enriquecer la metadata de los contenidos
- Implementar algoritmos híbridos explicables de recomendación de contenidos
- Diseñar una arquitectura escalable que se integre con microservicios
- Aplicar técnicas de XAI que permitan interpretar y justificar las recomendaciones generadas
- Evaluar el desempeño de los algoritmos de recomendación

2. Estado del arte

2.1 Evolución, tendencias y contexto argentino

En la última década, las plataformas de streaming han redefinido de manera profunda el consumo de contenido audiovisual a nivel global. Esta evolución ha estado impulsada por el avance tecnológico, la masificación de dispositivos conectados y la demanda creciente por acceder a contenidos de manera flexible, personalizada y bajo demanda. El mercado global de streaming supera actualmente los 100.000 millones de dólares anuales en ingresos y alcanza una base de usuarios cercana a los 1.800 millones. Se estima que esta cifra podría elevarse a 2.000 millones en 2029, con un crecimiento especialmente acelerado en Asia-Pacífico, mientras que mercados como el estadounidense comienzan a mostrar signos de saturación. A su vez, se proyecta que los ingresos por suscripciones alcancen los 190.000 millones para ese año, impulsados por la expansión internacional, la diversificación de modelos de negocio y la adopción de esquemas con publicidad (Broadcast PRO, 2024).

En Argentina, el consumo de contenidos por streaming continúa en aumento, consolidándose como un hábito transversal a distintos segmentos etarios. Según el informe *Inside Video Argentina 2024*, el 60 % de los argentinos consume video vía streaming, ya sea a través de servicios gratuitos o pagos. En promedio, dedican 5 horas y 43 minutos semanales a esta actividad, lo que evidencia el rol central que estas plataformas ocupan en las rutinas de entretenimiento (Kantar IBOPE, 2024).

El perfil de audiencia muestra una distribución diversa: mientras el 23 % corresponde al segmento de 25 a 34 años, el 24 % pertenece a personas mayores de 55 años, lo que revela que estas plataformas logran captar tanto a jóvenes adultos como a públicos mayores. (Kantar IBOPE, 2024). Estos datos refuerzan la necesidad de estrategias de personalización que puedan adaptarse a la heterogeneidad etaria y de consumo de los usuarios argentinos.

Uno de los aspectos más significativos del informe es el análisis de la penetración de las suscripciones a plataformas de streaming pagas. Según Kantar IBOPE (2024), el 46% de los argentinos cuenta con al menos una suscripción a un servicio de streaming pago. Además, destaca que los hogares argentinos tienen, en promedio, dos suscripciones a plataformas de streaming pagas. Este comportamiento sugiere que los usuarios no se limitan a una única

opción, sino que diversifican sus elecciones para acceder a un catálogo de contenidos más amplio y variado. Los principales factores que influyen en la decisión de suscribirse a una plataforma de streaming en Argentina son el precio y la amplitud del catálogo disponible. Esto implica que las plataformas que ofrezcan una propuesta de valor atractiva en términos de costos y una oferta de contenidos diversa y actualizada tendrán mayores posibilidades de captar y retener a los usuarios argentinos (Kantar IBOPE, 2024).

Kantar IBOPE también analiza el uso de dispositivos para el consumo. Como se observa en la Figura 1, la mayoría de los argentinos utilizan múltiples pantallas para acceder a contenidos audiovisuales. En concreto, el 91% de los encuestados afirma utilizar dos pantallas, mientras que el 79% utiliza tres pantallas y el 33% utiliza cuatro pantallas. (Kantar IBOPE, 2024). Este comportamiento multiplataforma refleja la flexibilidad y adaptabilidad de los usuarios argentinos a la hora de consumir contenidos de video, aprovechando la diversidad de dispositivos disponibles en el hogar y en movilidad. Estos datos tienen implicaciones importantes para el diseño de sistemas de recomendación, que deben ser capaces de ofrecer una experiencia de usuario coherente y personalizada en diferentes dispositivos.

Figura 1
Uso de dispositivos en Argentina



Nota: Adaptado de *Inside Video 2024*, Kantar IBOPE, 2023.

<https://kantariibopemedia.com/argentina/wp-content/uploads/sites/3/2024/04/Inside-Video-Argentina-2024.pdf>. Todos los Derechos Reservados.

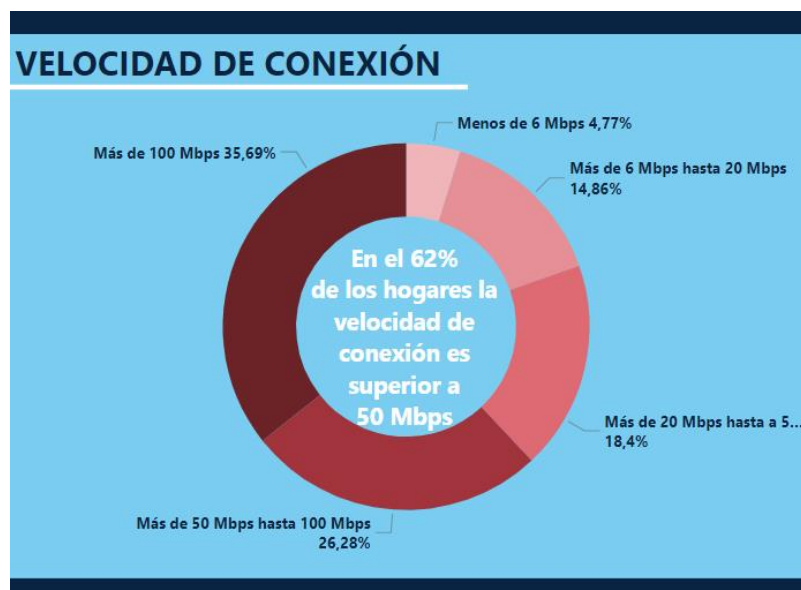
En paralelo, la infraestructura de conectividad en Argentina ha mejorado considerablemente. Según datos recientes de la Cámara Argentina de Internet, cuatro de cada diez hogares argentinos poseen conexiones que superan los 100 Mbps, duplicando el valor del año anterior real (CABASE, 2024). Este crecimiento en la velocidad de conexión no solo mejora la calidad de transmisión de contenidos audiovisuales, sino que también habilita nuevas

posibilidades para el funcionamiento eficiente de sistemas de recomendación basados en inteligencia artificial, los cuales requieren un procesamiento de datos dinámico y en tiempo.

En la Figura 2, se muestra cómo está compuesto el parque de velocidades en Argentina.

Figura 2

Distribución velocidades internet en hogares argentinos.



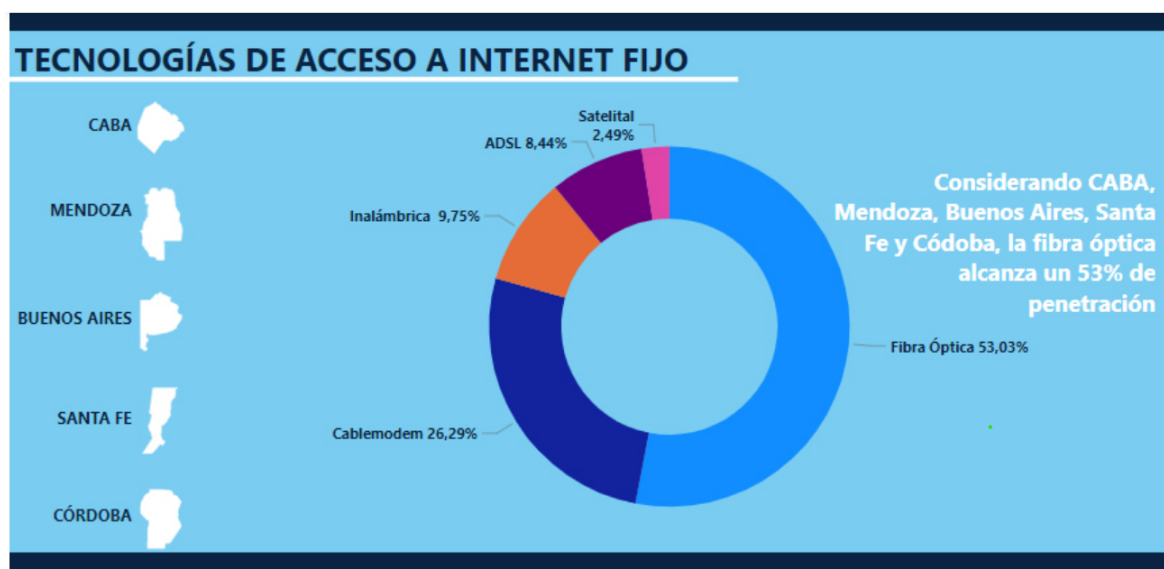
Nota: Adaptado de *Casi 4 de cada 10 conexiones a internet en hogares argentinos supera los 100 Mbps*, CABASE, 2024. <https://www.cabase.org.ar/casi-4-de-cada-10-conexiones-a-internet-en-hogares-argentinos-supera-los-100-mbps/>. Todos los Derechos Reservados.

Además del aumento en las velocidades de descarga, el informe de CABASE destaca un incremento en la penetración del acceso a internet a través de fibra óptica, alcanzando un 53 % del total de conexiones (CABASE, 2024). Esta tendencia no solo mejora la estabilidad del servicio, sino que también reduce los tiempos de latencia, lo cual es fundamental para ofrecer recomendaciones personalizadas de forma ágil y continua. En este sentido, el desarrollo de infraestructura digital actúa como un facilitador clave para la adopción de tecnologías emergentes en las plataformas de contenido audiovisual, permitiendo que el procesamiento masivo de datos del usuario sea más eficiente y menos costoso.

Esta mejora en la velocidad de conexión ha potenciado el acceso fluido a plataformas de streaming, permitiendo reproducir contenidos de alta calidad sin interrupciones o baja

calidad de imagen. Asimismo, la penetración del servicio de internet en los hogares argentinos alcanza el 88%, lo que representa una base sólida para la expansión de este tipo de servicios. Aunque la disponibilidad de fibra óptica es alta en muchas zonas del país, factores como el costo, la necesidad de instalaciones adicionales y la satisfacción con servicios ya contratados condicionan su adopción masiva. No obstante, más del 50% de los accesos en grandes centros urbanos ya se realizan a través de fibra óptica, lo que indica una consolidación progresiva de tecnologías que facilitan el desarrollo y consumo intensivo de contenido audiovisual por streaming (CABASE, 2024). En la Figura 3, podemos observar el resto de la composición de tipos de accesos que cuenta el país

Figura 3
Tecnologías de acceso a internet fijo



Nota: Adaptado de *Casi 4 de cada 10 conexiones a internet en hogares argentinos supera los 100 Mbps*, CABASE, 2024. <https://www.cabase.org.ar/casi-4-de-cada-10-conexiones-a-internet-en-hogares-argentinos-supera-los-100-mbps/>. Todos los Derechos Reservados.

CABASE también señala una creciente demanda de servicios digitales en el hogar, lo cual sugiere una transformación en los hábitos de consumo de los argentinos. A medida que el entretenimiento digital gana terreno frente a otros formatos tradicionales, las plataformas OTT (Over The Top) requieren fortalecer sus capacidades tecnológicas para responder a las crecientes y cada vez más exigentes expectativas del usuario moderno (CABASE, 2024). En este marco, contar con conexiones de alta velocidad se vuelve esencial para soportar motores de recomendación que dependan de modelos complejos de Machine Learning (ML) y Deep

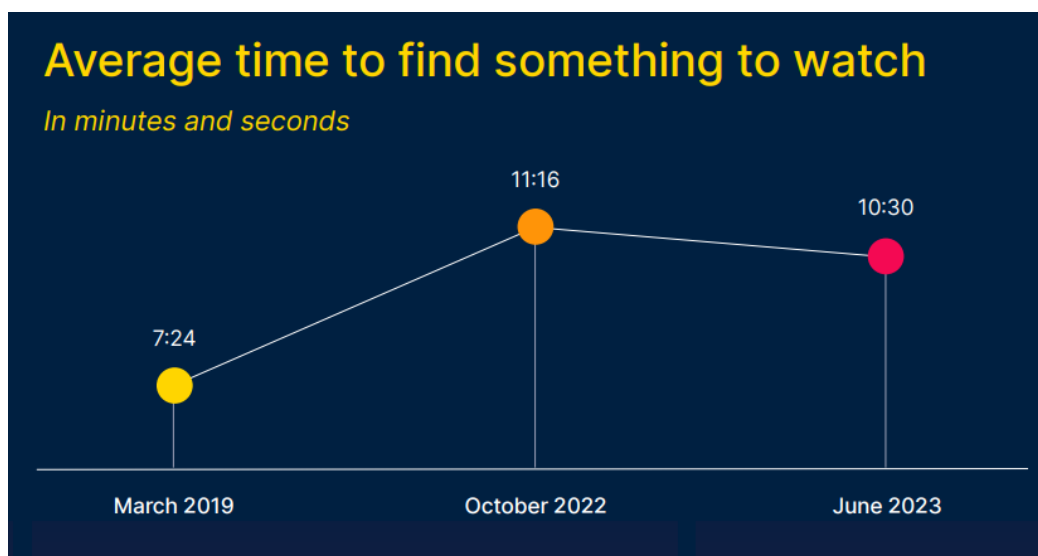
Learning (DP), reafirmando la importancia del entorno tecnológico en el desarrollo de experiencias personalizadas y efectivas.

2.2 Desafíos del descubrimiento de contenido y la importancia de la personalización

En los últimos años, el ecosistema del streaming ha vivido una expansión sin precedentes, tanto en la cantidad de plataformas disponibles como en la diversidad de contenidos que ofrecen. Según el informe *Data-driven personalization: The future of streaming content Discovery*, los usuarios ahora tienen acceso a más de 2,7 millones de títulos únicos a través de servicios de transmisión, cable y televisión abierta, lo que representa casi un millón más que hace apenas dos años (Nielsen, 2023). Esta abundancia ha fragmentado la audiencia y generado una creciente complejidad en el proceso de descubrimiento de contenido, forzando a las plataformas a optimizar sus sistemas de recomendación para facilitar la navegación en un entorno de alta competencia y sobreoferta.

Otro hallazgo relevante del estudio es el tiempo que los usuarios dedican a encontrar algo para ver. En promedio, una persona tarda 10 minutos y medio en decidir qué contenido visualizar, lo que representa un aumento significativo con respecto a los 7 minutos y medio reportados en 2019 (Nielsen, 2023). Esta evolución se puede observar en la Figura 4. Este fenómeno, impulsado por la saturación del mercado y la falta de personalización efectiva, demuestra que el diseño de experiencias de usuario centradas en la eficiencia y la relevancia es clave para sostener la atención y fidelización del público. Las plataformas que no logren resolver este desafío corren el riesgo de perder a sus usuarios frente a competidores que sí ofrezcan recomendaciones más precisas y personalizadas.

Figura 4
Promedio de tiempo para encontrar contenido



Nota: Adaptado de *Data-driven personalization: The future of streaming content discovery*, Gracenote, 2023.
<https://www.nielsen.com/es/insights/2023/data-driven-personalization-2023-state-of-play-report/>. Todos los Derechos Reservados.

Además del volumen creciente de contenidos y plataformas, el informe de Nielsen (2023) destaca otro desafío fundamental: la desconexión entre la oferta disponible y la capacidad de los usuarios para descubrir títulos relevantes. A pesar de la amplia disponibilidad, el 20% de los encuestados declaró que no encontró nada que quisiera ver en un momento determinado, frustrando al usuario quien en ocasiones decide no ver nada. Este dato revela una paradoja en el consumo digital: a mayor cantidad de opciones, mayor dificultad para tomar decisiones

La personalización basada en datos se consolida como uno de los elementos estratégicos más relevantes para las plataformas. Como explica Nielsen (2023), la inclusión de metadatos enriquecidos y descripciones contextuales permite construir experiencias más ajustadas a las preferencias individuales. Esta capacidad no solo mejora la relevancia de las recomendaciones, sino que también habilita nuevas formas de segmentación publicitaria, optimizando los ingresos a partir de la afinidad entre contenidos y audiencias. Por ello, el informe concluye que el éxito en la industria del streaming dependerá, en gran medida, de la

capacidad de cada plataforma para ofrecer experiencias personalizadas y dinámicas, ajustadas al perfil de cada usuario.

El informe de Ampere Analysis titulado *Streaming algorithms now beat word-of-mouth for TV and movie picks* revela un cambio significativo en los hábitos de consumo de contenido audiovisual: a nivel global, el 26 % de los usuarios de internet confía en las recomendaciones algorítmicas de las plataformas de streaming para decidir qué ver, superando al 23 % que aún se basa en sugerencias de amigos o familiares (Ampere Analysis, 2025). Este dato indica que los sistemas de recomendación personalizados han ganado una legitimidad comparable, e incluso superior, al tradicional boca a boca, consolidándose como una herramienta clave en la toma de decisiones del usuario.

Este fenómeno es especialmente pronunciado entre los usuarios más jóvenes y aquellos con niños pequeños en el hogar. Según el estudio, el 30 % de los espectadores de entre 25 y 34 años, y el 31 % de los suscriptores con hijos pequeños, dependen de las sugerencias algorítmicas de las plataformas. Sin embargo, en el segmento de 18 a 24 años, las redes sociales tienen un papel aún más destacado: el 31 % de estos jóvenes descubre contenido de larga duración a través de plataformas sociales, superando incluso a las recomendaciones basadas en actores o celebridades preferidas (30 %) y a las sugerencias de los propios servicios de streaming (27 %) (Ampere Analysis, 2025).

A nivel geográfico, el estudio de Ampere Analysis identifica variaciones significativas en la confianza hacia las recomendaciones algorítmicas. En el Reino Unido, el 30 % de los usuarios opta por estas sugerencias, mientras que en Estados Unidos la cifra es del 27 %. En contraste, en países como Francia e Italia, solo el 19 % se inclina por las recomendaciones de las plataformas, prefiriendo en su lugar las opiniones de amigos o reseñas positivas. En mercados emergentes como China e India, la afinidad por las sugerencias algorítmicas es alta, con un 30 % de usuarios confiando en ellas (Ampere Analysis, 2025).

La experiencia del usuario se ha transformado en el eje estratégico central para las plataformas OTT. De acuerdo con Fieiras-Ceide et al. (2024), llevaron a cabo entrevistas con directivos de las principales plataformas de España, los contenidos por sí solos ya no son suficientes para captar y fidelizar a las audiencias. Ahora se requieren capas adicionales de

valor que integren la interacción con el sistema, la navegación fluida entre dispositivos y la personalización ajustada al perfil individual de cada usuario.

2.3 Avances tecnológicos en la recomendación. Personalización y explicabilidad

2.3.1 Integración con modelos de lenguaje

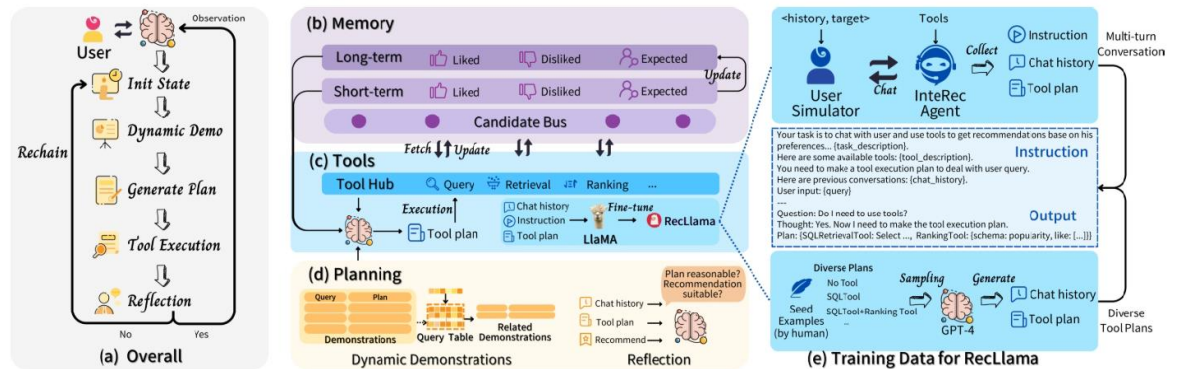
La posibilidad de ofrecer recomendaciones en tiempo real potencia la experiencia personalizada. Esto se debe a que permite una retroalimentación inmediata entre las acciones del usuario y las recomendaciones generadas por el sistema. En este sentido, Iglesias Albores (2022) resalta que “poder procesar a gran velocidad todos los datos en un tiempo muy corto es lo que posibilita que el análisis y la generación de nuevos datos sea casi en tiempo real” (p. 5), lo cual representa un diferencial competitivo para las plataformas que integran estas tecnologías.

En los últimos años, la integración de grandes modelos de lenguaje con sistemas tradicionales de recomendación ha marcado una evolución significativa en la búsqueda de interacciones más humanas y efectivas entre usuarios y plataformas digitales. Huang et al. (2023) proponen un marco denominado InteRecAgent, que combina el conocimiento específico de los modelos de recomendación tradicionales con las capacidades lingüísticas y de razonamiento de los Lenguajes de Modelos Grandes (LLM, por sus siglas en inglés). Este sistema permite interpretar consultas complejas del usuario en lenguaje natural, planificar acciones y ejecutar recomendaciones de forma conversacional y personalizada, superando las limitaciones de los enfoques clásicos basados exclusivamente en datos de comportamiento.

Uno de los aportes clave del InteRecAgent es la incorporación de mecanismos como el bus de candidatos y los perfiles de usuario de corto y largo plazo. Estos se muestran con mayor detalle en la Figura 5 y funcionan como memoria dinámica para adaptar las sugerencias en función del historial de consumo del usuario. Estos componentes permiten una continuidad coherente en las recomendaciones a lo largo de múltiples sesiones y contextos, lo cual representa un avance sustancial en la personalización sostenida en el tiempo. Además, el sistema emplea una estrategia de planificación denominada plan-first, que reduce

significativamente los costos computacionales en comparación con métodos tradicionales de interacción paso a paso, sin sacrificar precisión (Huang et al., 2023).

Figura 5
Estructura de InteRecAgent



Nota: Adaptado de *Recommender AI Agent: Integrating Large Language Models for Interactive Recommendations* (p. 4). por Huang et al., 2023.

Los experimentos realizados por los autores indican que InteRecAgent no solo ofrece mejores resultados en términos de precisión y satisfacción del usuario en escenarios conversacionales, sino que también logra adaptarse eficazmente a dominios menos representados, como el comercio electrónico de productos de belleza, donde los modelos generalistas presentan limitaciones. Esto valida la eficacia de los sistemas híbridos que combinan LLM con motores de recomendación entrenados en dominios específicos (Huang et al., 2023). El estudio refuerza la noción de que el futuro de los sistemas de recomendación reside en arquitecturas modulares, capaces de razonar, interactuar y aprender en función del contexto.

2.3.2 Mejora de la precisión aplicando IA Explicada

Govea et al. (2024) en su estudio *Transparency and precision in the age of AI: evaluation of explainability-enhanced recommendation systems*, destacan un aumento notable

en la precisión y fiabilidad de las recomendaciones cuando se integran métodos explicativos como LIME y SHAP. En particular, se observa una mejora en la precisión de las recomendaciones del sistema de hasta un 3%, lo que indica el valor añadido que la explicabilidad proporciona a los sistemas de Inteligencia Artificial (IA). Estos hallazgos resaltan la viabilidad de incorporar técnicas de explicabilidad en sistemas de recomendación en tiempo real y la posibilidad de una adopción más amplia en diversas plataformas digitales, impulsando la personalización y la confianza del usuario. El estudio aporta evidencia concreta sobre cómo la explicabilidad puede fortalecer la transparencia, la confianza y la eficacia de los sistemas de recomendación.

A partir del estudio de Govea et al. (2024) los modelos de recomendación y las explicaciones que se generaron se evaluaron utilizando métricas de rendimiento y técnicas específicas que aseguran su fiabilidad. Esta fase garantiza que las recomendaciones y sus explicaciones sean precisas, comprensibles y confiables para los usuarios finales. Se emplearon métricas estándar de la industria, como precisión, recall y F1-score, para evaluar el rendimiento de los modelos de recomendación. La precisión mide la proporción de recomendaciones relevantes entre todas las recomendaciones realizadas, mientras que el recall evalúa la proporción de recomendaciones relevantes que se identificaron correctamente. El F1-score combina la precisión y el recall en una sola métrica para ofrecer una visión equilibrada de la eficiencia del modelo. También se utilizó la métrica de exactitud de clasificación para evaluar la proporción de predicciones correctas en los casos (Govea et al., 2024).

Además del uso de métricas cuantitativas en el estudio de explicabilidad de Govea et al. (2024), se planteó un enfoque multifacético para validar la relevancia y claridad de las explicaciones generadas por las técnicas de interpretabilidad. Primero, se realizaron pruebas cualitativas, involucrando a los usuarios en la evaluación de las explicaciones en cuanto a su comprensibilidad y relevancia. Se utilizaron encuestas y entrevistas para recopilar comentarios directos sobre cómo los usuarios perciben y entienden las explicaciones proporcionadas por el sistema. Además, se llevaron a cabo sesiones de evaluación con expertos del dominio, quienes analizaron la consistencia y la precisión técnica de las explicaciones en relación con los datos y modelos subyacentes. Al emplear un enfoque integral de evaluación, este proceso fortalece la confianza en los modelos de recomendación y sus explicaciones, facilitando su aceptación y adopción por parte de los usuarios finales.

2.3.3 Explicabilidad aplicada en Spotify

Un caso reciente y relevante de aplicación de XAI en sistemas de recomendación es el de Spotify, que ha incorporado una herramienta denominada DJ Livi. Se trata de un sistema que consta de un sistema basado en inteligencia artificial que combina algoritmos de personalización con una voz generada por IA para acompañar la experiencia auditiva de los usuarios de habla hispana. Esta herramienta no solo selecciona canciones en función del historial del usuario, las preferencias de género y el momento del día, sino que también contextualiza las recomendaciones a través de una locución en tiempo real que explica por qué se eligieron ciertos temas. De este modo, se incorpora un componente de explicabilidad al proceso de recomendación, alineado con los principios de la XAI, lo que mejora la confianza y el compromiso del usuario con la plataforma (Spotify, 2024).

Este desarrollo forma parte de una estrategia más amplia de Spotify orientada a ofrecer experiencias con IA más interactivas, adaptativas y comprensibles para el usuario final. Según lo detallado en el blog oficial de la propia compañía, DJ Livi recurre a modelos de OpenAI para generar locuciones personalizadas, creando una narrativa contextual alrededor del contenido auditivo. Esta narrativa no solo guía al usuario, sino que permite comprender el razonamiento detrás de cada recomendación, reduciendo la opacidad algorítmica y promoviendo la transparencia (Spotify, 2024).

Esta funcionalidad integra algoritmos de recomendación personalizados con una voz generada por inteligencia artificial, ofreciendo así una experiencia auditiva guiada. De acuerdo con información publicada por Spotify (2023), esta herramienta emplea PLN para seleccionar pistas adaptadas al historial y contexto del usuario, al tiempo que ofrece comentarios hablados que explican la razón de cada selección. Esta dinámica no solo fortalece la transparencia del sistema de recomendación, sino que introduce un nuevo paradigma de interacción, en el cual la IA justifica sus decisiones utilizando lenguaje natural, aumentando la confianza del usuario.

La incorporación de la voz del DJ Livi para usuarios hispanohablantes refuerza el rol de la XAI en la construcción de experiencias personalizadas más cercanas y comprensibles. Spotify (2024) destaca en su blog que Livi no solo guía al usuario a través de su contenido preferido, sino que contextualiza las recomendaciones explicando por qué ciertas canciones o

artistas fueron seleccionados, considerando elementos como la hora del día, el estado de ánimo o hábitos recientes. Esta capacidad de narrar el criterio de recomendación demuestra cómo la explicabilidad puede integrarse de manera efectiva en experiencias cotidianas, potenciando el engagement mediante interacciones más humanas, transparentes y adaptativas.

2.3.4 Netflix y el uso de PLN e IA generativa en sus sistemas de recomendación

Netflix se consolida como líder en personalización dentro del ecosistema de streaming gracias a su enfoque avanzado en sistemas de recomendación, los cuales integran grandes volúmenes de datos como el historial de visualización, calificaciones y hábitos de consumo. Según el informe técnico publicado en el blog oficial de la compañía, la nueva estrategia se basa en un *Foundation Model* que permite centralizar representaciones semánticas tanto de los usuarios como del contenido audiovisual, lo cual habilita una personalización más coherente y precisa. Este enfoque, inspirado en los modelos de lenguaje a gran escala, también permite el ajuste fino para tareas específicas dentro de su arquitectura modular (Hsiao & Lamkhede, 2025).

En línea con esta evolución tecnológica, la compañía también ha comenzado a experimentar con un asistente conversacional basado en inteligencia artificial desarrollada por OpenAI. Este sistema, actualmente en etapa de pruebas en Australia y Nueva Zelanda, busca transformar la experiencia de búsqueda mediante interacciones en lenguaje natural. En lugar de navegar por menús tradicionales, los usuarios pueden expresarse con frases como “quiero ver una comedia liviana para después de cenar”, y recibir recomendaciones ajustadas no solo a su historial, sino también a su contexto emocional y temporal (Infobae, 2025).

Este avance constituye un paso decisivo hacia motores de recomendación más conversacionales, donde el PLN y la IA generativa desempeñan un rol central. Tal como explica Castillo (2025), este asistente es capaz de interpretar estados de ánimo, referencias culturales y preferencias expresadas informalmente, fortaleciendo la experiencia del usuario. Netflix también informó que los datos recolectados durante esta fase permitirán ajustar el modelo antes de su lanzamiento global. Estos desarrollos muestran una convergencia entre la XAI, el PLN y los modelos de lenguaje, reafirmando el rumbo que la industria del streaming

está tomando: avanzar hacia plataformas capaces de entender, contextualizar y justificar sus recomendaciones en un entorno de interacción más natural y adaptativo.

2.3 Conclusiones

El análisis desarrollado en este capítulo evidencia que la personalización de contenido, impulsada por el uso de inteligencia artificial, se ha consolidado como un eje central en la evolución de las plataformas de streaming. A nivel global, la expansión del mercado, la diversidad de audiencias y la sobreabundancia de contenidos han generado nuevas demandas en torno a la relevancia, la accesibilidad y la experiencia del usuario. En el caso argentino, el crecimiento sostenido del consumo digital, junto con la mejora en la conectividad y la multiplicación de dispositivos, refuerzan la necesidad de recomendaciones más precisas, ágiles y adaptadas al contexto local.

El desarrollo tecnológico ha permitido, además, la incorporación de enfoques avanzados como los LLM, el PLN y la XAI, que no solo incrementan la precisión de las sugerencias, sino que también mejoran su transparencia y comprensibilidad. Casos como DJ Livi, de Spotify, y el asistente conversacional de Netflix ilustran esta transición hacia motores de recomendación más humanos, interactivos y capaces de justificar sus decisiones en lenguaje natural.

Este nuevo paradigma no solo redefine el vínculo entre el usuario y el sistema, sino que también plantea desafíos técnicos y éticos relacionados con la explicabilidad, la confianza y la diversidad del contenido ofrecido. En este contexto, la integración de tecnologías que combinen personalización, lenguaje natural e interpretabilidad se perfila como la ruta estratégica para responder a las expectativas de una audiencia cada vez más exigente y heterogénea. Esta convergencia será clave en los próximos años para sostener la competitividad, maximizar el engagement y garantizar una experiencia audiovisual significativa.

3. Marco teórico

Este capítulo tiene como objetivo establecer los fundamentos teóricos necesarios para comprender el funcionamiento y la evolución de los sistemas de recomendación impulsados por IA. Se abordan los conceptos generales de ML, DL, PLN y XAI, así como la forma en que estas tecnologías se integran en el diseño de motores de recomendación. Dado que el diferencial de la propuesta radica en la incorporación de PLN para interpretar la metadata de los contenidos, y de XAI para justificar las recomendaciones ofrecidas, se profundiza en los aportes de ambas tecnologías al diseño de experiencias personalizadas, transparentes y comprensibles.

También se presenta un repaso de los antecedentes históricos de la IA en el ámbito del entretenimiento, con especial énfasis en el contexto argentino. Comprender estos conceptos resulta fundamental para contextualizar el problema abordado en este trabajo, identificar las brechas existentes en la aplicación de estas tecnologías en el país y justificar la necesidad de desarrollar soluciones innovadoras, adaptadas a las particularidades culturales, técnicas y de consumo del público local.

3.1 Fundamentos de la Inteligencia Artificial en sistemas de recomendación

La IA surge como una disciplina científica orientada no solo a comprender los procesos del pensamiento humano, sino también a replicarlos mediante la construcción de entidades capaces de razonar, percibir, aprender y actuar de forma autónoma en entornos complejos. Este campo, relativamente joven dentro de las ciencias, tuvo su origen en la década de 1950 y se ha consolidado como una de las áreas más prometedoras de la investigación interdisciplinaria. A diferencia de otras disciplinas tradicionales, en las que muchas teorías fundamentales ya han sido formuladas, la IA representa un terreno fértil con múltiples desafíos por resolver, lo que la convierte en un ámbito atractivo para científicos de diversas áreas. En la actualidad, abarca una amplia variedad de subcampos, desde áreas generales hasta aplicaciones específicas como los sistemas de diagnóstico médico, la generación automática de textos, la resolución de problemas matemáticos y la toma de decisiones en juegos estratégicos. Al automatizar procesos intelectuales, la IA se proyecta como una disciplina transversal, con potencial de impacto en

prácticamente todas las dimensiones de la actividad cognitiva humana (Russell & Norvig, 2004).

En su publicación, la UNESCO (2021) sostiene que “En la actualidad, estamos viviendo un renacimiento de la IA, con una gama cada vez mayor de sectores que adoptan el tipo de IA conocido como aprendizaje automático (AA), lo que implica que el sistema de IA analice enormes cantidades de datos” (pág. 8).

Uno de los objetivos principales de la IA es permitir que las máquinas aprendan de la experiencia, se adapten a nueva información y ejecuten tareas de manera autónoma. Según IBM (2021), existen dos grandes tipos: la IA débil y la IA fuerte.

La IA débil se refiere a sistemas diseñados para ejecutar tareas específicas y limitadas. No exhiben inteligencia comparable a la humana ni poseen la capacidad de aprender o adaptarse por sí mismos, aunque pueden ser altamente efectivos en su función asignada. Este tipo de IA está presente en muchas aplicaciones cotidianas. Por ejemplo, en los motores de recomendación, analiza patrones de consumo, tanto generales como individuales, para ofrecer resultados personalizados. También está en asistentes virtuales como Siri, que responden preguntas o ejecutan acciones como configurar alarmas o reproducir música. Otros casos incluyen el reconocimiento de voz, que convierte patrones sonoros en texto sin comprender su significado, o los filtros de spam, que identifican mensajes no deseados sin analizar su contenido semántico (IBM, 2021).

En contraste, la IA fuerte es un concepto aún hipotético que plantea la posibilidad de desarrollar sistemas con capacidades cognitivas similares a las humanas. Estos podrían razonar, aprender y resolver problemas en múltiples contextos, de forma autónoma. Sin embargo, esta forma de IA sigue siendo un objetivo teórico y lejano (IBM, 2021).

Russell y Norvig (2004) definen la IA como “el estudio de los agentes que reciben percepciones del entorno y llevan a cabo las acciones” (p. 21) destacando así la capacidad autónoma de los sistemas inteligentes para interactuar racionalmente con su entorno.

Aunque el álgebra lineal es fundamental en las ciencias y la ingeniería, no suele ser una fortaleza para muchos informáticos, debido a su naturaleza continua en contraste con las

matemáticas discretas. Sin embargo, para comprender y aplicar eficazmente numerosos algoritmos de ML, especialmente en el campo del DP, es esencial tener un dominio sólido de esta área matemática. (Goodfellow et al., 2016).

La materialización de la IA requiere de dos elementos fundamentales: la inteligencia en sí misma y un dispositivo que la albergue. El computador ha sido el artefacto seleccionado para tal propósito. La creación del computador electrónico digital moderno ocurrió de forma casi simultánea e independiente en tres naciones involucradas en la Segunda Guerra Mundial. En el Reino Unido, el equipo de Alan Turing construyó el Heath Robinson en 1940, el primer computador electromecánico operativo, diseñado específicamente para decodificar mensajes alemanes. Posteriormente, en 1943, el mismo grupo desarrolló el Colossus, una máquina más poderosa de propósito general, que utilizaba válvulas de vacío (Russell & Norvig, 2004).

Por otro lado, debido al interés de algunos investigadores no solo en lograr que las máquinas resolvieran problemas, sino también en comprender cómo el cerebro humano los aborda, muchas de las metodologías desarrolladas se propusieron también como aportes a las teorías sobre los procesos mentales humanos. Por esta razón, la investigación en psicología cognitiva y en inteligencia artificial se entrelazaron de forma significativa (Nilsson, 2009).

La lingüística resulta fundamental para facilitar la interacción entre las personas y los sistemas mediante PLN. Según Muñoz Guillena (2024), el PLN “es una emulación computacional de la capacidad humana para generar e interpretar textos en un idioma concreto (español, inglés, euskera, gallego, catalán, etc.). El sistema trata de generar o interpretar textos tal y como lo haría un ser humano” (p. 18).

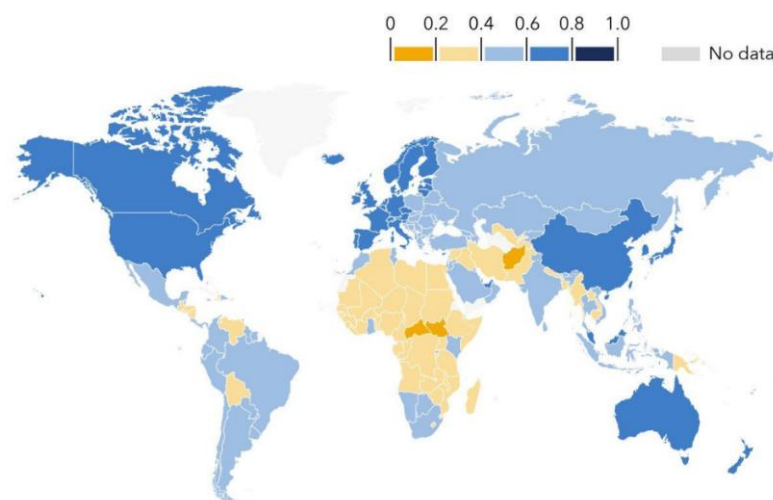
El PLN constituye una rama de la IA y la lingüística computacional, orientada a la interacción entre las computadoras y el lenguaje humano. Su objetivo principal es capacitar a los sistemas informáticos para comprender, interpretar y responder eficazmente al lenguaje natural. Este campo abarca todas las tareas relacionadas con dicha interacción, incluyendo tanto el análisis como la generación de texto, y entre sus aplicaciones más comunes se encuentran el análisis de texto (como la tokenización, lematización y etiquetado de partes del discurso), el reconocimiento de entidades nombradas, la traducción automática, el análisis de sentimiento y la clasificación de texto (Muñoz Guillena, 2024).

3.2 Contexto argentino

El *AI Preparedness Index* (AIPI) del Fondo Monetario Internacional (s.f.) es un indicador compuesto que evalúa el nivel de preparación de 174 países para adoptar y aprovechar tecnologías de inteligencia artificial. Este índice se calcula partir de cuatro dimensiones clave que son la infraestructura digital, capital humano y políticas laborales, innovación e integración económica, y marcos regulatorios y éticos. Proporciona una valoración integral, en una escala de 0 a 1, sobre las condiciones macroestructurales necesarias para un despliegue efectivo de la IA.

En el caso de Argentina, el AIPI para 2023 es de 0.47, una cifra inferior al promedio regional de América Latina y bastante por debajo de las economías avanzadas. Este dato revela desafíos en aspectos como conectividad e innovación. En la Figura 6 se visualiza el índice calculado para los 174 países. Este resultado, combinado con indicadores macroeconómicos y de capital humano, sugiere que el país aún tiene espacio para fortalecer su capacidad de adopción de IA, sobre todo en áreas como infraestructura digital y formación especializada (International Monetary Fund, 2023).

Figura 6
Índice de Preparación para la Inteligencia Artificial 2024

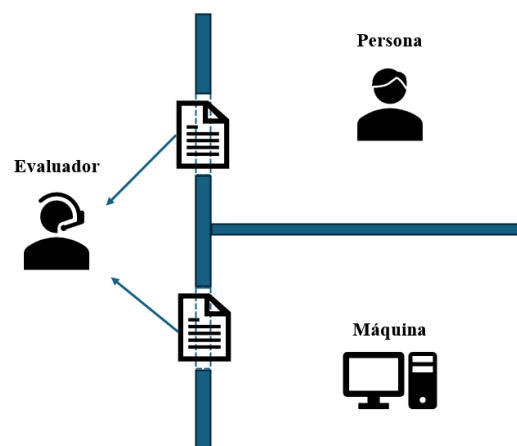


Nota: Adaptado de *AI Preparedness Index*, por International Monetary Fund, 2025, https://www.imf.org/external/datamapper/AI_PI@AIPI/ADVEC/EME/LIC. Todos los Derechos Reservados.

3.3 Prueba de Turing

La Prueba de Turing fue propuesta por el matemático y pionero de la computación Alan Turing en 1950 como un criterio para determinar si una máquina puede exhibir un comportamiento inteligente indistinguible al de un ser humano. En su ensayo *Computing Machinery and Intelligence*, plantea la pregunta: “¿Pueden pensar las máquinas?” (Turing, 1950, p. 2). La Prueba se basa en un experimento denominado juego de imitación, en el que participan tres actores: un evaluador humano, una persona y una máquina. El evaluador, que se encuentra aislado, mantiene una conversación escrita con ambos interlocutores sin saber cuál es cuál, con el objetivo de identificar quién es el ser humano. El desafío consiste en que la máquina logre imitar tan bien el comportamiento lingüístico humano que el evaluador no pueda distinguirla del verdadero humano. Si el evaluador falla en esta identificación con una frecuencia significativa, se considera que la máquina ha superado la prueba. De este modo, la prueba no busca demostrar que la máquina piense, sino que pueda comportarse de forma inteligentemente humana desde una perspectiva comunicacional (Turing, 1950). En la Figura 7 se representa conceptualmente la prueba y cómo se relacionan los actores.

Figura 7
Esquema conceptual de La Prueba de Turing



Nota: Esquema conceptual de los componentes de la Prueba de Turing. Elaboración propia

En la actualidad, aunque la Prueba de Turing ha sido objeto de críticas y revisiones, sigue siendo un referente simbólico en el campo de la IA, debido a que plantea cuestiones

fundamentales sobre la consciencia, la imitación del pensamiento humano y los límites de la inteligencia simulada. En su ensayo Turing (1950), anticipa una de las objeciones más recurrentes: la idea de que, para que una máquina realmente piense, debe ser consciente y experimentar sus propios pensamientos. Frente a esto, Turing responde que no es necesario que una máquina "sienta" el pensamiento, sino que basta con que su comportamiento sea indistinguible del de un ser humano para poder considerarla inteligente desde una perspectiva funcional.

3.4 Machine learning

3.4.1 Conceptos fundamentales

El ML se define, esencialmente, como la capacidad de adquirir la habilidad para realizar una tarea específica a partir de un conjunto de datos. Contrario a las representaciones comunes en la ciencia ficción, este proceso de aprendizaje no es enigmático, sino que a un nivel general implica la identificación y extracción precisa de patrones inherentes a los datos. Existen dos clasificaciones principales de ML: el aprendizaje automático supervisado y el aprendizaje automático no supervisado (Wolf, 2023).

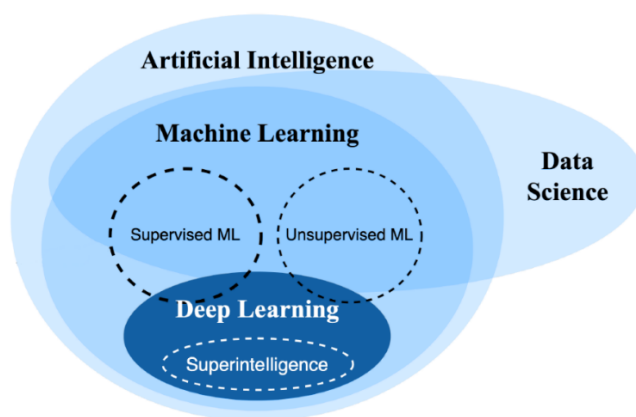
El aprendizaje supervisado aborda numerosos desafíos del mundo real donde, a partir de mediciones conocidas u observadas de un objeto, se busca predecir el valor de una variable objetivo desconocida. Los modelos de aprendizaje automático supervisado tienen la capacidad de pronosticar una variable objetivo categórica o una variable objetivo-numérica. El ámbito del aprendizaje automático supervisado se divide en dos categorías principales de problemas: los de clasificación y los de regresión. En los problemas de clasificación, la variable a predecir es de tipo categórico, como por ejemplo determinar si un correo electrónico es spam o si un cliente comprará un producto. Estos casos pueden abordarse mediante clasificadores binarios cuando hay solo dos categorías posibles, o clasificadores multiclase cuando existen más de dos. En cambio, los problemas de regresión implican la predicción de valores numéricos, como el precio de una vivienda según sus características (Wolf, 2023).

En cambio, el aprendizaje no supervisado aborda problemas complementarios al aprendizaje supervisado, ya que no requiere etiquetas de entrada generadas por humanos. Mientras que el ML supervisado infiere reglas a partir de los datos para predecir valores desconocidos, el no supervisado busca identificar patrones ocultos y agrupar los datos en clústeres o categorías distintivas. Un ejemplo práctico es el de una plataforma de comercio electrónico que desea comprender mejor a sus visitantes para ofrecer servicios personalizados según los distintos segmentos del público (Wolf, 2023).

3.4.1 Deep Learning

El DL es una subcategoría del aprendizaje automático supervisado, como se puede apreciar en la Figura 8, que se caracteriza por la utilización de modelos altamente complejos, capaces de abordar tareas sofisticadas que superan las capacidades de los enfoques tradicionales. Estos modelos, conocidos como redes neuronales profundas, se inspiran en el funcionamiento del cerebro humano para procesar información y han demostrado resultados sobresalientes en áreas como el reconocimiento de imágenes. Es importante destacar que, a pesar de su popularidad y logros, los modelos de DL no siempre superan a los métodos clásicos, ya que suelen requerir grandes volúmenes de datos, elevados recursos computacionales y ajustes técnicos complejos, lo que los vuelve inadecuados para ciertos contextos donde soluciones más simples pueden ofrecer mejores resultados. (Wolf, 2023).

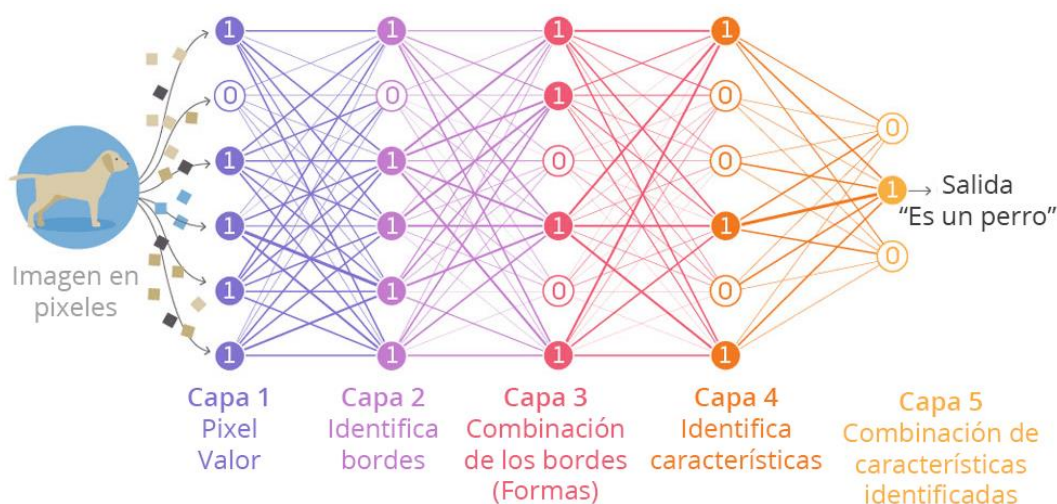
Figura 8
Áreas de la Inteligencia artificial



Nota: Adaptado de *Machine learning simplified* (p. 13), por Wolf, A. 2023, Themlsbook

A diferencia de los modelos tradicionales de ML, que suelen emplear arquitecturas con una o dos capas, el DL utiliza estructuras mucho más profundas, que pueden contar con cientos o incluso miles de capas como podemos ver en la Figura 9. Esta profundidad permite a los modelos procesar datos no estructurados¹ sin necesidad de etiquetas, mediante técnicas de aprendizaje no supervisado, lo que les facilita identificar patrones, características y relaciones de manera autónoma. Gracias a estas capacidades, el DP se ha convertido en un componente fundamental de muchas aplicaciones contemporáneas de inteligencia artificial, posibilitando el desarrollo de tecnologías como asistentes virtuales, comandos por voz, sistemas de detección de fraudes, vehículos autónomos y soluciones basadas en IA generativa (IBM, 2024).

Figura 9
Aprendizaje en base a la experiencia



Nota: Las redes neuronales profundas ajustan la fuerza de sus conexiones para mejorar la transmisión de señales de entrada a través de múltiples capas, hasta llegar a las neuronas asociadas con los conceptos generales correctos. Adaptado de *New Theory Cracks Open the Black Box of Deep Learning*, Quanta Magazine, 2017. <https://www.quantamagazine.org/new-theory-cracks-open-the-black-box-of-deep-learning-20170921/>. Todos los Derechos Reservados.

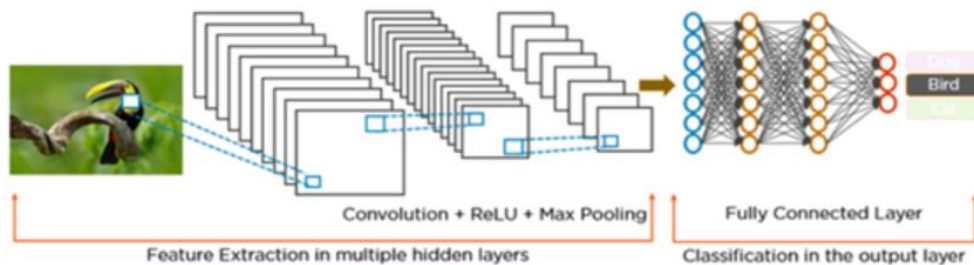
¹ Estos datos no están organizados de forma alguna, y se caracterizan porque no siguen un patrón predeterminado, o al menos este patrón no es evidente a simple vista. Los ejemplos más representativos de este tipo de datos son el audio, las imágenes y los videos.

Las redes neuronales artificiales buscan emular el funcionamiento del cerebro humano mediante la combinación de entradas de datos, ponderaciones y sesgos que actúan como neuronas de silicio, trabajando en conjunto para reconocer, clasificar y describir los objetos presentes en los datos. En este sentido, las redes neuronales profundas se estructuran en múltiples capas interconectadas, donde la capa de entrada recoge los datos y la de salida produce la predicción final. Emplean la propagación hacia adelante para ir refinando la información a medida que esta pasa de una capa a otra; además, utilizan la retro propagación, con algoritmos como el descenso de gradientes para calcular errores en las predicciones y ajustar las ponderaciones y los sesgos, haciendo que el modelo mejore su precisión con el paso del tiempo. En contra cara, el DL exige un elevado poder de cómputo, lo que nos lleva al uso de las GPUs de alto rendimiento y el cloud computing distribuido esenciales para gestionar el gran número de cálculos involucrados (IBM, 2024).

Como se mencionó anteriormente, el DL se distingue por el uso de redes neuronales profundas, modelos con múltiples capas que han demostrado una notable capacidad para abordar tareas complejas. En el contexto de los sistemas de recomendación, el DL permite analizar grandes volúmenes de datos de usuarios y contenido para identificar patrones intrincados y ofrecer sugerencias altamente personalizadas. Entre las diversas arquitecturas de redes neuronales relevantes para estos sistemas se destacan las Redes Neuronales Convolucionales (Convolutional Neural Networks, CNN). Esta tecnología está diseñada para procesar datos que presentan una topología en forma de cuadrícula, siendo un ejemplo típico las imágenes, que constituyen una matriz bidimensional de píxeles. Las CNN incorporan una operación matemática denominada convolución, que es un tipo particular de operación lineal utilizada en lugar de la multiplicación general de matrices en al menos una de sus capas (Goodfellow et al., 2016).

Como se muestra en la Figura 10, las CNN procesan imágenes mediante múltiples capas ocultas que extraen características visuales progresivamente más complejas. Este enfoque es útil en sistemas de recomendación donde los elementos visuales (como miniaturas de series o portadas de películas) influyen en las decisiones del usuario. Las características extraídas se utilizan luego en capas totalmente conectadas para realizar tareas de clasificación, mejorando la personalización del contenido ofrecido.

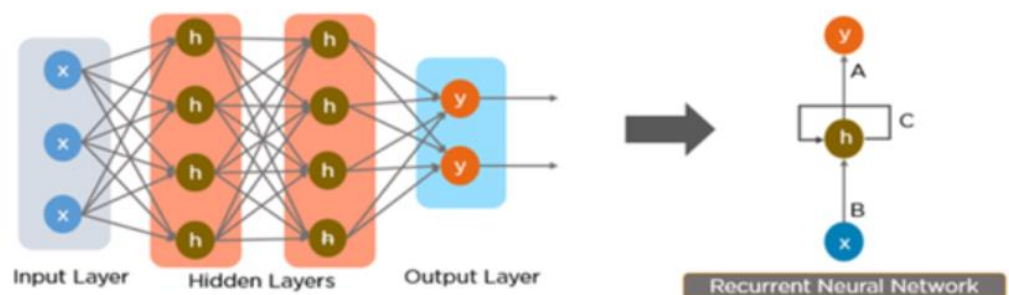
Figura 10
Redes Neuronales Convolucionales (CNN)



Nota: Adaptado de *What is The Main Difference between RNN and CNN*, Data Science Duniya, 2021. (<https://ashutoshtripathi.com/2021/07/12/the-main-difference-between-rnn-vs-cnn-nlp/>). Todos los Derechos Reservados.

Por otro lado, como vemos en la Figura 11, las Redes Neuronales Recurrentes (RNN) son especialmente útiles para analizar secuencias de datos, como el historial de visualizaciones de un usuario, lo que permite capturar dependencias temporales y predecir el comportamiento futuro (Goodfellow et al., 2016). Esta capacidad las convierte en una herramienta valiosa para analizar el historial de visualización de los usuarios, capturar dependencias temporales y predecir su comportamiento futuro de forma más precisa.

Figura 11
Redes Neuronales Recurrentes (RNN)



Nota: Adaptado de *What is The Main Difference between RNN and CNN*, Data Science Duniya, 2021. (<https://ashutoshtripathi.com/2021/07/12/the-main-difference-between-rnn-vs-cnn-nlp/>). Todos los Derechos Reservados.

3.5 Motor de recomendaciones

3.5.1 Contexto y fundamentos conceptuales

Un motor de recomendaciones es un sistema de software diseñado para filtrar información con el fin de sugerir productos o servicios que probablemente sean de interés para los usuarios de un sitio web. Las sugerencias pueden basarse tanto en la actividad individual de un cliente como en las acciones de otros compradores. Numerosas plataformas digitales y aplicaciones implementan sistemas de recomendación en sus funcionalidades, como se observa en Spotify, que ofrece sugerencias de canciones adaptadas a las preferencias de sus usuarios; Amazon, que presenta enlaces a productos basándose en el historial de compras de sus clientes; y diversos sitios web que exhiben publicidad personalizada fundamentada en el uso previo de internet por parte del usuario (Adobe for Business, 2021).

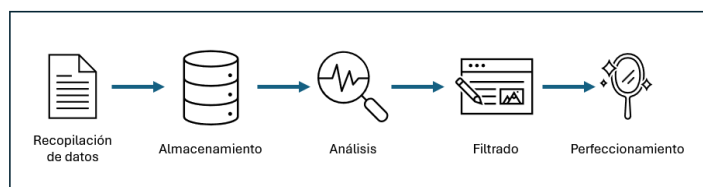
Los usuarios pueden descubrir contenido, productos o servicios que de otro modo no encontrarían gracias a los motores de recomendación. Estos sistemas son un elemento clave para las ventas y la interacción en muchas plataformas en línea, incluyendo el comercio electrónico, el streaming, los motores de búsqueda y las redes sociales. Un recomendador sugiere cuál es la próxima película por ver, una canción parecida, resultados de búsqueda relevantes o un producto que complementa una compra. La personalización de la experiencia del usuario depende en gran medida de las sugerencias generadas por estos sistemas de recomendación. (IBM, 2024).

En base a un informe de McKinsey & Company (2023) las experiencias personalizadas impulsan tanto la lealtad del cliente como las ventas brutas de la empresa. Además, la pandemia de COVID-19 ha intensificado la importancia de la personalización para las marcas: tres cuartas partes de los clientes cambiaron a una nueva tienda, producto o método de compra durante la pandemia, lo que evidencia que la lealtad a la tienda y al producto es cada vez menos frecuente.

3.5.2 Fases de un motor de recomendaciones

Un motor de recomendaciones funciona analizando datos relacionados con los usuarios y el contenido para identificar patrones y relaciones significativas. Consta de 5 etapas y se inicia con la recopilación de datos sobre las interacciones de los usuarios, tanto de forma explícita, como calificaciones, valoraciones y reseñas, como implícita, mediante el análisis de historiales de visualización, tiempos de permanencia y clics. Luego, en la fase de procesamiento y análisis de datos, se aplican técnicas de IA y ML para encontrar patrones de consumo, calcular similitudes entre usuarios y extraer características relevantes. Posteriormente, en la fase de generación de recomendaciones, se elabora una lista de contenidos calificados como relevantes para el usuario. Finalmente, en la etapa de evaluación y mejora, se monitorea el desempeño del motor de recomendación utilizando métricas como la tasa de clics, el tiempo de permanencia y la satisfacción del usuario, permitiendo ajustar los algoritmos para optimizar la calidad de las sugerencias (IBM, 2024). Las 5 fases del proceso se pueden observar en la Figura 12.

Figura 12
Fases de un motor de recomendaciones



Nota: Representación de las cinco fases de un motor de recomendaciones: recopilación de datos, almacenamiento, análisis, filtrado y perfeccionamiento. Fuente propia.

3.5.3 Modelos colaborativos, basados en contenido e híbridos

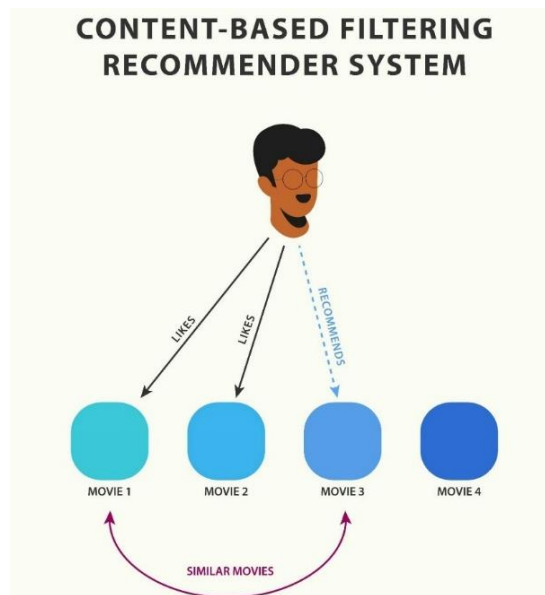
El filtrado colaborativo es una técnica de recomendación que se basa exclusivamente en el análisis de las interacciones pasadas entre usuarios e ítems, sin considerar las características específicas de los mismos. Este método registra los datos históricos de interacción en una matriz usuario-contenido, donde cada fila representa un usuario y cada columna un contenido, facilitando la agrupación de usuarios con comportamientos similares

para generar sugerencias personalizadas. Además, el filtrado colaborativo se clasifica en dos enfoques principales: uno basado en memoria y otro basado en modelos. El enfoque basado en memoria utiliza exclusivamente la matriz de interacción usuario-elemento junto con operaciones matemáticas para identificar usuarios similares y recomendar nuevos contenidos, sin recurrir a algoritmos de ML. Por otro lado, el enfoque basado en modelos introduce un modelo predictivo que anticipa las interacciones futuras; este modelo se ajusta a partir de los datos históricos y luego se emplea para clasificar y sugerir aquellos ítems que aún no han sido explorados por el usuario, priorizando los de mayor compatibilidad (Turing, 2022).

El filtrado basado en contenido en los sistemas de recomendación emplea algoritmos de ML para predecir y sugerir nuevos productos similares a los que el usuario ya ha mostrado interés. Este tipo de recomendación depende de contar con un conjunto definido de características de los productos y una lista de opciones disponibles para el usuario. El sistema construye un perfil personalizado a partir de interacciones previas, como clics, valoraciones y "me gusta", afinando sus predicciones a medida que el usuario acumula más actividad. Por ejemplo, si un cliente ha disfrutado de las dos primeras películas de un catálogo, el modelo recomendará una tercera película más parecida a las anteriores, en lugar de una cuarta opción menos relacionada. La similitud entre contenidos puede calcularse considerando aspectos como el elenco, el director, el género o la duración (Turing, 2022).

Este tipo de enfoque construye un perfil individualizado para cada usuario, utilizando datos explícitos o implícitos que reflejan sus preferencias previas, y lo compara con las características de nuevos ítems para determinar su grado de similitud. A medida que el usuario interactúa con más contenidos, el sistema refina sus predicciones y mejora su precisión. La Figura 13 ilustra de manera esquemática cómo opera un sistema de recomendación basado en filtrado de contenido, donde se identifican las similitudes entre ítems para sugerir aquellos que coinciden con los gustos previamente expresados por el usuario.

Figura 13
Sistema de Recomendación basado en Filtrado de Contenido



Nota: se dispone de cuatro películas y el sistema debe sugerir una al usuario que ya ha manifestado su preferencia por las dos primeras. Adaptado de *Guía para el filtrado basado en contenido en sistemas de recomendación*, por Turing.com, 2022. <https://www.turing.com/kb/content-based-filtering-in-recommender-systems>. Todos los Derechos Reservados.

Muchos motores de recomendación modernos, como por ejemplo Netflix, combinan técnicas de filtrado colaborativo y basado en contenido para aprovechar las fortalezas de ambos enfoques y proporcionar recomendaciones más precisas y diversificadas generando un nuevo tipo de sistemas de recomendación híbridos (Hong, 2024).

El filtrado colaborativo es una técnica clave en los sistemas de recomendación actuales, diseñada para ofrecer experiencias personalizadas. Se usa en plataformas de comercio electrónico, servicios de streaming y redes sociales, con el objetivo de mejorar la experiencia del usuario al sugerirle cosas que podrían gustarle. Su funcionamiento se basa en la idea de que, si los usuarios han tenido preferencias parecidas en el pasado, es probable que también las tengan en el futuro. De igual forma, si a usuarios con gustos similares les han gustado ciertos artículos, es probable que a otros con preferencias parecidas también les interesen. El filtrado colaborativo genera recomendaciones al analizar cómo interactúan los usuarios con diferentes elementos (Sanker, 2025).

Esta técnica se aplica en una gran variedad de campos para personalizar la interacción del usuario. Por ejemplo, en el comercio electrónico, sitios como Amazon lo utilizan para sugerir productos basándose en lo que has comprado o visto antes. Servicios de streaming como Netflix y Spotify lo emplean para recomendar películas, series o canciones, analizando lo que otros usuarios con gustos similares han visto o escuchado. En redes sociales como Facebook y TikTok, ayuda a sugerir amigos o contenido que se ajusta a tus intereses. Incluso en la educación, plataformas como Coursera y UdeMY lo usan para recomendar cursos, y en la salud, se explora para ofrecer tratamientos personalizados al comparar datos de pacientes con casos similares anteriores (Sanker, 2025).

3.6 Ética y seguridad de la información

3.6.1 Principios éticos en la inteligencia artificial

La UNESCO (2021) plantea un conjunto de principios éticos fundamentales para orientar el desarrollo, implementación y regulación de sistemas de IA, promoviendo su uso responsable, inclusivo y centrado en el bienestar humano. Entre estos valores se destacan la protección de la dignidad de cada ser humano, promover que ningún ser humano sufra daños o sometimientos físicos, económicos, político, cultural o mental durante toda la vida de un sistema de IA. Por otro lado, que los sistemas de IA puedan dar asistencia a personas vulnerables. Además, asegura que los derechos humanos y las libertades son fundamentales y deben ser respetados, protegidos y promovidos durante todo el ciclo de vida de cualquier sistema que trabaje con IA.

La privacidad es un derecho fundamental vinculado a la dignidad, autonomía y libertad de acción de las personas, y debe ser protegida en todas las etapas del ciclo de vida de los sistemas de inteligencia artificial. Para ello, es necesario que la recolección, uso, almacenamiento y eliminación de datos se realicen conforme al derecho internacional y los principios éticos establecidos, respetando también los marcos jurídicos vigentes a nivel nacional e internacional. En este sentido, se recomienda establecer mecanismos de gobernanza y marcos de protección de datos robustos, sustentados en principios internacionales y que contemplen derechos como el consentimiento informado y el uso legítimo de los datos

personales. Además, los sistemas algorítmicos deben someterse a evaluaciones rigurosas sobre su impacto en la privacidad, incluyendo aspectos sociales y éticos, aplicando enfoques como la privacidad desde el diseño. Los responsables del desarrollo e implementación de sistemas de IA deben asegurar la protección de la información personal durante todo su ciclo de vida. (UNESCO, 2021).

La transparencia y la explicabilidad en los sistemas de inteligencia artificial son condiciones esenciales para garantizar el respeto de los derechos humanos, las libertades fundamentales y los principios éticos. Su ausencia puede dificultar el funcionamiento adecuado de los marcos legales nacionales e internacionales en materia de responsabilidad, y limitar la capacidad de impugnar decisiones automatizadas. Esto, a su vez, puede poner en riesgo el derecho a un juicio justo y a un recurso efectivo, restringiendo así los contextos en los que estos sistemas pueden ser utilizados de manera legítima (UNESCO, 2021).

3.6.2 Buenas prácticas

La combinación de tecnología robusta y marcos normativos sólidos es clave para fortalecer la confianza de los usuarios y fomentar un entorno de innovación seguro y ético. La Organización para la Cooperación y el Desarrollo Económicos (OCDE) promueve un enfoque centrado en el ser humano para el desarrollo y uso de la IA. Esta tecnología ofrece importantes beneficios en ámbitos como la salud, la productividad y el avance científico, pero también implica riesgos significativos, como la desinformación, la inseguridad de los datos o la posible vulneración de derechos de autor. Dado que la IA trasciende las fronteras nacionales, estos principios propuestos por la OCDE establecen una base común para la cooperación internacional y la compatibilidad entre países, proporcionando lineamientos que buscan mantenerse vigentes frente al ritmo acelerado de evolución de la tecnología. (OECD, 2019).

Adoptar estos principios permite a los responsables de políticas públicas orientar el desarrollo de sistemas de IA de manera que se potencien sus beneficios y se reduzcan sus riesgos. Esto resulta esencial para aprovechar el potencial de la IA en términos de crecimiento económico, bienestar social y sostenibilidad ambiental, sin descuidar la protección de las personas y los valores colectivos. Hasta mayo de 2023, se registraron más de 1000 iniciativas

3.7 Inteligencia Artificial Explicada

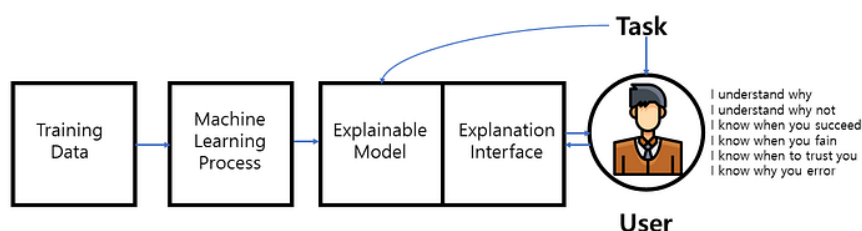
3.7.1 Conceptos de IA explicable

Según publicaciones de IBM en su portal *Think* define que la XAI se compone de procesos y métodos que permite a los usuarios comprender y confiar en los resultados que fueron creados por ML. La XAI es fundamental para comprender un modelo de IA, prever su impacto y detectar posibles sesgos. Permite caracterizar la precisión, imparcialidad y transparencia del modelo, así como los resultados de las decisiones tomadas por la IA. La XAI es crucial para que una organización construya confianza al implementar modelos de IA en entornos de producción. Además, la explicabilidad de la IA facilita que las organizaciones adopten un enfoque responsable en el desarrollo de la inteligencia artificial (IBM, s.f.).

En implementaciones de recomendaciones de contenidos basados en ML, se vuelve crucial que las empresas supervisen y gestionen continuamente sus modelos. Esto no solo promueve la explicabilidad de la IA, sino que también permite medir su impacto comercial. La XAI es fundamental porque fomenta la confianza del usuario final, mejora la auditabilidad del modelo y asegura un uso productivo de la IA. Asimismo, ayuda a mitigar riesgos importantes como los de cumplimiento normativo, legales, de seguridad y de reputación asociados con la IA en producción (IBM, s.f.).

En la Figura 15 se puede observar conceptualmente el modelo de XAI.

Figura 15
Concepto IA Explicada



Nota: Adaptado de *Transparent Intelligence: Lifting the Veil on AI's Decision Making with Explainable AI*, Medim, 2023. <https://itsudit.medium.com/transparent-intelligence-lifting-the-veil-on-ais-decision-making-with-explainable-ai-4a53ff440148>. Todos los Derechos Reservados.

3.7.2 Comparación entre IA y la XAI

La diferencia fundamental entre la IA y XAI radica en la transparencia del proceso de toma de decisiones. Mientras que la IA tradicional a menudo llega a un resultado utilizando un algoritmo de aprendizaje automático sin que sus creadores comprendan completamente cómo se llegó a esa conclusión, actuando como una "caja negra", la XAI implementa métodos específicos para asegurar que cada decisión sea rastreable y explicable. Esta capacidad de explicar el razonamiento detrás de las decisiones de la IA es crucial. Sin ella, resulta difícil verificar la exactitud de los resultados y se pierde el control, la responsabilidad y la capacidad de auditar el sistema (IBM, s.f.).

3.7.3 Métodos de explicabilidad

Existen diferentes enfoques para lograr la interpretabilidad en los modelos de inteligencia artificial. Uno de ellos es emplear modelos intrínsecamente interpretables, como los árboles de decisión o la regresión lineal, que por su naturaleza permiten entender fácilmente cómo se llega a una predicción. Otra estrategia se centra en generar explicaciones locales, que revelan por qué un modelo tomó una decisión específica para un caso particular; ejemplos populares de esto son las técnicas LIME y SHAP. Finalmente, las explicaciones globales buscan desentrañar el funcionamiento general del modelo, analizando las interacciones entre todas las características de entrada y las predicciones resultantes, como se ve en los métodos de reglas de asociación o los mapas autoorganizados (Educaopen, s.f.).

Profundizando en estas dos técnicas antes mencionadas, podemos decir que LIME simplifica modelos complejos para explicar predicciones individuales, siendo útil para interpretaciones rápidas, aunque puede ser inestable al asignar la importancia de las variables. Por otro lado, SHAP ofrece una cuantificación precisa de la contribución de cada variable a una predicción, basándose en la teoría de juegos, lo que la hace matemáticamente más sólida, pero también más exigente computacionalmente. Por ejemplo, en un diagnóstico médico, tanto LIME como SHAP podrían señalar la edad y la presión arterial como factores clave para un alto riesgo de infarto, aunque SHAP iría más allá al detallar la contribución exacta de cada uno. Uno de los mayores desafíos en XAI es lograr un equilibrio entre la capacidad predictiva del

modelo y su explicabilidad, lo que a menudo lleva a la implementación de enfoques híbridos. Así, un banco podría usar un sistema de aprendizaje profundo para detectar fraudes y, al mismo tiempo, emplear valores SHAP para auditar sus decisiones y asegurar que no haya discriminación (Portal de datos abiertos del Gobierno de España, 2025).

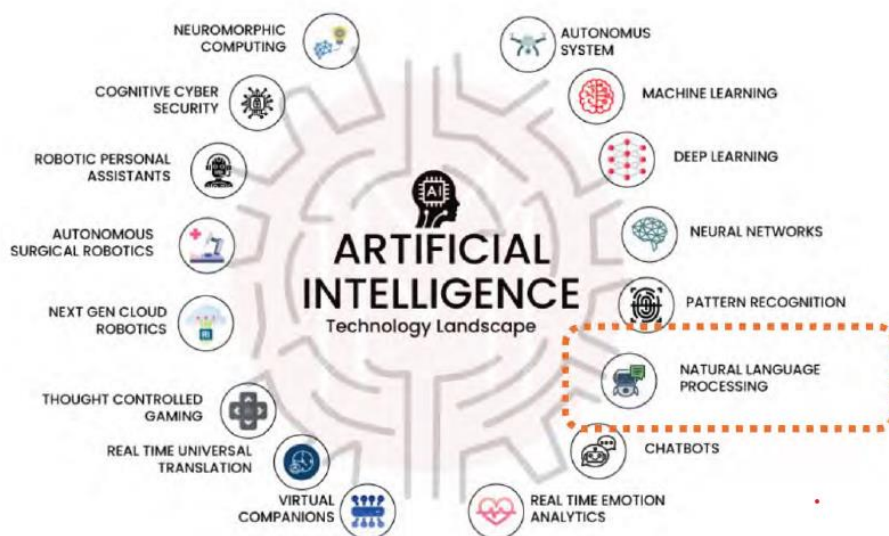
3.8 Procesamiento de lenguaje natural

3.8.1 Fundamentos del procesamiento de lenguaje natural

Según Muñoz Guillena (2024), el procesamiento de lenguaje natural comprende los aspectos de la inteligencia artificial que permiten la interacción entre humanos y máquinas mediante lenguas naturales. Esta área se centra en desarrollar capacidades computacionales que posibiliten a las máquinas interpretar, comprender y generar lenguaje humano, integrando metodologías provenientes de distintas disciplinas en un enfoque eminentemente interdisciplinario.

El PLN es solo una de las múltiples disciplinas que integran el ecosistema de la inteligencia artificial. Estas áreas, que abarcan desde el aprendizaje profundo hasta la robótica quirúrgica autónoma, conforman un panorama tecnológico diverso e interconectado. La Figura 16 ofrece una visión general de este paisaje, situando al PLN como un componente esencial en la evolución de las interfaces hombre-máquina, junto con otras disciplinas como el reconocimiento de patrones, los chatbots o los sistemas autónomos.

Figura 16
Disciplinas dentro de la inteligencia artificial



Nota: Adaptado de Curso de experto en Procesamiento del Lenguaje Natural de la Universidad de Alicante. En Procesamiento de lenguaje natural. (p. 20) por Muñoz Guillena, R., 2024.

Según Muñoz Guillena (2024), un sistema de PNL busca reproducir computacionalmente la capacidad humana para generar e interpretar textos en un idioma determinado. Si bien estas técnicas pueden estar inspiradas en modelos lingüísticos o cognitivos propios del ser humano, no necesariamente replican sus mecanismos mentales. En definitiva, el objetivo es lograr que la producción o interpretación del lenguaje por parte de una máquina sea indistinguible de la realizada por una persona.

Muñoz Guillena (2024) argumenta que los sistemas de procesamiento de lenguaje natural pueden operar en tres grandes escenarios funcionales. En primer lugar, cuando el texto o la voz actúan como entrada del sistema, se trata de un proceso de interpretación automática, como ocurre en herramientas de análisis de opiniones, detección de emociones o generación de imágenes a partir de descripciones textuales. En segundo lugar, existen sistemas en los que el texto es la salida, donde el procesamiento se enfoca en la creación de contenido, como en la descripción automática de imágenes o la producción de textos creativos. Finalmente, también se encuentran sistemas donde tanto la entrada como la salida son textos, lo cual implica

procesos simultáneos de comprensión y generación, como en la traducción automática, el resumen de documentos o los sistemas de diálogo interactivo.

3.8.2 Aplicaciones específicas del procesamiento de lenguaje natural

Debido a la complejidad del lenguaje natural, su procesamiento requiere abordar tanto tareas intermedias como aplicaciones específicas. Las tareas intermedias comprenden operaciones fundamentales desde el punto de vista lingüístico que permiten dotar a los sistemas de información útil para desarrollar funciones más avanzadas. Entre ellas se incluyen el análisis morfológico y sintáctico, la desambiguación léxica, la resolución de correferencias y el análisis semántico. Estas tareas actúan como base para el desarrollo de múltiples aplicaciones generales del PLN, que abarcan una amplia gama de enfoques y finalidades (Muñoz Guillena, 2024).

Una de las aplicaciones más representativas del PNL es la traducción automática, cuyo origen se remonta al experimento de Georgetown en 1954, cuando IBM y la Universidad de Georgetown lograron traducir automáticamente más de sesenta oraciones del ruso al inglés. Este tipo de procesamiento implica el uso de complejas técnicas matemáticas y algorítmicas, ya que realizar una traducción precisa requiere no solo conocimientos de morfología, sintaxis y semántica, sino también una comprensión profunda de los aspectos culturales asociados a los idiomas involucrados (Jurafsky & Martin, 2000, como se citó en Muñoz Guillena, 2024, p. 28).

Otra de las aplicaciones relevantes del procesamiento de lenguaje natural es la Recuperación de Información (IR), cuya importancia creció con el auge de Internet y la expansión exponencial de los contenidos digitales. Los sistemas de IR tienen como objetivo principal facilitar el acceso de las personas a la información más adecuada, útil y en el momento preciso, dentro de grandes volúmenes de datos. Estos sistemas enfrentan el desafío de clasificar documentos según su relevancia respecto de una consulta específica, como ocurre en los motores de búsqueda. Las búsquedas pueden basarse en el texto completo o en mecanismos de indexación basados en contenido. En este contexto, la recuperación de información se ocupa no solo de encontrar datos dentro de un documento, sino también de identificar documentos relevantes, metadatos descriptivos y contenido multimedia. Un ejemplo paradigmático de sistema de IR es el buscador Google (Muñoz Guillena, 2024).

La extracción de información constituye otra aplicación fundamental del procesamiento de lenguaje natural, orientada a identificar y estructurar datos relevantes que se encuentran explícita o implícitamente en los textos. Estos sistemas permiten obtener entidades, relaciones, eventos, fechas y hechos clave, que luego pueden ser almacenados en bases de datos relacionales o, en etapas más avanzadas, en estructuras semánticas. Este tipo de procesamiento fue especialmente relevante entre las décadas de 1990 y 2000, impulsado por iniciativas como la Message Understanding Conference (MUC), organizada por la agencia DARPA para evaluar distintos enfoques y tareas asociadas a esta disciplina. Entre las actividades más comunes se encuentran el reconocimiento de entidades nombradas, la extracción de eventos y la identificación de relaciones entre elementos del texto (Muñoz Guillena, 2024).

La búsqueda de respuestas es una aplicación avanzada del procesamiento de lenguaje natural que permite recuperar información precisa a partir de preguntas formuladas en lenguaje natural. A diferencia de los sistemas tradicionales de recuperación de documentos, esta técnica busca extraer fragmentos concretos que respondan directamente a una consulta específica. Está estrechamente relacionada con los motores de búsqueda web y contempla diversos tipos de preguntas, como listas, definiciones o interrogantes del tipo “cómo”, “cuándo”, “dónde” o “por qué”. Los sistemas de respuesta automática comparten similitudes con técnicas de resumen y extracción de información, ya que identifican palabras, frases u oraciones relevantes dentro de un documento y las presentan de manera coherente en función de la solicitud planteada (Muñoz Guillena, 2024).

La generación de resúmenes es una tarea clave dentro del PNL que consiste en condensar uno o varios documentos en una versión breve que conserve las ideas centrales y la información esencial. Esta técnica resulta especialmente valiosa en contextos donde se requiere visualizar rápidamente grandes volúmenes de contenido, agilizar procesos de búsqueda o facilitar la toma de decisiones informadas mediante la presentación sintética de datos relevantes (Muñoz Guillena, 2024).

Una de las aplicaciones más extendidas del PNL es la clasificación de textos, que consiste en asignar documentos a categorías predefinidas. Esta tarea tiene una amplia gama de usos, desde el análisis de sentimientos hasta la categorización temática de noticias o comentarios de usuarios. Aunque inicialmente se abordaba con algoritmos clásicos de

aprendizaje automático, los avances en redes neuronales han permitido mejoras significativas. (Muñoz Guillena, 2024).

El análisis de sentimientos, también conocido como minería de opiniones o emociones, es una técnica del procesamiento de lenguaje natural que permite identificar, extraer y estudiar de manera sistemática expresiones afectivas y contenidos subjetivos presentes en textos. Esta técnica se aplica frecuentemente sobre opiniones vertidas en redes sociales, reseñas en línea y otros espacios donde los usuarios formulan juicios explícitos o implícitos. Su utilidad es amplia y se extiende a campos como el marketing, la atención al cliente y la medicina clínica. En el ámbito hispanohablante, uno de los principales impulsores del desarrollo de esta disciplina es el Taller de Análisis Semántico (TASS), promovido por la Sociedad Española de Procesamiento del Lenguaje Natural (Muñoz Guillena, 2024).

3.8.3 Aplicación del PLN e IA en sistemas de recomendación

Los sistemas de recomendación constituyen una subcategoría dentro de los sistemas de filtrado de información, diseñados con el objetivo de anticipar la valoración o preferencia que un usuario podría otorgar a un determinado contenido. Estos sistemas combinan diversas técnicas para mejorar la precisión de sus predicciones. Por un lado, utilizan enfoques estadísticos como el filtrado colaborativo, el filtrado basado en contenido, la factorización de matrices y los modelos de regresión logística multinivel. Por otro lado, incorporan arquitecturas de redes neuronales especializadas: los autoencoders para la reducción de dimensionalidad, las CNN para procesar contenido visual y RNN para capturar patrones en secuencias de interacción entre el usuario y la plataforma (Muñoz Guillena, 2024)

4. Propuesta a implementar

La presente propuesta plantea el diseño e implementación de un sistema inteligente de recomendación de contenidos audiovisuales, orientado a su integración en plataformas de streaming activas en el mercado argentino. El sistema se propone adaptarse a las características específicas del consumo local, incorporando técnicas de XAI con el objetivo de aumentar la precisión de las recomendaciones, reducir los tiempos de búsqueda y mejorar la experiencia de los usuarios.

El enfoque estratégico de esta solución combina capacidades tecnológicas con una arquitectura algorítmica híbrida, capaz de integrar múltiples fuentes de datos y modelos de recomendación. Esta hibridación permitirá capturar tanto patrones individuales como colectivos de consumo, enriquecer el perfilado del usuario y generar sugerencias más contextualizadas y transparentes.

La planificación se articula sobre dos pilares fundamentales. Por un lado, los recursos tecnológicos, que incluyen lenguajes de programación, *frameworks*, herramientas de infraestructura y servicios en la nube, seleccionados por su capacidad de asegurar escalabilidad, interoperabilidad y eficiencia en el manejo de grandes volúmenes de datos. Por otro, los recursos metodológicos, que comprenden los enfoques algorítmicos, técnicas de modelado, PLN, ML y mecanismos de explicabilidad. Estos elementos son claves para dotar al sistema de inteligencia, transparencia y capacidad de adaptación.

Ambos ejes se integran mediante una arquitectura modular basada en microservicios, lo que facilitará su implementación en diversos entornos técnicos y garantizará su evolución futura. A través de esta estrategia, se busca no solo responder a las necesidades de personalización, sino también promover una experiencia de usuario más fluida, ética y centrada en la persona.

4.1 Recursos tecnológicos

4.1.1 Lenguaje de programación

El lenguaje de programación elegido para el desarrollo del sistema es Python, ampliamente reconocido como la tecnología líder en proyectos de IA, ML y ciencia de datos. Su predominancia se refleja en que recientemente desbancó a JavaScript como el lenguaje más utilizado en GitHub, impulsado principalmente por el auge de la IA y los frameworks de *DL* (GitHub, 2024).

Según la encuesta realizada por Developer Nation (2023), el 75,5 % de los desarrolladores de IA y ML elige Python como su lenguaje principal, destacando su sintaxis simple, su ecosistema robusto y la disponibilidad de múltiples bibliotecas especializadas. Le sigue JavaScript con un 58,8 %. Esta encuesta, dirigida a desarrolladores de todo el mundo, recopiló respuestas de más de 26.000 personas provenientes de 166 países.

Además, Python cuenta con un ecosistema particularmente sólido para el procesamiento de lenguaje natural, siendo spaCy una de las bibliotecas más destacadas por su enfoque práctico y su facilidad de uso. Esta herramienta permite realizar tareas como la segmentación de textos, el análisis sintáctico, la lematización, el reconocimiento de entidades nombradas y la clasificación de textos, entre otras funciones fundamentales para el desarrollo de sistemas inteligentes centrados en el lenguaje. Su arquitectura eficiente, pensada tanto para investigadores como para aplicaciones industriales, la convierte en una opción ideal para proyectos que requieren analizar grandes volúmenes de datos textuales de forma rápida y precisa. Esta capacidad técnica resulta clave para el enfoque metodológico planteado en esta propuesta, orientado al uso de PLN y explicabilidad (Cuéllar Hidalgo, 2024).

4.1.2 Infraestructura en la nube

Para garantizar la ejecución eficiente, escalabilidad y mantenimiento del sistema inteligente de recomendaciones, se ha seleccionado Google Cloud Platform (GCP) como proveedor de infraestructura en la nube. Esta elección responde a su reconocida robustez y adaptabilidad para entornos de ML, DL y procesamiento de datos en tiempo real. GCP permite

desplegar soluciones basadas en IA con requerimientos computacionales intensivos, ofreciendo recursos virtualizados con alta disponibilidad y arquitectura serverless, lo cual elimina la necesidad de gestionar servidores físicos, optimizando así tiempos y costos de operación (Google Cloud, s.f.).

En el contexto de esta propuesta, la plataforma ofrece funcionalidades esenciales para implementar los enfoques de PLN e XAI. Vertex AI proporciona un entorno completamente gestionado para la construcción, entrenamiento y despliegue de modelos de aprendizaje automático, incluyendo soporte nativo para frameworks como TensorFlow y PyTorch. Este servicio permite a los desarrolladores experimentar, ajustar y poner en producción modelos con mayor agilidad y eficiencia. Una de las herramientas clave dentro de este entorno es Vertex Explainable AI, que facilita la comprensión del comportamiento de los modelos mediante técnicas de atribución de características. Estas explicaciones ayudan a identificar cuáles atributos de entrada influyen más en las predicciones, y también permiten generar ejemplos contrastivos para mejorar la interpretación del modelo. Este tipo de explicabilidad resulta especialmente útil en aplicaciones donde se busca aumentar la transparencia y generar confianza en los resultados, como en los sistemas de recomendación personalizados (Google Cloud, 2025).

Para asegurar la eficiencia operativa de la arquitectura modular basada en microservicios, se utilizará Google Kubernetes Engine (GKE), una plataforma gestionada que permite automatizar la implementación, el escalado y la administración de aplicaciones contenerizadas. Según el portal de Google (Google Cloud, s.f.), GKE proporciona una solución robusta para ejecutar cargas de trabajo en contenedores con alta disponibilidad y eficiencia operativa. En particular, su modalidad Autopilot permite a los desarrolladores centrarse en el desarrollo de aplicaciones, mientras la plataforma gestiona automáticamente los recursos de infraestructura, lo cual resulta fundamental en entornos que procesan grandes volúmenes de datos en tiempo real, como los sistemas de recomendación basados en PNL.

demás, GKE incorpora herramientas nativas de observabilidad como Cloud Monitoring y Cloud Logging, las cuales permiten recolectar métricas, registrar eventos del sistema y supervisar el rendimiento de las aplicaciones. Estas herramientas, tal como indica Google Cloud (2024), ofrecen funcionalidades clave para detectar anomalías, mantener la estabilidad

operativa y asegurar una experiencia confiable para el usuario final, lo que es especialmente relevante en aplicaciones que requieren procesamiento continuo y aprendizaje automático. En el apartado 4.1.5 Monitoreo y visualización, se profundizará en el tema.

4.1.3 Almacenamiento y bases de datos

Se requerirá de la gestión de grandes volúmenes de datos heterogéneos, provenientes de múltiples fuentes como ser interacciones de usuarios, catálogos de contenidos, metadatos semánticos y eventos en tiempo real. Para abordar este desafío, se ha definido una solución de almacenamiento que equilibra eficiencia, escalabilidad y capacidad analítica, basada en BigQuery como sistema principal de análisis y en Elasticsearch como motor de recuperación avanzada.

BigQuery, la solución analítica de GCP, será el núcleo de la infraestructura de datos. Se trata de una base de datos columnar y serverless, diseñada para ejecutar consultas SQL altamente paralelizables sobre volúmenes de datos del orden de petabytes, con baja latencia. Esta arquitectura permite realizar análisis complejos sin necesidad de administrar servidores ni hardware dedicado, lo que facilita la escalabilidad y reduce los costos operativos (Google Cloud, s.f.). En la Tabla 1 podemos observar un detalle de tipos de datos que se pueden almacenar en este tipo de arquitectura.

Tabla 1
Detalle de tipos de datos a almacenar

Tipo de dato	Descripción
Registros de interacción de los usuarios	Información sobre clics, reproducciones y tiempo de visualización
Información de perfilado	Preferencias de contenido, historial de consumo
Datos enriquecidos de contenido	Género, actores, descripciones, etiqueta
Métricas de rendimiento del sistema	Click-Through Rate (CTR), precisión, tiempo medio de respuesta

Nota: Datos estructurados y semiestructurados propuestos que se almacenaran para utilizarse como inputs de recomendaciones futuras. Fuente: Elaboración propia.

Además, BigQuery destaca por su facilidad de uso, su bajo costo y su eficiencia para procesar datos estructurados y no estructurados provenientes de múltiples fuentes. Como remarcan Ali et al. (2021), esta herramienta permite realizar consultas analíticas complejas sobre grandes volúmenes de datos en tiempo real utilizando sintaxis SQL estándar, lo que la convierte en una alternativa ágil y accesible frente a soluciones *on-premise* más costosas y difíciles de escalar.

Por su parte, Elasticsearch se utilizará como motor de búsqueda avanzada y recuperación de información en tiempo real. Este sistema distribuido permite almacenar, buscar y analizar grandes volúmenes de datos no estructurados con alta velocidad y precisión, gracias a su arquitectura basada en índices invertidos, *shards*, réplicas y escalabilidad horizontal (Elastic, s.f.). Estas características lo convierten en una opción eficaz para entornos que requieren respuestas inmediatas y confiables, como los motores de recomendación personalizados.

Para capturar la dimensión semántica de los textos asociados a cada contenido (por ejemplo, descripciones o comentarios), se emplearán embeddings generados mediante modelos de lenguaje natural como DistilBERT. Este modelo, desarrollado mediante una técnica de *knowledge distillation*, conserva aproximadamente el 97 % del rendimiento de BERT, pero con solo la mitad de sus capas, lo que permite una reducción significativa del uso de memoria y una mejora en la velocidad de inferencia, sin comprometer su precisión en tareas clave de procesamiento de lenguaje natural (Gulati, 2025). Gracias a estas características, DistilBERT se presenta como una solución ideal para sistemas que deben operar en tiempo real y con recursos computacionales limitados, como los motores de recomendación personalizados.

4.1.4 Procesamiento en tiempo real

Uno de los pilares fundamentales del sistema propuesto es la capacidad de generar recomendaciones en tiempo real, adaptándose de forma dinámica al comportamiento actual del usuario. Para alcanzar este objetivo, se implementará un pipeline de procesamiento de eventos basado en Google Cloud Dataflow, un servicio gestionado de procesamiento de datos en

streaming y batch, construido sobre el modelo de programación de Apache Beam (Google Cloud, s.f.).

Este componente permitirá capturar y procesar eventos generados por la interacción del usuario con la plataforma, como clics, visualizaciones, búsquedas, pausas, abandonos, cambios de dispositivo o tiempo de reproducción, y transformarlos en señales útiles que alimenten en tiempo real tanto los perfiles individuales como los modelos de recomendación. El uso de Dataflow asegura baja latencia, escalabilidad automática y procesamiento tolerante a fallos, cualidades esenciales para brindar una experiencia responsiva en entornos de alta demanda (Google Cloud, s.f.). Entre sus ventajas clave se encuentra la posibilidad de aplicar transformaciones avanzadas sobre los flujos de datos, como el uso de ventanas temporales (windowing), que permiten agrupar eventos por franjas horarias o duración personalizada, adaptando las recomendaciones al momento específico del día.

Asimismo, se implementarán uniones de flujos de datos entre las interacciones de los usuarios y los metadatos de los contenidos, junto con agregaciones dinámicas por usuario, género o dispositivo (Google Cloud, s.f.). Estas operaciones permitirán detectar patrones de comportamiento con mayor granularidad y generar sugerencias más precisas y contextualizadas.

Además, el procesamiento en tiempo real facilitará la actualización dinámica tanto de los perfiles de usuario como de los modelos de recomendación. Al identificar cambios recientes en los hábitos de visualización o en las tendencias colectivas, el sistema podrá ajustar su lógica de inferencia de forma adaptativa, alineándose con las preferencias actuales del usuario y minimizando el desajuste entre las recomendaciones ofrecidas y los intereses reales.

Integrado con BigQuery como repositorio analítico y con Vertex AI como entorno de inferencia, el pipeline de procesamiento en tiempo real proporcionará una infraestructura robusta para mantener actualizado el sistema de recomendación, generar contenidos relevantes en contexto y ofrecer una experiencia de usuario fluida y personalizada (Google Cloud, s.f.).

Tabla 2*Pipeline de procesamiento en tiempo real para recomendaciones personalizadas*

Etapa del Pipeline	Descripción
Captura de eventos	Registro en tiempo real de interacciones del usuario: clics, pausas, búsquedas, cambios de dispositivo, etc.
Procesamiento de datos	Uso de Apache Beam en Dataflow para transformaciones como windowing, stream joins y agregaciones
Enriquecimiento de señales	Combinación de eventos con metadatos del contenido (género, actor, duración, etc.).
Actualización de perfiles	Modificación dinámica del perfil del usuario en base a su comportamiento reciente.
Inferencia en tiempo real	Generación de recomendaciones con modelos alojados en Vertex AI
Almacenamiento analítico	Persistencia de eventos y resultados en BigQuery para análisis histórico y supervisión del sistema

Nota. Etapas propuestas para el procesamiento en tiempo real de datos de usuarios, con el fin de alimentar un sistema de recomendación personalizado. Fuente: Elaboración propia.

4.1.5 Monitoreo y visualización

Para asegurar el funcionamiento óptimo del sistema de recomendación y posibilitar mejoras continuas basadas en evidencia, se implementará una capa robusta de monitoreo y visualización. Esta estará soportada por los servicios nativos de GCP, específicamente Cloud Monitoring y Cloud Logging, los cuales ofrecen capacidades avanzadas para recopilar, registrar y analizar métricas clave en tiempo real (Google Cloud, s.f.).

Entre los indicadores a monitorear se incluyen la latencia de respuesta del motor de recomendaciones, la disponibilidad del servicio, el Click Through Rate (CTR), la precisión de las sugerencias, el tiempo medio de visualización posterior a una recomendación, y métricas indirectas de satisfacción del usuario. La observación sistemática de estos indicadores permitirá identificar cuellos de botella, detectar anomalías operativas y anticipar posibles degradaciones en el rendimiento (Google Cloud, 2023).

Los datos recopilados se integrarán en tableros personalizados contruidos con Looker Studio, lo que permitirá a los equipos técnicos y de negocio interpretar el estado del sistema con facilidad. Estos dashboards facilitarán la toma de decisiones basadas en datos y permitirán configurar alertas automáticas ante comportamientos anómalos, como picos de latencia o disminuciones repentinas en el CTR.

Asimismo, se propone una estrategia de retroalimentación cruzada entre el monitoreo técnico y los análisis de comportamiento del usuario. Por ejemplo, una caída sostenida en la tasa de clics podría activar ajustes automáticos en los parámetros del modelo de recomendación o en los umbrales semánticos empleados en la selección de contenidos. De este modo, el monitoreo no solo actúa como herramienta operativa, sino que se integra al ciclo de mejora continua del sistema, reforzando su adaptabilidad y eficacia.

4.1.6 Arquitectura de microservicios e integración mediante APIs

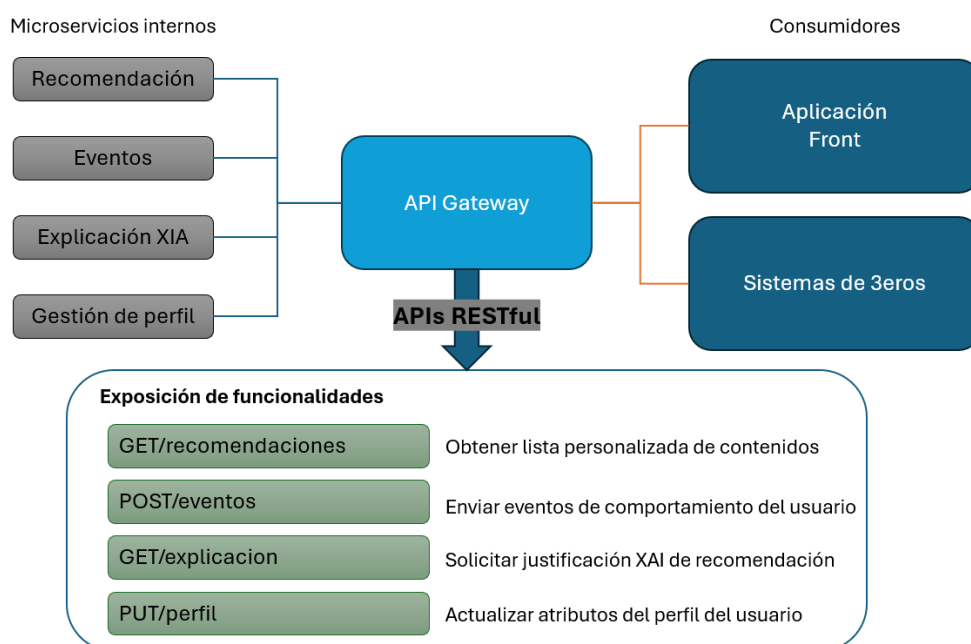
El sistema de recomendación será diseñado bajo una arquitectura de microservicios desacoplada, donde cada función crítica, como la ingesta de datos, el perfilado de usuario, el motor de inferencia y el módulo de XAI, se implementará como un servicio autónomo, desplegado en contenedores Docker y orquestado mediante GKE. Este enfoque permite que cada microservicio escale, se actualice o se despliegue de forma independiente, lo cual mejora la agilidad del desarrollo y la resiliencia operativa. Tal como plantea Ambassador Labs (2024), los microservicios promueven una separación clara de responsabilidades, facilitando que equipos distintos trabajen en paralelo, lo que acelera la innovación y reduce riesgos en sistemas complejos.

La comunicación entre los microservicios y con sistemas externos se implementará a través de APIs RESTful, respetando principios de interoperabilidad, uso explícito de recursos y separación entre cliente y servidor. Esta aproximación favorece la escalabilidad y la evolución independiente de cada servicio sin comprometer la estabilidad del sistema completo. Según Microsoft (2025), una arquitectura REST bien estructurada permite exponer funcionalidades mediante endpoints uniformes, aplicar métodos HTTP estándar y mantener una organización que promueva el bajo acoplamiento y la alta cohesión entre componentes distribuidos.

En la Figura 17 se presenta un esquema que ilustra cómo se relacionan entre sí los microservicios internos y los consumidores externos, como front-end o sistemas de terceros. Este diseño facilita la integración del sistema en plataformas web, móviles, Smart TVs y decos Android y permite futuras extensiones sin comprometer la infraestructura original. Al operar

de forma desacoplada y con GKE como soporte, el sistema logra alta disponibilidad, balanceo de carga y tolerancia a fallos, condiciones esenciales para una experiencia de usuario fluida y confiable.

Figura 17
Comunicación de Microservicios y Plataformas mediante APIs RESTful



Nota: Mapa conceptual que representa cómo se consumirán los servicios desde las aplicaciones de los fronts o sistemas de terceros. Fuente: Elaboración propia

Esta arquitectura desacoplada basada en microservicios y APIs RESTful permitirá que el sistema de recomendación se integre con facilidad en plataformas de streaming existentes, sin necesidad de modificaciones profundas en su infraestructura. Gracias al principio de independencia entre servicios y a la interoperabilidad que ofrecen las APIs, será posible desplegar el sistema como un componente autónomo, que puede evolucionar, escalar y adaptarse a distintos entornos y dispositivos, como PCs, celulares y pantallas grandes, sin afectar al resto del ecosistema. Además, este enfoque facilita la incorporación futura de nuevas funcionalidades, como algoritmos alternativos, módulos conversacionales o herramientas analíticas, garantizando flexibilidad y sostenibilidad a largo plazo (Ambassador Labs, 2024).

4.2 Recursos metodológicos

Desde el punto de vista metodológico, la propuesta se fundamenta en un enfoque integral que combina distintas técnicas y modelos de recomendación con el objetivo de ofrecer una experiencia personalizada, precisa y transparente para los usuarios. Este enfoque contempla tanto métodos tradicionales como innovaciones recientes en IA, ML y PNL. La integración de estas herramientas permite abordar desafíos habituales en sistemas de recomendación, como la escasez de datos, la sobre especialización o la falta de interpretabilidad. A su vez, se busca garantizar una arquitectura flexible y escalable que facilite la mejora continua del sistema a partir de datos reales de uso y retroalimentación del usuario.

4.2.1 Inteligencia Artificial Explicable XAI

En el marco de esta propuesta, se integrarán mecanismos de XAI con el fin de dotar al sistema de recomendación de mayor transparencia, trazabilidad y confiabilidad. Esta decisión responde a la creciente necesidad de que los usuarios comprendan por qué se les sugiere determinado contenido, especialmente en entornos donde las recomendaciones tienen un impacto directo en la experiencia de uso y en métricas de negocio como la retención o el engagement.

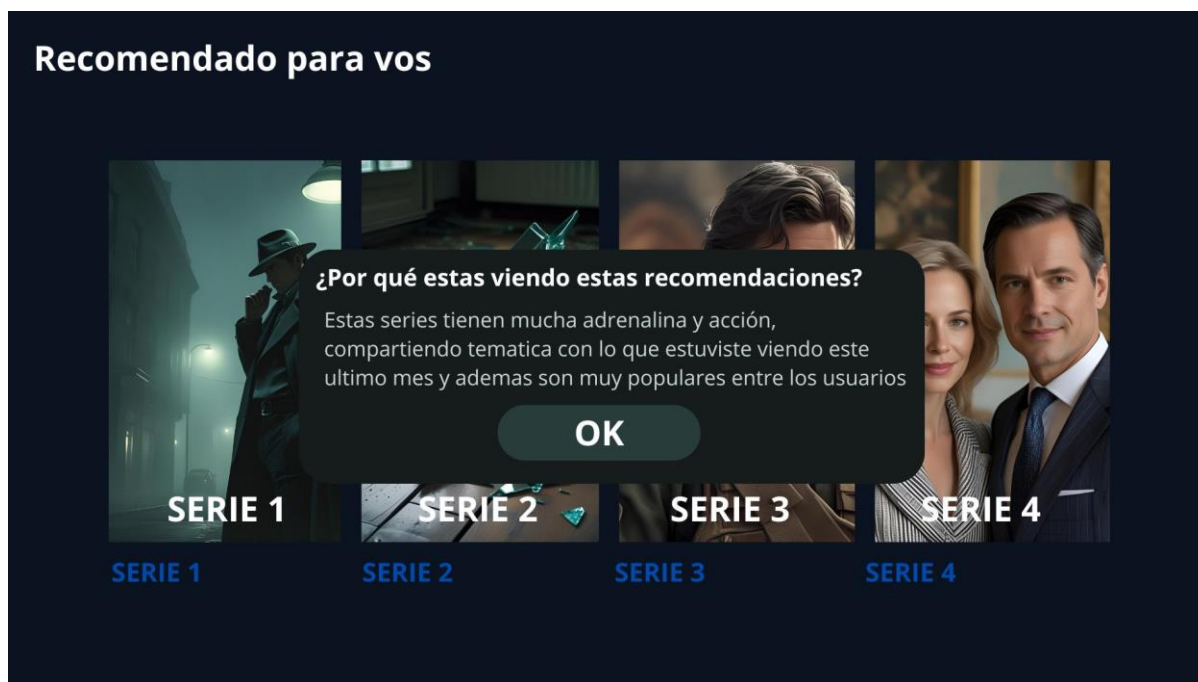
Como parte de la implementación, se utilizarán técnicas modelo-agnósticas y de explicabilidad local como SHAP (SHapley Additive Explanations) y LIME (Local Interpretable Model-Agnostic Explanations). En particular, LIME fue desarrollado como una solución flexible para explicar predicciones individuales de cualquier modelo de ML, y su eficacia fue validada en distintos dominios, como texto e imágenes. En su paper fundacional, Ribeiro et al. (2016) argumentan que la posibilidad de generar explicaciones interpretables fortalece la confianza del usuario, facilita la comparación entre modelos, ayuda a detectar errores y proporciona información valiosa sobre las decisiones del sistema. Esta capacidad resulta especialmente valiosa en contextos como los sistemas de recomendación, donde comprender por qué se sugiere un contenido puede marcar la diferencia en la percepción del usuario. Las herramientas de explicabilidad permitirán identificar los factores de mayor peso en la predicción, como la similitud semántica de contenidos, patrones de consumo horario o

coincidencias con perfiles similares, y traducir esa información en mensajes comprensibles para el usuario.

Complementariamente, se empleará SHAP, una técnica basada en la teoría de juegos que permite atribuir a cada variable de entrada su contribución marginal a la predicción de manera consistente y localmente precisa. Esta herramienta ofrece una forma robusta de descomponer la decisión del modelo en componentes comprensibles, proporcionando una explicación cuantitativa del impacto de cada característica. Tal como señalan Lundberg y Lee (2014), el marco SHAP surge como respuesta a la creciente tensión entre precisión e interpretabilidad en los modelos de ML, proponiendo un enfoque unificado que integra múltiples métodos previos bajo principios comunes. Esta perspectiva no solo aporta claridad conceptual, sino que sienta las bases para futuras mejoras metodológicas orientadas a lograr explicaciones más eficientes, precisas y adaptables a distintos tipos de modelos.

Desde el punto de vista técnico, los modelos explicativos serán entrenados en paralelo a los modelos principales de recomendación, utilizando como insumos las variables más relevantes definidas por el pipeline de procesamiento y el perfilado multidimensional del usuario. Esta estrategia busca generar explicaciones precisas para cada sugerencia presentada, sin afectar el rendimiento del sistema. Las salidas generadas por SHAP y LIME se integrarán a la interfaz mediante componentes visuales no intrusivos, como ventanas emergentes, leyendas contextuales o secciones informativas dentro del carrusel de recomendaciones. Estas explicaciones permitirán al usuario comprender los factores que justifican cada sugerencia, como la similitud con contenidos previamente vistos, coincidencias de horario de consumo o afinidad con otros perfiles similares, sin interrumpir su experiencia de navegación. Un ejemplo conceptual de esta funcionalidad puede observarse en la Figura 18.

Figura 18
Recomendaciones con IA Explicada



Nota: Mockup propuesto de recomendación aplicando inteligencia artificial explicada, donde da contexto de qué criterios se utilizaron las recomendaciones expuestas. Imágenes generadas con IA de Canva. Fuente propia

Además, se desarrollarán dashboards internos interactivos en Looker Studio, orientados a equipos técnicos y de negocio, que permitirán auditar el funcionamiento del sistema de recomendación y detectar posibles sesgos, como la sobreexposición de ciertos tipos de contenido o la falta de diversidad en los resultados. Estos tableros incluirán métricas como la distribución de recomendaciones por género, la tasa de repetición de contenidos, la diversidad de recomendaciones por usuario o la representatividad por segmentos, visualizadas mediante gráficos de torta, líneas o Indicadores Clave de Rendimiento (KPI), tal como se ejemplifica en la Tabla 3. Este enfoque no solo responde a una necesidad ética y normativa emergente, sino que también fortalece la relación de confianza entre los usuarios y la plataforma, convirtiendo a la explicabilidad en un eje estratégico de la innovación.

Tabla 3
Indicadores para auditoría de recomendaciones

Indicador	Descripción	Visualización sugerida
Distribución de géneros recomendados	Proporción de géneros en los contenidos sugeridos	Gráfico de torta o barra apilada
Repetición de contenidos	Frecuencia con la que se repiten contenidos para un mismo usuario	Indicador numérico y serie temporal
Diversidad por usuario	Cantidad de géneros distintos recomendados por usuario	Histograma por percentiles
Representatividad por segmento	Cobertura de recomendaciones por grupo demográfico o dispositivo	Mapa de calor o gráfico de burbujas
Niveles de engagement posterior	Porcentaje de recomendaciones que generaron visualización efectiva	KPI destacado y gráfica comparativa

Nota. Tabla ilustrativa que resume métricas e indicadores posibles a incluir en dashboards de auditoría interna del sistema de recomendaciones. Fuente propia.

4.2.2 Perfilado de usuario

El armado de perfilados tiene como objetivo modelar de manera precisa y dinámica los intereses, hábitos y comportamientos de cada usuario, para habilitar una experiencia personalizada y contextual. La construcción del perfil de cada usuario es un proceso fundamental que se realiza mediante la integración de diversas fuentes de datos. Por un lado, se consideran los datos explícitos, que corresponden a las preferencias declaradas directamente por el usuario, como la selección de géneros preferidos o la creación de listas de favoritos. Por otro lado, y de manera complementaria, se incorporan los datos implícitos, los cuales se obtienen a partir del registro de su comportamiento dentro de la plataforma, incluyendo el historial de visualización, el tiempo de permanencia por título, la frecuencia de uso, el tipo de dispositivo utilizado, la franja horaria de consumo y las secuencias de contenido reproducido. Esta combinación de información explícita e implícita permite generar un perfil de usuario completo y dinámico, esencial para la personalización de las recomendaciones.

La representación del perfil se basará en un vector multidimensional enriquecido con información contextual y semántica. La similitud entre el perfil del usuario y nuevos contenidos se calculará mediante funciones vectoriales y métricas como la distancia coseno, apoyadas en embeddings generados por modelos de PLN. El perfilado será dinámico y se actualizará en función del uso, incorporando nuevas interacciones. Cada evento será procesado mediante

funciones serverless en Cloud Functions, lo que permitirá modificar el perfil casi en tiempo real. Además, se aplicarán reglas de ponderación para dar mayor relevancia a los contenidos recientes.

Además de los datos de comportamiento y contenido, se incluirán metadatos contextuales para enriquecer el perfil del usuario y afinar las recomendaciones. Estos datos abarcan información como el dispositivo utilizado (Decodificador Android TV, Smart TV, celular, tablet), la franja horaria en que se consume el contenido, el día de la semana e incluso la localización geográfica, si esta información está disponible. La integración de estas variables contextuales permitirá una segmentación más precisa de los usuarios, posibilitando una personalización aún más profunda de las recomendaciones al considerar el momento y el entorno de consumo.

La Tabla 4 ilustra un ejemplo representativo de perfilado de usuario dentro del sistema propuesto, evidenciando cómo se combinan atributos conductuales, contextuales y semánticos para enriquecer la precisión de las sugerencias ofrecidas. Además, se incorpora una dimensión específica vinculada al análisis de metadatos del contenido, como sinopsis, etiquetas, actores o temas predominantes, procesados mediante técnicas de PLN. Esta capa semántica permitirá generar una representación más rica del perfil del usuario, basada no solo en lo que visualiza, sino también en el contenido profundo de lo que consume.

Tabla 4
Ejemplo de perfilado de usuario

Dimensión	Valor
Usuario	ID1234
Géneros Preferidos	Comedia (0.85), Thriller (0.70), Documental (0.65)
Franja horaria frecuente	20:00–23:00
Dispositivo habitual	Smart TV
Embedding semántico promedio	[0.12, -0.03, ..., 0.21]
Cluster asociado	"Jóvenes adultos – series nocturnas"
Metadatos procesados por PLN	Etiquetas: "Viajes en el tiempo", Actor: "Pedro Pascal", Genero: "acción"
Últimos contenidos vistos	"Stranger Things", "Luis Miguel: La Serie", "The last of us"

Nota: Dimensiones posibles para un perfil ejemplo. Fuente: Elaboración propia.

4.2.3 Integración metodológica con la arquitectura del sistema

La propuesta metodológica se integra de manera armónica con la arquitectura tecnológica definida, estructurada en tres capas: la capa de ingesta y perfilado, de modelado y recomendación, y de entrega e interacción con el usuario. Esta articulación garantiza que los procesos analíticos, los modelos de inteligencia artificial y las herramientas de explicabilidad funcionen de forma orquestada y eficiente, favoreciendo tanto la escalabilidad técnica como la personalización centrada en el usuario.

En la capa de ingesta y perfilado, los eventos generados por el comportamiento del usuario como ser clics, visualizaciones, búsquedas, pausas, se procesan en tiempo real y se almacenan en una base analítica centralizada. Simultáneamente, se generan representaciones semánticas enriquecidas del contenido a través de modelos de procesamiento de lenguaje natural, que permiten alimentar perfiles de usuario dinámicos, sensibles al contexto y a la evolución de los intereses individuales.

La capa de modelado y recomendación contiene los algoritmos híbridos definidos en la propuesta: filtrado colaborativo, modelos basados en contenido y segmentación por clusters. Estos modelos operan sobre la información centralizada y actualizada, y se encuentran encapsulados en servicios que permiten su entrenamiento, evaluación y despliegue automatizado. Además, se integran mecanismos de explicabilidad que generan justificaciones automáticas de las recomendaciones, facilitando la interpretación por parte del usuario.

En la capa de entrega e interacción, el motor de recomendación expone sus resultados a través de APIs RESTful, asegurando la integración fluida con distintas plataformas de consumo de los fronts. Esta capa también recopila retroalimentación implícita del comportamiento del usuario para retroalimentar el ciclo de aprendizaje. Las explicaciones generadas se presentan junto a las recomendaciones en la interfaz, mediante elementos visuales no intrusivos, mejorando la transparencia y confianza del usuario en el sistema.

Esta estructura modular y desacoplada busca permitir adaptar o reemplazar componentes individuales sin afectar el funcionamiento global. Esta integración metodológica con la arquitectura garantiza un sistema robusto, ágil, mantenible y orientado a ofrecer una experiencia personalizada de calidad.

La Figura 19 presenta un esquema conceptual de la integración metodológica con la arquitectura del sistema. Este diagrama sintetiza cómo interactúan las distintas capas de ingesta y perfilado, modelado y recomendación, y entrega al usuario, para habilitar un flujo continuo y orquestado de datos, aprendizaje y personalización. Cada componente refleja el recorrido de la información desde que es generada por el usuario hasta que retorna en forma de recomendaciones explicables y contextualizadas, consolidando un sistema modular, escalable y centrado en la experiencia.

Figura 19
Integración metodológica con la arquitectura del sistema



Nota: Diagrama conceptual de capas y su integración con el sistema. Fuente: Propia

4.3 Validación de resultados

La validación constituye una etapa clave para garantizar su efectividad, precisión y utilidad práctica. Dado que el sistema se apoya en algoritmos de inteligencia artificial para generar recomendaciones personalizadas y explicables, su evaluación debe abordar tanto métricas cuantitativas tradicionales como indicadores cualitativos vinculados a la experiencia del usuario.

En primer lugar, la validación del sistema se apoyará en métricas cuantitativas que permiten medir con objetividad el desempeño de los algoritmos de recomendación. Una de las principales métricas será la Precisión@K, que representa la proporción de contenidos relevantes dentro de los primeros K resultados recomendados por el sistema. Esta métrica es especialmente útil cuando el objetivo es minimizar la entrega de resultados irrelevantes, lo que se traduce en una mejor experiencia para el usuario. Por ejemplo, si se presentan 10 recomendaciones y el usuario considera que 7 de ellas fueron relevantes, la precisión será del 70 %. Este indicador se asocia directamente con la tasa de falsos positivos: una alta precisión significa que el sistema evita recomendar contenido que el usuario probablemente no desee consumir (Google Developers, 2025).

Por otro lado, se utilizará también la Recall@K, una métrica que evalúa la capacidad del sistema para identificar la mayor cantidad posible de contenidos relevantes dentro del conjunto total disponible. En este caso, se enfoca en minimizar los falsos negativos, es decir, aquellos contenidos que el usuario hubiese valorado como relevantes pero que no fueron recomendados. Un Recall alto implica que el sistema tiene una mayor cobertura de las preferencias del usuario, aunque puede venir acompañado de una disminución en la precisión si se incluyen más resultados potencialmente irrelevantes (Google Developers, 2025).

Dado que en muchos casos existe una tensión entre estos dos indicadores (mejorar uno puede implicar empeorar el otro), se empleará el F1-score, una métrica que combina precisión y Recall en una sola medida armónica. El F1-score es especialmente valioso cuando se requiere un equilibrio entre ambos aspectos, ya que penaliza fuertemente los escenarios en los que uno de los dos indicadores es muy bajo. Esta métrica será utilizada para realizar comparaciones globales del rendimiento de los distintos modelos y configuraciones del sistema, permitiendo seleccionar aquellas versiones que logren un mejor compromiso entre exactitud y cobertura (Google Developers, 2025).

Además, se incorporarán métricas más sofisticadas que consideran no solo la relevancia de los contenidos recomendados, sino también su posición dentro del ranking. Una de ellas es la Precisión Media Promedio (MAP), utilizada para evaluar la calidad del ordenamiento de las recomendaciones. Esta métrica analiza cuántos elementos relevantes se presentan y en qué lugar de la lista aparecen, brindando una medida más fina del rendimiento. El MAP se calcula

como el promedio de la precisión en cada posición donde se encuentra un contenido relevante, considerando múltiples usuarios o búsquedas. Su valor oscila entre 0 y 1, siendo 1 el resultado ideal, donde todos los contenidos relevantes están ubicados en los primeros lugares. Cuanto más alto sea el MAP, mayor será la efectividad del sistema para ordenar de forma óptima sus recomendaciones (Evidently AI, 2025).

Por su parte, el Normalized Discounted Cumulative Gain (NDCG) es una métrica que evalúa la calidad del ordenamiento de las recomendaciones en relación con un ranking ideal, en el cual todos los contenidos relevantes figuran en las primeras posiciones. Para calcularlo, se compara la Ganancia Acumulada Descontada (DCG), que mide la relevancia total de los ítems recomendados, aplicando un factor de penalización a los que se encuentran más abajo en la lista, con su valor ideal (IDCG), es decir, una clasificación perfecta. El resultado se expresa como un valor normalizado entre 0 y 1, donde 1 representa una coincidencia exacta con el orden óptimo y valores inferiores reflejan menor calidad en la priorización de resultados. Esta métrica puede agregarse a nivel global para obtener una visión general del rendimiento del sistema sobre todo el conjunto de usuarios (Evidently AI, 2025).

Estas métricas permitirán evaluar el comportamiento del sistema en diferentes escenarios de uso y ajustar los algoritmos para lograr recomendaciones más certeras, priorizando tanto la relevancia como la posición estratégica de los contenidos sugeridos.

Además, se incluirá el análisis del CTR, es decir, el porcentaje de recomendaciones efectivamente seleccionadas por los usuarios, y la tasa de retención, entendida como el tiempo de visualización acumulado posterior a la interacción con las sugerencias. Estas métricas serán evaluadas inicialmente sobre datasets históricos anonimizados y, posteriormente, a través de pruebas controladas en entornos simulados.

Complementariamente, se aplicará una estrategia de evaluación cualitativa centrada en la percepción de los usuarios sobre las recomendaciones recibidas. Esta fase incluirá encuestas breves para indagar en el nivel de satisfacción, la percepción de relevancia y personalización del contenido sugerido, y la utilidad de las explicaciones provistas por el sistema. Esto se puede observar en el mockup propuesto en la Figura 20, que ilustra una interfaz de usuario diseñada para recopilar feedback cualitativo. En ella, se consulta al usuario sobre el nivel de satisfacción

con la recomendación recibida, la percepción de su relevancia y la utilidad de la explicación provista. Esta herramienta busca complementar las métricas algorítmicas con una evaluación subjetiva del impacto real de las recomendaciones, permitiendo detectar oportunidades de mejora desde la perspectiva del usuario final.

Figura 20
Interfaz de usuario para encuestas

Recomendado para vos

SERIE 1

SERIE 2

SERIE 3

Evaluación de recomendaciones

Nos gustaría conocer tu opinion sobre las recomendaciones. Por favor, responde las siguientes preguntas:

¿Que tan satisfecho estás con esta recomendación?

Muy insatisfecho ☐ ☐ ☐ ☐ ☒ Muy satisfecho

¿Que relevante es esta recomendación?

Muy insatisfecho ☐ ☐ ☐ ☒ Muy satisfecho

¿Te resultó útil la explicación?

Nada útil ☐ ☐ ☐ ☐ ☒ Muy útil

Nota: Esta interfaz recopila la valoración de los usuarios sobre las recomendaciones y sus explicaciones. Imágenes generadas con IA de Canva. Fuente: Propia.

Asimismo, dentro del componente de validación cualitativa, se tendrá en cuenta el impacto de los mecanismos de PLN aplicados a la interpretación de metadatos de contenido como sinopsis, etiquetas, nombres de actores y categorías temáticas, los cuales son fundamentales para enriquecer las recomendaciones personalizadas. Estos elementos semánticos no solo refuerzan la precisión algorítmica, sino que también inciden directamente en la percepción de coherencia y relevancia por parte del usuario.

Por otro lado, se buscará conocer el grado de confianza generado por el sistema de recomendación, especialmente en lo relativo a su capacidad de justificar por qué se sugiere determinado contenido, uno de los ejes fundamentales del componente de explicabilidad. Para

enriquecer esta perspectiva, se llevarán a cabo entrevistas semiestructuradas con un grupo reducido de usuarios, con el objetivo de comprender en mayor profundidad la experiencia de interacción con las recomendaciones y sus explicaciones.

Por último, se contempla la realización de A/B testings como parte del proceso experimental. Esta estrategia permitirá comparar el rendimiento de diferentes versiones del sistema en grupos de usuarios equivalentes. De este modo, se podrá observar el impacto diferencial de cada componente sobre variables como la satisfacción, el uso o la retención. Este enfoque mixto de validación, que integra elementos técnicos y experienciales, busca no solo demostrar la eficacia del sistema desde un punto de vista algorítmico, sino también confirmar su valor agregado para los usuarios finales en términos de confianza, transparencia y experiencia personalizada.

5. Propuesta de cronograma y recursos para la implementación

5.1 Consideraciones generales

La implementación del sistema de recomendación propuesto implica una planificación estratégica que articula tiempos, recursos técnicos y humanos. Dado que se trata de una solución diseñada para integrarse a plataformas de streaming operativas en el contexto argentino, se priorizará una arquitectura modular y escalable, tanto en lo metodológico como en lo tecnológico.

En esta etapa del proyecto no se contempla el desarrollo funcional completo del sistema, sino la elaboración de un diseño detallado y factible, que pueda servir como base para una implementación futura. Por lo tanto, el cronograma y los recursos que aquí se presentan deben entenderse como un plan referencial, orientado a demostrar la viabilidad técnica y metodológica del sistema. Se considera un ciclo de trabajo dividido en fases, donde cada una responde a un conjunto de objetivos específicos: análisis de requerimientos, diseño técnico, preparación de datos, modelado, evaluación, integración y documentación final. Este enfoque secuencial permitirá validar los supuestos conceptuales, identificar desafíos técnicos anticipadamente y garantizar la alineación entre los objetivos del sistema y los resultados esperados.

Además, se detallarán los recursos necesarios para cada etapa, incluyendo herramientas, entornos de desarrollo, servicios en la nube y perfiles técnicos recomendados. Esta planificación, basada en las decisiones metodológicas del capítulo anterior, busca garantizar la coherencia entre el diseño propuesto y su eventual implementación en una plataforma de streaming local.

5.2 Recursos requeridos

5.2.1. Recursos humanos

Para llevar adelante el desarrollo de la solución, se requerirá un equipo interdisciplinario con perfiles técnicos y de gestión. A continuación, se describen los roles necesarios:

- **Product Owner (PO):** representa los intereses del usuario y de la empresa. Será quien defina los requerimientos funcionales del sistema, priorice tareas en base a valor de negocio, valide entregables y garantice que la solución cumpla con los objetivos de experiencia de usuario y personalización.
- **Tech Lead:** responsable de coordinar técnicamente al equipo, tomar decisiones sobre la arquitectura, tecnologías y metodologías a emplear, y asegurar la calidad técnica del desarrollo. Supervisará la implementación de buenas prácticas, revisiones de código, testing automatizado y el cumplimiento de los plazos desde una perspectiva técnica. Actuará como puente entre los desarrolladores y el Product Owner, facilitando la planificación y la resolución de obstáculos técnicos.
- **QA Automation Engineer:** responsable de garantizar la calidad y estabilidad del sistema mediante la implementación de pruebas automatizadas orientadas a microservicios, APIs RESTful y flujos de recomendación. Este perfil deberá diseñar y mantener suites de pruebas funcionales, de regresión y de performance. También será clave en la validación de las salidas generadas por los modelos de XAI, asegurando la coherencia y claridad de las explicaciones ofrecidas al usuario. Además, participará en pruebas manuales exploratorias en etapas críticas, colaborando con el equipo de desarrollo para detectar, reportar y hacer seguimiento de incidencias antes de cada despliegue.
- **Desarrollador backend:** encargado del desarrollo del motor de recomendación, integración de APIs y lógica de negocio. Debe manejar frameworks como FastAPI o Flask, y tener experiencia en entornos de microservicios y Python.
- **Especialista en Machine Learning y PLN:** Responsable de la implementación de algoritmos de recomendación híbridos (colaborativos, basados en contenido y conocimiento), evaluación de métricas como precisión y recall, y gestión del ciclo de entrenamiento de modelos. Debe contar con experiencia en bibliotecas de aprendizaje automático, manejo de pipelines de modelado y herramientas de explicabilidad como SHAP o LIME.
- **Especialista en PLN:** será el encargado de interpretar y transformar los metadatos textuales del contenido (sinopsis, etiquetas, géneros, nombres de actores, etc.) en representaciones semánticas útiles para el perfilado de usuarios y la recomendación personalizada. Deberá aplicar técnicas como embeddings generados con modelos

preentrenados (por ejemplo, DistilBERT), análisis de similitud semántica y extracción de entidades clave, contribuyendo al enriquecimiento contextual del sistema de recomendación.

- **Ingeniero de datos:** encargado de la ingesta, procesamiento, limpieza y estructuración de datos. Debe manejar servicios como Google BigQuery, Cloud Storage, y herramientas de procesamiento como Dataflow.
- **Ingeniero DevOps:** responsable de la infraestructura en GCP, incluyendo el despliegue con Docker, orquestación con GKE, y monitoreo con herramientas como Cloud Monitoring y Logging. Este perfil garantizará la escalabilidad, resiliencia y disponibilidad del sistema, facilitando una integración continua efectiva.
- **Analista UX Researcher:** Este perfil será clave en la evaluación centrada en el usuario. Tendrá a su cargo el diseño e implementación de encuestas, entrevistas semiestructuradas y pruebas A/B para validar la utilidad, transparencia y nivel de confianza generado por el sistema de recomendación. Además, colaborará con el equipo técnico para analizar resultados cualitativos, identificar oportunidades de mejora y asegurar que el sistema cumpla con los principios de explicabilidad desde la perspectiva del usuario final.

5.2.2. Recursos técnicos

La propuesta fue diseñada para optimizar el uso de tecnologías open source y servicios escalables en la nube, con el objetivo de reducir las barreras de entrada y permitir una implementación gradual. Dada la naturaleza intensiva en datos y procesamiento del sistema de recomendación propuesto, se contempla una arquitectura robusta que permita manejar grandes volúmenes de información, responder en tiempo real y garantizar la escalabilidad futura.

Se tomó como referencia un escenario de uso correspondiente a una plataforma con aproximadamente 500.000 usuarios activos, cada uno de los cuales consume en promedio 4 horas diarias de contenido. Bajo esta hipótesis, se estimaron las capacidades mínimas necesarias para asegurar el funcionamiento fluido del sistema.

En la tabla 5, se detallan los componentes y una estimación mínima requerida para la implementación del sistema.

Tabla 5
Capacidades técnicas estimadas según recurso

Recurso	Función principal	Capacidad estimada
Vertex AI	Inferencias en tiempo real y ejecución de modelos de recomendación y XAI (SHAP, LIME)	Hasta 240 millones de inferencias mensuales (4/hora × 4 horas × 500.000 usuarios)
BigQuery	Análisis de datos estructurados y semiestructurados (logs, perfiles, eventos)	Procesamiento de hasta 5 TB diarios
Cloud Storage	Almacenamiento de historiales, metadatos, etiquetas, embeddings y backups	Volumen inicial estimado: 1 TB mensual
Dataflow	Procesamiento de eventos en tiempo real (clics, búsquedas, abandonos, etc.)	Configuración de 16 workers activos 24/7
Google Kubernetes Engine	Orquestación de microservicios (recomendación, perfilado, XAI, APIs)	Base de 20 nodos Autopilot con escalamiento automático
GPU (T4)	Entrenamiento de modelos de PLN y aprendizaje profundo (BERT, DistilBERT)	200 horas mensuales para generación de embeddings y análisis semántico de sinopsis y etiquetas
Monitoring & Logging	Observabilidad: métricas de desempeño, registros de eventos, detección de anomalías	Tableros operativos, alertas, trazabilidad de componentes críticos

Nota: Estimaciones de capacidad calculadas en función de un escenario con 500.000 usuarios activos, 6000 títulos y 4 horas diarias promedio de consumo. Basado en la documentación técnica de Google Cloud Platform. Fuente: Elaboración propia.

Las capacidades presentadas fueron estimadas en función de un escenario proyectado de 500.000 usuarios activos, un catálogo aproximado de 6.000 títulos entre series y películas, con un promedio de 4 horas diarias de consumo, según patrones de consumo analizados por Kantar Ibope (2024). Para el caso de Vertex AI se proyectaron 240 millones de inferencias mensuales, considerando 4 recomendaciones por hora activa en 500.000 usuarios.

Respecto al volumen proyectado para Cloud Storage, se consideraron múltiples fuentes de datos: embeddings semánticos generados por modelos como DistilBERT (≈ 3 KB por representación), historiales de navegación (≈ 300 bytes por evento), perfiles enriquecidos de usuario, backups periódicos y modelos entrenados. Solo el registro mensual de sesiones de navegación genera aproximadamente 600 millones de eventos, que comprimidos representan cerca de 60 GB. Al agregar los perfiles (≈ 2 GB), backups, logs y modelos (≈ 15 GB), se obtiene un volumen base de almacenamiento comprimido cercano a los 80 GB mensuales. Aplicando

un factor de proyección para escalabilidad, redundancia y retención histórica, se estima un volumen operativo de hasta 1 TB mensual.

Para Dataflow, se estimó una configuración de 16 workers activos 24/7 en base al volumen proyectado de eventos. Con 60 millones de sesiones mensuales y un promedio de 10 eventos por sesión, el sistema procesaría aproximadamente 600 millones de eventos por mes.

En cuanto a GKE, se planteó una base de 20 nodos Autopilot con escalamiento automático. Esto contempla una arquitectura desacoplada compuesta por microservicios especializados (motor de recomendación, perfilado, explicación XAI, APIs, autenticación, logging, monitoreo). Considerando un promedio de 2–3 pods por servicio crítico y la necesidad de asegurar alta disponibilidad y balance de carga, se definió esta configuración mínima como base operativa.

El uso de GPU NVIDIA T4 se justifica por la necesidad de procesar texto con modelos de procesamiento de lenguaje natural como DistilBERT. Cada contenido puede generar múltiples textos relevantes para análisis semántico: título, sinopsis, etiquetas, nombres de actores, géneros y versiones en distintos idiomas. Para un catálogo de 6.000 títulos, se estiman al menos 10 entradas de texto por contenido, lo que equivale a 60.000 textos únicos. Considerando nuevas incorporaciones mensuales, reentrenamientos parciales, pruebas experimentales y la ejecución periódica de variantes del modelo, se estima un volumen total de procesamiento de aproximadamente 250.000 textos mensuales. Procesar 50.000 textos con DistilBERT en una GPU T4 puede requerir entre 5 y 8 horas, dependiendo del tamaño de lote y las épocas de entrenamiento. Escalando a 250.000 textos y sumando overhead operativo, se proyecta un uso mensual de entre 150 y 250 horas de GPU. Se adopta como estimación conservadora una reserva de 200 horas mensuales de GPU T4, suficiente para cubrir la demanda base y asegurar margen para escalabilidad.

Finalmente, la capa de Monitoring & Logging incluye tableros de métricas, trazabilidad de eventos, alertas operativas y registros distribuidos para auditoría. Si bien su peso en almacenamiento es marginal comparado con los datos de usuario, su implementación es crítica para mantener la estabilidad del sistema. Se utilizarán herramientas nativas como Cloud

Logging, Cloud Monitoring que permiten recolectar y visualizar logs estructurados y KPIs técnicos en tiempo real.

5.3 Cronograma de implementación

El proceso de implementación del sistema de recomendación inteligente se estructurará en cinco etapas consecutivas, distribuidas a lo largo de cinco meses. Este cronograma responde a una planificación iterativa e incremental, diseñada para validar tempranamente funcionalidades clave, minimizar riesgos técnicos y garantizar la alineación con los objetivos generales del proyecto, en el marco del escenario definido en la sección anterior donde se contemplan 500.000 usuarios activos y 6.000 contenidos a modo de ejemplo.

Durante el primer mes se realizará un relevamiento exhaustivo de los requisitos técnicos y funcionales, con foco en el análisis de los datos disponibles, las restricciones de integración con plataformas de streaming locales y la definición definitiva del stack tecnológico. En paralelo, se diseñará la arquitectura general de la solución, incluyendo la estructura de microservicios, la estrategia de almacenamiento y el flujo de datos que integrará los distintos componentes del sistema.

En el segundo mes se configurarán los entornos técnicos en GCP, incluyendo la creación de buckets en Cloud Storage, instancias en BigQuery para almacenamiento y análisis, entornos de desarrollo en GKE y canalizaciones de procesamiento inicial con Dataflow. También se generarán datasets simulados de usuarios y contenidos para pruebas tempranas sin comprometer datos reales.

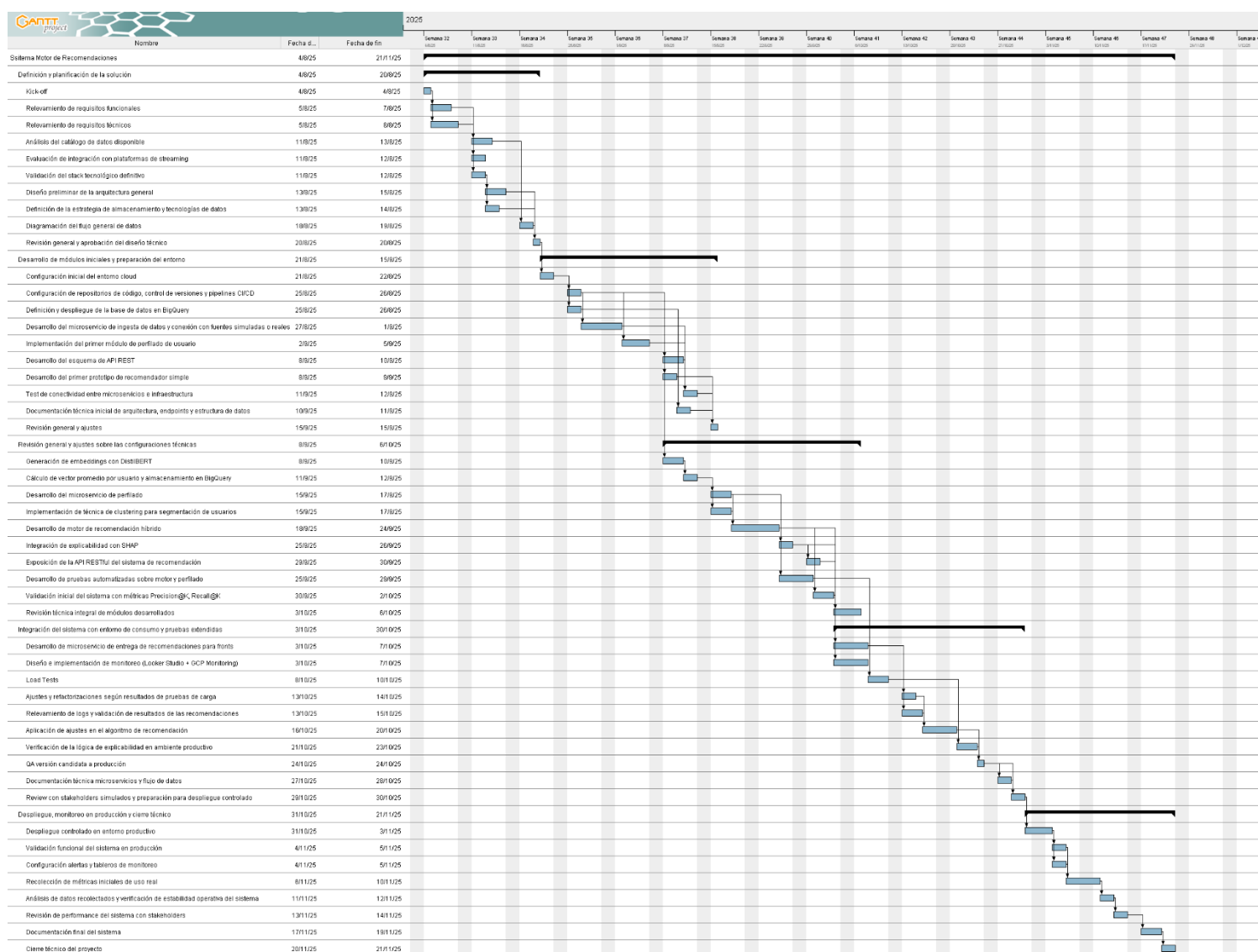
Durante el tercer mes se desarrollará el núcleo del sistema de recomendación, utilizando un enfoque híbrido que combine filtrado colaborativo, análisis de contenido y embeddings semánticos entrenados con DistilBERT. En esta fase se implementará el sistema de perfilado de usuarios, se desplegarán las APIs necesarias para la interacción con las interfaces, y se construirán los módulos de recomendación adaptativa.

El cuarto mes estará enfocado en la incorporación de técnicas de XAI, como SHAP y LIME, en la capa de inferencia del sistema. Estas herramientas permitirán brindar explicaciones comprensibles al usuario sobre las recomendaciones recibidas, fortaleciendo la transparencia y confianza. Se evaluará esta funcionalidad en términos de claridad, utilidad percibida y tiempos de respuesta, integrándola a la experiencia de usuario.

Finalmente, en el quinto mes se integrarán todos los microservicios, se entrenarán los modelos sobre los datos simulados y se validará el sistema completo. Se aplicarán métricas como Precision@K, Recall@K y F1-score para evaluar el desempeño del motor de recomendación. Además, se probarán los endpoints RESTful en múltiples entornos (web, móvil, Smart TV, decodificadores Android TV), garantizando su interoperabilidad. También se desplegarán los módulos de observabilidad con Cloud Monitoring y Logging, y se elaborarán dashboards de métricas clave en Looker Studio. El proyecto culminará con un informe técnico y recomendaciones para su implementación futura en producción.

Teniendo como referencia este desglose de etapas, se procede a detallar las tareas en el diagrama de Gantt que a continuación se expone:

Figura 21
Cronograma tentativo



Nota: Diagrama de Gantt del proyecto hasta la implementación productiva. Fuente: Elaboración propia

5.4 Propuesta económica de la solución

Desde el enfoque propuesto como modelo llave en mano, se contempla que el proveedor del sistema de recomendación asuma la totalidad de los costos asociados a infraestructura, licencias y recursos humanos, liberando a las plataformas de streaming de estas responsabilidades. Para ello, se ha estructurado una estimación detallada del presupuesto.

5.4.1. Costos de capital humano

La estimación de costos de recursos humanos se realizó en base a datos actualizados del mercado argentino en 2025. Los valores corresponden a perfiles en relación de dependencia, trabajando en el país y con experiencia acorde a las responsabilidades asignadas en el proyecto.

De acuerdo con un informe publicado por iProfesional (2025) sobre los sueldos en el sector de software en Argentina en 2025, el salario mensual de un PO oscila entre \$3.000.000 y \$5.000.000, dependiendo del nivel de experiencia y responsabilidad. Tomando como referencia un valor medio de \$4.000.000 mensuales, y considerando una cotización del dólar MEP de \$1.250, el salario mensual estimado asciende a USD 3.200.

El mismo informe señala que los diseñadores de producto, categoría que incluye a perfiles de UX Research, perciben sueldos que van desde \$1.800.000 hasta \$3.500.000 mensuales, en función del seniority. Considerando un valor promedio de \$2.650.000 mensuales, y aplicando la misma cotización del dólar MEP, el salario estimado es de aproximadamente USD 2.120 (iProfesional, 2025).

De acuerdo con otra publicación de iProfesional, los salarios para posiciones de liderazgo técnico en el sector IT, incluyendo jefes, gerentes y directores, parten desde los USD 3.750 actualizadas a la fecha teniendo en cuenta inflación y poder adquisitivo (iProfesional, 2024).

Según un artículo de Teclab (2025) sobre salarios en Argentina, en promedio un sueldo de un QA ronda entre \$2.400.000 y los \$3.500.000, dependiendo de experiencia y responsabilidad. Tomando un valor medio de \$2.950.000 mensuales, y aplicando una cotización del dólar MEP de \$1.250, el salario estimado alcanza los USD 2.360.

De acuerdo con iProfesional (2025), los sueldos promedio de programadores en Argentina alcanzan los ARS 2.289.000 mensuales, mientras que los perfiles senior, como desarrolladores backend y especialistas DevOps, superan los ARS 2.650.000 mensuales. Considerando una cotización del dólar MEP de ARS 1.250, esto equivale a aproximadamente

USD 1.830 mensuales para el desarrollador backend, y USD 2.120 mensuales para el perfil DevOps.

Según iProfesional (2025), en referencia a búsquedas abiertas para Quilmes Argentina los puestos en áreas tecnológicas (soporte, cloud, ciberseguridad, análisis de negocio y datos, inteligencia artificial y machine learning) para perfiles senior pueden alcanzar salarios de hasta \$4.000.000 mensuales. Considerando una cotización del dólar MEP de \$1.250, esto equivale a un salario en torno a USD 3.200 mensuales. Se adopta este valor como referencia para los perfiles de ML+PLN Specialist, PLN Specialist y Data Engineer, dado que sus responsabilidades (preprocesamiento de datos, entrenamiento de modelos híbridos, embeddings y pipelines de análisis semántico) entran de lleno en ese nivel técnico y salarial.

En la Tabla 6, se realiza el cálculo de los costos mensuales y totales del personal que se requiere para el proyecto.

Tabla 6
Costos del personal del proyecto

Perfil	Cantidad	USD/mes	Meses asignados	Subtotal (USD)
Product Owner (PO)	1	3.200	5	16.000
UX Researcher	1	2.120	5	10.600
Tech Lead	1	3.750	5	18.750
QA Automation Engineer	1	2.360	5	11.800
Desarrollador Backend	1	1.830	5	9.150
DevOps Engineer	1	2.120	5	10.600
ML+PLN Specialist	1	3.200	5	16.000
PLN Specialist	1	3.200	5	16.000
Data Engineer	1	3.200	5	16.000
Total				124.900

Nota: Los salarios fueron calculados en base a información publicada por iProfesional (2025) y Teclab (2025), con una cotización de referencia del dólar MEP de ARS 1.250. Se consideraron perfiles en relación de dependencia con experiencia senior, en el mercado argentino. Fuente. Elaboración propia a partir de datos del mercado laboral argentino 2025.

5.4.2. Costos técnicos

La estimación de los costos técnicos de la solución se realizó en base a precios públicos de GCP vigentes en 2025. El motor de inferencias y explicabilidad estará desplegado sobre

Vertex AI, cuya tarifa para nodos dedicados a predicción en tiempo real ronda los USD 0.75 por hora (Google Cloud, 2025). Dado el volumen proyectado de hasta 240 millones de inferencias mensuales, incluyendo el uso de técnicas como SHAP y LIME, se estimó un costo mensual de aproximadamente USD 800.

El análisis de logs, eventos de navegación y perfiles de usuario se realizará mediante BigQuery, con un costo estimado de USD 23.55 por terabyte de almacenamiento activo (Google Cloud, 2025). Considerando un procesamiento diario de hasta 5 TB y un volumen activo mensual de 25 TB, el presupuesto mensual asignado a este servicio es de USD 600.

Para el almacenamiento de datos no estructurados como sinopsis enriquecidas, historiales de consumo y backups periódicos, se utilizará Cloud Storage en clase estándar, cuyo precio ronda los USD 0.02 por GB al mes (Cloudwards, 2025). Se proyecta un volumen de entre 6 y 10 TB mensuales, justificando un gasto estimado de USD 120 mensuales.

El entrenamiento de modelos de procesamiento de lenguaje natural, como DistilBERT, requerirá el uso de GPUs NVIDIA T4. Este tipo de recurso presenta un costo promedio de USD 0.35 por hora de uso en GCP (DEV Community, 2024). Se estiman 200 horas de cómputo mensual, equivalentes a un costo de USD 350.

El pipeline de datos en tiempo real se implementará utilizando Google Cloud Dataflow. Si bien el costo varía según la cantidad y tipo de workers activos, se adoptó una estimación conservadora de USD 400 mensuales, basada en escenarios recomendados por la propia documentación de GCP para arquitecturas con alto caudal de eventos (Google Cloud, s.f.).

Los microservicios del sistema serán orquestados con GKE en modalidad Autopilot, que cobra por los recursos utilizados y el mantenimiento del clúster. Para el tamaño proyectado de este sistema, se estimó un costo mensual de USD 600 (Cloudchipr, 2025).

Finalmente, los servicios de observabilidad, como monitoreo de rendimiento, trazabilidad de eventos y alertas de error, se cubrirán con las herramientas nativas de Google Cloud Operations Suite. Estos servicios ofrecen planes gratuitos y escalables según uso, es por eso que, durante la etapa de implementación, se usará esta modalidad.

Tabla 7
Costos técnicos estimados de infraestructura

Servicio	Descripción	Costo mensual (USD)	Costo total 5 meses (USD)
Vertex AI	Inferencias en tiempo real y ejecución de modelos XAI	800	4.000
BigQuery	Análisis de logs, eventos y perfiles de usuario	600	3.000
Cloud Storage	Almacenamiento de sinopsis, historiales y backups	120	600
GPU NVIDIA T4	Entrenamiento de modelos PLN (DistilBERT)	350	1.750
Dataflow	Procesamiento en tiempo real de eventos de usuario	400	2.000
Google Kubernetes Engine (GKE)	Orquestación de microservicios en modalidad Autopilot	600	3.000
Total estimado			14.350

Nota. Los costos estimados corresponden a valores de referencia públicos provistos por Google Cloud Platform en 2025. Se calculó un uso intensivo pero optimizado de cada servicio, considerando el volumen estimado de usuarios y eventos. Fuente: Elaboración propia con base en Google Cloud (2025), Cloudwards (2025), DEV Community (2024) y Cloudchpr (2025).

5.4.3. Evaluación integral de costos y modelo comercial

El análisis de costos permite dimensionar los recursos necesarios para la implementación del sistema de recomendación inteligente propuesto. La inversión principal se concentra en los perfiles técnicos especializados requeridos para el desarrollo, despliegue y validación del sistema, con un total estimado de USD 124.900 en cinco meses. A esto se suman los gastos derivados de la infraestructura técnica en GCP, optimizada para un escenario de 500.000 usuarios activos, con un costo proyectado de USD 14.350. Considerando ambos rubros, el subtotal asciende a USD 139.250.

Para asegurar la viabilidad comercial del proyecto y contemplar un margen de rentabilidad adecuado, se aplica un recargo del 30 %, equivalente a USD 41.775. De esta manera, el costo total del proyecto para la empresa contratante asciende a USD 181.025. Esto se puede observar con mejor detalle en la Tabla 8.

Tabla 8*Resumen total del costo del proyecto*

Concepto	Costo (USD)
Recursos humanos	124.900
Infraestructura técnica	14.350
Subtotal	139.250
Margen comercial (30 %)	41.775
Costo total del proyecto	181.025

Nota. El margen comercial aplicado es del 30 % sobre el subtotal de recursos humanos e infraestructura. Fuente: Elaboración propia.

La modalidad de pago propuesta consiste en el desembolso del 50 % del monto total al inicio del proyecto, es decir, USD 90.512,50, y el resto distribuido en cuotas mensuales durante los cinco meses de ejecución. Esta estructura busca garantizar previsibilidad financiera para ambas partes y facilitar el seguimiento de avances y entregables en cada etapa del desarrollo.

5.5 Analisis de riesgos

5.5.1. Factores externos condicionantes

La implementación de un sistema de recomendación inteligente para plataformas de streaming en Argentina está sujeta a diversos factores externos que pueden incidir directa o indirectamente en su viabilidad, adopción y efectividad. Estos factores deben ser tenidos en cuenta en la planificación y en la propuesta de valor para las empresas del sector.

Uno de los principales condicionantes es el nivel de conectividad y calidad de acceso a internet en los hogares argentinos. Si bien los datos más recientes de CABASE (2024) indican que casi 4 de cada 10 conexiones fijas superan los 100 Mbps, todavía persiste una brecha digital significativa en algunas regiones del país. Esta heterogeneidad puede impactar en la experiencia del usuario final, especialmente en lo que refiere a la capacidad de las plataformas para procesar y entregar recomendaciones en tiempo real, y a la velocidad de respuesta de los servicios basados en la nube.

Otro factor externo clave es el nivel de madurez tecnológica de las plataformas de streaming locales, muchas de las cuales operan con arquitecturas heredadas que pueden

dificultar la integración de microservicios avanzados o APIs externas. La propuesta debe ser compatible con una amplia gama de entornos técnicos, lo que implica diseñar un sistema modular y desacoplado, preparado para interactuar con infraestructuras heterogéneas.

En tercer lugar, las regulaciones locales en materia de protección de datos personales constituyen un factor crítico. La Ley 25.326 de Protección de Datos Personales (Ministerio de Justicia de la Nación, 2000) establece principios y obligaciones para el tratamiento de datos en Argentina, incluyendo el consentimiento informado, la finalidad del uso de datos y la confidencialidad. La solución propuesta incorpora medidas para garantizar la privacidad y trazabilidad del uso de datos, en conformidad con dicha normativa, lo que también constituye un valor diferencial frente a otras propuestas menos adaptadas al contexto legal argentino.

Por último, deben considerarse las dinámicas competitivas del mercado local. El crecimiento de nuevas plataformas OTT, la entrada de actores internacionales con inversiones agresivas y la creciente exigencia de los consumidores en cuanto a personalización, hacen que la capacidad de diferenciarse mediante recomendaciones inteligentes y explicables se convierta en un factor estratégico clave. En este sentido, la solución propuesta no solo apunta a mejorar la experiencia del usuario, sino también a dotar a las plataformas locales de una herramienta de competitividad frente a servicios globales.

5.5.2. Factores internos condicionantes

Además de los factores externos, existen condicionantes internos que pueden afectar la implementación y éxito del sistema de recomendación inteligente propuesto. Estos aspectos están relacionados con la capacidad operativa, técnica y organizacional del equipo desarrollador, así como con las decisiones estratégicas asumidas en la planificación del proyecto.

Uno de los principales factores internos es la composición y experiencia del equipo técnico. El éxito del sistema depende en gran medida de la correcta coordinación entre los distintos perfiles especializados dentro del equipo. La complejidad técnica de la solución requiere una comunicación fluida entre áreas, una clara definición de roles y una gestión técnica

efectiva liderada por el Tech Lead, lo cual fue contemplado en el diseño organizativo del proyecto. Además, esta coordinación debe estar alineada con los requerimientos y priorizaciones que son responsabilidad pura del PO del equipo.

Otro aspecto relevante es la madurez metodológica del proceso de desarrollo. La implementación contempla una planificación iterativa, con validaciones parciales en cada fase, lo que contribuye a minimizar riesgos. Sin embargo, el cumplimiento del cronograma depende de la disciplina en la ejecución de tareas, la correcta gestión del backlog y la adecuada priorización de funcionalidades. El uso de metodologías ágiles, como SCRUM, será clave para asegurar entregas incrementales y adaptabilidad a posibles cambios.

El tercer factor interno está vinculado a la gestión eficiente de los recursos tecnológicos, especialmente en lo relativo a la infraestructura en la nube. Aunque se optó por servicios escalables y mayormente open source, puede implicar costos operativos relevantes si no se realiza un monitoreo proactivo del consumo. La implementación de alertas automáticas y dashboards de uso será una herramienta esencial para mantener el control presupuestario.

Por último, un elemento condicionante fundamental es el alineamiento estratégico entre el diseño del sistema y los objetivos del negocio. El sistema de recomendación no es una funcionalidad aislada, sino un componente que debe integrarse con la lógica comercial, editorial y operativa de las plataformas de streaming. Para lograrlo, será clave involucrar tempranamente a las áreas de producto, UX y contenido, asegurando que las recomendaciones sean coherentes con la identidad y estrategia de cada plataforma.

6. Conclusiones

A lo largo de este trabajo se ha abordado la problemática del descubrimiento de contenido en plataformas de streaming, resaltando cómo la inteligencia artificial, especialmente los sistemas de recomendación personalizados, puede mejorar la experiencia del usuario y, al mismo tiempo, aportar valor estratégico para las empresas del sector. En un ecosistema digital saturado de opciones, donde la atención del usuario es limitada y volátil, los motores de recomendación representan una herramienta clave para conectar de manera eficiente la oferta con la demanda de contenidos.

El recorrido teórico incluyó un análisis de los fundamentos de la IA, el aprendizaje automático, el DL, PLN, XAI, y los distintos enfoques de recomendación (colaborativos, basados en contenido e híbridos). También se exploraron los aspectos éticos, regulatorios y contextuales, tanto a nivel global como en el escenario argentino, donde se evidencian oportunidades, pero también desafíos significativos en términos de infraestructura, políticas públicas y formación de talento especializado.

Desde una perspectiva técnica, se propuso una arquitectura basada en tecnologías de Google Cloud Platform, capaz de procesar datos en tiempo real y generar recomendaciones personalizadas con explicaciones comprensibles para el usuario. La propuesta metodológica se apoyó en la construcción de perfiles dinámicos, el uso de técnicas de clustering, PLN para análisis semántico, y modelos XAI para justificar las recomendaciones.

Como valor agregado, se subrayó la necesidad de combinar precisión algorítmica con transparencia y respeto por los derechos de los usuarios, entendiendo que la confianza es un componente central en la adopción de sistemas de IA. En este sentido, la incorporación de principios éticos, privacidad diferencial y mecanismos de explicabilidad se plantea no solo como una obligación normativa, sino también como una oportunidad para construir sistemas más humanos, confiables y sostenibles.

Finalmente, si bien este trabajo propone una arquitectura y una metodología concreta, no se implementó de forma práctica en el entorno actual de la empresa. Esta decisión se tomó para mantener un enfoque teórico-metodológico con valor proyectivo. Por tanto, se deja abierta la posibilidad de desarrollar prototipos o pilotos en futuras investigaciones, así como explorar

mejoras vinculadas al perfilado emocional del usuario, la integración con tecnologías de voz o la generación automática de metadatos a través de interacciones conversacionales.

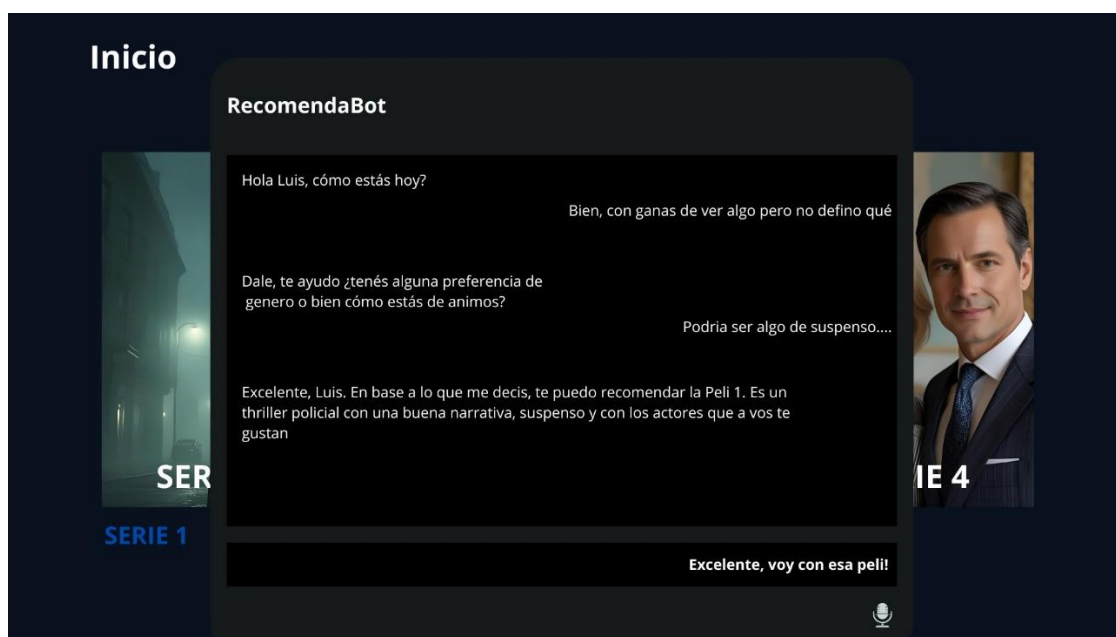
7. Trabajos futuros

7.1 Interacción conversacional basada en lenguaje natural

Una posible línea de desarrollo futuro es la incorporación de una interfaz conversacional que permita a los usuarios interactuar con el sistema de recomendación utilizando lenguaje natural. Esta funcionalidad, basada en técnicas avanzadas de PLN, habilitaría un diálogo dinámico con el sistema, facilitando la expresión de gustos, estados de ánimo o necesidades específicas sin depender de filtros manuales o búsquedas tradicionales.

Su incorporación requeriría nuevas capas de modelado semántico, mecanismos de interpretación de intención del usuario y generación de respuestas contextuales. Sin embargo, representa una oportunidad para enriquecer la experiencia de uso, reduciendo la fricción en la exploración del contenido y capturando señales cualitativas difíciles de modelar con enfoques clásicos. En la Figura 22 se presenta un ejemplo simulado de cómo podría verse este tipo de interacción en una interfaz conversacional integrada en una plataforma de streaming. Este recurso visual busca ilustrar conceptualmente la experiencia esperada, dejando abierta esta línea de trabajo para futuras implementaciones o investigaciones aplicadas.

Figura 22
Simulación de una interfaz conversacional



Nota: Mockup propuesto donde se podría implementar una conversación con la plataforma, usando lenguaje natural con el fin de recomendar contenidos. Imágenes generadas con IA de Canva. Fuente propia

7.2 Privacidad diferencial y preservación de datos personales en sistemas de recomendación

Otra línea relevante para futuras propuestas y desarrollos en sistemas de recomendación es la incorporación de enfoques avanzados de preservación de la privacidad, particularmente mediante técnicas de privacidad diferencial. Dado que estos sistemas requieren recopilar, procesar y analizar grandes volúmenes de datos de usuarios, incluyendo historiales de visualización, patrones de comportamiento y preferencias explícitas e implícitas, surge la necesidad de garantizar que dicha información no pueda ser utilizada de forma indebida o reidentificada.

Para preservar la privacidad de los datos, la respuesta verdadera se modifica mediante la incorporación de ruido aleatorio generado a partir de una distribución seleccionada estratégicamente. El resultado que recibe el usuario es la combinación de dicha respuesta con el ruido añadido (Dwork et al., 2006). Esto protege la identidad y comportamiento del usuario sin comprometer de manera sustancial la calidad del modelo.

Si bien la propuesta actual no contempla la implementación de esta técnica, se reconoce que su aplicación representa un desafío técnico y computacional importante, pero también una vía necesaria para aumentar la confianza del usuario, el cumplimiento regulatorio y la adopción sostenible de soluciones basadas en inteligencia artificial. Por tanto, se sugiere explorar en trabajos futuros la integración de mecanismos de anonimización robusta, privacidad diferencial y otras prácticas de seguridad computacional que permitan conciliar personalización y privacidad, minimizando los riesgos inherentes al manejo de datos sensibles en entornos digitales de alta escala.

8. Acrónimos

API	AI Preparedness Index
API	Application Programming Interface
CNN	Redes Neuronales Convolucionales
CTR	Click Through Rate
DCG	Ganancia acumulada descontada
DP	Deep Learning
GCP	Google Cloud Platform
GKE	Google Kubernetes Engine
IA	Inteligencia Artificial
IR	Recuperación de Información
KPI	Indicadores Clave de Rendimiento
LIME	Local Interpretable Model-Agnostic Explanations
LLM	Lenguajes de Modelos Grandes
MAP	Precisión Media Promedio
ML	Machine Learning
MUC	Message Understanding Conference
NDCG	Normalized Discounted Cumulative Gain
OCDE	Organización para la Cooperación y el Desarrollo Económicos
OTT	Over The Top
PLN	Procesamiento de Lenguaje Natural
PO	Product Owner
RNN	Redes Neuronales Recurrentes
SHAP	SHapley Additive Explanations
TASS	Taller de Análisis Semántico

UX	User Experience
XAI	Inteligencia Artificial Explicable

9. Referencias

- Accenture. (2022). *El streaming del futuro*. <https://newsroom.accenture.es/es/news/2022/el-streaming-del-futuro>
- Adobe for Business. (2021). *Recommendation Engine*. <https://business.adobe.com/blog/basics/recommendation-engine>
- Ali, M., Hosain, M., & Hossain, M. (2021). Big data analysis using BigQuery on cloud computing platform. *Australian Journal of Engineering and Innovative Technology*, 3(1), 1-9. <https://doi.org/https://doi.org/10.34104/ajeit.021.0109>
- Ambassador Labs. (2024). *How APIs Power Microservices Architectures 101 Guide*. <https://www.getambassador.io/blog/apis-microservices-architectures-guide>
- Ampere Analysis. (2025). *Streaming algorithms now beat word-of-mouth for TV and movie picks*. https://www.ampereanalysis.com/insight/streaming-algorithms-now-beat-word-of-mouth-for-tv-and-movie-picks?utm_content=328588583&utm_medium=social&utm_source=twitter&hss_channel=tw-2832088119
- Broadcast PRO. (2024). *Global OTT subscription revenues to touch \$190bn with over two billion subscribers by 2029*. <https://www.broadcastprome.com/news/global-ott-subscription-revenues-to-touch-190bn-with-over-two-billion-subscribers-by-2029/>
- CABASE. (2024). *Casi 4 de cada 10 conexiones a internet en hogares argentinos supera los 100 Mbps*. <https://www.cabase.org.ar/casi-4-de-cada-10-conexiones-a-internet-en-hogares-argentinos-supera-los-100-mbps/>
- Castillo, J. (2025). *La Opinion*. <https://laopinion.com/2025/04/11/netflix-apuesta-por-la-inteligencia-artificial-de-openai-para-revolucionar-su-sistema-de-busqueda/>
- Cloudchipr. (2025). *GKE Pricing Explained: What You Need to Know for Cost-Effective Kubernetes*. <https://cloudchipr.com/blog/gke-pricing>

- Cloudwards. (2025). *Google Cloud Review*. <https://www.cloudwards.net/google-cloud-review/>
- Cuéllar Hidalgo, R. (2024). *spaCy: una herramienta para realizar Procesamiento del Lenguaje Natural de manera amigable*. <https://bdcv.hypotheses.org/5507>
- DEV Community. (2024). *Understanding Google Cloud Platform Pricing | AI/ML Pricing on Google Cloud Platform*. <https://dev.to/ddeveloperr/understanding-google-cloud-platform-pricing-gcp-pricing-59h4>
- Developer Nation. (2023). *Pulse Report*. <https://www.developernation.net/developer-reports/dn24/>
- Dwork, C., McSherry, F., Nissim, K., & Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. In S. Halevi, & T. Rabin, *Theory of Cryptography Conference, TCC 2006* (Vol. 3876, pp. 265-284). Berlin, Germany: Springer. https://doi.org/https://doi.org/10.1007/11681878_14
- Educaopen. (s.f.). *Explicabilidad*. <https://www.educaopen.com/digital-lab/metaterminos/e/explicabilidad>
- Elastic. (s.f.). *Elasticsearch*. Elastic: <https://www.elastic.co/es/elasticsearch>
- Evidently AI. (2025). *Ranking and Recommendation Metrics Guide. Mean Average Precision (MAP) in ranking and recommendations*. <https://www.evidentlyai.com/ranking-metrics/mean-average-precision-map>
- Fieiras-Ceide, C., Ufarte-Ruiz, M., & Murcia-Verdú, F. (2024). *Transformando la experiencia de usuario en televisión en streaming: prioridades de innovación y tecnología en plataformas OTT*. Ediciones Complutense. <https://doi.org/https://doi.org/10.5209/emp.96367>
- GitHub. (2024). *Octoverse: AI leads Python to top language as the number of global developers surges*. https://github.blog/news-insights/octoverse/octoverse-2024/?utm_source=chatgpt.com

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.

Google Cloud. (2023). *Cloud Logging*. <https://cloud.google.com/logging?hl=es-419>

Google Cloud. (2024). *Aprovecha la escalabilidad horizontal*. <https://cloud.google.com/architecture/framework/reliability/horizontal-scalability?hl=es-419>

Google Cloud. (2024). *Observabilidad de GKE*. https://cloud.google.com/kubernetes-engine/docs/concepts/observability?utm_source=chatgpt.com&hl=es-419

Google Cloud. (2025). *BigQuery pricing*. <https://cloud.google.com/bigquery/pricing>

Google Cloud. (2025). *Introducción a Vertex Explainable AI*. <https://cloud.google.com/vertex-ai/docs/explainable-ai/overview?hl=es-419>

Google Cloud. (2025). *Vertex AI pricing*. https://cloud.google.com/vertex-ai/pricing?utm_source=chatgpt.com&hl=es-419

Google Cloud. (s.f.). *Cloud Monitoring*. <https://cloud.google.com/monitoring?hl=es-419>

Google Cloud. (s.f.). *Dataflow pricing*. <https://cloud.google.com/dataflow/pricing?hl=en>

Google Cloud. (s.f.). *Dataflow. Inteligencia de datos en tiempo real*. <https://cloud.google.com/products/dataflow?hl=es-419>

Google Cloud. (s.f.). *De almacén de datos a plataforma de IA y datos autónomos*. <https://cloud.google.com/bigquery?hl=es-419>

Google Cloud. (s.f.). *Google Kubernetes Engine (GKE)*. <https://cloud.google.com/kubernetes-engine?hl=es-419>

Google Cloud Platform. (2024). *Pricing calculator*. Pricing calculator: <https://cloud.google.com/products/calculator>

Google Cloud. (s.f.). *Serverless computing*. <https://cloud.google.com/serverless?hl=es-419>

- Google Developers. (2025). *Clasificación: Exactitud, recuperación, precisión y métricas relacionadas*. <https://developers.google.com/machine-learning/crash-course/classification/accuracy-precision-recall?hl=es-419>
- Govea, J., Gutierrez, R., & Villegas-Ch, W. (2024). *Transparency and precision in the age of AI: evaluation of explainability-enhanced recommendation systems*. *Frontiers in Artificial Intelligence*. <https://doi.org/https://doi.org/10.3389/frai.2024.1410790>
- Gulati, J. (2025). *Using DistilBERT for resource-efficient natural language processing*. KDNuggets: <https://www.kdnuggets.com/distilbert-resource-efficient-natural-language-processing>
- Hong, Z. (2024). *Personalized Recommendations: How Netflix and Amazon Use Deep Learning to Enhance User Experience*. <https://medium.com/@zhonghong9998/personalized-recommendations-how-netflix-and-amazon-use-deep-learning-to-enhance-user-experience-e7bd6fcd18ff>
- Hsiao, K.-J., & Lamkhede, S. (2025). *Building a Foundation Model for Recommendations*. <https://netflixtechblog.com/foundation-model-for-personalized-recommendation-1a0bd8e02d39>
- Huang, X., Lian, J., Lei, Y., Yao, J., Lian, D., & Xie, X. (2023). *Recommender AI Agent: Integrating Large Language Models for Interactive*. University of Science and Technology of China. <https://doi.org/https://arxiv.org/pdf/2308.16505>
- IBM. (s.f.). *¿Qué es la IA explicable?* <https://www.ibm.com/es-es/think/topics/explainable-ai>
- IBM. (2021). *¿Qué es una IA fuerte?* IBM Think: <https://www.ibm.com/mx-es/think/topics/strong-ai>
- IBM. (2024). *¿Qué es el deep learning?* IBM Think: <https://www.ibm.com/es-es/think/topics/deep-learning>
- IBM. (2024). *¿Qué es un motor de recomendación?* <https://www.ibm.com/mx-es/think/topics/recommendation-engine>

- Iglesias Albores, E. L. (2022). *Netflix: análisis comparativo del consumo de los usuarios antes y durante la pandemia*. Universidad Autónoma de Barcelona. Universidad del Rosario.
- Infobae. (2025). *Netflix cambiaría completamente: están probando una IA de OpenAI para todos los usuarios*. <https://www.infobae.com/tecno/2025/04/14/netflix-cambiaria-completamente-estan-probando-una-ia-de-openai-para-todos-los-usuarios/>
- International Monetary Fund. (2023). *AI Preparedness Index*. https://www.imf.org/external/datamapper/AI_PI@AIPI/ADVEC/EME/LIC
- International Monetary Fund. (s.f.). *AI Preparedness Index (AIPI)*. <https://www.imf.org/external/datamapper/datasets/AIPI>
- iProfesional. (2024). *Los sueldos del mercado IT en Argentina: cuánto ganan por mes jefes, gerentes y directores*. <https://www.iprofesional.com/management/402524-sueldos-del-mercado-it-cuanto-ganan-jefes-gerentes-y-directores>
- iProfesional. (2025). *Los buenos sueldos que pagan las empresas de software en Argentina y los tres perfiles más buscados*. <https://www.iprofesional.com/management/422473-excelente-sueldo-que-pagan-empresas-software-en-argentina-2025>
- iProfesional. (2025). *Quilmes busca empleados en Argentina y ofrece atractivo sueldo*. <https://www.iprofesional.com/negocios/422052-quilmes-busca-empleados-argentina-ofrece-atractivo-sueldo>
- iProfesional. (2025). *Sueldos: ¿cuánto cobra un programador en la Argentina?* <https://www.iprofesional.com/management/422431-sueldos-cuanto-cobra-un-programador-en-la-argentina>
- Kantar IBOPE. (2024). *Inside Video 2024 Argentina*. <https://kantaribopemedia.com/argentina/wp-content/uploads/sites/3/2024/04/Inside-Video-Argentina-2024.pdf>

- Lundberg, S., & Su-In, L. (2014). A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*, 30, pp. 4765–4774. <https://doi.org/10.48550/arXiv.1705.07874>
- McKinsey & Company. (2023). *What is personalization?* <https://www.mckinsey.com/featured-insights/mckinsey-explainers/what-is-personalization>
- Microsoft. (2025). *Best practices for RESTful web API design*. <https://learn.microsoft.com/es-mx/azure/architecture/best-practices/api-design>
- Ministerio de Justicia de la Nación. (2000). *Ley de Datos Personales*. <https://www.argentina.gob.ar/justicia/derechofacil/leysimple/datos-personales>
- Muñoz Guillena, R. (2024). *Procesamiento de lenguaje natural*. Editorial Académica Española.
- Nielsen. (2023). *Data-driven personalization: The future of streaming content discovery*. <https://www.nielsen.com/es/insights/2023/data-driven-personalization-2023-state-of-play-report/>
- Nilsson, N. (2009). *The Quest for artificial intelligence. A history of ideas and achievements*. Cambridge University Press. <https://doi.org/http://www.cambridge.org/us/0521122937>
- OECD. (2019). *AI principles*. <https://www.oecd.org/en/topics/ai-principles.html>
- Portal de datos abiertos del Gobierno de España. (2025). *La inteligencia artificial explicable (XAI): cómo los datos abiertos pueden ayudar a entender los algoritmos*. <https://datos.gob.es/es/blog/la-inteligencia-artificial-explicable-xai-como-los-datos-abiertos-pueden-ayudar-entender-los>
- Ribeiro, M., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144. <https://doi.org/https://doi.org/10.1145/2939672.2939778>

- Russell, S., & Norvig, P. (2004). *Inteligencia artificial. Un enfoque moderno*. Pearson Prentice Hall.
- Sanker, A. (2025). *Filtrado Colaborativo: Tu guía para recomendaciones más inteligentes*.
<https://www.datacamp.com/es/tutorial/collaborative-filtering>
- Simon-Kucher. (2024). *Unlock the future of streaming*. <https://www.simon-kucher.com/en/insights/streaming-study-2024>
- Spotify. (2023). Spotify debuts a new AI DJ, right in your pocket.:
<https://newsroom.spotify.com/2023-02-22/spotify-debuts-a-new-ai-dj-right-in-your-pocket/>
- Spotify. (2024). ¡Hola, DJ Livi! Nuestra voz de IA ahora habla español:
<https://newsroom.spotify.com/2024-07-17/spanish-ai-dj-livi-voice/>
- Spotify. (2024). *Spotify presenta AI DJ en español para los fans de la música en todo el mundo*.
<https://newsroom.spotify.com/2024-07-17/ai-dj-espanol-livi/>
- Teclab. (2025). *¿Qué hace y cuánto gana un Quality Assurance?*
<https://teclab.edu.ar/general/que-hace-y-cuanto-gana-un-quality-assurance/>
- Turing. (2022). *A Guide to Content-Based Filtering In Recommender Systems*.
<https://www.turing.com/kb/content-based-filtering-in-recommender-systems>
- Turing, A. (1950). *Computing machinery and intelligence*. Mind.
<https://doi.org/https://doi.org/10.1093/mind/LIX.236.433>
- UNESCO. (2021). *Inteligencia artificial y educación. Guía para las personas a cargo de formular políticas*. <https://unesdoc.unesco.org/ark:/48223/pf0000379376>
- UNESCO. (2021). *Recomendación sobre la ética de la inteligencia artificial*.
https://unesdoc.unesco.org/ark:/48223/pf0000380455_spa/PDF/380455spa.pdf.multi
- Wolf, A. (2023). *Machine Learning Simplified*. Themlsbook.

