UNIVERSIDAD NACIONAL MAYOR DE SAN MARCOS

Facultad de Ingeniería de Sistemas e informática

DISEÑO Y EVALUACIÓN DE UN SISTEMA DE APRENDIZAJE AUTOMÁTICO DISTRIBUIDO PARA MEJORAR LA EFICIENCIA EN SISTEMAS DE RECOMENDACIONES DE PRODUCTOS A CLIENTES DE ECOMMERCE

Tesis para obtener el Título de:

Ingeniero de Sistemas

Presentado por

Ñacari Elescano Alan Jesus Valentino

Lima – Perú

Marzo-2023

Contenido

LISTA DE FIGURAS

LISTA DE TABLAS

CAPÍTULO 1: INTRODUCCIÓN

Situación Problemática

En la era digital actual, la cantidad de datos generados por las organizaciones ha aumentado exponencialmente, Según un informe de IBM, se generan aproximadamente 2.5 quintillones de bytes de datos cada día, y se espera que la cantidad de datos generados por las organizaciones se duplique cada dos años. [1]

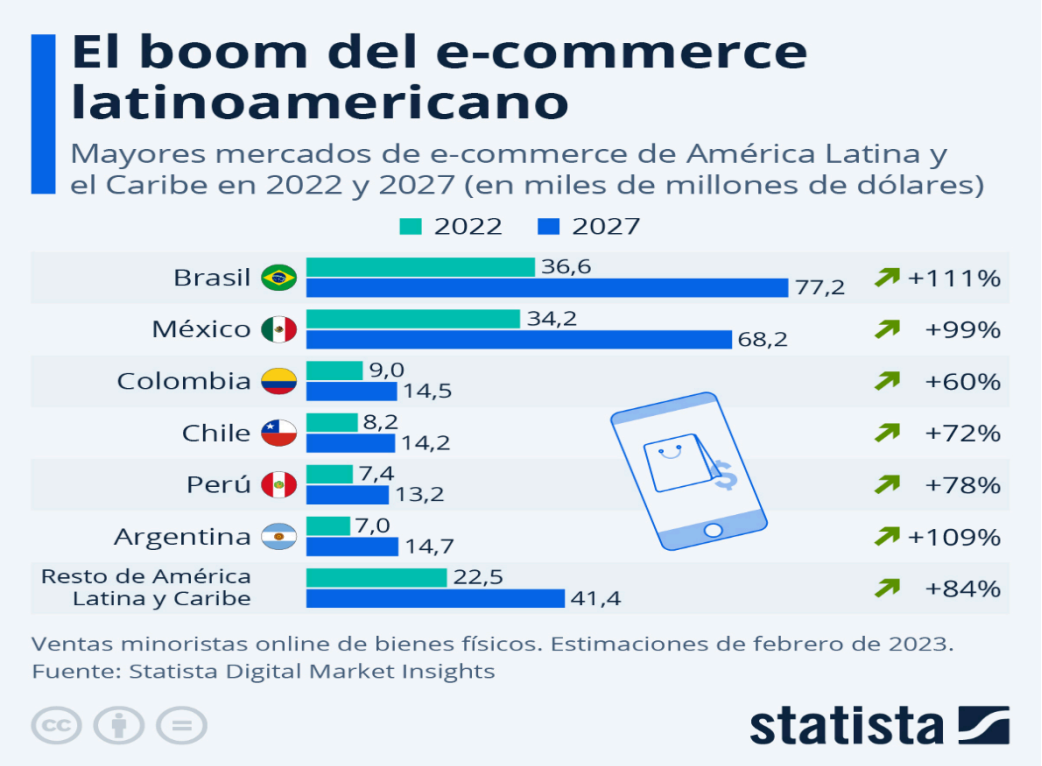


Figura 1: *Estimación del Crecimiento en el comercio electrónico en Latinoamérica (2023)*

Fuente: *Statista Digital Market Insights*

La cantidad de datos generados en tiempo real en las plataformas de comercio electrónico es enorme y puede ser difícil de manejar y analizar de manera efectiva. Además, los clientes en línea esperan una experiencia personalizada y relevante en la plataforma de comercio electrónico, lo que requiere un análisis en tiempo real de los datos del cliente y la generación de recomendaciones precisas de productos. Sin embargo, el análisis y la generación de recomendaciones a menudo se realizan de forma centralizada, lo que puede llevar a una mayor latencia y a una experiencia de usuario insatisfactoria.

Los sistemas tradicionales de análisis de datos pueden no ser suficientes para procesar los grandes conjuntos de datos en tiempo real. Además, los sistemas tradicionales de análisis de datos son inadecuados para procesar los grandes conjuntos de datos en tiempo real, lo que hace necesario el uso de soluciones de aprendizaje automático distribuido [2].

Por lo tanto, hay una necesidad de diseñar y evaluar un sistema de aprendizaje automático distribuido para el análisis de grandes conjuntos de datos en tiempo real en el comercio electrónico, que permita la generación de recomendaciones precisas y personalizadas de productos para los clientes. Este sistema podría ayudar a las ecommerce de comercio electrónico a mejorar la experiencia del usuario y aumentar las ventas, al tiempo que permite un análisis más efectivo y eficiente de los grandes conjuntos de datos generados en tiempo real en sus plataformas.

Formulación del Problema

Problema General

¿Cómo diseñar y evaluar un sistema de aprendizaje automático distribuido que permita analizar grandes conjuntos de datos en tiempo real y ofrecer recomendaciones de productos personalizadas a clientes de una ecommerce, con el fin de mejorar la satisfacción del cliente y aumentar las ventas de un ecommerce?

Problemas específicos

- ¿Cuáles son los algoritmos de aprendizaje automático distribuido más adecuados para recomendar productos a clientes en tiempo real?
- ¿Cómo se pueden abordar los problemas de escalabilidad y disponibilidad del sistema de recomendación de productos en tiempo real?
- ¿Cuál es el impacto de la implementación del sistema de aprendizaje automático distribuido en la satisfacción de los clientes y en las ventas de la ecommerce?
- ¿Cómo se pueden abordar los problemas de privacidad y seguridad de los datos de los clientes en el sistema de aprendizaje automático distribuido?

Descripción del problema

El problema de investigación se centra en las recomendaciones que ofrecen los distintos sistemas de recomendaciones para la gran cantidad de usuarios de los ecommerce .Los sistemas de recomendación ayudan a los usuarios dando sugerencias personalizadas sobre la gran cantidad de productos, servicios o contenido que podrían ser de su interés dentro de los ecommerce.En general los sistemas de recomendaciones procesan una gran cantidad de datos y requieren de grandes cantidades de recursos computacionales para su correcto funcionamiento ,en este contexto se identificaron diversos problemas para la correcta implementación , como los datos iniciales limitados (1% del total de ítems cuentan con calificaciones), ademas estos sistemas realizar trabajan con una gran cantidad de datos por lo que necesitan una gran capacidad de procesamiento por lo que sistemas monolíticos (1 maquina local) en los sistemas de recomendación afectan de manera negativa la velocidad en la generación de las recomendaciones .Esto genera la deficiencia en sistemas de recomendaciones de productos para cliente de ecommerce , como consecuencia se tiene una escasez de datos para realizar el filtrado colaborativo (85%) por lo que el sistema mostrara resultados poco fiables ,además del tiempo de procesamiento excesivo(36 minutos) para la generación de respuestas

(Variable 1: Tiempo de respuesta del sistema para 100 mil registros: valor (50-60 minutos))

http://research.unir.net/unesco-congreso/wp-content/uploads/sites/76/2016/06/u2016-ROJASAndres.pdf

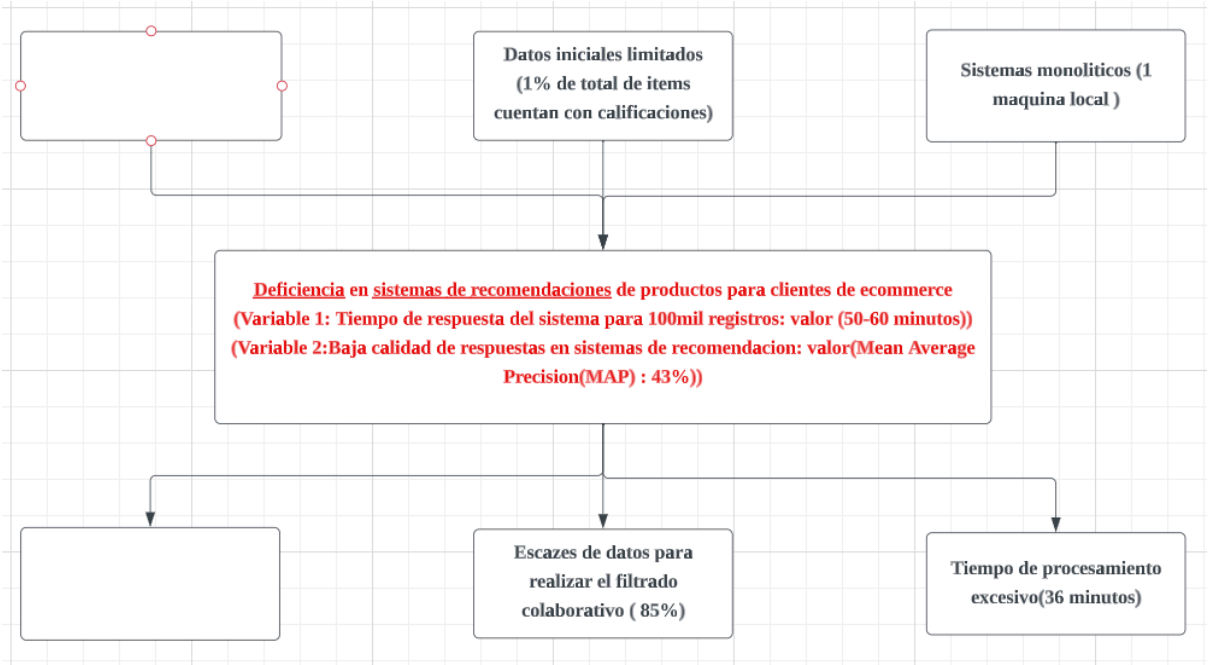
(Variable 2:Baja calidad de respuestas en sistemas de recomendación: valor(Mean Average Precision(MAP) : 43%))

https://ieeexplore.ieee.org/document/10123455

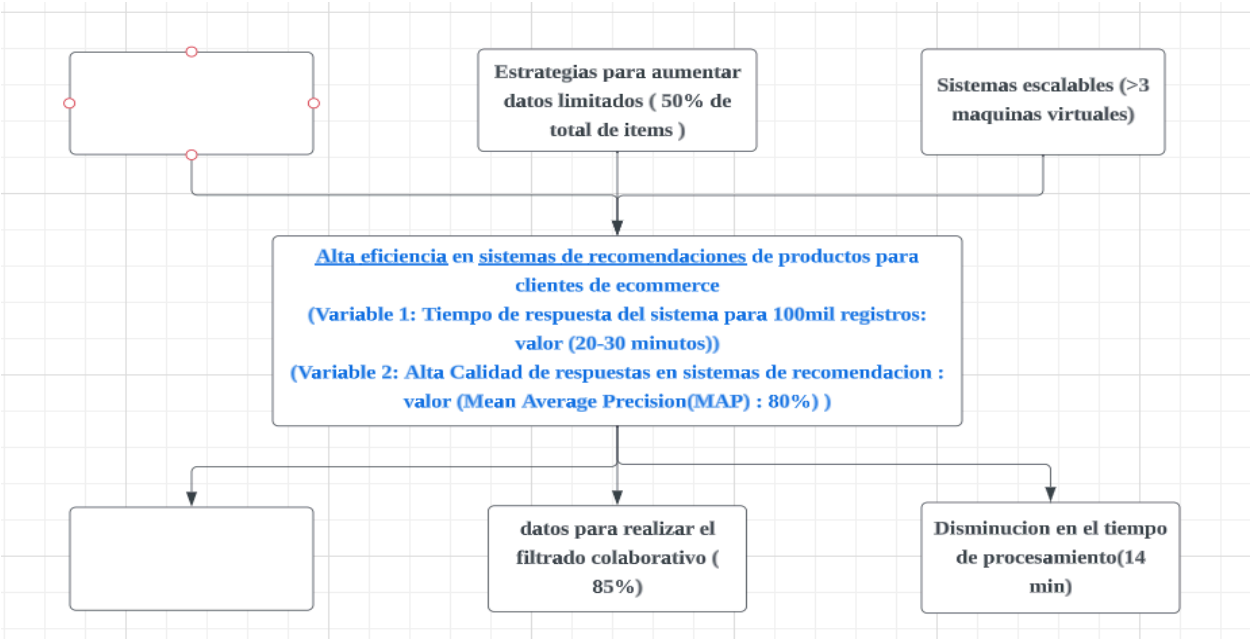
Objetivos de la Investigación

Marco lógico

Árbol del problema



Árbol de objetivo



Objetivo general

Diseñar y evaluar un sistema de aprendizaje automático distribuido para el análisis de grandes conjuntos de datos en tiempo real, que permita dar recomendaciones de productos a clientes en el comercio electrónico, utilizando estrategias para aumentar la cantidad de datos limitado (50 % de total de ítems) y un sistemas escalable (> 3 máquinas virtuales) para mejorar la eficiencia del sistema de recomendación en ecommerce y tener como consecuencia una mayor cantidad de datos para realizar el filtrado colaborativo (85 %) y una disminución considerable en el tiempo de procesamiento (14 minutos)

Objetivos específicos

- Identificar los conjuntos de datos relevantes para el estudio y análisis de las preferencias de los clientes en el comercio electrónico.
- Identificar las herramientas y tecnologías más adecuadas para implementar el sistema.
- Evaluar el rendimiento del sistema de aprendizaje automático distribuido en términos de velocidad y precisión.
- Analizar técnicas y soluciones para abordar los problemas de privacidad y seguridad de los datos de los clientes en sistemas de aprendizaje automático distribuido

Justificación de la Investigación

Justificación teórica

- Existe un gran interés en el desarrollo de sistemas de aprendizaje automático distribuido, ya que estos permiten procesar grandes volúmenes de datos en tiempo real, mejorando la eficiencia y escalabilidad del procesamiento de datos en comparación con los sistemas centralizados.
- El estudio de técnicas de aprendizaje automático es relevante en el campo de la inteligencia artificial y es una de las áreas de mayor crecimiento en la actualidad. Su aplicación en la industria puede ser muy beneficiosa para optimizar procesos y mejorar la calidad de los productos y servicios ofrecidos.
- La investigación en el diseño de sistemas de recomendación en el comercio electrónico es importante debido al gran volumen de datos que se generan y a la necesidad de proporcionar a

los clientes recomendaciones personalizadas y relevantes de productos para mejorar su experiencia de compra.

Justificación práctica

- Las ecommerce que se dedican al comercio electrónico se enfrentan al desafío de proporcionar a sus clientes recomendaciones personalizadas y relevantes de productos para mejorar su experiencia de compra, lo que puede aumentar la lealtad y las ventas.
- La implementación de un sistema de aprendizaje automático distribuido para el análisis de grandes conjuntos de datos en tiempo real puede proporcionar a las ecommerce la capacidad de procesar grandes volúmenes de datos de manera eficiente, lo que les permitiría tomar decisiones más rápidas y precisas.
- La aplicación de técnicas de aprendizaje automático en el comercio electrónico puede tener un impacto significativo en la satisfacción del cliente y, por tanto, en la rentabilidad de la ecommerce.

CAPÍTULO 2: MARCO TEÓRICO

Marco Filosófico o epistemológico de la investigación

El estudio se basará en el enfoque positivista. El objetivo principal de la investigación es diseñar y evaluar un sistema de aprendizaje automático distribuido para el análisis de grandes conjuntos de datos en tiempo real. Se utilizará un enfoque cuantitativo para recopilar y analizar los datos, utilizando herramientas estadísticas y matemáticas para identificar patrones y relaciones entre las variables. El objetivo final del estudio es producir un sistema de aprendizaje automático distribuido que pueda utilizarse en aplicaciones prácticas.

2.2Antecedentes de investigación

Se identificaron estudios internacionales que permitieron conocer enfoques teóricos y metodológicos que han sido utilizados en investigaciones similares.

2.2.1 "Sistema de recomendación para el comercio electrónico aplicado a una tienda de libros" [3] (SSI: 1665-0654) (18 citaciones) (Fuente: redalyc)

- En el artículo se propone un sistema de recomendación basado en contenido y filtrado colaborativo para una tienda de libros en línea. El objetivo es mejorar la experiencia del usuario al

proporcionar recomendaciones personalizadas y relevantes basadas en sus gustos y comportamientos de navegación.En el artículo se mencionan los casos para páginas que utilizan sistemas de recomendación de distinto tipo ,como el caso de “MovieLens” que utiliza un S.R que utiliza valoraciones de otros usuarios de acuerdo a una categoría y recomendará algunas en las que el usuario podría estar interesado pero luego cambio al modelo de recomendaciones basado en producto y no en usuarios, también está el caso de “Zagat Survey” que también usa el modelo basado en recomendaciones colaborativas al igual que “Last” .Tambien se menciona los sistemas basados en conocimientos que a diferencia de los colaborativos , no necesitan un historial del usuario para hacer las recomendaciones , basta con un listado de preguntas para que el sistemas pueda recomendar los productos.Por último menciona los sistemas híbridos de recomendación que combina el filtrado colaborativo y el basado en contenido para aumentar la precisión de las recomendaciones. El sistema de recomendación se evaluó mediante un experimento en el que se compararon las recomendaciones del sistema con las de los expertos en el campo, el sistema guarda la información de usuario activo y los combina con objetos relevantes para generar recomendaciones,Amazon utiliza este tipo de sistemas y es una de las razones de su éxito , incluso las frases "Los clientes que están viendo este producto también compraron"," ¿Qué otros productos compran los clientes tras ver este producto?" son resultado del sistema de recomendaciones híbrido ,esto es posible gracias al perfil del usuario, ya que se almacena con el historial de navegación, los productos adquiridos y las preferencias marcadas en el perfil. Es un sistema que se encarga de calcular la similitud entre los vecinos más cercanos a él y además recomendar de acuerdo al historial de cada usuario.En el artículo se implementó un sistema de recomendación que utilizó el algoritmo SLOPE ONE que es un sistema de clasificación colaborativo que sigue un proceso para predecir cómo un usuario clasificará un objeto a partir de las clasificaciones que han dado otros usuarios.Los criterios para la calificación de los usuarios fueron tres, los cuales sirvieron como referencia para el planteamiento de las preguntas a realizar a los clientes. La calificación que hace el usuario para un libro se basa a su vez en tres criterios, se califica la prosa del autor con la pregunta: ¿Qué te ha parecido del libro?,se califica la estructura del libro con la pregunta: ¿Cómo te ha favorecido el tipo de fuente, tamaño y vocabulario del libro? Y por último la secuencia lógica del libro: ¿Te parece lógico el avance entre capítulos del libro?.Los resultados mostraron que el sistema propuesto constituye una vía innovadora para las empresas al ingresar al entorno digital, ya que estas metodologías de recomendación representan un tipo único de asesor virtual. Esto se debe a que estas técnicas no solo abren nuevas perspectivas para el mercado en línea, sino que también operan como un consejero digital personalizado.

Utilidad del artículo para la tesis

Gracias a este artículo se logró identificar una forma para lograr recomendaciones para usuarios que no tengan un historial de usuario con sus preferencias, utilizando los sistemas basados en conocimientos , también se vio la posibilidad de mejorar la precisión de la recomendación utilizando

sistemas híbridos de recomendación que combinan el filtro colaborativo y el basado en contenido ,por último se destacó la importancia de los criterios para la calificación de los usuarios que servirán como base para la recomendación de los productos ,esta información se utilizará para brindar recomendaciones a usuarios que no tengan un historial de usuario , mejorar la precisión de los resultados con sistemas híbridos y tener datos fiables mediante buenos criterios para la calificación de usuarios de sistema de recomendación

2.2.2"Diseño e implementación de un prototipo experimental para aprendizaje automático distribuido" [4]

- En el artículo Describe el diseño e implementación de un prototipo de sistema de aprendizaje automático distribuido utilizando la plataforma Apache Spark. El objetivo del prototipo es permitir el entrenamiento de modelos de aprendizaje automático utilizando grandes conjuntos de datos distribuidos en un clúster de computadoras.Resalta la importancia de considerar la arquitectura que se implementará. Para determinar esto, es necesario evaluar diversos aspectos, incluyendo el tipo de algoritmo a emplear y el grado de distribución requerido.

Estructura de servidor de parámetros: En este esquema, los datos se dispersan entre los nodos clientes, mientras que los nodos servidor retienen los parámetros fundamentales de la arquitectura, los cuales son compartidos de manera global.

Enfoque de peer-to-peer: En este modelo, todos los nodos se enlazan directamente entre sí, y cada uno de ellos guarda la información relacionada con los parámetros del modelo. Este enfoque tiene ventajas como su escalabilidad sencilla y su eficaz manejo de fallos en caso de que uno de los nodos falle.

Estructura de árboles: Basada en una jerarquía por niveles, esta estructura se caracteriza por su capacidad escalable y administrativa, ya que la comunicación entre los nodos se lleva a cabo entre nodos padres e hijos.

Estructura de anillos: En este tipo de estructura, la comunicación es directa y sencilla, dado que un nodo sólo se conecta a los nodos adyacentes, permitiendo intercambiar información únicamente con estos. Se emplea cuando la interacción entre los nodos es mínima .

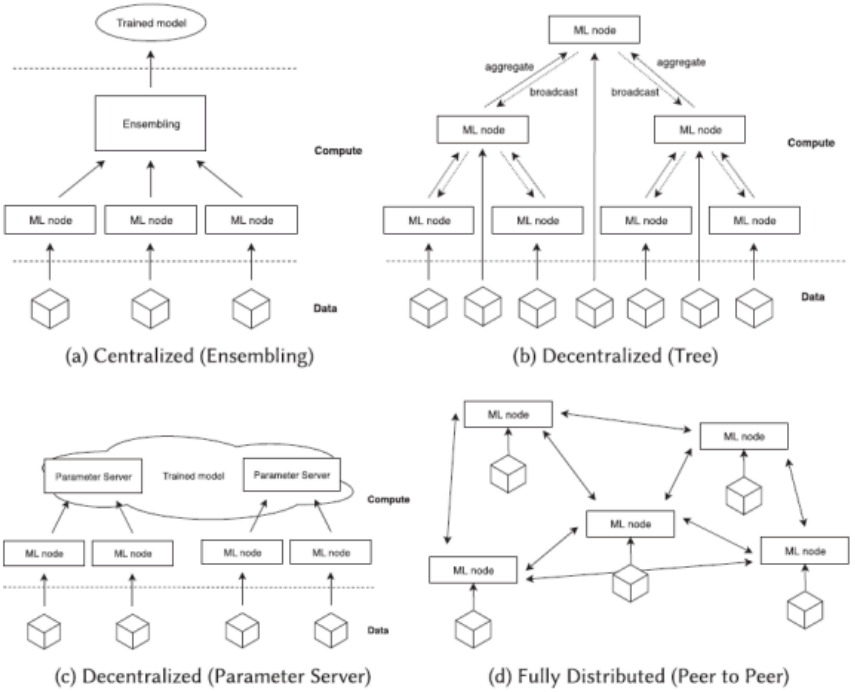


Figura 2: Estructura de nodos

Fuente:[4]

.También menciona la importancia de los frameworks como TensorFlow que ofrece dos estrategias con características distintivas: MultiWorkerMirroredStrategy y ParameterServerStrategy. Ambas estrategias incorporan un mecanismo eficaz de manejo de fallos, permitiendo que el progreso del entrenamiento del modelo se conserve después de cada iteración a través de funciones personalizables definidas por el usuario.Cuando se identifican múltiples dispositivos, como varios CPUs o GPUs, TensorFlow lleva a cabo un proceso denominado "ubicación de nodos". Mediante este procedimiento, se evalúa el costo computacional estimado de ejecutar una operación en todos los dispositivos disponibles. Esto permite asignar tareas adecuadas a cada dispositivo según su capacidad y recursos disponibles. Luego describe la arquitectura del prototipo, que consiste en varios componentes, incluyendo un gestor de recursos, un servidor de metadata y un conjunto de nodos de computación que se encargan del procesamiento de datos. El artículo también describe en detalle el proceso de implementación del prototipo, incluyendo la configuración de la plataforma Apache Spark y la programación de los diferentes componentes del sistema. Finalmente, se presentan los resultados de las pruebas realizadas en el prototipo, que demuestran la escalabilidad y eficiencia del sistema en el entrenamiento de modelos de aprendizaje automático distribuido. Este artículo presenta un enfoque práctico para el diseño e implementación de un sistema de aprendizaje automático distribuido. Este prototipo experimental demostró que puede ser utilizado como base para el desarrollo de sistemas de aprendizaje automático distribuido a gran escala en diferentes aplicaciones, en este caso para el análisis de datos en el sector ecommerce real.

Utilidad del artículo para la tesis

El artículo sirvió como guía para conocer las arquitecturas aplicadas a sistemas distribuidos en machine learning , en este caso la estructura de Estructura de anillos fue la que más se adapta a la estructura arquitectónica requerida , también se analizó los diferentes framework que facilitan el desarrollo de los sistemas distribuidos como lo es el framework TensorFlow que tiene una gran comunidad de soporte , con esta información se lograra realizar un estructura de anillos para la configuración de los nodos en el sistema distribuido y con el framework TensorFlow se facilitará el desarrollo del mismo ya que es un framework especializado en machine learning distribuido

2.2.3“Revisión de las principales metodologías para la construcción de aplicaciones distribuidas en la nube” [5] (ISSN:1659-0775) (1 citación)

-En el artículo tuvo como objetivo la revisión de las principales metodologías utilizadas para la construcción de aplicaciones distribuidas en la nube, destaca sus características, ventajas, desventajas y aplicabilidad en diferentes contextos, se destaca la importancia de las aplicaciones distribuidas en la nube en el contexto actual de la tecnología de la información, y cómo estas aplicaciones permiten una mayor escalabilidad, disponibilidad y rendimiento en comparación con las aplicaciones monolíticas tradicionales, también describe varias metodologías utilizadas para construir aplicaciones distribuidas en la nube, incluyendo arquitecturas basadas en microservicios, contenedores, servicios sin servidor.En el artículo se establecen una serie de obstáculos y oportunidades respecto de la computación en la nube. Tal y como se indica en la siguiente tabla:

	Obstáculo	Oportunidad
1	Disponibilidad	Continuidad del negocio Uso de múltiples proveedores de la nube
2	Datos cerrados	API (application programming interface: interfaz de programación de aplicaciones) estándares; compatible con software para habilitar Surge o computación en la nube híbrida
3	Confidencialidad y auditoría de los datos	Implementación de cifrado, redes virtuales y firewalls
4	Cuellos de botella en la transferencia de datos	Discos rápidos; switches de alto nivel
5	Imprevisibilidad del rendimiento	Soporte a máquinas virtuales mejorados; Memoria flash; horarios de máquinas

		virtuales
6	Almacenamiento escalado	Invención de tiendas escalables
7	Errores en grandes sistemas distribuidos	Depuradores innovadores que dependan de las máquinas virtuales distribuidas
8	Escalar rápidamente	Innovaciones autoescalable s que dependen del lenguaje de programación ; snapshots para la conservación
9	Reputación del destino compartido	Ofrecer servicios de protección de la reputación como los de correo electrónico
10	Licenciamiento de software	Licencias de pago por uso

Tuvo como conclusiones que las aplicaciones distribuidas en la nube permiten una mayor escalabilidad, disponibilidad y rendimiento en comparación con las aplicaciones monolíticas tradicionales. Sin embargo, existen desafíos y riesgos asociados con la construcción de estas aplicaciones, como la complejidad de la infraestructura y la gestión de datos, la seguridad y la privacidad de los datos.

Utilidad del artículo para la tesis

El artículo proporciona información acerca de los desafíos y riesgos asociados con la construcción de los sistemas distribuidos , estos se centran en la infraestructura , gestión y seguridad , también se tomó en cuenta los 10 obstáculos y oportunidades relacionadas con datos , hardware , software , tomando en cuenta esta información se pudo identificar los desafíos y riesgos comunes en la creación de sistemas distribuidos y además los obstáculos y oportunidades relacionadas a estos se utilizaran para mejorar la distribución del tiempo en la creación del sistema

2.2.4 “Collaborative filtering and deep learning-based recommendation system for cold start items” [6] (DOI:10.1016/j.eswa.2016.09.040) (486 citaciones) (Fuente: Scopus)

2.2.5 “Distributed processing using cosine similarity for mapping Big Data in Hadoop” [7] (DOI: 10.1109/TLA.2016.7555265) (10 citationes) (Fuente : Scopus)

El artículo se centra en el uso de un sistema de recomendación que utiliza el algoritmo de similitud coseno para analizar registros de una base de datos plana. Estos registros contienen información sobre cien mil películas de la base de datos MovieLens, que es de acceso público y gratuito en el sitio web movielens.org. Esta base de datos se considera un ejemplo de Big Data debido a su volumen y su carácter abierto. Dado el alto costo computacional asociado con el procesamiento de esta cantidad de información, se recurre a la computación distribuida como solución. El artículo menciona que se utiliza EMR (Elastic Map Reduce), un servicio de Amazon Web Services diseñado para facilitar la computación distribuida en la nube , este ofrece un entorno autoadministrado basado en Hadoop que facilita la escalabilidad dinámica de instancias llamado Elastic Cluster (EC2). EMR se encarga de manera segura y confiable de diversos casos de uso de Big Data, como análisis de registros, indexación web, almacenamiento de datos, aprendizaje automático, análisis financiero, simulación científica y bioinformática.

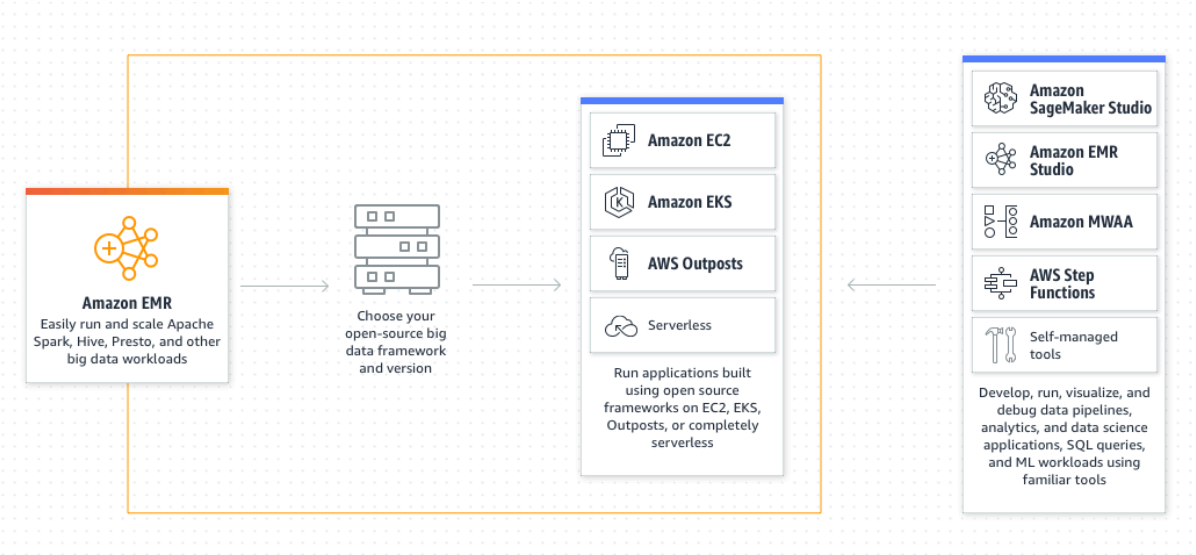


Figura 3: Entorno del funcionamiento EMR

Fuente: [8]

También menciona la similitud del coseno que es una medida de similitud entre 2 vectores utilizada para medir la similitud entre 2 palabras o documentos representados como vectores .En el artículo se muestra como ejemplo del uso de Map Reduce en AWS con EMR utilizando la data de 100 mil registros cuya información está en un documento de texto plano (esto significa que no es

necesario recopilar, clasificar y organizar información) , en este caso se realizó los siguientes pasos preprocesamiento de la información, distribución, procesamiento, resultados y análisis .

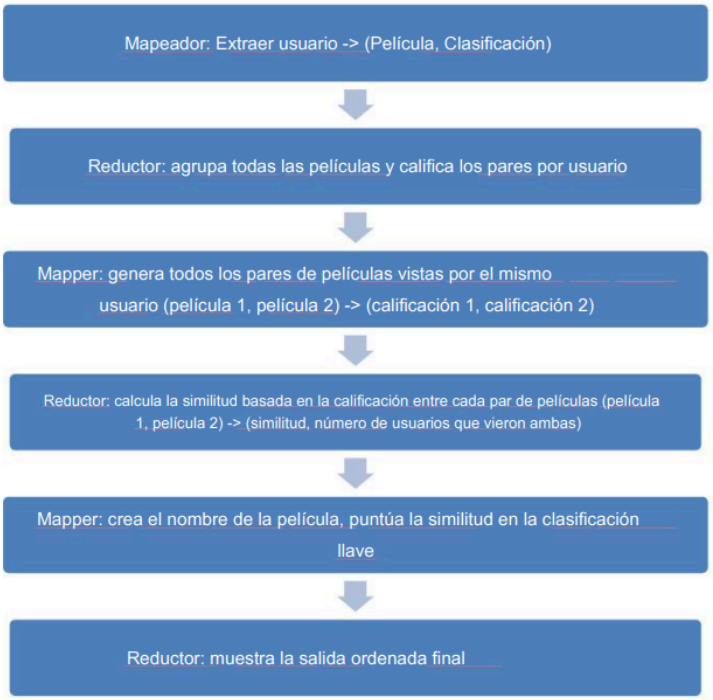


Figura 4: Pasos para la prueba de correlación

Fuente: [7]

Los resultados que se obtuvieron muestran la correlación entre 2 peliculas y esta delimitada entre 0 y 1.

Tabla 1: Muestra de archivo de correlación final

```
"Star Wars (1977)"    ["Nell (1994)", 0.95440149533124929, 75]
"Star Wars (1977)"    ["Jumanji (1995)", 0.95454005841416234, 91]
"Star Wars (1977)"    ["Shine (1996)", 0.95478403680312518, 104]
"Star Wars (1977)"    ["Mary Poppins (1964)",
```

Fuente:[7]

Y el resultado de las métricas de rendimiento entre las distintas configuraciones de los clusters

Tabla 2: Pasos en tiempo de ejecución para cada configuración de clúster

Paso (tiempo) / Conf. clúster.	Local Máquina	(1) Control Máquina	(2) Clúster con 1 esclavo	(3) Clúster con 2 esclavos	(4) Clúster con 3 esclavos
Paso 1	--	2 minutos	2 minutos	2 minutos	3 minutos
Paso 2	--	37 minutos	35 minutos	21 minutos	14 minutos
Paso 3	--	2 minutos	2 minutos	2 minutos	2 minutos
Total	50-60 minutos	50 minutos	48 minutos	34 minutos	27 minutos

Fuente:[7]

Utilidad del artículo de tesis

Este artículo se muestra información acerca de EMR(Elastic Map Reduce) que sirve para el procesamiento de big data en sistemas distribuidos , además proporcionan múltiples herramientas para las gestión del sistemas , almacenamiento ; utiliza Map Reduce lo cual permitirá dividir la carga de trabajo entre los distintos nodos ,además se mostraron las métricas de rendimiento entre los distintos tipos de configuración de clusters con esta información se podrá construir el sistema distribuido con la ayuda de EMR para crear el sistema de recomendación para ecommerce en la nube y utilizar las métricas de rendimiento para ver que configuración favorece más al e commerce

2.2.6 "A Scalable Recommendation System Approachfor a Companies | Seniors Matching" [8] (DOI: 10.1109/AI4I54798.2022.00008) (Fuente: IEEE)

Este artículo propone la implementación de un sistema de recomendación en la plataforma "WisdomOfAge", una plataforma web en desarrollo que tiene como objetivo conectar a trabajadores seniors que se jubilarán o ya lo han hecho con empresas que buscan habilidades específicas. El sistema de recomendación se utilizará para generar listas de trabajadores seniors que coincidan con las necesidades de las empresas al crear nuevas ofertas de misiones. El sistema de recomendación se enfrenta a varios desafíos, incluyendo la necesidad de estar operativo desde la primera oferta de misión en el sitio web y la capacidad de adaptarse al crecimiento de usuarios registrados de manera escalable. Además, el sistema debe ser capaz de evaluar la relevancia de la experiencia y habilidades de los trabajadores seniors, otorgando mayor peso a las experiencias recientes y pertinentes. El artículo menciona la importancia de preprocesar el texto y muestra 5 pasos : tokenización ,

eliminación de palabras vacías , lematización , representación de documentos de texto y extracción de caracteres , esto es importante para evitar una disminución del rendimiento.

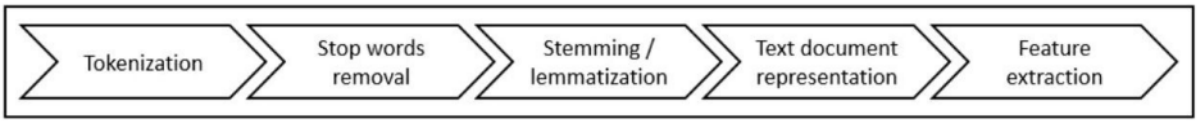


Figura 5: Pasos de preprocesamiento

Fuente:[8]

también hace un análisis entre los tipos de filtrado colaborativo para ver cuál funciona mejor con el sistema , el filtrado basado en ítems no tendría buenos resultados ya que si una empresa disfruta trabajando con un senior y tiene una necesidad similar más adelante, hay una probabilidad muy alta de que la empresa no utilizará la plataforma para solicitar una lista de recomendación de personas mayores , entonces el filtrado por usuario sería la mejor opción ya que si una empresa ha disfrutado trabajar con un senior, existe una alta probabilidad de que este señor sea una buena opción para una empresa similar con una necesidad similar. Además, propone un subsistema de índice de similitud que compara los perfiles con las misiones ofrecidas por las empresas para proporcionar una lista de personas con mayor similitud a las misiones, en estos sistemas cada elemento del perfil senior se compara uno a uno con la misión ofrecida. El artículo aborda el problema de arranque en frio que sufren los sistemas de filtrado colaborativo, para solucionar esto propone un sistema híbrido entre la similitud de índice y el filtrado colaborativo, proponiendo una suma ponderada para combinar la producción de cada subsistema

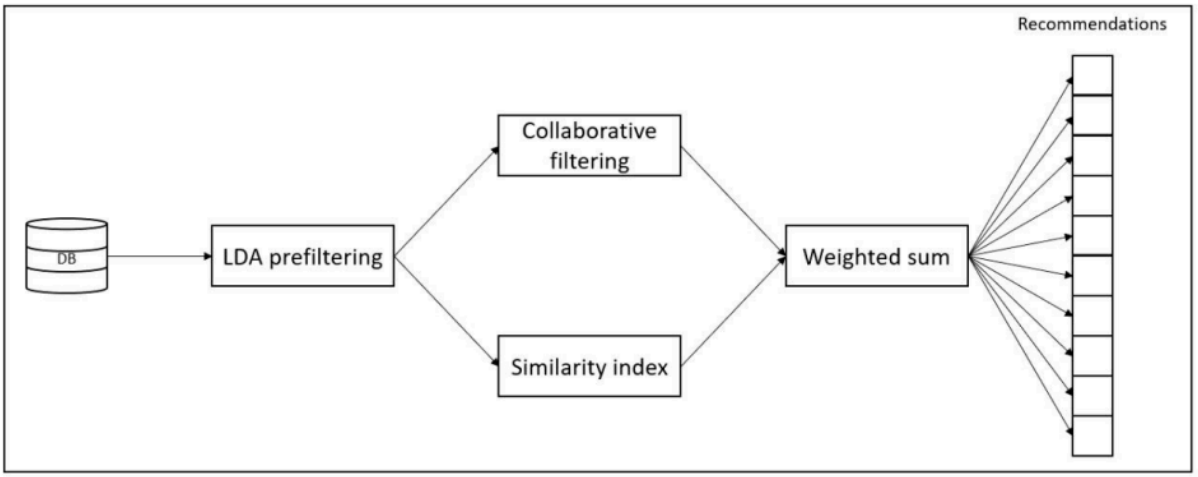


Figura 6: Sistema híbrido

Fuente:[8]

Utilidad del artículo de tesis

La información de este artículo aportó los pasos para realizar el preprocesamiento del texto que se encuentra en los mensajes que dejan los usuarios acerca de los ítems que en el caso de un enormes serian sus productos, también se mostró una posible solución al arranque en frío que sufren los sistemas de filtrado colaborativo que sería un sistema híbrido utilizando el sistema de similitud de índice , con esta información se podrá realizar el preprocesamiento de las reseñas que dejan los usuarios y se obtendrá unas recomendaciones más precisas , también se solucionara el problema de arranque en frio ,causado por la falta de datos iniciales,con la ayuda del sistema de similitud de índice que es menos preciso pero que no requiere de datos iniciales

2.2.7 "Machine learning based recommender system for e-commerce." [9] (DOI: 10.11591/ijai.v12.i4.pp1803-1811) (Fuente: Scopus)

El artículo menciona que el comercio electrónico se ha vuelto esencial en los negocios debido a su simplicidad, disponibilidad, variedad de productos, métodos de pago flexibles y conveniencia de comprar desde cualquier lugar. Por ende, los sistemas de recomendación son cruciales para el éxito de las empresas de comercio electrónico, ya que personalizan la experiencia del cliente, aumentando la participación y las tasas de compra. El articulo defines a los sistemas de recomendación como un mecanismo que toma en cuenta la información del usuario (U), un conjunto de productos (P) y una función de utilidad (h) para determinar el nivel de interés de un usuario en un producto específico. La elección de la función de utilidad (h) depende del enfoque utilizado y se ve influenciada por el tipo de información que se utiliza, ya sea implícita o explícita. Por ejemplo, en un enfoque de filtrado colaborativo, la función de utilidad se basa frecuentemente en la similitud entre usuarios o elementos. En cambio, en un enfoque basado en contenido, la función de utilidad se fundamenta en las características o atributos de los productos. También buscaron mejorar la precisión de estos sistemas es mediante el desarrolló un algoritmo de recomendación personalizado basado en reglas de asociación, que demostró resultados prometedores al aumentar la probabilidad de compra de los productos sugeridos. Utilizaron un sistema de recomendación basado en reglas de asociación, específicamente el algoritmo FP-Growth, debido a su alta precisión y facilidad de implementación y explicación.

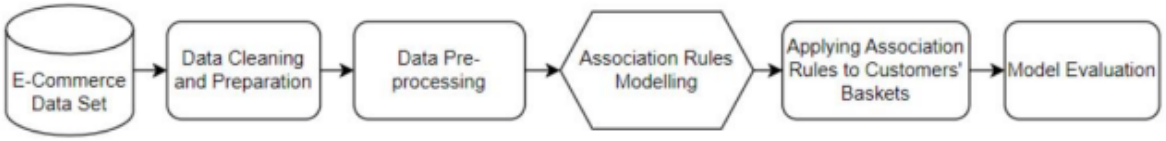


Figura 7. Flujo de trabajo del sistema de recomendación propuesto

Fuente:[9]

Visualización de Datos: Se utilizó un conjunto de datos transaccionales que contenía información sobre clientes, facturación, productos comprados, etc. Se mostraron las primeras 10 filas del conjunto de datos.

Preprocesamiento de Datos: Se eliminaron los productos que eran obsequios de la empresa a los clientes y se agruparon los productos comprados por cada cliente en un nuevo conjunto de datos de transacciones.

Modelado de Reglas de Asociación: Se empleó el algoritmo FP-Growth para determinar las reglas de asociación más frecuentes en el conjunto de datos. Se mencionaron dos hiperparámetros clave: minSupRatio (soporte mínimo) y minConf (confianza mínima).

Evaluación del Modelo y Resultados: Se evaluó el desempeño del sistema de recomendación calculando la probabilidad promedio del próximo producto que un cliente compraría (Paverage). También se calcularon los ingresos esperados de los productos recomendados y sugeridos. Se presentaron los resultados en una tabla.

Utilidad del artículo de tesis

Se pudo recopilar información acerca de que tipos de algoritmos funcionan mejor para distintos enfoques y la influencia que tendrán en la calidad de las respuestas del sistema, por ende, la elección de la función de utilidad dentro de los sistemas de recomendación será una parte fundamental para que se alineen con las preferencias del usuario , además el uso del algoritmo FP-Growth basado en reglas de asociación permiten , a sistemas de recomendaciones para eccomerce , tener una alta precisión y facilidad de implementación , esta información se utilizara para implementar una buena función de utilidad que sea correspondiente a las necesidades de un eccomerce y usar el algoritmo FP-Growth que ya demostró tener buenos resultados en eccomerce

2.2.8”A link prediction-based recommendation system using transactional data” [10]

(DOI: 10.1038/s41598-023-34055-5) (Fuente: Scopus)

Este articulo habla sobre la creciente importancia de recomendar elementos relevantes a los usuarios debido al aumento de la cantidad de datos generados. Para lograr esto, se utilizan conjuntos de datos de transacciones, como registros de tarjetas de crédito y registros de compras en línea, para comprender los intereses de los usuarios mediante el análisis de las interacciones entre usuarios y

productos. El enfoque propuesto en este estudio es un sistema de recomendación basado en la predicción de enlaces, que combina algoritmos de aprendizaje de representación gráfica y clasificadores de aumento de gradiente en conjuntos de datos de transacciones.

2.2.9 "A fuzzy recommendation system for predicting the customers interests using sentiment analysis and ontology in e-commerce" [11] (DOI: 10.1016/J.ICTE.2018.05.003) (Fuente: Scopus)

El articulo menciona que la mayoría de los sistemas de recomendación existentes tienen en cuenta las reseñas de los clientes, el historial de compras del usuario y la calificación del producto para predecir el producto recomendado. Pero que los intereses de los usuarios varían con el tiempo, entonces los sistemas de recomendación existentes carecen de la capacidad de encontrar los elementos actualmente relevantes para los clientes. En el artículo proponen un nuevo sistema de recomendación de productos basado en lógica difusa que predice dinámicamente los productos más relevantes para los clientes en compras en línea según los intereses actuales de los usuarios, proponen un algoritmo novedoso para calcular la puntuación sentimental del producto con la categoría objetivo del usuario final asociada. Para esto se mencionan factores clave con para superar los problemas existentes y mejorar el rendimiento general del sistema de recomendación:

Consideración del Usuario Final: En lugar de enfocarse en el cliente que va a comprar, se toma en cuenta al usuario final real para quien se está realizando la compra. Las recomendaciones se basan en la puntuación de calificación del producto asociada a ese usuario final real.

Cálculo de Puntuaciones para Diferentes Categorías de Usuarios: Se calculan puntuaciones de calificación para diferentes grupos de usuarios categorizados según sus intereses y preferencias. Esto permite adaptar las recomendaciones a diferentes niveles de interés.

Inclusión de Nuevos Productos: Además de considerar productos similares, se incorpora una parte significativa de nuevos productos en la lista de recomendaciones.

Seguimiento del Historial de Recomendaciones: Se realiza un seguimiento del historial de recomendaciones previas para proponer un nuevo conjunto de productos recomendados para el cliente.

Uso de Información Demográfica: Se emplea información demográfica, como el grupo de edad del usuario final y la ubicación de entrega, para mejorar aún más la precisión de las predicciones y ofrecer productos más relevantes.

Enfoque Inteligente con Ontología: Se utiliza un enfoque inteligente que se basa en la representación de conocimiento y el razonamiento basado en ontología para mejorar la precisión de las decisiones en el desarrollo de sistemas de recomendación de productos.

El articulo propuso un sistema de siete módulos principales, a saber: cliente (usuario de compras en línea), módulo de interfaz de usuario, gestor de decisiones, Sistema de Recomendación Difusa, base de datos ontológica, base de reglas y gestor de reglas difusas.

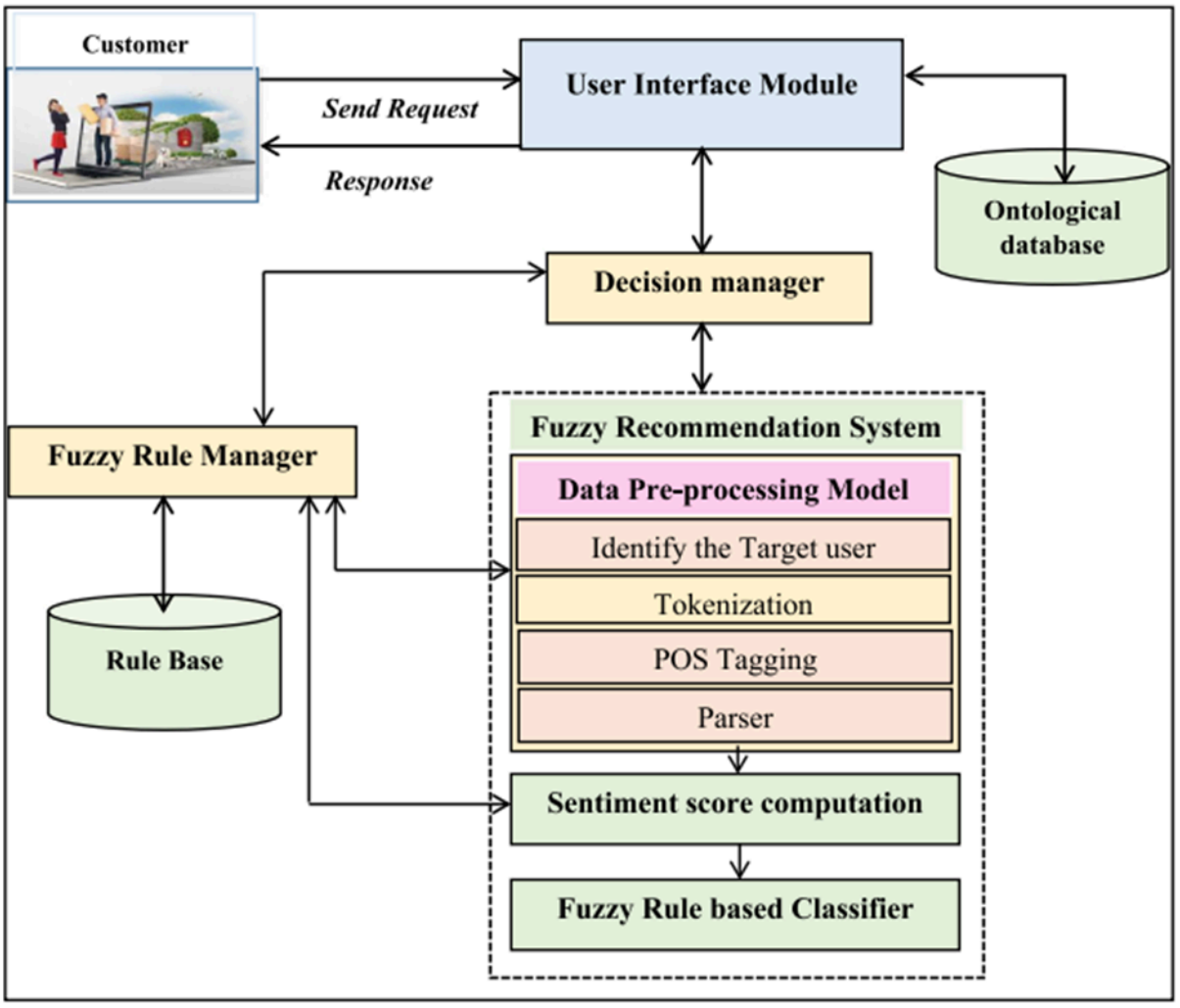


Figura 8: Arquitectura del sistema

Fuente: [11]

El sistema de recomendación consta de tres submódulos principales, como el preprocesamiento de datos, el cálculo de la puntuación sentimental y el clasificador basado en reglas difusas. Los pasos utilizados en el proceso de toma de decisiones son los siguientes:

Paso 1: Calcular la puntuación de revisión del producto con el grupo de edad objetivo asociado a partir de las revisiones de clientes preprocesadas y almacenar la puntuación en la base de datos.

Paso 2: Obtener los productos conocidos por el usuario a partir del historial de compras, lista de deseos y lista de búsqueda.

Paso 3: Extraer los productos con calificaciones altas y calcular la puntuación de similitud.

Paso 4: Utilizar información demográfica como el grupo de edad y la ubicación de entrega, así como el alineamiento ontológico para mejorar la lista de recomendaciones.

Paso 5: Utilizando la lógica difusa, generar una lista de recomendaciones altamente recomendadas y una lista de recomendaciones probables, y mostrar la lista de recomendaciones final.

Para el sistema difuso se han formulado reglas difusas considerando la calificación general del producto, la puntuación del producto objetivo calculada, la similitud con el historial de compras y el período de compra. Estas reglas difusas se aplican para tomar decisiones efectivas sobre los comentarios de revisión de cada usuario en línea y sus comentarios. Estas diversas reglas difusas se almacenan en la base de datos ontológica y se utilizan cuando se presenta la situación adecuada.

Tabla 3: Factores internos difusos

Factor	Base 1	Base 2	Outcome
Sentiment score	0	3	very negative
	2.5	5.5	negative
	5	8	positive
	7	10	very positive
Product rating score	0	2.5	poor
	2.3	3	average
	2.8	4	good
	3.8	5	excellent
Similarity score	0	3	poor
	2.5	5.5	average
	5	8	good
	7	10	excellent

Fuente: [11]

Por ejemplo, si un cliente está buscando comprar un producto para su hijo, el sistema de recomendación propuesto recomendará productos basándose en el nivel de recomendación. Todos los productos altamente recomendados se ofrecen al usuario, seguidos de los productos recomendados. Esto garantiza que el sistema recomiende productos que se adapten a las necesidades y preferencias del usuario en función de las reglas difusas definidas.

Utilidad del artículo de tesis

Este artículo mostro información sobre como implementar un sistema difuso en un sistema de recomendación para la mejora de las respuestas del sistema, de esto se pudo rescatar que un sistema de recomendación requiere 3 submódulos necesarios, como el de preprocesamiento, cálculo de puntuación sentimental y de reglas difusas , además da un ejemplo de reglas difusas que utiliza en el sistema para mejorar el rendimiento del sistema , con esta información se podrá realizar una correcta arquitectura del sistema utilizando los 3 submódulos necesario para incorporar el sistema difuso y saber que reglas utilizar mediante el ejemplo de reglas difusas que otorga el articulo

2.2.10 "Recommender systems: An overview of different approaches to recommendations" [12]

(DOI: 10.1016/J.ICTE.2018.05.003) (Fuente: Scopus)

El articulo menciona que los sistemas de recomendación se han convertido en un área activa de investigación desde la aparición de los primeros trabajos sobre filtrado colaborativo a mediados de la década de 1990. En los últimos años, muchos sitios web utilizan ampliamente sistemas de recomendación. Estos sistemas proporcionan sugerencias a los usuarios según sus necesidades y se utilizan en una variedad de aplicaciones, desde la recomendación de productos hasta la recomendación de noticias y música. También menciona la evaluación de la precisión de los sistemas de recomendación, utilizando métricas como el Error Absoluto Medio (MAE) y el Error Cuadrático Medio (RMSE). El MAE evalúa la magnitud promedio de los errores en un conjunto generado de predicciones, sin tener en cuenta su dirección. El RMSE es una regla de puntuación cuadrática que también evalúa la magnitud promedio del error. Es la raíz cuadrada del promedio de las diferencias al cuadrado entre la predicción y la observación real.Además, se aborda el problema de encontrar el mejor elemento o una lista de los mejores elementos para un usuario activo.

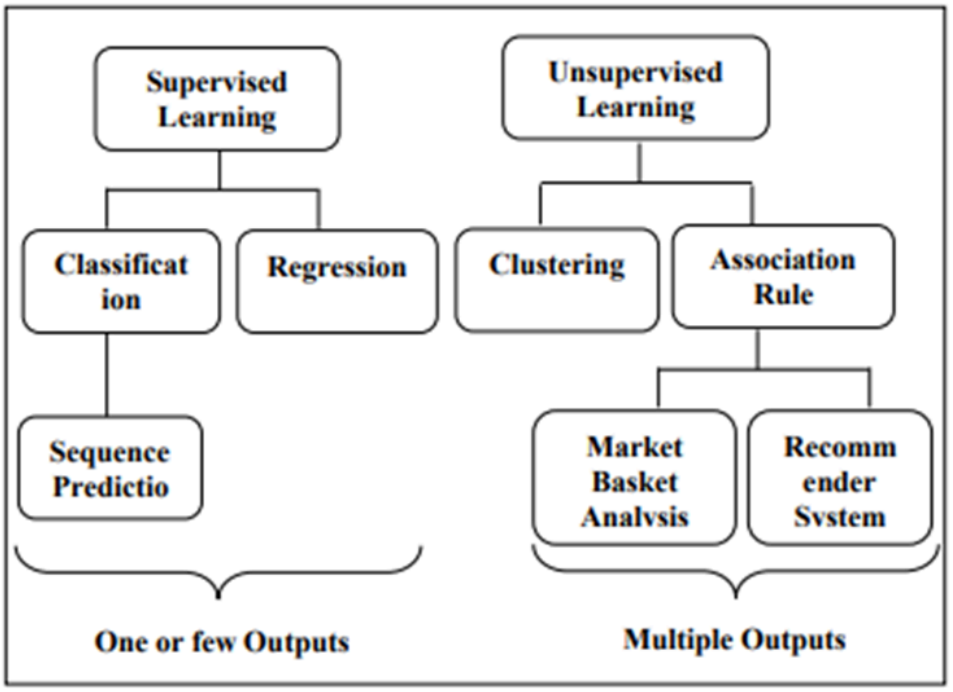


Figura 9: Sistemas de recomendación en aprendizaje automático

Fuente:[13]

Las ventajas de la técnica de filtrado colaborativo incluyen la facilidad de implementación en el caso del filtrado colaborativo basado en memoria y la capacidad de manejar nuevos datos de manera incremental. Además, los enfoques basados en modelos a menudo ofrecen un rendimiento de predicción mejorado.Sin embargo, también existen desventajas en el filtrado colaborativo, como el problema del inicio en frío, que se produce cuando se trata de nuevos usuarios y el sistema no sabe qué recomendar, así como desafíos de escalabilidad debido a la necesidad de manejar grandes cantidades de datos y problemas de dispersión debido a la falta de calificaciones para muchos elementos.

Ventajas del filtrado basado en contenido:

- Proporciona independencia del usuario al utilizar calificaciones exclusivas que el usuario utiliza para construir su perfil.
- Ofrece transparencia al usuario al explicar cómo funciona el filtro basado en contenido.
- Es útil para recomendar elementos que aún no han sido calificados o vistos por ningún usuario, lo que beneficia a los nuevos usuarios.

Limitaciones del filtrado basado en contenido:

- Es difícil generar características precisas para un elemento en el filtrado basado en contenido.

- Tiende a sufrir problemas de sobre-especialización, ya que tiende a recomendar los mismos tipos de elementos.

- Es más difícil obtener retroalimentación de los usuarios en el filtrado basado en contenido, ya que los usuarios normalmente no ordenan los elementos y, por lo tanto, no es posible determinar si la recomendación es correcta.

Utilidad del artículo de tesis

El artículo menciona las métricas necesarias para definir la calidad de un sistemas de recomendación y como estas funcionan en este caso menciona las métricas MAE Y RMSE ,además muestra las ventajas y limitaciones que tiene el uso del filtrado basado en contenido , con esta información se logrará tener unas métricas comúnmente utilizadas en sistemas de recomendación como el MAE Y RMSE para la evaluación final del sistema de recomendación y ver si es viable aplicar el filtrado basado en contenido para un sistema de recomendación en ecommerce

2.2.11 "Augmenting e-commerce product recommendations by analyzing customer personality" [13] (DOI: 10.1109/CICN.2017.8319380) (Fuente: Scopus)

2.2.12 "Improved recommendation system with review analysis" [14] (DOI: 10.1109/ICGTSPICC.2016.7955273) (Fuente: Scopus)

- En el artículo realizado por Hazem et al (2022) al llamado "**A distributed real-time recommender system for big data streams**". Se propuso un sistema de recomendación distribuido en tiempo real para grandes flujos de datos. Este sistema utiliza técnicas de aprendizaje automático para analizar y procesar grandes cantidades de datos en tiempo real y generar recomendaciones personalizadas para los usuarios. Utiliza un mecanismo de división y replicación para hacer que los sistemas de recomendación de transmisión sean escalables. Se estudia su capacidad con dos algoritmos de transmisión diferentes (ISGD y similitud de coseno incremental) y se aplica a dos conjuntos de datos diferentes, Netflix y Movielens25M. Además, se aplican técnicas de olvido (LRU y LFU) para hacer frente a la falta de límites de la corriente. El mecanismo adopta la arquitectura de nada compartido y se implementa utilizando las API de Flink. Los resultados muestran que el mecanismo de división y replicación puede escalar los sistemas de recomendación de transmisión y mejorar su precisión en términos de recuperación. también mencionan que se pueden agregar más mejoras en el futuro, como el reequilibrio de carga y la prueba del mecanismo con otros algoritmos de recomendación y diferentes técnicas de olvido.

2.3. Bases Teóricas

En el marco de investigación sobre soluciones de aprendizaje automático distribuido, los enfoques teóricos relevantes son:

2.3.1 Teoría de la inteligencia artificial

La teoría de la inteligencia artificial (IA) es un conjunto de enfoques teóricos y prácticos que se enfocan en la creación de sistemas informáticos capaces de realizar tareas que, si fueran realizadas por humanos, requerirían inteligencia. La IA busca crear máquinas que puedan aprender, razonar y tomar decisiones como lo haría un ser humano. La inteligencia artificial busca replicar en máquinas el proceso cognitivo humano, incluyendo la percepción, el razonamiento, la toma de decisiones y el aprendizaje (Nilsson, 1998).

El aprendizaje automático es un enfoque que se enfoca en crear algoritmos y modelos que puedan aprender a partir de datos para realizar tareas específicas, como la clasificación de imágenes o el análisis de datos. Una de las técnicas de análisis de datos más comunes es el análisis estadístico, que utiliza herramientas como la regresión, la correlación y la desviación estándar para encontrar patrones y tendencias en los datos. El análisis de datos puede ser utilizado en una amplia variedad de campos, desde la investigación científica hasta la gestión ecommercerial, y puede involucrar diferentes técnicas, como la minería de datos, la inteligencia ecommercerial y el aprendizaje automático"[12]. Además, el análisis de datos también puede involucrar técnicas de aprendizaje automático, como la agrupación y la clasificación, que pueden ayudar a identificar patrones más complejos y relaciones no lineales en los datos.

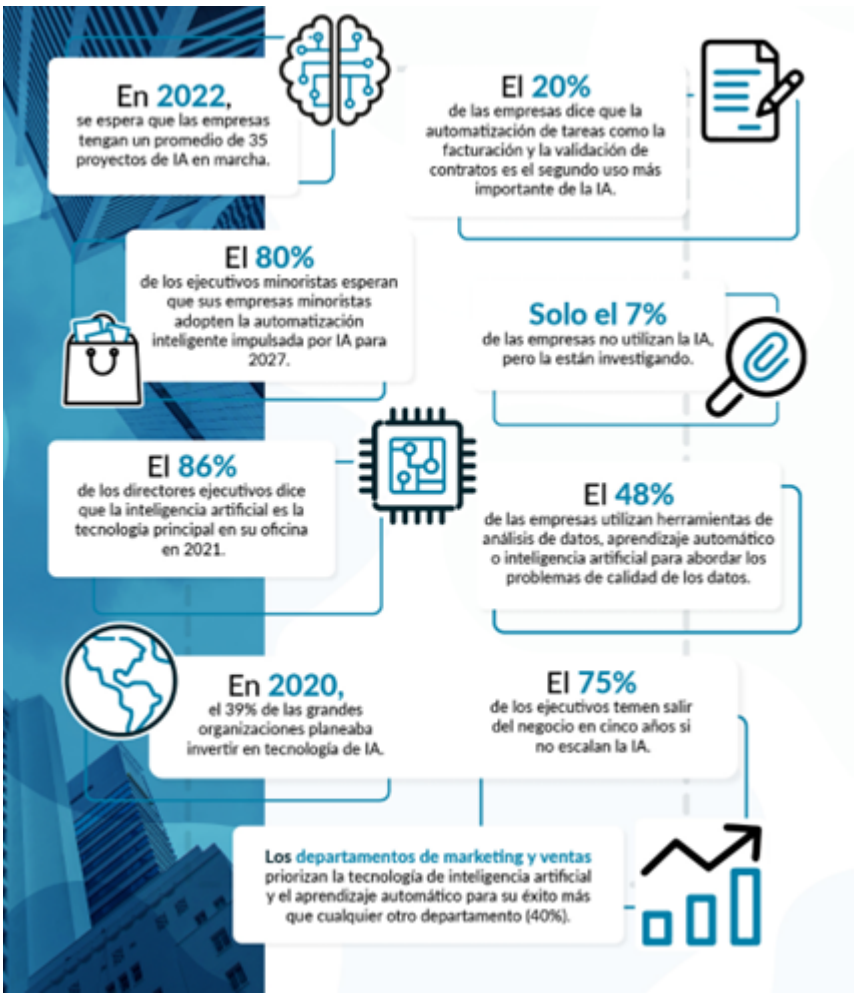


Figura 7: *Estadísticas de adopción de A.I. (2021)*
Fuente: *Inbest cloud*

El análisis de datos puede ser una tarea desafiante debido a la gran cantidad de datos que se deben procesar y analizar. Por lo tanto, se utilizan herramientas y técnicas especiales, como software de análisis de datos y sistemas de bases de datos, para manejar grandes conjuntos de datos y realizar análisis más complejos. El análisis de datos puede revelar información oculta sobre las preferencias de los clientes, las páginas populares del sitio web, el comportamiento de navegación, los comentarios de los clientes y la interacción con los formularios del sitio web [13]

En el contexto del análisis de grandes conjuntos de datos en tiempo real, la teoría de la IA es relevante para el diseño del sistema de aprendizaje automático distribuido que se propone en la investigación, ya que el aprendizaje automático es una técnica fundamental para el análisis de grandes conjuntos de datos. Además, la IA también es relevante para el diseño de sistemas informáticos capaces de procesar grandes cantidades de datos en tiempo real y tomar decisiones en función de esos datos. La IA también es esencial para la toma de

decisiones en tiempo real basada en datos, ya que puede analizar grandes cantidades de información y proporcionar recomendaciones o acciones en función de esa información [14]

2.3.2 Teoría de la optimización

La teoría de la optimización es esencial para mejorar la precisión y eficiencia de los modelos de aprendizaje automático. La optimización se utiliza para ajustar los parámetros del modelo con el fin de minimizar una función de pérdida que mide el error entre las predicciones del modelo y los valores reales. La optimización es una técnica clave en el aprendizaje automático, ya que permite ajustar los parámetros del modelo para minimizar una función de pérdida y mejorar su precisión en la tarea específica [15].

Existen diversas técnicas de optimización utilizadas en el aprendizaje automático, como el descenso de gradiente estocástico, el método de Newton, el algoritmo de Levenberg-Marquardt, entre otros. Cada técnica tiene sus propias ventajas y desventajas, y su elección depende del problema específico y de las características de los datos.

La teoría de la optimización también se aplica en otros campos relacionados con la informática y la ciencia, como la ingeniería de sistemas, la estadística y la economía. La teoría de la optimización es importante para mejorar la eficiencia y eficacia en la solución de problemas complejos. En la economía, la optimización se utiliza para tomar decisiones estratégicas y maximizar las ganancias (Nocedal y Wright, 2006).

2.3.3 Teoría de la arquitectura de software distribuido

La teoría de la arquitectura de software distribuido se centra en el diseño y la implementación de sistemas de software que permiten que múltiples dispositivos trabajen juntos de manera eficiente y efectiva. En el contexto de la investigación sobre el diseño y evaluación de un sistema de aprendizaje automático distribuido para el análisis de grandes conjuntos de datos en tiempo real, la teoría de la arquitectura de software distribuido es importante porque el sistema propuesto debe ser capaz de procesar grandes cantidades de datos en tiempo real a través de múltiples dispositivos que se comunican entre sí. Es importante considerar la seguridad y la privacidad de los datos transmitidos a través del sistema distribuido y asegurar que el sistema sea capaz de manejar las posibles inconsistencias que puedan surgir debido a la distribución de los datos y el procesamiento en múltiples dispositivos. (Omatu et al, 2017)

Implica la utilización de diversos componentes y servicios que trabajan juntos para proporcionar una funcionalidad global del sistema. Esto incluye la utilización de tecnologías de red, protocolos de comunicación, algoritmos de distribución de carga y servicios de almacenamiento de datos distribuidos. Los sistemas de aprendizaje automático distribuido pueden utilizar diferentes arquitecturas de software distribuido, como la arquitectura cliente-servidor, la arquitectura de red punto a punto o la arquitectura de procesamiento por lotes. Cada arquitectura tiene sus propias ventajas y desventajas y puede ser más adecuada para diferentes tipos de aplicaciones.

Es importante en la investigación sobre el diseño y evaluación de un sistema de aprendizaje automático distribuido para el análisis de grandes conjuntos de datos en tiempo real porque permite la creación de un sistema eficiente y escalable que pueda manejar grandes cantidades de datos en tiempo real a través de múltiples dispositivos que se comunican entre sí.

2.4. Marco Conceptual

Es necesario hacer énfasis en términos como Apache Spark, Hadoop, procesamiento en tiempo real, aprendizaje automático, informática distribuida, entre otros.

Apache Hadoop

Hadoop se enfoca principalmente en el almacenamiento y procesamiento de grandes volúmenes de datos en sistemas de archivos distribuidos (HDFS), y utiliza el marco de procesamiento MapReduce para ejecutar tareas de análisis en paralelo en un clúster de computadoras. Hadoop también ofrece una variedad de herramientas complementarias para el procesamiento de datos, como Pig y Hive.

Apache Spark

Apache Spark se enfoca en el procesamiento de datos en memoria y ofrece una plataforma de análisis de datos más rápida y flexible que Hadoop. Spark utiliza un modelo de programación basado en RDD (Resilient Distributed Datasets) para el procesamiento de datos, lo que permite la realización de operaciones complejas en memoria y la ejecución de tareas de análisis de datos en tiempo real. Spark también ofrece una amplia gama de herramientas para el análisis de datos, incluyendo Spark SQL, Spark Streaming y MLlib.

Escalabilidad

En el contexto de los sistemas informáticos, la escalabilidad se refiere a la capacidad de un sistema para adaptarse a un aumento en la carga de trabajo o en la cantidad de usuarios sin afectar su rendimiento. Se busca que un sistema escalable mantenga una eficiencia y calidad en el servicio incluso en situaciones de alta demanda. Por lo tanto, es importante considerar la escalabilidad desde el inicio del proceso de diseño y desarrollo de un sistema informático. Esto implica la elección adecuada de la arquitectura, el uso de tecnologías escalables, la implementación de técnicas de optimización de rendimiento y la planificación para el crecimiento futuro.

- Computación en la nube: permite el acceso a recursos informáticos escalables y flexibles a través de internet.
- Bases de datos distribuidas: permiten la gestión de grandes cantidades de datos en múltiples servidores, lo que aumenta la capacidad y el rendimiento del sistema.
- Balanceo de carga: distribuye la carga de trabajo entre varios servidores, lo que ayuda a evitar la sobrecarga y mejora el rendimiento.
- Caché: almacena temporalmente datos y recursos en memoria para reducir el tiempo de acceso y aumentar la velocidad de respuesta del sistema.
- Microservicios: descomponen el sistema en pequeñas partes independientes y escalables, lo que facilita la adaptación al crecimiento y cambios en la demanda.

Latencia: En el procesamiento de datos, la latencia se refiere al tiempo que tarda un sistema en procesar una solicitud y devolver una respuesta. La latencia puede ser afectada por varios factores, como la velocidad del hardware, la cantidad de datos que se procesan, la complejidad del algoritmo utilizado y la latencia de red. En aplicaciones de análisis de datos en tiempo real, se busca minimizar la latencia para obtener resultados lo más rápido posible.

Procesamiento en tiempo real: El procesamiento en tiempo real se refiere al procesamiento de datos a medida que se generan, en lugar de procesarlos en lotes o después de que se hayan generado. Esto permite obtener resultados y respuestas en tiempo real, lo que es importante en aplicaciones que requieren monitoreo continuo o la toma de decisiones en tiempo real. Algunos ejemplos de aplicaciones de procesamiento en tiempo real incluyen la detección de fraudes en tarjetas de crédito, la monitorización de la calidad del aire en ciudades y la detección de terremotos.

CAPÍTULO 3: HIPOTESIS Y VARIABLES

3.1 Hipótesis general:

Un sistema de aprendizaje automático distribuido mejora la satisfacción del cliente y aumenta las ventas de una ecommerce al ofrecer recomendaciones de productos personalizadas en tiempo real.

3.2 Hipótesis específicas:

H1: El sistema de aprendizaje automático distribuido procesa grandes conjuntos de datos en tiempo real de manera eficiente y efectiva.

H2: El sistema de aprendizaje automático distribuido ofrece recomendaciones de productos personalizadas precisas y relevantes a los clientes.

H3: Los clientes que reciben recomendaciones personalizadas están más satisfechos con la experiencia de compra y son más propensos a comprar más productos.

H4: La implementación del sistema de aprendizaje automático distribuido mejora la eficiencia operativa de la ecommerce al reducir el tiempo necesario para ofrecer recomendaciones personalizadas.

3.3 Identificación de variables:

Variable independiente: sistema de aprendizaje automático distribuido.

Variable dependiente: satisfacción del cliente y ventas de la ecommerce.

Variables intervinientes: Tamaño del conjunto de datos, comportamiento del cliente, complejidad del modelo de aprendizaje automático y recursos de hardware disponibles

Variables cualitativas: Tipo de datos, tipo de modelo de aprendizaje automático y grado de precisión.

3.4 Operacionalización de variables:

Variable 1: Satisfacción del cliente

- Dimensión 1: Calidad del servicio al cliente
  - Item 1: Tiempo de respuesta a las consultas de los clientes
- Dimensión 2: Experiencia de compra
  - Item 1: Facilidad para encontrar el producto deseado
  - Item 2: Proceso de pago sencillo y seguro
- Dimensión 3: Precisión de la recomendación

Item 1: Encuesta al cliente después de recibir una recomendación de producto

Variable 2: Ventas de la ecommerce

Dimensión 1: Volumen de ventas

Item 1: Número de ventas totales en un periodo determinado

Item 2: Cantidad de productos vendidos por cliente en un periodo determinado

Dimensión 2: Rentabilidad

Item 1: Costos de producción y operación del negocio

Item 2: Retorno de inversión en el sistema de aprendizaje automático distribuido

Variable 3: Cantidad de datos

Dimensión 1: Precisión del modelo de recomendación

Item 1: Nivel de acierto en las recomendaciones ofrecidas a los clientes

Item 2: Porcentaje de clientes satisfechos con las recomendaciones ofrecidas

Variable 4: Tiempo de respuesta

Dimensión 1: Velocidad de procesamiento

Item 1: Tiempo de respuesta del sistema al procesar grandes conjuntos de datos

Item 2: Tiempo de entrega de las recomendaciones al cliente

Item 3: Eficiencia del sistema al trabajar con múltiples solicitudes simultáneas

3.5 Matriz de consistencia y Matriz de Operacionalizad:

Matriz de Consistencia

Problema	Objetivo	Hipótesis	Variables	Metodología
¿Cómo diseñar y evaluar un sistema de aprendizaje automático distribuido que permita analizar grandes conjuntos de datos en tiempo real y ofrecer recomendaciones de productos personalizadas a clientes de una ecommerce, con el fin de mejorar la satisfacción del cliente y aumentar las ventas de la ecommerce?	Diseñar y evaluar un sistema de aprendizaje automático distribuido para el análisis de grandes conjuntos de datos en tiempo real, que permita dar recomendaciones de productos a clientes en el comercio electrónico	Un sistema de aprendizaje automático distribuido mejora la satisfacción del cliente y aumenta las ventas de una ecommerce al ofrecer recomendaciones de productos personalizadas en tiempo real.	Variable independiente: sistema de aprendizaje automático distribuido. Variable dependiente: satisfacción del cliente y ventas de la ecommerce. Variables intervinientes: Tamaño del conjunto de datos, comportamiento del cliente, complejidad del modelo de aprendizaje automático y recursos de hardware disponibles. Variables cualitativas: Tipo de datos, tipo de modelo de aprendizaje automático y grado de precisión.	Tipo de investigación: aplicada Nivel de investigación: aplicada Técnicas de recolección de datos: Encuestas y análisis de datos de transacciones de ventas de la ecommerce

Tabla2 Matriz de Consistencia

Matriz de Operacionalidad:

Variable	definición	dimensión	Indicador
Cantidad de datos	Número de datos procesados por el sistema de aprendizaje automático distribuido en un periodo de tiempo determinado	-Precisión del modelo de recomendación	Registro del número de datos procesados en un archivo de registro del sistema
Tiempo de respuesta	Tiempo que tarda el sistema en proporcionar una recomendación de producto personalizada después de recibir una solicitud del cliente	-Velocidad de procesamiento	Registro del tiempo transcurrido entre la solicitud del cliente y la respuesta del sistema en un archivo de registro del sistema
Satisfacción del cliente	Percepción del cliente sobre la calidad de las recomendaciones de productos personalizadas proporcionadas por el sistema	-Calidad del servicio al cliente - Experiencia de compra - Precisión de la recomendación	Encuesta al cliente después de recibir una recomendación de producto
Aumento de ventas	Cambio en la cantidad de ventas de la ecommerce después de la implementación del sistema de aprendizaje automático distribuido	-Volumen de ventas -Rentabilidad	Análisis de las ventas antes y después de la implementación del sistema en un período de tiempo determinado

Tabla3 Matriz de Operacionalidad

7. Referencias bibliográfica

[1] "Big Data Analytics | IBM," 2018. https://www.ibm.com/analytics/big-data-analytics (accessed Sep. 05, 2023).

[2] "¿Qué es el análisis de datos? - Explicación del análisis de datos - AWS." https://aws.amazon.com/es/what-is/data-analytics/ (accessed Sep. 05, 2023).

[3] F. Ocampo-Botello, F. Felipe-Durán, and R. De Luna-Caballero, "Sistema de recomendación para el comercio electrónico aplicado a una tienda de libros," vol. 18, no. 2, pp. 55–62, 2014, Accessed: Sep. 05, 2023. [Online]. Available: www.filmaffinity.com,

[4] J. Sebastián Lucero Olalla, "Diseño e implementación de un prototipo experimental para aprendizaje automático distribuido," 2022.

[5] M. B. Roldán, "Revisión de las principales metodologías para la construcción de aplicaciones distribuidas en la nube," *Tecnología Vital*, vol. 2, no. 4, 2018, Accessed: Sep. 13, 2023. [Online]. Available: https://revistas.ulatina.ac.cr/index.php/tecnologiavital/article/view/43/33

[6] J. Wei, J. He, K. Chen, Y. Zhou, and Z. Tang, "Collaborative filtering and deep learning based recommendation system for cold start items," *Expert Syst Appl*, vol. 69, pp. 29–39, 2016, doi: 10.1016/j.eswa.2016.09.040.

[7] A. F. Rojas Hernandez, N. Yaneth, and G. Garcia, "Distributed processing using cosine similarity for mapping Big Data in Hadoop," 2016, Accessed: Sep. 05, 2023. [Online]. Available: http://hadoop.apache.org

[8] K. C. Guyard and M. Deriaz, "A scalable recommendation system approach for a companies - seniors matching," *2022 5th International Conference on Artificial Intelligence for Industries (AI4I)*, pp. 5–9, Sep. 2022, doi: 10.1109/AI4I54798.2022.00008.

[9] M. Loukili, F. Messaoudi, and M. El Ghazi, "Machine learning based recommender system for e-commerce," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 12, no. 4, pp. 1803–1811, Dec. 2023, doi: 10.11591/ijai.v12.i4.pp1803-1811.

[10] E. A. Yilmaz, S. Balcisoy, and B. Bozkaya, "A link prediction-based recommendation system using transactional data," *Scientific Reports 2023 13:1*, vol. 13, no. 1, pp. 1–14, Apr. 2023, doi: 10.1038/s41598-023-34055-5.

[11] R. V. Karthik and S. Ganapathy, "A fuzzy recommendation system for predicting the customers interests using sentiment analysis and ontology in e-commerce," *Appl Soft Comput*, vol. 108, p. 107396, Sep. 2021, doi: 10.1016/J.ASOC.2021.107396.

[12] H. Lee and J. Lee, “Scalable deep learning-based recommendation systems,” *ICT Express*, vol. 5, no. 2, pp. 84–88, Jun. 2019, doi: 10.1016/J.ICTE.2018.05.003.

[13] A. Marwade, N. Kumar, S. Mundada, and J. Aghav, “Augmenting e-commerce product recommendations by analyzing customer personality,” *Proceedings - 9th International Conference on Computational Intelligence and Communication Networks, CICN 2017*, vol. 2018-January, pp. 174–180, Mar. 2018, doi: 10.1109/CICN.2017.8319380.

[14] V. B. Savadekar and M. E. Patil, “Improved recommendation system with review analysis,” *Proceedings - International Conference on Global Trends in Signal Processing, Information Computing and Communication, ICGTSPICC 2016*, pp. 79–82, Jun. 2017, doi: 10.1109/ICGTSPICC.2016.7955273.