Universidad de Lima
Facultad de Ingeniería y Arquitectura
Carrera de Ingeniería de Sistemas

ECOM ZONABIO

Trabajo de suficiencia profesional para optar el Título Profesional de Ingeniero de Sistemas

Angela Lourdes Enriquez Chavez
Código 19932246
Katya Muñoz Caballero
Código 19920521

Asesor

Mg. Enrique Palacios López

Lima – Perú
Noviembre de 2020

UNIVERSIDAD DE LIMA
MCMLXII
SCIENTIA ET PRAXIS

ECOM ZONABIO

TABLA DE CONTENIDO

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

ÍNDICE DE ANEXOS

RESUMEN

El año 2020 es un año sin precedentes, la presencia de la Covid 19 ha revolucionado el modo de hacer negocios, es así como vemos el mayor aceleramiento en el uso de las tecnologías de información a nivel mundial y también de un mejor aprovechamiento de las tecnologías emergentes tales como: Inteligencia Artificial en la Nube, Machine Learning, IoT, Analítica avanzada, Ecosistemas digitales, realidad aumentada, etc.

Si bien es cierto que ya contábamos con tiendas online como Amazon, Linio, Alibaba, etc. la situación actual nos obliga a dar un giro de 180 grados en lo que a comercio electrónico se refiere, debido a que actualmente las empresas y emprendedores que promocionan productos/servicios utilizan como estrategia de crecimiento de ventas, sistemas de recomendación mediante el uso de algoritmos inteligentes - Machine Learning, lo cual permite una mejor orientación de los productos a los consumidores de los sectores interesados.

Para el Perú es una gran oportunidad que se implemente un sistema de recomendación verde e inteligente de productos orgánicos, debido a que en la actualidad no existen negocios online con esa clasificación de sistemas de recomendaciones. Su implementación tendrá un impacto positivo, orientando a los clientes a adquirir un estilo de vida saludable, mejorando la calidad de vida de los consumidores e incrementando las ventas. Según la OMS (2020) Una alimentación saludable es muy importante durante la pandemia de COVID-19. Lo que comemos y bebemos puede afectar a la capacidad de nuestro organismo para prevenir y combatir las infecciones y para recuperarse de ellas. Por ello a nivel del globo terráqueo hay una creciente tendencia al consumo de productos ecológicos de acuerdo con el estilo de vida de las personas.

Nuestra creación es una plataforma digital innovadora llamada ECOM ZONABIO de fácil visualización en los distintos dispositivos tales como: laptops, tablets, smartphones etc.; cuya diferenciación se encuentra en los algoritmos de recomendación, lo que permitirá a los blogueros ecológicos expandirse con mayor rapidez para una mejor difusión de los productos orgánicos.

Nuestra solución también permitirá a nuestros consumidores obtener información de los productos que fueron calificados por clientes que tienen intereses o expectativas similares. Así lograremos brindar una gran experiencia al momento de buscar un producto orgánico para la compra, el cliente no deberá tener conocimientos previos de tecnologías de información avanzados como requisito para interoperar con la plataforma ECOM ZONABIO.

Palabras clave:

Consumidores, productos orgánicos, Inteligencia Artificial, Machine Learning, sistema de recomendación.

ABSTRACT

2020 has been a year without precedents, Covid 19 has revolutionized the ways of doing business, this is how we see the greatest acceleration in the use of information technologies worldwide and also a better use of emerging technologies such as: Cloud-based Artificial intelligence, Machine learning, IoT, Advanced analytics, Digital ecosystems, Augmented reality, etc.

Although it is true that we already had online stores such as Amazon, Linio, Alibaba, etc. The current situation forces us to take a 180 degree turn in terms of electronic commerce, because currently companies and entrepreneurs that promote products and services use recommendation systems, through the use of intelligent algorithms - Machine Learning, as a sales growth strategy which allows them a better customer orientation in regard their products in different business sectors.

For Peru, it is a great opportunity to implement a green and smart recommendation system for organic products, because currently there are no online businesses with that classification of recommendation systems. Its implementation will have a positive impact, guiding customers to acquire a healthy lifestyle, improving quality of life for consumers, and increasing entrepreneurs sales. According to WHO (2020) a healthy diet is very important during the COVID-19 pandemic. What we eat and drink can affect our bodies' ability to prevent, fight and recover from infections. For this reason, there is a growing trend towards the consumption of organic products according to people's lifestyles all over the world.

Our creation is an innovative product called ECOM ZONABIO that is easy to view in different devices such as: laptops, tablets, smartphones etc.; whose differentiation is in the recommendation algorithms, allowing ecological bloggers to expand their online presence more quickly for a better diffusion of organic products.

Our solution will also allow our consumers to obtain information on products and customers who have similar interests or expectations, as well as a great experience when looking for an organic product for purchase, customers will not have to have prior knowledge of advanced information technologies as a requirement to interoperate with the ECOM ZONABIO platform.

Keywords:

Consumers, organic products, artificial intelligence, Machine Learning, recommendation system.

CAPÍTULO I: INTRODUCCIÓN

Es innegable que esta pandemia ha tenido y seguirá teniendo un impacto significativo en nuestro comportamiento a nivel socio cultural económico. Repentinamente, el mundo se tuvo que acomodar a un nuevo estilo de vida y que no, precisamente, se estaba preparado. Lo cual este confinamiento con el consecutivo nerviosismo y miedo al contagio "obligó a migrar a canales digitales -en algunos casos por primera vez- para poder realizar actividades básicas como la compra de alimentos y artículos de primera necesidad" (Orams, 2020, p. 2).

Es así como muchos consumidores se incorporan al mundo digital casi por obligación, por lo que Suito (2020) señala que es importante que las marcas sepan cómo pueden conectar de nuevo con los consumidores en el tiempo de cuarentena. Desde el inicio del confinamiento los consumidores han desarrollado nuevos hábitos de consumo, nuevos comportamientos que las marcas deben considerar al momento de realizar una estrategia publicitaria, es el mejor momento para captar la atención de las personas por el medio digital a fin de incrementar la confianza en las plataformas de compra online, para que esta sea sostenible en el tiempo.

Una experiencia de cliente positiva puede ser mejorada con la implementación de un sistema de recomendaciones inteligente que aporte la valoración de los mejores productos y los más vendidos. Nuestra idea apunta a demostrar que su aplicación en nuestro marketing digital significa mucho más que un algoritmo de recomendación, que es una herramienta que puede brindar grandes experiencias y ayudar a crecer las ventas de los productos ecológicos del startup ECOM ZONABIO.

En lo referente al sistema de recomendaciones inteligentes, podemos dividir un recomendador en 4 partes asignándole un peso a cada una: la "base de conocimiento 25% (la información, los datos), el procesamiento de la base de conocimientos 5% (tecnología, algoritmos, filtros), la analítica y control de negocio 20% (medir todo, estrategia de negocio) y finalmente la interface del usuario 50%" (González, 2020)

Los sistemas de recomendación inteligentes son los sistemas de Inteligencia Artificial más utilizados en modelos de aplicación en industrias y especialmente en el comercio electrónico. Consideramos que este enfoque tiene un papel muy importante

a la hora de orientar la elección de la etiqueta ecológica por parte del consumidor. La Inteligencia Artificial en pleno boom en el desarrollo de aplicaciones nos abre el potencial para encontrar soluciones efectivas con un nuevo enfoque que estamos iniciando y que proponemos llamarlo "consumo verde e inteligente".

La tecnología de machine learning y la personalización están logrando que algunos negocios prosperen e incrementen su promedio de ventas y lo novedoso de la solución es que sintoniza con la información que difunden los blogueros acerca de las nuevas tendencias en estilos de vida saludable.

Por ello, ECOM ZONABIO utilizará sistemas de recomendación con IA que nos permita sugerir los productos que son de interés de cada uno de nuestros clientes consumidores de productos orgánicos, la plataforma empleará lo último en tecnologías Cloud, adicionalmente, ofrecerá una alternativa de compra simple, ágil, con conocimiento de los gustos y preferencias de los consumidores y también permitirá recomendar los productos novedosos, económicos y más vendidos, construyendo una relación cercana y generando una gran experiencia con cada uno de los compradores.

Para el presente trabajo planteamos el uso de Machine Learning con la expectativa de obtener excelentes resultados como es el caso de otros negocios digitales ranqueados de la música, películas, libros, etc.

Tomaremos como referencia la utilización de las mejores prácticas de innovación que generen crecimiento sostenible y cumpliendo nuestra misión de llevar a nuestros clientes hacia una nueva experiencia de consumo para una vida saludable que incluye a deportistas, diabéticos, personas con cáncer, síndrome metabólico, etc.

CAPÍTULO II: FUNDAMENTOS TEÓRICOS

En este capítulo explicaremos los conceptos que estamos empleando para la construcción de nuestro producto de innovación tecnológica basado en tecnologías Cloud, inteligencia artificial, machine learning, etc.

Comenzaremos explicándoles acerca de las tecnologías Cloud, para luego ir desarrollando los conceptos y fundamentación de cada una de las tecnologías que hemos estudiado para el presente trabajo.

Desarrollaremos con mayor detalle los conceptos relacionados a Inteligencia Artificial con machine learning, la misma que se implementa en la mayoría de las plataformas online en la actualidad como: Amazon, Spotify, Netflix, eBay, las cuales hemos analizado por sus excelentes resultados en el crecimiento de participación de mercado y por su gran potencial de explotación de información de los productos a promocionar.

Los sistemas de machine learning también están presentes en otros ámbitos, como noticias (Twitter) y redes sociales (YouTube, Instagram), etc.

A continuación, el desarrollo de los conceptos y fundamentos:

2.1 Cloud Computing, Big Data y Machine learning

El ecosistema en la nube está cambiando hacia las aplicaciones de big data. La computación en la nube, los sensores de IoT, las bases de datos y las tecnologías de visualización son indispensables para analizar big data. Estas tecnologías desempeñan un papel fundamental en los servicios cognitivos, la inteligencia empresarial, el aprendizaje automático o machine learning, el reconocimiento facial, el procesamiento del lenguaje natural, etc. (Talend.com, s.f.)

Los servicios en la nube, hoy en día, proporcionan los recursos necesarios para poder desarrollar de manera eficiente y rentable un sistema de inteligencia artificial. La nube permite procesar grandes cantidades datos, lo cual es necesario en el desarrollo de algoritmos de aprendizaje automático. Otro aspecto importante, además es que los servicios cloud proporcionan escalabilidad y elasticidad permitiendo que los recursos se adapten

a las necesidades de cada momento o en función de lo que se necesita y a costos accesibles.

2.2 Algoritmos de recomendación

Nuestra solución incluirá algoritmos de recomendación para ofrecer al cliente una experiencia personalizada en cuanto a los productos que más respondan a sus necesidades de acuerdo con un perfil especifico.

En la actualidad los algoritmos de recomendación se han vuelto muy populares en los sitios de comercio electrónico o tiendas en línea, ya que permiten a los proveedores optimizar las ventas de sus productos, pero también ayudan a los clientes a minimizar sus esfuerzos en la búsqueda de un producto ante una inmensa cantidad de información.

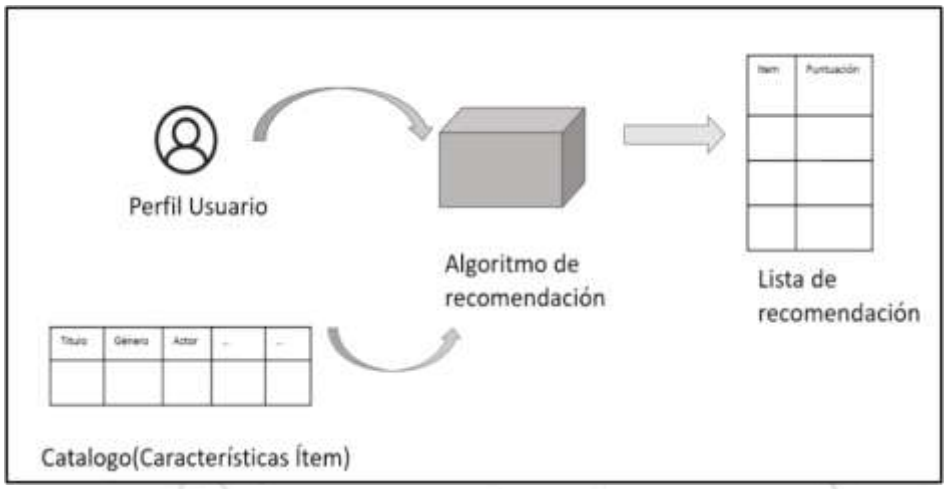
Según un informe de Gartner, citado en una página web especializada, menciona que este año "el 60% de los comercios electrónicos incluirían soluciones de Inteligencia Artificial. [Dicha] Cifra pudiera resultar ser mayor debido al efecto de la pandemia que ha elevado las compras por esta vía" (Torres, 2020).

Existen diversos algoritmos de recomendación, la mayoría de estos se basan principalmente en los siguientes modelos:

2.2.1 Modelo basado en el contenido

Consiste en recomendar al usuario lo más parecido a lo que ya había consumido o visualizado anteriormente, es decir, similitud entre ítems. (Velez-Langs & Santos, 2006)

Figura 2. 1
Modelo Basado en el Contenido



Nota. De "Sistemas Recomendadores: Un enfoque desde los algoritmos genéticos". Por Vélez y Santos. *Ind. Data*, 9 (1), p. 27. Madrid, 2006.

- **Ventajas:**

1 Recomienda ítems similares a los que los usuarios les han gustado en el pasado;
2 Se basa en el perfil del usuario para las recomendaciones;
3 No hay ningún problema de ¨arranque en frio¨ cuando se agrega un nuevo ítem al catálogo, ya que trata de coincidir con las preferencias del usuario y las características del ítem;
4 Es posible recomendar nuevos ítems o aquellos que no son muy populares.

- **Desventajas:**

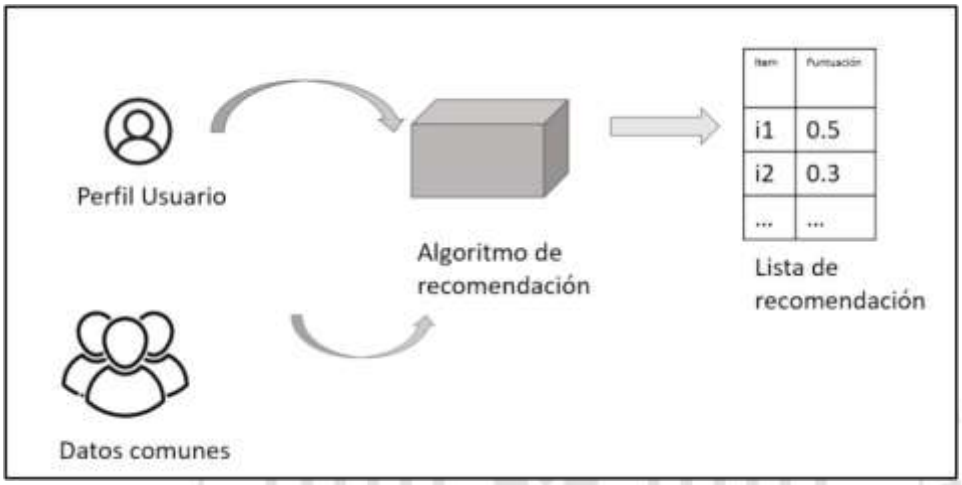
1 Cuando un cliente es nuevo, no hay historial;
2 Los clientes que han visto una gran cantidad de ítems pueden representar un problema, dada la inmensa cantidad de información en su perfil para coincidir con las características de los ítems;
3 Riesgos de sobre especialización, es decir limitación a ítems similares, por lo tanto, la respuesta será homogéneas.

2.2.2 Filtrado colaborativo

El filtrado colaborativo consiste en generar recomendaciones calculando la similitud entre las preferencias de un usuario y la de otros usuarios. Estos algoritmos no intentan

analizar o comprender el contenido o las características de los elementos recomendados. (Galan, 2007).

Figura 2.2
Filtrado Colaborativo



Nota. De "Sistemas Recomendadores: Un enfoque desde los algoritmos genéticos". Por Vélez y Santos. *Ind. Data,* 9 (1), p. 27. Madrid, 2006.

Según Breese, Heckerman & Carl (1998) mencionaron que los algoritmos de filtrado colaborativo pueden dividirse en 2 categorías:

a. Enfoque basado en la memoria

También llamados basados en vecinos, contiene 2 tipos de relaciones de modelo colaborativo usuario-usuario y modelo basado en ítem-ítem:

- Usuario-Usuario: En estos algoritmos, la predicción para un cliente se basa en las evaluaciones anteriores que han tenido el ítem y el grado de similitud de los usuarios que lo evaluaron, con el usuario actual.
- Ítem-ítem: El algoritmo predice la evaluación de un ítem en base a las evaluaciones que han tenido los ítems más cercanos a este; y al grado de similitud entre ellos. Este enfoque calcula la similitud entre los ítems en función de las evaluaciones de los usuarios.

b. Enfoque basado en el modelo

En este enfoque, se requiere primeramente crear modelos que se asemejen al comportamiento de los usuarios y así utilizar estos modelos para hacer las recomendaciones.

Entre las ventajas e inconvenientes del modelo de filtro colaborativo, podemos mencionar:

Ventajas:

1. Utilizan las puntuaciones o clasificaciones de otros usuarios para evaluar la utilidad del ítem.
2. Encuentran usuarios o grupos cuyos intereses coinciden con el usuario actual.
3. Cuantos más usuarios, se puede calcular mejor las puntuaciones y los resultados para las recomendaciones serán mejores.

Desventajas:

1 Dificultad de encontrar usuarios o grupos de usuarios similares.

2 No contar con rating cuando se agrega un nuevo ítem.

3 No contar con historial cuando se agrega un nuevo usuario, por lo que se vuelve difícil la recomendación.

2.2.3 Método Hibrido

Combina los 2 modelos precedentes, por lo tanto, este método considera las características de diferentes contenidos, así como los perfiles de los diversos usuarios. El objetivo es apoyarse en todas las fuentes de conocimiento, eligiendo las opciones más adecuadas para una determinada tarea, con el fin de utilizarlas de la manera más eficaz posible.

Dado que este enfoque se basa en la combinación de los 2 primeros modelos, se toman las ventajas de estos y al mismo tiempo limita sus inconvenientes o desventajas.

Hay tres categorías principales de sistemas de recomendación para diseñar un sistema de recomendación híbrido: la combinación monolítica, la combinación en paralelo y la combinación tubular. Por otro lado, es importante mencionar que para construir un modelo de recomendación se necesitan datos. Los datos a su vez se recopilan de forma explícita o implícita.

- Explicita: el usuario indica explícitamente su interés, ya sea atribuyendo una clasificación o rating al producto o indica su apreciación (con un ¨like¨, por ejemplo)
- Implícita: Se basa en la observación y en el análisis del comportamiento implícito del usuario en la aplicación o página web, por ejemplo: su historial de búsqueda, los clics realizados en la página, el tiempo de visita en la página, etc.

2.3 Aprendizaje automático (Machine learning)

"El Machine Learning es una disciplina del campo de la Inteligencia Artificial que, a través de algoritmos, dota a las computadoras la capacidad de identificar patrones en datos masivos para hacer predicciones**.** Este aprendizaje permite a las computadoras realizar tareas específicas de forma autónoma, es decir, sin necesidad de ser programados" (Iberdrola.com, 2020).

Los algoritmos de Machine learning son, por lo tanto, algoritmos que aprenden (y a menudo predicen) a partir de su experiencia con datos. En lugar de formalizar las reglas manualmente, un algoritmo de aprendizaje automático aprenderá a partir de estos datos sin procesar que se le proporcione para tomar mejores decisiones en el futuro sin intervención humana.

2.3.1 Categorías de Machine Learning

Dependiendo de la naturaleza del problema que se está tratando, existen diferentes enfoques que varían según el tipo y volumen de datos:

2.3.1.1 Algoritmos de aprendizaje automático supervisado

En general comienza con un conjunto de datos bien definidos y cierta compresión de cómo se clasifican estos datos. El objetivo del aprendizaje supervisado es descubrir patrones en los datos y aplicarlos a un proceso analítico.

El aprendizaje automático supervisado se puede utilizar para hacer predicciones sobre datos futuros (hablamos de "modelado predictivo"). El algoritmo intenta desarrollar una función que predice con precisión la salida de las variables de entrada. “Esta técnica es útil cuando se sabe cuál debe ser el resultado” (Microsoft Azure, 2020).

Entre los principales algoritmos de aprendizaje automático supervisados están considerados los siguientes: bosques aleatorios, árboles de decisión, método de k vecino más cercano (k-NN), regresión lineal, máquina de vectores de soporte (SVM), regresión logística, etc.

2.3.1.2 Algoritmos de aprendizaje automático no supervisados

Se utilizan cuando el problema requiere una gran cantidad de datos sin clasificar o no está etiquetada. Para entender el significado de estos datos, es necesario utilizar algoritmos que clasifiquen los datos según las tendencias o clústeres que detectan. “Esta técnica es útil cuando no se sabe cuál debe ser el resultado” (Microsoft Azure, 2020).

Los principales algoritmos para el aprendizaje automático no supervisado son: K-Means, agrupación en clústeres / agrupación jerárquica y reducción de la dimensionalidad.

2.3.1.3 Los algoritmos de aprendizaje automático de refuerzo

Es un modelo de aprendizaje conductual. El algoritmo recibe información del análisis de datos y guía al usuario hacia el mejor resultado. El aprendizaje por refuerzo se diferencia de otros tipos de aprendizaje supervisado en que el sistema no se entrena con un conjunto de datos de muestra. En cambio, el sistema aprende a través de un método de prueba y error. Este método permite que las máquinas determinen automáticamente el comportamiento ideal en un contexto específico para optimizar su rendimiento. “Es una buena técnica para usarla en sistemas automatizados que tienen que tomar muchas

decisiones pequeñas sin indicaciones por parte de humanos" (Microsoft Azure, 2020, párr. 26).

En el artículo de García-Ramírez, Morales & Escalante (2019) mencionan que los principales algoritmos de aprendizaje automático de refuerzo son: Q-learning, Deep Q Network (DQN) y SARSA (Estado-Acción-Recompensa-Estado-Acción).

El aprendizaje automático requiere de un conjunto de datos adecuados para el proceso de aprendizaje. La Big data por lo tanto contribuye a mejorar la precisión de los modelos de aprendizaje automático.

Como podemos apreciar, existe un gran número de algoritmos de machine learning que pueden ser efectivos según el problema que se desea resolver. El modelo que se elija por lo tanto depende en gran medida del problema y los datos que se está analizando.

2.3.2 Pruebas y validación

El objetivo de un modelo de aprendizaje automático es identificar patrones de un conjunto de datos. Una vez identificado los patrones, estos se usan para realizar predicciones con datos nuevos.

Un método muy usado para entrenar un modelo consiste, según Recuero de los Santos (2020), en la división de los datos en 2 conjuntos:

- Datos de entrenamiento (training set), que se utilizará para entrenar el modelo
- Datos de prueba (test set), que se usará para probar el modelo entrenado.

Usualmente los datos se dividen en 70% o 80% para entrenamiento y 30% o 20% para las pruebas, sin embargo, esta división dependerá del tamaño del conjunto de datos. Lo más importante es evitar el sobreajuste (un modelo sobreentrenado) o sub-ajuste (ocurre cuando el conjunto de datos de entrenamiento es insuficiente).

Se pueden usar diferentes tipos de métricas para medir si un modelo está aprendiendo o no. Para los algoritmos de aprendizaje supervisado, muchas medidas de rendimiento miden el número de errores de predicción.

Según Géron (2019) menciona que la tasa de error en los nuevos casos se denomina error de generalización o error fuera de muestra, y al evaluar el modelo con los datos de prueba (test set) se obtiene una estimación de este error. Este valor indica que tan bien funcionara el modelo en instancias no vistas antes.

Otra técnica muy usada para evaluar un modelo de machine learning es la validación cruzada o cross validación. La validación cruzada es un método estadístico que permite medir la capacidad de generalización de un modelo.

La variante más común de validación cruzada es el K-fold cross-validation, que consiste en dividir todos los datos disponibles en k partes iguales (pliegues), con los cuales entrenaremos y probaremos el modelo durante k iteraciones. En cada iteración, el modelo se entrena con k-1 pliegues y se prueba con el pliegue restante. Una vez finalizadas las k iteraciones se calcula la precisión y el error para cada uno de los modelos y para obtener la precisión y el error final se calcula el promedio de los K modelos entrenados.

El modelo de machine learning seleccionado por lo tanto será aquel que produzca el mejor valor de precisión y menor error promedio.

2.3.3 Selección del algoritmo

Actualmente, Sheng (2020) menciona que dentro del enfoque de machine learning se encuentran 3 tipos de enfoques de algoritmos: supervisados, no supervisados y de reforzamiento. Dado que lo que se busca es implementar un sistema de recomendación inteligente, el tipo de algoritmo cuyo uso será bajo un enfoque supervisado nos permitirá predecir de forma más certera, con los datos etiquetados y a través de un proceso analítico, lo que le resultaría interesante al usuario.

Dentro de estos algoritmos supervisados podemos citar:

- Algoritmo de árbol de decisión

Se utiliza un árbol de decisión para clasificar las observaciones futuras dado un conjunto de observaciones ya etiquetadas. Con este conjunto de datos que se

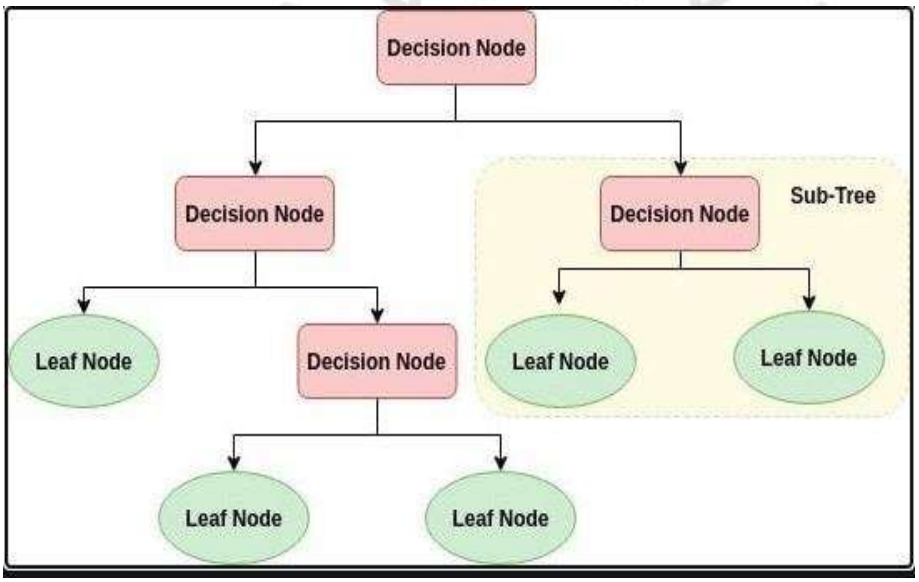
tiene, “se fabrican diagramas de construcciones lógicas que sirven para categorizar una serie de condiciones que ocurren de forma sucesiva, para la resolución de un problema”. (Sierra, Arbelaitz, Armañanzas, Arruti, & Bahamonde, 2006, p. 11).

El árbol de decisión tiene una estructura similar a un diagrama de flujo, está compuesto de nodos, cada uno de los cuales tiene un cierto número de variables. El árbol comienza con un nodo (parte superior del árbol), donde se toman decisiones en base a diferentes atributos disponibles para descartar opciones que no cumplen y así sucesivamente, hasta llegar al resultado ideal para dichas condiciones.

Aguirre (2019) menciona que:

El árbol cuenta con nodos internos que generan una división o Split en los datos que recibe, y los nodos externos que ya no realizan divisiones, sino que cuentan con una predicción y es el final de la rama o hojas (párr. 19).

Figura 2.3
Árbol de decisión



Nota. De “Decision tree classification in Python”. Por Navlani A. *Data Camp*. 2018. (https://www.datacamp.com/community/tutorials/decision-tree-classification-python)

- Algoritmo K-NN

El algoritmo K-NN, también conocido como vecinos cercanos, es un método de aprendizaje supervisado que puede ser utilizado tanto para la regresión como

para la clasificación. Para poder hacer una predicción, este algoritmo se basa principalmente en clasificar los nuevos datos de entrada en la misma clase que tenga la mayor cantidad de vecinos más parecidos a ellos en el conjunto de entrenamiento. (Ray, 2020)

Para predecir la clase de un nuevo dato de entrada, buscará por lo tanto sus K vecinos más cercanos (usando la distancia euclidiana u otra) en el espacio de las características identificadas por aprendizaje.

El algoritmo K-NN necesita una función para calcular la distancia entre dos observaciones (cuanto más cerca están dos puntos, más similares son), Existen varias funciones de cálculo de distancia como la distancia euclidiana, la distancia de Manhattan, la distancia de Hamming, la distancia del coseno, error cuadrático medio, etc. La distancia euclidiana es muy utilizada, por ejemplo, para atributos numéricos y la distancia de Hamming para atributos nominales. (Bedoya P., 2011).

Figura 2.4
Modelo K-Nearest Neighbor (KNN)

Nota. De "Aprende Machine Learning en Español. Teoría + Práctica Python". Por Ignacio, J. 2020. Basado en el blog https://www.aprendemachinelearning.com/clasificar-con-k-nearest-neighbor-ejemplo-en-python/

En este capítulo se han presentado los principales conceptos relacionados a los algoritmos de recomendación y de aprendizaje automático (Machine learning). Hemos

descrito los enfoques y las medidas de evaluación en base a pruebas y al cálculo del error.

Entender el problema que se desea resolver, así como saber que datos tenemos disponibles (suficientes y relevantes), permitirá seleccionar y construir un modelo de machine learning que mejor responda a nuestras necesidades y objetivos.

Entre los modelos de Machine learning que hemos descrito previamente, se puede deducir que el algoritmo K-NN es el más adecuado para el desarrollo de un sistema de recomendación.

CAPÍTULO III: FUNDAMENTACIÓN DEL PROYECTO

“Ante un presuroso incremento de las herramientas digitales en algunos sectores importantes del mercado peruano, hemos presenciado cómo el comercio electrónico ha crecido, haciendo que muchas empresas que tradicionalmente no consideraban el e-commerce como una alternativa, tengan ahora una presencia bastante sólida con un crecimiento importante en la internet”. (Fuentes, 2020).

Existen aplicaciones en plataformas de comercio electrónico como la de Amazon (Ménard, 2017), que utilizan algoritmos que identifican las características del producto y los historiales de compra de acuerdo con los intereses de los clientes. En nuestro proyecto se implementarán esas características dentro de la primera fase, para una segunda fase estamos ideando ofrecer publicidad para atraer a los clientes e invitarlos a comprar nuestros productos, lo cual hará más atractivo nuestro sistema de recomendación.

Nuestra propuesta de innovación tecnológica la hemos llamado “ECOM ZONABIO” la misma que tiene un predictor cuya función es actuar como “eco asesor inteligente' dígase en otras palabras, un recomendador inteligente de productos orgánicos, el cual se basa en los datos de 'consumidores' y datos de 'productos saludables' para constituir una base de datos exploratoria y así utilizar el aprendizaje que permita realizar propuestas y recomendar el mejor producto de acuerdo al rating otorgado por los clientes. (McKenna, Richardson, & Thomson, 2012).

Varios enfoques de Machine Learning pueden satisfacer la necesidad de sistemas de recomendación como Reinforcement Learning o incluso algoritmos de árbol de decisión (Decision Tree). que también podrían ser considerados, pero para nuestro caso hemos elegido trabajar con el algoritmo KNN como sistema de recomendación basado en filtrado colaborativo (Nilashi, Ibrahim, & Bagherifard, 2018; Golovin & Rahm, 2005; Pazzani & Billisus, 2007).

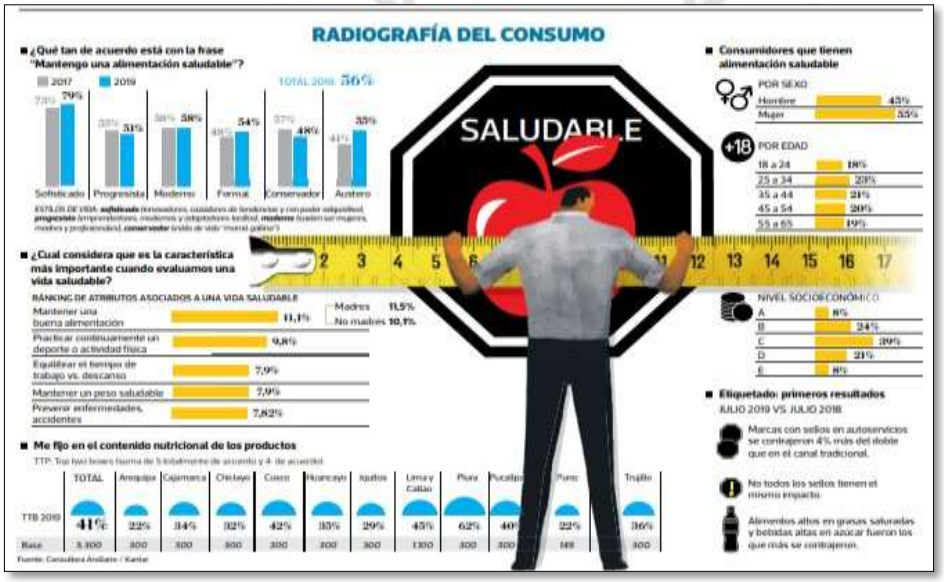
3.1 Deseabilidad del proyecto

De acuerdo con los últimos datos obtenidos, el consumo de productos orgánicos continuará creciendo, consolidándose en todos los países del mundo. (Helga & Lemond, 2019).

La agricultura orgánica está destinada a producir alimentos nutritivos y de alta calidad que contribuyan a la atención médica preventiva y al bienestar de sus consumidores permitiéndoles mantener un estilo de vida saludable.

Con el surgimiento del COVID 19, el comercio electrónico de productos orgánicos ha repuntado en nuestra región, debido al masivo cierre de eco tiendas, mercados y bioferias por prevención al contagio. Sin embargo, la coyuntura ha impulsado la masificación del teletrabajo y mayor atención en el cuidado de la salud, promoviendo el consumo de alimentos con nutraceúticos y priorizando la calidad de vida, la tranquilidad, la prevención, favoreciendo un mejor estilo de vida y consumo de Healthy Food.

Figura 3. 1
Radiografía del consumo



Nota: De "La migración al consumo saludable". Por Inga M., Claudia. En *Suplemento Día 1,* p. 2, por el Diario El Comercio. 2019.

Por otro lado, los productos ecológicos se ofrecerán a través de una solución innovadora que mostrará los beneficios, el ranking de acuerdo con las compras realizadas, también realizará las recomendaciones de los productos que sean de interés del comprador, lo cual impulsará el crecimiento de compras orgánicas con las predicciones que reciben los clientes de los algoritmos de recomendación.

Podemos inferir, que el comportamiento del consumidor peruano respecto a sus preferencias alimenticias ha cambiado, pues está buscando nuevos hábitos de consumo, el cual quedó comprobado, ya que "el 90% de las personas prefieren adquirir alimentos saludables" por internet, por ello, están dispuestos a pagar en la compra de estos productos, en comparación al gasto en comidas no saludables. La especialista Martha Neves sostiene que antes "el interés estaba enfocado en bajar de peso, pero ahora se ha re direccionado hacia una alimentación sana que se convierte en un nuevo estilo de vida" (Neves, 2017).

Dada la coyuntura, las empresas tienen una alta responsabilidad de comprender cómo los cambios ocurridos tras la pandemia seguirán afectando a las respuestas de sus compradores. De ahí surge la importante labor de analizar el crecimiento del e-commerce mientras dure el coronavirus. Los negocios están expandiendo sus ventas mediante la incursión al comercio electrónico para satisfacer las nuevas necesidades de los clientes. Este cambio está descontrolando a muchos otros negocios. "Los consumidores comienzan a comprar determinadas clases de producto online que no se había previsto" (Antevenio, 2020, párr. 4).

Según la web Andina (2020) menciona que "el porcentaje anual de cambio del número de transacciones realizadas online comparadas con las transacciones realizadas en persona fue 5.8 veces mayor en mayo del 2020 que el promedio de enero y febrero del 2020", mostrando una aceleración en el uso del e-commerce por los tarjetahabientes de la región, durante el periodo analizado de la pandemia de COVID-19. De esta situación se puede concluir que:

> Hay impacto positivo en las categorías de bienes digitales (juegos, aplicaciones, libros, plataformas de contenido y streaming, TV en línea), pagos recurrentes (TV por cable o satelital, radio, servicios básicos como electricidad y gas, subscripciones), y servicios profesionales (servicios de limpieza,

servicios a domicilio, almacenamiento, mudanzas, servicios financieros y gastos legales, gastos de educación). (Andina, 2020, párr. 10).

Para definir la deseabilidad hemos revisado diversos papers que se han publicado en la pandemia, también se ha realizado una investigación de mercados y entrevistas a profundidad realizadas a profesionales de diferentes edades; donde se concluyó que debido a la pandemia las personas buscan confort y productos recomendados, además que el público prefiere personalización.

3.1.1 Definición de Modelo de Negocio

Para la definición del modelo de negocio a implementar se tomó como guía el modelo Canvas (Business Model Canvas), que fue creado por Alex Osterwalder, siendo publicado en el libro "Business Model Generation" (Osterwalde & Pigneur, 2011). El modelo Canvas consta de 9 aspectos a tomar en cuenta como modelo de negocio y se podrá observar en la figura 7.1. más adelante.

3.1.2 Segmento de clientes

En cuanto al segmento de clientes el presente plan de negocio considera como mercado efectivo a hombres y mujeres de edades entre los 25 y 64 años, que se encuentren en el nivel socio económico A, B y C1, y que vivan en la Zona 7 de Lima Metropolitana. Y personas que buscan estilos de vida saludable, deportistas, vegetarianos, veganos, o también que busquen personalización o un estilo de vida sofisticado.

3.1.3 Propuesta de valor

Ofrecer un algoritmo de recomendación dentro de la plataforma ECOM ZONABIO que propicie identificar nuevas propuestas de productos saludables que satisfagan los gustos o las necesidades de nuestros consumidores lo que contribuirá a crear relaciones duraderas. Nuestro sistema se alimenta y aprende de la información recibida de los clientes que tienen intereses similares, lo que permitirá formular estas recomendaciones de los productos en base a la información aprendida, direccionándolos hacia un estilo de vida saludable.

Adicionaremos otros beneficios en la plataforma donde los clientes podrán encontrar también recetas, webinars, actividades, talleres etc. donde se promueven una vida con buena salud.

3.1.4 Canales

Estamos proponiendo realizar la atención mediante los siguientes canales:

- Página web
- Redes sociales
- App
- Chatboots
- Alianzas con: Gimnasios, Influencers de vida saludable, eventos relacionados, ferias, etc.

Los pedidos serán atendidos virtualmente, siendo la zona de reparto cualquier distrito correspondiente a la Zona 7 de Lima Metropolitana.

Para mantener una fluida comunicación con nuestros clientes publicaremos los catálogos de los productos en las redes sociales, estimamos que los medios de primer orden serán las redes sociales tipo Facebook, WhatsApp e Instagram, blogs o sites.

3.1.5 Relación con clientes

Se enfoca en generar experiencias satisfactorias, por lo cual consideramos importante la valoración que ellos realizan acerca de nuestros productos, el resultado debe conducirlos y estimularlos a mantener un estilo de vida saludable.

Nuestros medios digitales permitirán una fácil interacción con el usuario, en relación a los productos que se ofrecen ellos contarán con fotografías en 3D y además ofrecerán la opción de pago en línea, adicionalmente ofreceremos asesoría profesional.

- **Chat en línea**: Se ofrecerá el servicio de chat en línea para brindar orientación a todos los usuarios frente a las dudas que puedan surgir.
- **Redes sociales**: Mantendremos una actividad de permanente acercamiento con los clientes mediante publicaciones con las novedades y ofertas del mes o de la semana.

3.1.6 Fuentes de ingresos

Nuestra principal fuente de ingresos se origina por el incremento de ventas, el cual depende de la implementación del modelo IA dentro de nuestra plataforma ecommerce. Contaremos con una variedad de productos orgánicos, naturales y de diferentes marcas. Proponemos que las formas de pago serán efectuadas por canales digitales o tarjetas de crédito.

3.1.6.1 Recursos claves

Contaremos con expertos en desarrollo de tecnologías emergentes, tecnologías cloud, comunity manager con experiencia en contenidos y beneficios de productos orgánicos para mantener un estilo de vida saludable., que cuenten con potencial innovador, enfocados en el trabajo en equipo, trato al cliente, comprometidos con el bien social y el medio ambiente.

Nuestro sistema se alojará en una plataforma cloud computing de alta disponibilidad en clúster, contaremos con dispositivos tecnológicos de última generación que permitan realizar las compras y ventas en línea sin problemas de conectividad.

3.1.6.2 Actividades claves

Mantendremos actualizados los contenidos, catálogos de productos y precios, promoveremos la venta online de nuestros productos 100% orgánicos, con calidad comprobada. Actualización permanente de la plataforma y distribución eficaz. Reforzaremos nuestras ventas brindando servicios de preventa y post venta que recepcionen y brinden soporte a los clientes.

Contaremos con un modelo de IA que incorpore la mejora continua, adicionalmente consideramos realizar publicidad de nuestra marca y productos a ofertar, mediante redes sociales, página web interactiva, que en conjunto logren transmitir una experiencia de vida.

3.1.6.3 Asociaciones claves

Como socios clave contaremos con organizaciones que promuevan estilos de vida saludable, proveedores orgánicos, proveedores de servicios de internet,

de plataformas tecnológicas cloud, empresas de delivery, bancos y empresas que ofrezcan pagos digitales certificados.

3.1.7 Estructura de costos

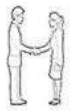
La proyección de ingresos, costos y gastos se realizan en base a las estimaciones de la demanda, información cuantitativa del estudio de mercado (encuestas) y costos requeridos para la operatividad continuidad del negocio.

- Tecnologías: Plataforma e-commerce, algoritmos machine learning, medios de pago certificados, también hemos considerado invertir en marketing digital en redes sociales, desarrollo de contenidos, herramientas y equipos de última generación.
- Producción: se adquirirán variedades de productos orgánicos, como alimentos y nutraceúticos. También se realizará la contratación de maquila y packing.

3.1.8 Modelo del Negocio.

Figura 3.2

Modelo Canvas para ECOM ZONABIO

<table>
<tr><th>Asociaciones Clave</th><th>Actividades clave</th><th>Propuesta de valor</th><th>Relaciones con los clientes</th><th>Segmentos de mercado</th></tr>
<tr><td rowspan="2">● Proveedores estratégicos como: servicios y plataforma de internet, plataformas cloud.
● Bancos
● Proveedores y Consumidores de productos orgánicos
● Empresas de delivery
● Organizaciones que promueven la vida saludable
● Medios de pago digitales certificados.
 
  </td><td>● Promover ventas online de productos orgánicos.
● Actualización de contenidos, catálogos y precios.
● Actualización permanente de la plataforma y distribución eficaz.
● Servicio de pre venta y post venta
● Ofertar por redes sociales,
● Página web interactiva.
● Delivery
  </td><td>● Crear un algoritmo de recomendación dentro de la plataforma ECOM ZONABIO que permita la identificación de nuevos productos saludables que satisfagan los gustos o necesidades de los consumidores con la finalidad de crear relaciones duraderas mediante el sistema machine learning.
  </td><td>● Conocer los gustos y preferencias de los consumidores
● Ofrecer productos y Servicios de calidad
● Chat en línea
● Redes sociales
● Webinars
● Feria virtual
  </td><td rowspan="2">● Hombres y mujeres de edades entre los 25 y 64 años.
● NSE: A, B y C1
● Factor geográfico: Zona 7 de Lima Metropolitana
● Emprendedores y productores
● Deportistas, niños
● Ancianos
● Pacientes síndrome metabólico
● Vegetarianos
● Veganos
● Clientes que busquen estilo de saludable
● Personalización
 
  </td></tr>
<tr><td>Recursos clave
● Desarrollador de tecnologías emergentes
● Desarrollador de contenidos.
● Internet
● Plataforma cloud
● Comunity Manager
● Disponibilidad de Clúster</td><td>  </td><td>Canales Distribución
● Página de internet
● Social media
● App
● Chatboots
● Alianzas estratégicas</td></tr>
<tr><td colspan="2">Estructura de costos
● Tecnologías y Producción
● Administrador web
● Community manager
● Marketing y comercial online
● Gestión de plataforma tecnológica cloud
● Costos fijos y variables</td><td colspan="3">Flujos de ingresos
● Ventas directas de los productos a través del Marketplace.
● Publicidad que se obtendrá por la divulgación de los productos en el medio digital como lo hace Google, YouTube o las otras redes sociales con énfasis en la llamada publicidad nativa.</td></tr>
</table>

3.2 Factibilidad del Negocio

Dentro del mundo, actualmente hay una alta demanda por productos que tengan personalización, el público desea encontrar recomendaciones de los productos que va a consultar antes de comprar queremos dirigirnos a este mercado que es un público entre 25 a 65 años con nuestra propuesta que tiene esta característica que va a ser través de medios digitales, la intención es tratar en la medida del tiempo hacer crecer y se convierta en una comunidad virtual de tal manera que la relación con los clientes sea muy cercana muy uno a uno, y esto lo vamos a conseguir en función a que podamos tener una comunicación constante mediante mailings, novedades y también haciendo recomendaciones a través de los algoritmos de recomendación y ofreciendo productos frescos a través de nuestras alianzas con diferentes proveedores de productos naturales y expertos de tecnologías.

Por otro lado, también nos apoyamos en información obtenida de la Asociación Peruana de Empresas de Inteligencia de Mercados ([APEIM], 2019) se obtuvo el gasto promedio mensual en soles de los peruanos en los Niveles socio económicos A, B, C, D, E en los rubros conservación de la salud, cuidados personales y servicios médicos.

Tabla 3.1

Gasto promedio mensual en soles por hogar para cuidado, conservación de la salud y servicios médicos - Lima Metropolitana por zonas.

Gasto PROMEDIO mensual en soles	NSE A	NES B	NSE C	NSE D	NSE E	Total
Grupo 1: Alimentos	1,530	1,541	1,273	1,046	805	1,245
Grupo 2: Vestido y Calzado	379	258	165	125	103	185
Grupo 3: Alquiler de vivienda, Combustible, Electricidad y Conservación de la Vivienda	1,180	705	415	300	205	482
Grupo 4: Muebles, Enseres y Mantenimiento de la vivienda	889	267	133	93	73	190
Grupo 5: Cuidado, Conservación de la Salud y Servicios Médicos	680	363	220	150	102	253
Grupo 6: Transportes y Comunicaciones	1,357	695	297	144	82	396

(continúa)

(continuación)

Gasto PROMEDIO mensual en soles	NSE A	NES B	NSE C	NSE D	NSE E	Total
Grupo 7: Esparcimiento, Diversión, Servicios Culturales y de Enseñanza	1,443	832	415	217	140	502
Grupo 8: Otros bienes y Servicios	490	287	201	142	120	217
PROMEDIO GENERAL DE GASTO FAMILIAR MENSUAL	7,949	4,858	3,119	2,218	1,629	3,470
PROMEDIO GENERAL DE INGRESO FAMILIAR MENSUAL	13,253	7,181	4,080	2,775	1,965	4,841

Nota. Los datos son extraídos de la Asociación peruana de Empresas de inteligencia de Mercados (2018).

3.3 Beneficios esperados

Finalmente, podemos inferir de las investigaciones de algoritmos de recomendación y personalización que nuestro modelo brindará los siguientes beneficios:

- Cubrir un nicho de mercado significativo con la introducción del algoritmo de recomendación en nuestra plataforma digital, logrando una mayor promoción del consumo orgánico y reducción de precios.
- Mejorar la calidad de vida y salud de los consumidores de productos orgánicos, a través de las recomendaciones que realizará el algoritmo de personalización permitiendo un mayor volumen de ventas en función a las recomendaciones.

CAPÍTULO IV: DEFINICIÓN DEL PROYECTO

4.1 Definición del proyecto

Desarrollar una plataforma de comercio electrónico para la adquisición de productos orgánicos y naturales basado en un sistema de recomendación mediante la técnica de filtrado colaborativo utilizando el algoritmo KNN. Este algoritmo permitirá que de acuerdo con los datos recopilados (en nuestro caso las calificaciones o ratings) que los clientes otorgaron a ciertos productos, predecir para un nuevo producto no calificado y dentro del conjunto de ítems similares que calificación o rating el cliente le daría y así poder recomendarlo.

Nuestro sistema de recomendación como núcleo esencial en nuestro sitio web nos posibilitará, producir una lista personalizada de sugerencias de productos en relación con los intereses de nuestros clientes. El enfoque basado en ítems permite, además, recomendar productos completamente diferentes a los que ya un cliente puede haber adquirido antes y no enfocarse solamente en recomendar el mismo tipo o categoría de productos. Esto suscitará el interés de nuestros clientes en nuevos productos de tendencia actual que les posibilitará mantener un estilo de vida más saludable.

4.2 Objetivos del proyecto

4.2.1 Objetivo general

Ofrecer una experiencia personalizada a nuestros clientes, a través de nuestro sitio web ECOM ZONA BIO, recomendando productos que serán de su interés para mantener o adoptar un estilo de vida saludable y posicionar a los emprendedores de productos orgánicos en el contexto del mercado actual.

4.2.2 Objetivos específicos

- Realizar un algoritmo de predicción que permita ofrecer productos con alta probabilidad de compra.
- Analizar las ventajas y el rendimiento de los diferentes algoritmos K-NN en nuestro sistema de recomendación.

- Analizar el algoritmo K-NN utilizando diferentes medidas de similitud (Coseno, MSD y Pearson).
- Evaluar entre los dos enfoques del filtrado colaborativo, basado en usuario y basado en Ítems, y seleccionar el que menor error nos proporcione (MAE).
- Afinar el algoritmo con parámetros que mejoren la predicción de la calificación de nuestros productos orgánicos.

4.3 Beneficios esperados

Posicionar ECOM ZONABIO como la principal plataforma on-line de venta de productos orgánicos y naturales captando, a través de nuevas tecnologías emergentes como la inteligencia artificial, las necesidades y gustos de los clientes en la búsqueda de un estilo de vida saludable. Así mismo, incrementar las ventas de estos productos en los diferentes canales digitales y contribuir al crecimiento y a la expansión del comercio justo de productos orgánicos.

4.4 Segmento de Mercado

Tabla 4.1

Segmento de Mercado

Usuarios/Clientes	Necesidad	Expectativas/Preocupaciones	Impacto
Compradores de productos orgánicos	Encontrar un canal de compra ágil de productos orgánicos	Adquirir productos organicos y recibir recomdaciones de los productos mas rankeados	Adquirir productos 24*7 a precios competitivos; Recibir oportunamente los productos en el sitio donde se encuentran.
		Navegación personalizada y respuestas rápidas.	Recibir actualizaciones de las promociones y nuevos productos.
		Esperan campañas y ofertas de productos	Brindar orientación de servicio al cliente ofreciendo productos personalizados; según estilo de vida.
		Encontrar propuestas, recetas, etc. de acuerdo con cada perfil	Adoptar un estilo de vida saludable.
Nuevos clientes	Buscar un estilo de vida saludable	Encontrar los productos que necesito con poco esfuerzo	Adquirir productos en un tiempo razonable en un mismo lugar y con la comodidad de recibirlos en donde el cliente se encuentre.
		Encontrar recomendaciones de productos relevantes que faciliten mi elección	Realizar mis compras sin esfuerzo.

4.5 Roles y responsabilidades del equipo del proyecto

Tabla 4.2
Roles y responsabilidades

- PO = Product Owner
- SM = Scrum Master
- ES = Equipo Scrum/Desarrolladores

Fase	Responsabilidades	Roles
1. INICIO	1. Crear la Visión del Proyecto	PO
	2. Identificar al Scrum Master y a los interesados	PO , SM
	3. Formar el Equipo Scrum	PO , SM , ES
	4. Desarrollo de Epicas	PO , SM , ES
	5. Crear la Lista Priorizada de Pendientes del algoritmo	PO , SM , ES
	6. Realizar la Planificación del Lanzamiento del algoritmo	PO , SM , ES
2. PLANIFICACION Y ESTIMACION	7. Crear Historias de Usuarios	PO , SM , ES
	8. Aprobar, estimar y asignar las Historias de Usuarios al equipo	PO , SM , ES
	9. Crear las Tareas	PO , SM , ES
	10. Estimar las Tareas	PO , SM , ES
	11. Crear la Lista de Pendientes del Sprint	PO , SM , ES
3. IMPLEMENTACION	12. Crear Entregable	PO , SM , ES
	13. Realizar un Standup Diario	SM , ES
	14. Mantenimiento Priorizado de los Pendientes del Producto	PO , SM , ES

(continúa)

(continuaión)

Fase	Responsabilidades	Roles
4. REVISION Y RETROSPECTIVA	15. Convocar al Scrum de Scrum	SM , ES
	16. Demostrar y validar el Sprint	PO , SM , ES
	17. Retrospectiva del Sprint	SM , ES
5. LANZAMIENTO	18. Envío de los Entregables	PO
	19. Retrospectva del Proyecto	PO , SM , ES

Nota. Los datos son extraídos de Scrum Body of Knowledge (SBOK GUIDE)

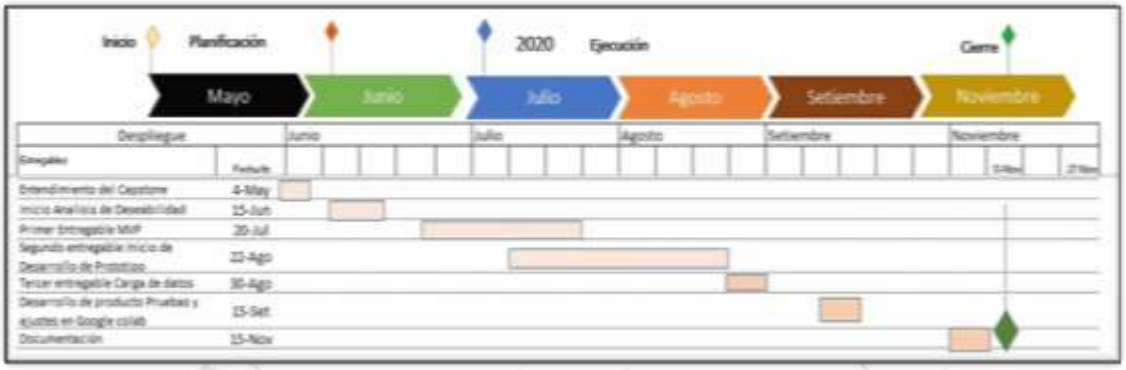
En el siguiente cuadro se muestran las principales actividades que se realizaran en la primera parte del proyecto:

Tabla 4.3
Matriz RACI

Actividades	PO	SC	ES Programador	ES Certificador
Concepción y planificación del proyecto	C,A	R	C	C,I
Redactar la documentación del proyecto	I/A	R	R,I	R,I
Análisis y Diseño	I	C,I	R	R
Prototipo/ Desarrollo	I	A	R	R
Pruebas Integrales	I	A	R	R
Go live	A	C,I	R	R

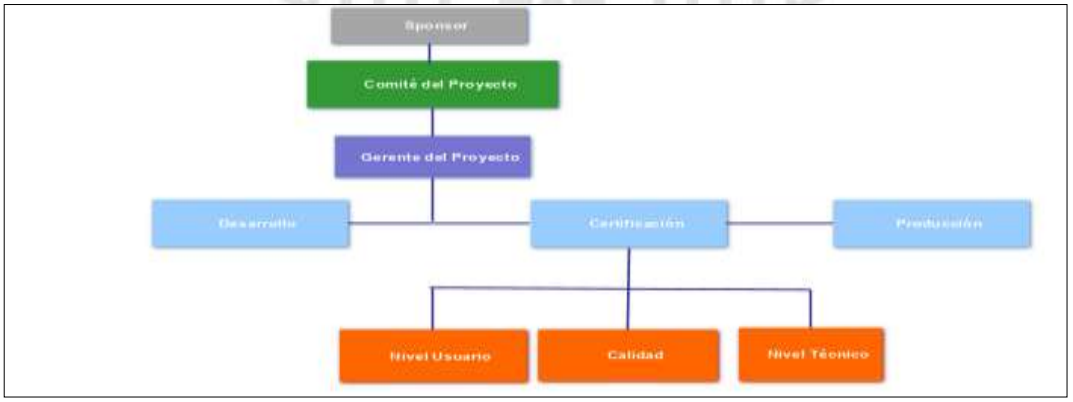
4.6 Cronograma

Figura 4.1
Línea de Tiempo



4.7 Equipo de trabajo

Figura 4.2
Organigrama



4.8 Medidas de control (Indicadores)

Tabla 4.4
Indicadores

Indicadores	Fórmula	Observación
Índice de facturación	Facturación = Tráfico **x** C onversión **x** Ticket promedio	La tasa de conversión es una métrica que explica el porcentaje de visitas que se convirtieron en transacciones en un periodo de tiempo.
Volumen de tx totales	%tasa de conversión = transacciones / visitantes	
Valor actual neto de la inversión (VAN)	$\mathrm{VAN} = \sum_{t=1}^{n} \frac{\mathrm{V_t}}{(1+k)^t} - \mathrm{I_0}$	Vt representa los flujos de caja en cada periodo t. Io es la cantidad de dinero inicial de la inversión. n indica el número de períodos que se consideran. k se refiere al tipo de interés definido.
Error medio absoluto (MAE)	$\mathrm{MAE} = \frac{1}{\lvert\hat{R}\rvert} \sum_{\hat{r}_{ui} \in \hat{R}} [r_{ui} - \hat{r}_{ui}]$	Es la métrica que permitirá la evaluación del modelo, entre más bajo el número, mejor. Este calcula el promedio absoluto existente entre las predicciones realizadas y las calificaciones o ratings reales del usuario.
Retorno de inversión (ROI)	ROI=(Ingresos-Inversión) /Inversión	Métrica financiera para validar la inversión realizada.

Nota. Formulas extraídas del libro "Gestión de proyectos con Project, Excel y Visio (Bajo enfoque PMBOK", por Angulo, p. 25, 2014.

4.9 Recursos y presupuesto

Tabla 4.5
Recursos y presupuesto

Mano de Obra	Cantidad	Haber básico	Sueldo Mensual	Costo empresa (+50%)	Costo MO (S/.)
Administración Desarrollador 1 web	1.0	4500.00	4500	2250	6750
Desarrollador AI admin BD	1.0	2400.00	2400	1200	3600
Community Manager	1.0	1600.00	1600	800	2400
Total de Mano de Obra					12750

De acuerdo con el objetivo principal del proyecto, así como los objetivos específicos, los beneficios y el segmento del mercado en este capítulo justifica el desarrollo del algoritmo de recomendación, dado que la solución permitirá responder a las necesidades y expectativas de los clientes que buscan tener personalización de los productos.

CAPÍTULO V: DESARROLLO DEL MVP

5.1. Deseabilidad

Para tener un mejor conocimiento de la deseabilidad hemos analizado el mercado y definido las prioridades de las necesidades de los clientes que los ayude y oriente en las compras virtuales.

5.1.1 Mercado

La información del mercado se obtiene por medio de la realización de un estudio de mercado (cualitativo y cuantitativo). Para ello, hemos considerado un público entre 25 y 64 años de los niveles socioeconómicos A, B, C, D, E que tienen preferencia por estilos de vida saludable y dentro de la encuesta también hemos considerado variables como el gasto mensual familiar y otras variables; a continuación, explicamos las bases que analizamos:

- Conocer la oferta y la demanda de los productos orgánicos y servicios por Internet.
- Conocer los medios digitales más adecuados para la comercialización de productos y servicios.
- Determinar lo que busca el público objetivo.
- Crear una estrategia de mercado que tenga éxito entre los clientes actuales y potenciales.
- Conocer la evolución de la industria y las tendencias actuales del mercado digital.
- Mejorar la experiencia del cliente.
- Promocionar y distribuir los productos, ajustándolos al precio justo y de acuerdo con un ranking.
- Conocer a las empresas competidoras, sus debilidades y fortalezas.

Para la investigación cualitativa se realizó un focus group con el objetivo de conocer los hábitos de compra de productos orgánicos, la opinión del concepto e imagen que representan, como los atributos de valor más importantes.

En la investigación cuantitativa, el resultado de las encuestas confirma la información obtenida en el análisis cualitativo. Respecto a los objetivos específicos se confirma: la intención de compra digital, personalización; estimación de precios, preferencias, entre otros que contribuyan a definir la estrategia de marketing.

5.1.2 Entrevistas a profundidad

En este punto se realizó la entrevista a 5 expertos (ver Anexo I) En el grupo de entrevistados contamos con una Medico, Economista, Administradora, Psicóloga holística y una Dentista. A continuación, se detallan los resultados obtenidos, ya que la entrevista a profundidad permite conocer en detalle las opiniones de los profesionales que se encuentran en la búsqueda de una compra verde (orgánica), inteligente y ágil.

Tabla 5.1
Resultados a expertos entrevistados

Ítems	Resultados
Identificar como contribuyen el uso de los productos orgánicos en la mejora de la calidad de vida.	- El consumo de los productos orgánicos eleva la calidad de vida y generan bienestar de las personas. En algunos casos se usa como tratamientos preventivos como algunos alimentos: chía, cañihua, frutos secos, etc.
Identificar el nivel de importancia que le da el cliente a la compra por internet de los productos orgánicos.	-Los compradores de productos orgánicos tienen mucho interés de compra digital debito a la coyuntura del Covid 19, prefieren estar lejos del riesgo de adquirir el virus. -Los compradores indican que les gustaría recibir personalización sobre sus compras y recomendaciones que los orienten a conocer nuevos productos y que se encuentren de oferta. -Confort del comprador, porque adquiere los productos en casa.
Identificar UX suelen ser más valorados por el cliente en general	- Contar con sistemas respondedores automáticos - Contar con personalización en sus compras
Identificar perfiles de clientes asiduos	- Aquellos que tienen interés en la ecología - Deportistas, profesionales con estilo de vida sofisticado - Cosmética natural y spas
Identificar posibles valores agregados adicionales a los que se proponen con el algoritmo KNN	- Contar con postcads - Incorporar IA artificial a la plataforma que permita disponer de aprendizaje - Ser respondidos inmediatamente, en el punto de necesidad, dentro de la plataforma

5.1.3 Investigación Cuantitativa.

La presente investigación cuantitativa se realizó en la zona 7 de Lima Metropolitana del 04 al 06 de Setiembre del 2018, siendo los distritos donde se llevó a cabo las encuestas los siguientes: La Molina, San Isidro, Miraflores, San Borja y Santiago de Surco. Las encuestas fueron realizadas en diferentes tiendas ecológicas y mercados bioferias. Para dar inicio a cada encuestado se le realizaban preguntas filtro que determinará el público objetivo.

- **Desarrollo de las Encuestas:**

La encuesta se realiza como si se tratara de una entrevista, la principal ventaja es que el encuestador controla y guía las preguntas que hace a las personas encuestadas, por lo que, se aumenta la calidad y veracidad de la información obtenida.

Se aplicó un cuestionario estructurado con preguntas cerradas y temáticas relacionadas a los objetivos de la investigación.

La población objetivo de estudio está conformada por personas del NSE A, B y C1 que compran productos orgánicos. Dentro los productos orgánicos que son solicitados encontramos: los aceites esenciales, alimentos, nutraceúticos, artículos de aseo etc. (Ver anexo A).

- **Diseño de la muestra**

En esta sección se detallan los pasos para calcular el tamaño de la muestra del presente estudio:

1. **Tipo de muestreo:** El tipo de muestreo es no-probabilístico por conveniencia, los elementos a encuestar se seleccionan debido a su fácil disponibilidad (Kinner y Taylor, 1998, p.405), para una mejor representatividad de la muestra se ha estratificado por los distritos. Las expresiones de confianza del presente estudio deben ser tomadas con cautela.
2. **Tamaño de muestra:** El tamaño de la muestra es de 317 encuestas, el nivel de confianza del 95.0% y margen de error del 5.5%; además, el factor de probabilidad éxito/fracaso es de 50.0%, qué es lo recomendado en estudios

de este tipo, Se aplican preguntas filtro a los encuestados antes de realizar la entrevista, para asegurar que pertenezcan a la población objetivo de estudio.

3. **Distribución de la muestra:** La muestra se distribuye entre la zona 7 de Lima en forma proporcional al tamaño de la población de los distritos.

Tabla 5.2
Distribución muestral

Nº	Distrito	Población de 25 a 64 años proyectada al 2015 /1	Distribución % de la Población	Distribución Muestral
		X	Y = (X / X1)	Z = (Z1 x Y)
1	**Total**	**421,274**	**100.0%**	**317**
2	La Molina	94,150	22.3%	71
3	Miraflores	47,292	11.2%	36
4	San Borja	61,665	14.6%	46
5	San Isidro	30,361	7.2%	23
6	Santiago de Surco	187,806	44.6%	141

Tabla 5.3 Análisis de resultados

Ítems	Resultados
Conocer el perfil de compradores digitales potenciales.	- De acuerdo con los resultados obtenidos en el presente estudio, el perfil de los clientes digitales potenciales orgánicos por e-commerce es en similar proporción del género masculino y femenino, principalmente de las edades de 35 a 54 años y en la mayoría de los casos del nivel socioeconómico B.
Estimar la proporción de personas que usan productos orgánicos	- De las encuestas se dio a conocer que de las personas que consumen productos naturales el 54,6% son mujeres y el 45,4% hombres. - Además del total de personas encuestadas, el 76.7% manifestó que acostumbran a usar productos orgánicos y el 23.3% que no usan productos no orgánicos.
Conocer hábitos en la compra de productos orgánicos y/o productos naturales.	- La importancia del consumo de productos orgánicos y/o productos naturales en el bienestar de su salud, resultando que el Top Two Box es 91.5%, lo que significa que, de cada 10 personas, 9 señalaron que es importante y muy importante
Estimar las preferencias de compra online de productos orgánicos	- Se observa que la mayor proporción (82.6%) realiza compras virtuales; mientras que, el resto de las personas compran con frecuencia bimensual (12.4%), trimestral (4.1%) y quincenal (0.8%). - Se observa que la mayor proporción de personas compra con frecuencia mensual por e-commerce tanto productos naturales en general como productos de alimentación (62.1%); mientras que, el resto (37.9%) compra con frecuencia quincenal, bimensual, trimestral y semestral consumo de estos productos
Determinar el nivel UX de los productos orgánicos en la salud y estilo de vida.	- El nivel de UX del consumo de productos orgánicos para su salud resulto Top Two Box= 95.6%, lo que significa que, de cada 100 personas, 9 han experimento una buena compra y satisfecha.
Evaluar la aceptación del Sistema de recomendación dentro de la herramienta ECOM ZONABIO	- Finalmente, se percibe un alto nivel de aceptación por el uso de un sistema Recomendador en las personas encuestadas, debido a que la mayoría han consumido estos productos y busca conocer los nuevos productos, los más vendidos y las ofertas.
Medir el interés de compra de una marca de productos orgánicos en e-commerce	- El 30.3% manifestó que definitivamente si los comprase y el 41.6% que probablemente si los comprase; mientras que, el 17.4% se mostraron indecisos (tal vez sí o no), el 7.9% señaló que probablemente no los compraría y el 2.8% que definitivamente no los compraría.
Saber los medios de comunicación preferidos para recibir información de ECOM ZONABIO	- Del total de personas interesadas la mayoría (61.0%) prefiere usar portales de internet y app para recibir noticias y recomendaciones, el 28.0% prefiere redes sociales y el 11.0% correo electrónico.
Determinación del tamaño del mercado	- Se determinó que los distritos de la Zona 7 son los que más invierten en el cuidado, conservación de la salud y estilos de vida saludable. Haciendo la revisión de las encuestas y aplicando a las respuestas la metodología Top 2 Box, encontramos que el 71.92% respondieron definitivamente si comprasen los productos orgánicos por e-commerce. Esto significa que el público objetivo estaría conformado por 87,312 personas, y el específico por 4,366 personas.

Los ítems se han validado de acuerdo con el deseo de los consumidores de contar el algoritmo de recomendación y personalización porque facilita la compra y porque los lleva a descubrir productos con alto rating de ventas y ofertas.

5.2 Factibilidad

Tenemos considerado realizar las implementaciones de forma gradual, considerando las siguientes etapas:

- Primero: Desarrollaremos el algoritmo de recomendación porque es lo más difícil de desarrollar y donde hay mayor incertidumbre. En esta etapa se va a analizar, desarrollar y evaluar el algoritmo de recomendación que mejor responda a las necesidades anteriormente identificadas de nuestros clientes.
- Segundo Realizaremos el portal donde ofreceremos todos los productos orgánicos y el cual se podrá visualizar en cualquier dispositivo, smartphones, laptops, etc.
- Tercero: Luego implementaremos Algoritmos que clasifiquen los productos por categorías.

Para este paper nos vamos a enfocar en el primer punto que es el desarrollo del algoritmo de recomendación utilizando el modelo KNN.

5.2.1 Diseño funcional de la solución

Para la elaboración del algoritmo de recomendación utilizando el modelo KNN (identificación de vecinos cercanos), hemos utilizado el lenguaje Python que posee una licencia de código abierto y se caracteriza por ser un lenguaje interpretado, dinámico y multiplataforma.

La plataforma que estamos utilizando para la ejecución del Algoritmo de recomendación de Ecom ZonaBio se encuentra en la suite Google denominada “Google Colab”. El Google Colaboratory, permite ejecutar y programar en Python brindando las siguientes ventajas:

- ✓ No requiere ser configurado

- ✓ Brinda acceso gratuito a GPUs

La ubicación del link donde tenemos alojado el producto es: https://colab.research.google.com/drive/1Onnu3CG5iVxDkbDZ7o6A7Bu-mZhQI2lJ

Por otro lado, el archivo en Python se ha nombrado: PrototipoRecomendador.ipynb

1. Prototipo del Producto

En esta sección mostramos la interface, que consta de las siguientes características:

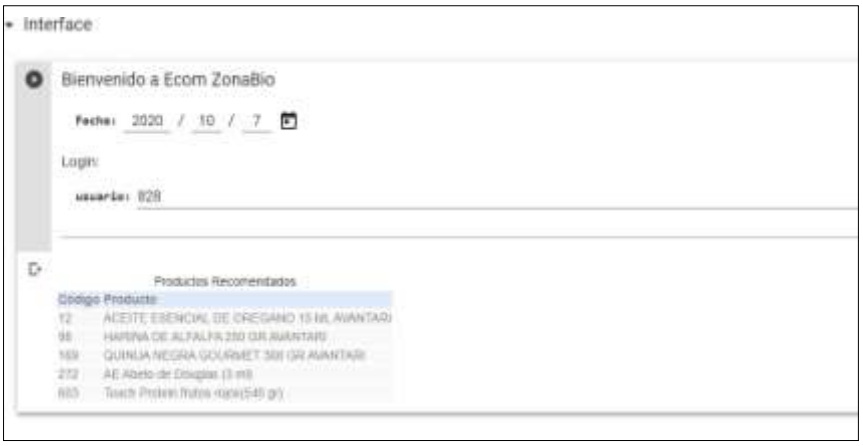
a. La fecha actual: (de acuerdo con la fecha que se encuentre),

b. El Login: Donde se introduce el código asignado al cliente.

c. Productos recomendados: Es el resultado del algoritmo KNN

El algoritmo utiliza la Categoría Memory Based**,** aplica básicamente técnicas de estadísticas al dataset para el cálculo de predicciones. **KNN** es el algoritmo elegido para nuestro sistema.

Figura 5.1
Prototipo Interface



- **Pseudocódigo del Algoritmo de recomendación**

1. Cargar tabla "ratings" en el array dataset_ratings.

2. Leer dataset_ratings y realizar el cómputo de las similitudes entre todos los pares de ítems (vectores) usando la fórmula de Similitud del MSD (Mean Square Difference).

2a. Dado item 1, ítem 2, ……, ítem n:

Estimar la MSD por cada par de ítems usando la fórmula:

$$\text{msd}(i,j) = \frac{1}{|U_{ij}|} \cdot \sum_{u \in U_{ij}} (r_{ui} - r_{uj})^2$$

Donde:

rui: el rating del usuario u al ítem i.

ruj: el rating del usuario u al ítem j.

Uij: el conjunto de todos los usuarios que han otorgado rating a ambos ítems i y j.

2b. Dado msd (item1, item2), ……, msd (i, j):

Estimar la Similitud del MSD por cada par de ítems usando la fórmula:

$$\text{msd\_sim}(i,j) = \frac{1}{msd(i,j) + 1}$$

Nota: La constante +1 es solo es para evitar dividir entre 0.

2c. Guardar todas las similitudes msd_sim (i, j) en la matriz sim (i, j)

3. Estimar ratings:

3a. Dados todos los pares usuario-ítem "ui" en los cuales el usuario "u" no ha otorgado rating al ítem "i":

Leer la matriz sim (i, j)

Estimar el rating que el usuario "u" otorgaría al ítem "i" usando la fórmula:

$$\hat{r}_{ui} = \frac{\sum_{j \in N_u^k(i)} \text{sim}(i,j) \ . \ \text{r}_{\text{uj}}}{\sum_{j \in N_u^k(i)} \text{sim(i, j)}}$$

Donde:

$$N_u^k(i) :$$

Es el conjunto de los "k" vecinos más cercanos (o más similares) al ítem "i" que han recibido rating del usuario "u"

k (int): el máximo número de vecinos a tomar en cuenta.

4. Guardar los ratings estimados en una tabla MatrizDeRatings en el servidor.

5. Usar la tabla MatrizDeRatings para hacer las 5 o 10 mejores recomendaciones a un usuario que ha ingresado al sistema.

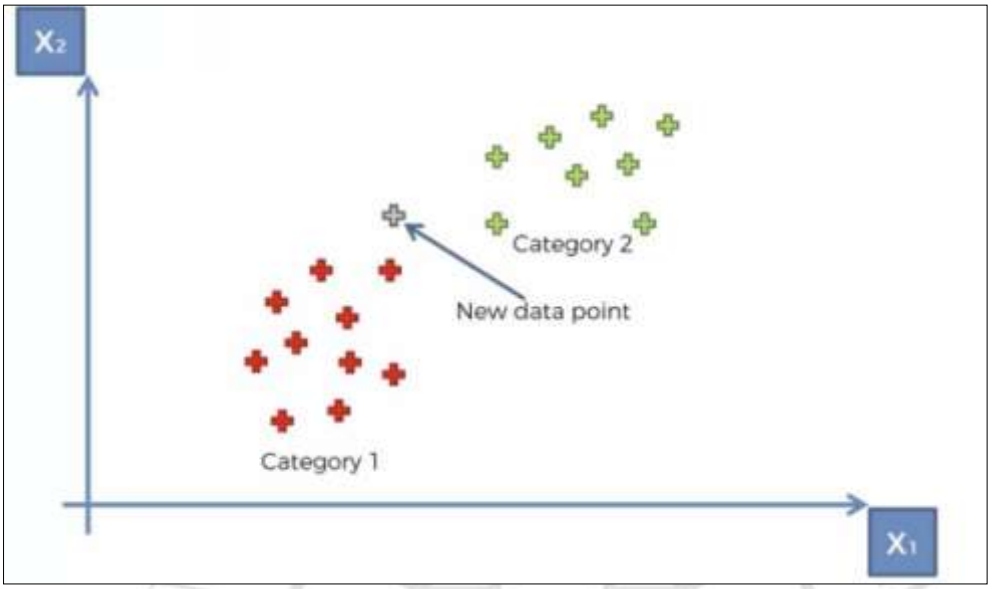
2. Servicios del producto

Vamos a explicar que consiste el Sistema Recomendador basado en filtros colaborativos:

El sistema Recomendador será desarrollado usando la librería Surprise (Librería para sistemas recomendadores) el cual corre bajo Python. De esta librería vamos a evaluar varios tipos de algoritmos basados en KNN).

Para nuestro sistema de recomendación hemos escogido el modelo de vecinos cercanos o K-NN basado en el modelo Item Item, dado que mediante la calificación que los usuarios otorgaron a los productos o ítems, podremos analizar, correlacionar y contrastar con otros elementos que se encuentran "próximos" y por ende proporcionarnos mejores resultados de predicción.

Figura 5.2
Algoritmo KNN



Nota. De "K Nearest Neighbor Algorithm In Python", por Cory Maklin, 2019. (https://towardsdatascience.com/k-nearest-neighbor-python-2fccc47d2a55)

Pasos Comprendidas en Filtros Colaborativos (Memory Based):

Paso 1: Encontrar usuarios o ítems similares

La técnica usada para encontrar ítems similares basados en los ratings que ellos reciben, se denomina ítem-based o ítem-ítem (Filtro Colaborativo basado en el ítem)

La técnica usada para encontrar usuarios similares basados en los ratings que ellos otorgan se denomina user-based o user-user (Filtro Colaborativo basado en el usuario).

Paso 2: Predicción de Ratings

User-Based Vs Item-Based

Ambos enfoques son matemáticamente casi similares pero la diferencia se encuentra en el cálculo del rating en cada enfoque:

Figura 5.3
Formula Predicción de Ratings

The prediction $\hat{r}_{ui}$ is set as:

$$\hat{r}_{ui} = \frac{\sum_{v \in N_i^k(u)} \mathrm{sim}(u,v) \ . \ \mathrm{r}_{v\mathrm{i}}}{\sum_{v \in N_i^k(u)} \mathrm{sim}(\mathrm{u},v)}$$

Nota. "Comparison of collaborative and content-based automatic recommendation approaches in a digital library of Serbian PhD dissertations". Por Azzopardi, J., Ivanovic, D. y Kapitsaki, G. 2016, *KEYSTONE*, 10151, pp. 100–111. (https://doi.org/10.1007/978-3-319-53640-8_9)

a. Predicción de Ratings Bajo la Técnica de User-based:

El rating para un ítem **I** (que aún no tiene rating) se halla seleccionando N usuarios de la lista de usuarios similares que han otorgado rating al ítem **I**. Luego se calcula el rating basado en esos N ratings.

b. Predicción de Ratings Bajo la Técnica de Item-based:

El rating para un ítem **I** (que aún no tiene rating) se halla seleccionando N ítems de la lista de ítems similares que han obtenido rating del usuario **U**. Luego se calcula el rating basado en esos N ratings.

El objetivo de nuestro sistema Recomendador es llenar la Matriz de Ratings, es decir predecir aquellos ratings que aún no se conocen.

Figura 5.4
Matriz Ratings

	i_1	i_2	i_3	i_4	i_5
u_1	5		4	1	
u_2		3		3	
u_3		2	4	4	1
u_4	4	4	5		
u_5	2	4		5	2

Nota. "Comparison of collaborative and content-based automatic recommendation approaches in a digital library of Serbian PhD dissertations". Por Azzopardi, J., Ivanovic, D. y Kapitsaki, G. 2016, *KEYSTONE*, 10151, pp. 100–111. (https://doi.org/10.1007/978-3-319-53640-8_9)

Paso 3: Evaluación del Modelo

Como métrica de evaluación del modelo vamos a realizar el cómputo del Mean Absolute Error (MAE), que viene a ser el promedio de los valores absolutos de todas las diferencias entre el rating conocido y el rating previsto o pronosticado (predicted rating A menor sea el MAE, mejor será el grado de precisión que tendrá el Sistema Recomendador para predecir los ratings.

Fórmula del MAE:

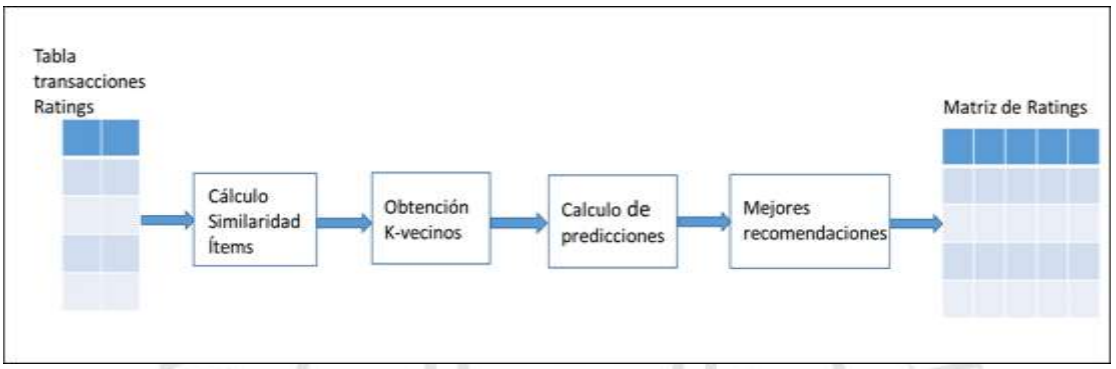
Figura 5.5
Fórmula del MAE

$$\text{MAE} = \frac{1}{\left|\hat{R}\right|} \sum_{\hat{r}_{ui} \in \hat{R}} [r_{ui} - \hat{r}_{ui}]$$

Nota. "Comparison of collaborative and content-based automatic recommendation approaches in a digital library of Serbian PhD dissertations". Por Azzopardi, J., Ivanovic, D. y Kapitsaki, G. 2016, *KEYSTONE*, 10151, pp. 100–111. (https://doi.org/10.1007/978-3-319-53640-8_9)

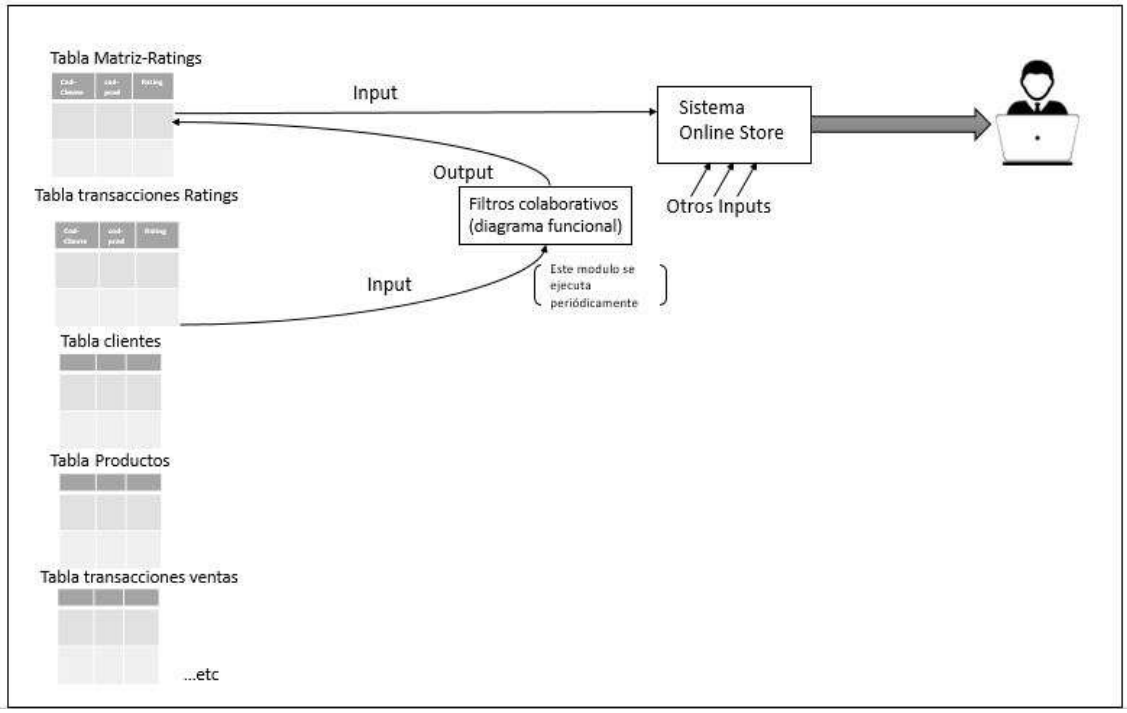
3. Diagrama Funcional de la solución

Figura 5.6
Diagrama Funcional



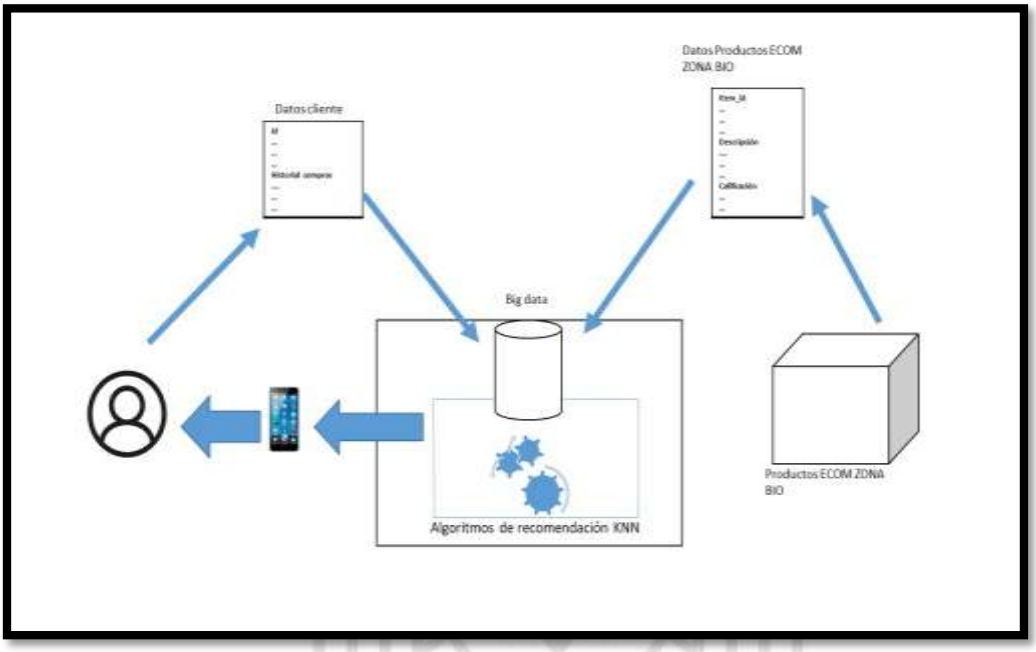
5.2.2 Diseño de la Arquitectura

Figura 5.7
Arquitectura



5.2.3 Diseño técnico de la solución

Figura 5. 8
Diseño técnico de la solución

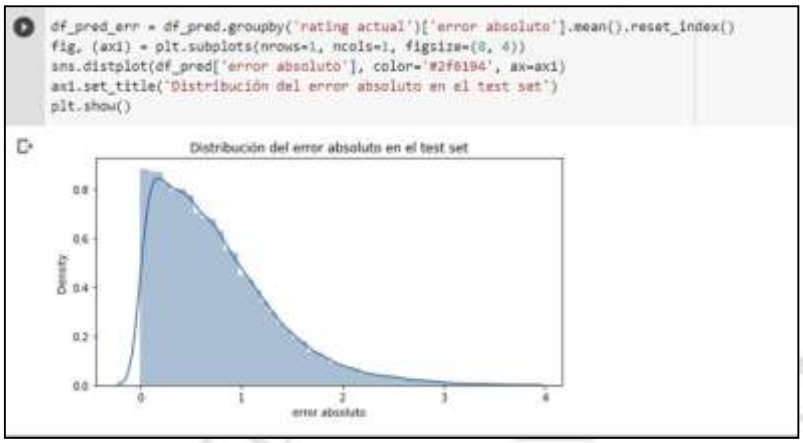


Nota. De "Simple app architecture with Amazon Dynamodb". Por Rao, P. 2015. *Amazon Web Services*

5.2.4 Resultados de las pruebas del modelo

Evaluación del modelo: para la evaluación de nuestro modelo estamos utilizando el Mean absolute Error (MAE), el cual mide la diferencia (como valor absoluto) entre la estimación del algoritmo y el valor real del rating. El cómputo se realiza sobre el subconjunto de datos apartados para el proceso de evaluación usando la siguiente fórmula: El MAE arrojado por nuestro modelo es de aproximadamente 0.73, el cual consideramos un valor moderado y bueno si tenemos en cuenta que los valores del rating van del 1 al 5. Veamos el error absoluto de 5 muestras aleatorias en la siguiente tabla Rating Actual Vs. Rating Estimado. En esta muestra aleatoria vemos 3 errores absolutos por debajo de 0.2 y 2 muestras por debajo de 1.5. Ahora veamos cómo está distribuido el error absoluto en el subconjunto de datos apartados para la evaluación (test set) Finalmente podemos concluir de la gráfica que la mayoría de los errores es pequeña: entre 0 y 1. A la derecha observamos una larga cola que nos indica que la minoría de errores fluctúa entre 1 y 5. (Ver anexo B)

Figura 5.9
Distribución Error absoluto



5.3 Viabilidad

Presentamos el análisis económico y financiero de "ECOM ZONABIO". Para dichos fines se han realizado estimaciones de la demanda, ventas y costos según los objetivos propuestos en la estrategia del negocio.

5.3.1 Supuestos y consideraciones generales

Para la evaluación de la plataforma ECOM ZONABIO se han tomado los siguientes supuestos relevantes:

- Horizonte de evaluación de 5 años.
- Inflación anual del 2% y el tipo de cambio de S/3.50 por dólar de acuerdo con estimaciones del BCRP
- Para la estructura de capital se ha considerado que el 70% será financiado por los accionistas y el 30% con financiamiento bancario del Interbank con una tasa de 16.96%, de acuerdo con datos tomados de la Superintendencia de Banca y Seguros.
- La tasa de impuestos considerada es 29.5%.
- La participación en el mercado meta asciende a 5% en el primer año, crecimiento de 5% para el segundo y tercer año, y de 2% para el cuarto y quinto año.

- La depreciación anual corresponde al 10% para el caso de equipos y mobiliario y no habrá valor residual, y 25% en lo que se refiere a equipos de cómputo y afines.

5.3.2 Proyecciones

La proyección de ingresos, costos y gastos se realizaron en base a las estimaciones de la demanda de compradores en e-commerce, así como la información cuantitativa del estudio de mercado (encuestas) y costos requeridos para la operatividad del negocio.

5.3.3 Estado de Ganancias y Pérdidas

Con la información de ingresos, costos y gastos se determinó el Estados de Ganancias y Pérdidas para los 5 años del proyecto. Ver Tabla 5.4 a continuación:

Tabla 5.4
Estado de Resultados

Descripción	Año 1	Año 2	Año 3	Año 4	Año 5
	S/	S/	S/	S/	S/
Ingresos	657,843	704,550	754,573	785,058	816,775
-Costos	-207,231	-221,945	-237,703	-247,306	-257,297
Utilidad Bruta	450,612	482,606	516,871	537,752	559,477
-Gastos de comunity manager	-109,800	-111,996	-114,236	-116,521	-118,851
-Gastos de cloud computing	-50,250	-51,255	-52,280	-53,326	-54,392
-Gastos de programadores	-43,200	-44,064	-44,945	-45,844	-46,761
-Gastos de ventas online (comisión)	-13,609	-13,881	-14,159	-14,442	-14,731
-Gasto de marketing digital	-117,046	-105,630	-108,707	-100,760	-102,775
-Gastos generales	-36,631	-37,363	-38,110	-38,873	-39,650
-Depreciación y amortización	-52,745	-3,290	-3,290	-3,290	-1,119
Utilidad (Pérdida) Operativa	27,331	115,126	141,143	164,697	181,198
-Impuestos	-8,063	-33,962	-41,637	-48,586	-53,454
Utilidad (Pérdida) después de Impuestos S/	19,268	81,164	99,506	116,111	127,745

5.3.4 Análisis Financiero

Para este análisis se elabora el flujo de caja operativo y de inversiones. Luego de realizado los análisis respectivos, se evidencia que el flujo de caja económico muestra un flujo positivo a partir del primer año de operación. Ver Tabla 11 a continuación:

Tabla 5.5
Flujo de caja económico

Descripción	Año 0 S/	Año 1 S/	Año 2 S/	Año 3 S/	Año 4 S/	Año 5 S/
Ingresos		657,843	704,550	754,573	785,058	816,775
-Costos		-207,231	-221,945	-237,703	-247,306	-257,297
Utilidad Operativa		450,612	482,606	516,871	537,752	559,477
-Gastos de Community manager		-109,800	-111,996	-114,236	-116,521	-118,851
-Gastos de cloud computing		-50,250	-51,255	-52,280	-53,326	-54,392
-Gastos de programadores		-43,200	-44,064	-44,945	-45,844	-46,761
-Gastos de ventas online		-13,609	-13,881	-14,159	-14,442	-14,731
-Gasto de marketing digital		-117,046	-105,630	-108,707	-100,760	-102,775
-Gastos generales		-36,631	-37,363	-38,110	-38,873	-39,650
-Depreciación y amortización		-52,745	-3,290	-3,290	-3,290	-1,119
Utilidad antes de impuestos		27,331	115,126	141,143	164,697	181,198
-Impuestos		-8,063	-33,962	-41,637	-48,586	-53,454
+Depreciación		52,745	3,290	3,290	3,290	1,119
Flujo de Caja Operativo S/		72,014	84,454	102,796	119,402	128,863
Flujo de Caja de Inversión						
Fijos tangibles	-14,280					
Fijos intangibles	-11,657					
Inversión en Pre operativos	-37,798					
Inversión en capital de trabajo	-207,231	-14,713	-15,758	-9,603	-9,991	257,297
Flujo de Caja de Inversiones S/	-270,966	-14,713	-15,758	-9,603	-9,991	257,297
Flujo de Caja Económico	**-270,966**	**57,300**	**68,696**	**93,193**	**109,410**	**386,161**
VAN	88,542					
TIR	29.79%					
Tasa de Descuento	19.80%					

El proyecto genera valor para sus accionistas según el flujo de caja económico. El análisis realizado da como resultado un VAN de S/ 88,542 y una TIR de 29.79% respecto a una evaluación de flujos económicos a 5 años y considerando una tasa de descuento de 19.80%.

CONCLUSIONES

- Los diferentes modelos de recomendación y algoritmos de Machine Learning, nos han permitido comprender, analizar y desarrollar nuestro producto ECOM ZONABIO, bajo un modelo de recomendación basado en filtros colaborativos usando el algoritmo KNN.
- Este algoritmo de aprendizaje supervisado nos posibilitará ofrecer o recomendar a nuestros clientes productos de su interés antes que este haya expresado su necesidad, generando una experiencia de cliente personalizada.
- La propuesta de valor y desarrollado el modelo de negocio nos muestra que existe gran potencial si utilizamos tecnologías emergentes, porque facilitan el comercio digital de los productos orgánicos, permitiendo a los interesados satisfacer las necesidades de llevar un estilo de vida saludable.
- Mediante entrevistas a clientes y consumidores de productos orgánicos, se confirma la deseabilidad de adquirir estos productos a través de un canal de ventas digital que incluya un sistema de recomendación y/o personalización.
- Validamos que el rendimiento de los diferentes algoritmos K- NN; el modelo KNN Baseline tuvo el mejor performance en términos de MAE.
- Dentro de las medidas de similitud validadas la distancia MSD (Mean square distance) es la más apropiada para clasificar según la data, para lo cual el menor error obtenido por el algoritmo fue el basado en ítems.
- El algoritmo ofrecerá los productos que han sido mejor calificados por otros compradores el cual generará un rating y propondrá los 5 mejores productos orgánicos a los potenciales clientes.
- El MVP ha obtenido resultados positivos del prototipo desarrollado el cual nos ha permitido comprobar que es posible recomendar aquellos productos orgánicos que son de interés de nuestros clientes y que también podrán recibir orientación aquellos que están buscando adaptarse a un estilo de vida saludable.

RECOMENDACIONES

Se han identificado algunos puntos de mejora para el curso de “titulación de proyecto integrador en innovación tecnológica”:

- Utilizar valoraciones o calificaciones implícitas de los usuarios, incorporando por ejemplo un sistema de análisis de tráfico para medir el bounce rate, tiempo de duración de la sesión total y por producto, número de páginas o productos vistos por cada sesión, etc. Esto permitirá integrar rápidamente un nuevo usuario que aún no ha calificado explícitamente ningún producto y recomendarle desde el inicio productos que podrían ser de su interés.
- Guardar o registrar el comportamiento de compra de un usuario frente a los productos recomendados. Comparar si un usuario compro o no compro algún o algunos de los productos recomendados nos permitirá seguir evaluando la eficacia de nuestro algoritmo o modelo.
- Permitir a los usuarios comentar acerca del producto y al mismo tiempo incorporar un sistema de análisis de texto. Por ejemplo, podríamos utilizar un clasificador de texto para analizar los comentarios de miles de productos con respecto a una marca y clasificar el sentimiento de cada mensaje como “positivo”, “negativo” o “neutral”.
- Pensar en el desarrollo de un modelo híbrido usando tanto un sistema basado en el contenido, así como también usando un sistema basado en filtros colaborativos. Lo cual significaría incorporar información acerca de cada producto que incluya atributo (ingredientes, principales compuestos, grado de concentración de los compuestos claves, sistema de preservación, etc.), como información de los clientes (edad, profesión, lugar de residencia, etc.), así será posible de recomendar nuevos productos basado en las preferencias de los usuarios con respectos a estos atributos sin tener en cuenta solo la calificación.

GLOSARIO DE TÉRMINOS

Término	Concepto
PRODUCTO ORGÁNICO:	Producto limpio de residuos químicos, con un mayor valor nutricional que los otros productos comunes y con un sabor más intenso.
PANDEMIA:	Enfermedad epidémica que se extiende a muchos países o que ataca a casi todos los individuos de una localidad o región.
COVID-19:	Enfermedad del Coronavirus 2019, es una enfermedad causada por un virus de la familia SARS-CoV-2.
IA:	Abreviatura de inteligencia artificial, que tiene por objeto la reproducción artificial de las facultades cognitivas de la inteligencia humana para crear sistemas o máquinas capaces de realizar funciones que normalmente le pertenecen.
RATING:	Evaluar o realizar un juicio o valor a un producto.
K-fold:	Es una técnica de validación cruzada para evaluar el rendimiento de los modelos de aprendizaje automático. Garantiza que cada observación del conjunto de datos original tenga la posibilidad de aparecer en el conjunto de entrenamiento y prueba.
SOBREAJUSTE:	El sobre ajuste se produce cuando un sistema de aprendizaje automático se entrena demasiado lo que hace que el algoritmo pierda la capacidad de generalizar.
SUBAJUSTE:	Se refiere a un modelo que no puede modelar los datos de entrenamiento y puede no generalizar a nuevos datos, esto ocurre cuando el modelo de Machine Learning es muy simple.
VENTA DIGITAL:	Venta a través de medios digitales, sin necesidad presencial del cliente.

GOOGLE COLAB:	Servicio en la nube, ofrecido por Google gratuitamente, basado en Júpiter Notebook y destinado a la formación e investigación en aprendizaje automático. Esta plataforma permite entrenar modelos de aprendizaje automático directamente sin necesidad de instalar nada en nuestro ordenador salvo un navegador.
SURPRISE:	Librerías de Python para crear y analizar sistemas de recomendación que tratan con datos explícitos.

REFERENCIAS

Aguirre, C. (26 de Mayo de 2019). *ML Part 1: Introducción a los arboles de decisión*. Obtenido de https://www.cristobal-aguirre.com/arboles-de-decision

Andina. (13 de Julio de 2020). *Comercio electrónico muestra mayor dinamismo en América Latina y el Caribe*. Obtenido de https://andina.pe/agencia/noticia-comercio-electronico-muestra-mayor-dinamismo-america-latina-y-caribe-805522.aspx

Angulo A., L. (2014). *Gestión de proyectos con Project, Excel y Visio (Bajo enfoque PMBOK).* Lima: Macro Ed.

Antevenio. (28 de Mayo de 2020). *Descubre el crecimiento del ecommerce durante el coronavirus*. Obtenido de https://www.antevenio.com/blog/2020/05/crecimiento-del-ecommerce-durante-el-coronavirus/

Asociación Peruana de Empresas de Inteligencia de Mercados. ([APEIM], 2019). *Niveles Socioeconómicos 2019*. Obtenido de http://apeim.com.pe/wp-content/uploads/2019/12/NSE-2019-Web-Apeim-2.pdf

Bedoya P., J. (Setiembre de 2011). Aplicación de distancias entre términos para datos planos y jerárquicos. *Máster en ingeniería de software, métodos formales y sistemas de información*. Valencia, España: Universidad Politécnica de Valencia.

Breese, J., Heckerman, D., & Carl, K. (1998). Empirical Analysis of Predictive Algorithms for Collaborative. *Microsoft Research,* 43-52.

Fuentes, H. (06 de Julio de 2020). *Crecimiento del e-commerce en el Perú en época de pandemia*. Obtenido de data.: https://www.datatrust.pe/ecommerce/ecommerce-en-el-peru-en-epoca-de-pandemia/

Galan, S. (2007). Filtrado Colaborativo y Sistemas de Recomendación. *Inteligencia en Redes de Comunicaciones,* 20-26.

García-Ramírez, J., Morales, E., & Escalante, H. (2019). *Transferencia de Conocimiento utilizando múltiples tareas en Algoritmos de Aprendizaje por Refuerzo Profundo.* Puebla, México: Instituto Nacional de Astrofísica Óptica y Electrónica.

Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems.* Gravenstein Highway North, Sebastopol, USA: O'Reilly Media, Inc.

Golovin, N., & Rahm, E. (2005). Optimización automática de recomendaciones web mediante comentarios y gráficos de ontología. *Congreso Internacional de Ingeniería Web*, 375-386.doi: https://doi.org/10.1007/11531371_49

Gonzalez, A. (2020). *Sistemas de recomendación de contenido con Machine Learning*. Obtenido de Clever.io: https://cleverdata.io/sistemas-recomendación-machine-learning/

Helga, W., & Lemond, J. (2019). *El mundo de la agricultura orgánica: estadísticas y tendencias emergentes 2019.* Ginebra, Suiza: Instituto de Investigación de Agricultura Orgánica FiBL, IFOAM - Organics International.

Higuchi, A. (2015). Características de los consumidores de productos orgánicos y expansión de su oferta en Lima. *Apuntes*, *42*(77). doi:ISSN 0252-1865

Iberdrola.com. (2020). *Descubre los principales beneficios del 'Machine Learning'*. Obtenido de https://www.iberdrola.com/innovacion/machine-learning-aprendizaje-automatico

McKenna, E., Richardson, I., & Thomson, M. (2012). Smart meter data: Balancing consumer privacy concerns with legitimate applications. *Energy Policy, 41*, 807-814.

Ménard, A. (2017). *How can we recognize the real power of the Internet of Things?* London: McKinsey & Company.

Microsoft Azure. (2020). *Algoritmos de aprendizaje automático*. Obtenido de https://azure.microsoft.com/es-es/overview/machine-learning-algorithms/

Neves, M. (11 de julio de 2017). Entrevista a Martha Neves de Nutrición con Apego. (Nutrición y Apego, Entrevistador)

Nilashi, M., Ibrahim, O., & Bagherifard, K. (2018). A recommender system based on collaborative filtering using ontology and dimensionality reduction techniques. *Expert Systems with Applications, 92*, 507-520.

Orams, E. (2020). *El eCommerce en Perú, con "E" de Experiencia.* Lima: Ipsos Apoyo.

Osterwalde, A., & Pigneur, Y. (2011). *Generación de modelos de negocios: Un manual para visionarios, revolucionarios y retadores.* Barcelona, España: Deusto S.A.

Pazzani, M., & Billisus, D. (2007). Content-Based Recommendation Systems. *The Adaptive Web*, 325-341.

Ray, S. (2020). An analysis of computational complexity and accuracy of two supervised machine learning Algorithms—K-nearest neighbor and support vector machine. *Avances en Computación y Sistemas Inteligentes, 1174*, 335-347.

Recuero de los Santos, P. (28 de Abril de 2020). *Datos de entrenamiento vs datos de test*. Obtenido de Think Bigs: https://empresas.blogthinkbig.com/datos-entrenamiento-vs-datos-de-test/

Scrum Body of Knowledge (SBOK GUIDE) (2018)

Sheng, B., Moosman, O., Del Pozo, J., Alfonso-Rosa, R., & Zhang, J. (2020). A comparison of different machine learning algorithms, types and placements of activity monitors for physical activity. *Measurement: Journal of the International Measurement Confederation*.

Sierra, B., Arbelaitz, O., Armañanzas, R., Arruti, A., & Bahamonde, A. (2006). *Aprendizaje automático: conceptos básicos y avanzados: aspectos prácticos utilizando el software WEKA.* México: Pearson.

Suito, J. (06 de Abril de 2020). *¿Es viable la publicidad en tiempos de pandemia?* Obtenido de Revista Mercado Negro : https://www.mercadonegro.pe/marketing/es-viable-la-publicidad-en-tiempos-de-pandemia/

Talend.com. (s.f.). *Qu'est-ce que le machine learning?* Obtenido de Talend: https://www.talend.com/fr/resources/what-is-machine-learning/

Torres, M. (30 de Abril de 2020). *Cómo usar inteligencia artificial en un ecommerce*. Obtenido de https://www.posicionamientoweb.systems/tecnologia/como-usar-inteligencia-artificial-en-un-ecommerce/

Velez-Langs, O., & Santos, C. (2006). Sistemas Recomendadores: Un enfoque desde los algoritmos genéticos. *Ind. data, 9*(1), 23-31.

ANEXOS

ANEXO A: ENCUESTA

ENCUESTA
Recomendación Productos Orgànicos
Setiembre 2018

NRO CUEST: [_____]

INTRODUCCIÓN:
Buenos días / tardes / noches. Mi nombre es _____ soy encuestador de la empresa.... y estamos llevando a cabo una investigación acerca del conocimiento de aceites esenciales. Le agradecería que nos de algunos minutos su tiempo.

DATOS DE CONTROL

F1. Genero:
(Por observación)

1. Hombre
2. Mujeres

F2. ¿En qué distrito vive usted?

1. Miraflores
2. San Isidro
3. Surco
4. San Borja
5. La Molina
6. Otro **(TERMINAR)**

F2. ¿Cuál es su edad?
(Circule solo una alternativa)

1. De 25 a 34
2. De 35 a 44
3. De 45 a 54
3. De 55 a 64
4. Otra edad **(TERMINAR)**

F3. Nivel Socioeconómico:
(Llenar la ficha filtro de NSE, luego circule la alternativa correspondiente)

1. A 2. B 3. C1 4. Otro **(TERMINAR)**

F4. ¿Suele usar productos naturales como medicina alternativa y para el cuidado? (Al menos una vez cada semestre)
(Circule solo una alternativa)

1. Si
2. No **(TERMINAR)**

A. HÁBITOS DE COMPRADORES DIGITALES POTENCIALES DE PRODUCTOS ORGÁNICOS

P1. ¿Suele usar productos orgánicos y/o naturales?
(Circule solo una alternativa)

1. Si
2. No ➔ **Pase a la pregunta 4**

P2. ¿Qué tipo de productos orgánicos y/o naturales usa?
(Puede circular más de una alternativa)

1. Aceite de coco
2. Panela
3. Café
4. Sal de maras
5. Cosmética Natural
6. Aromaterapia
7. Chocolates de cacao
8. Otra (especifique): _____

P3. ¿Qué productos consumes con más frecuencia? (VER P1)
(Circule solo una alternativa)

1. Aceite de coco
2. Panela
3. Café
4. Sal de maras
5. Cosmética Natural
6. Aromaterapia
7. Chocolates de cacao
8. Otra (especifique): _____

P4. ¿Qué tipo de canal visitas para la compra de estos de productos orgánicos y/o naturales?
(Puede circular más de una alternativa)

1. App
2. Facebook
3. Instagram
4. Páginas web
5. Vía Online
6. Marketplace
7. Otra (Especifique): _____

P5. ¿Desde qué distrito realizas la compra con mayor frecuencia? (Circule solo una alternativa)

1. La Molina
2. Miraflores
3. Barranco
4. San Isidro
5. San Borja
6. Surco
7. San Miguel
8. Otro (especifique): _____

P6. ¿Con que frecuencia compras los productos orgánicos y/o naturales?
(Circule solo una alternativa)

1. Semanal
2. Quincenal
3. Mensual
4. Bimensual
5. Trimestral
6. Semestral

P7. Según su respuesta anterior, en promedio, ¿Cuánto es el gasto que realiza cada vez que compra productos orgánicos y/o naturales?
(Circule solo una alternativa)

1. Menos de S/. 50
2. Entre S/. 51 y S/. 100
3. Entre S/. 101 y S/. 200
4. Entre S/ 201 y S/. 300
5. Más de S/. 300

P8. En una escala del 1 al 5, siendo 1 nada importante y 5 muy importante, en general, ¿Qué tan importante es para su bienestar y salud el uso de productos orgánicos y/o naturales?
(Circule solo una alternativa)

1. Nada importante
2. Poco importante
3. Moderadamente importante
4. Importante
5. Muy importante

P9. En una escala del 1 al 5, siendo 1 muy insatisfecho y 5 muy satisfecho, en general, ¿Qué tan satisfecho está con los beneficios que percibe de los productos orgánicos y/o naturales que consume?

1. Muy insatisfecho
2. Insatisfecho
3. Ni satisfecho ni satisfecho
4. Satisfecho
5. Muy satisfecho

P10. ¿Cuáles son los TRES aspectos que más valora o decide para elegir una determinada marca de productos orgánicos y/o naturales? (Ordene del 1 al 3, donde 1 es el más importante)

Aspectos	Orden
1. El precio	
2. Disponibilidad	
3. Promociones	
4. Recomendaciones	
5. Tiempo de respuesta	
6. La ubicación/ zona del establecimiento si en caso desea recoger	
7. Asesoramiento (recetas, cuidado de la piel, prevención)	
8. Información detallada de los productos	
9. Otros: _____	

P11. ¿Con que frecuencia visita las tiendas online de productos orgánicos y/o naturales? (VER P4)
(Circule solo una alternativa)

1. Diario
2. Semanal
3. Quincenal
4. Mensual
5. Bimensual
8. En eventos virtuales
9. No asiste a centros de aromaterapia

P12. En promedio, ¿Cuántas horas permanece en estas tiendas online de productos orgánicos y/o naturales cada vez que los visita? (Anote el número de horas)

[ ] Horas

P13. ¿A quién recomendaría estas tiendas online de productos orgánicos y/o naturales?
(Puede circular más de una alternativa)

1. Amigos
2. Familiares
3. Pareja
4. Otros (especifique): _____

B. EVALUACIÓN DE IDEA DE NEGOCIO

ENCUESTADOR LEA LO SIGUIENTE:
Ahora les presentaremos nuestra idea de desarrollar una solución que contenga un sistema de recomendación de productos orgánicos y/o naturales, que le brinde una experiencia personalizada de los productos que les permitan adoptar un estilo de vida saludable.

P14. En una escala del 1 al 5, donde 1 es definitivamente no y 5 es definitivamente sí, ¿Qué tan dispuesto estaría usted de utilizar la solución que le acabamos de presentar? (Circule solo una alternativa)

1. Definitivamente no
2. Probablemente no
3. Tal vez sí o no (indeciso) → **TERMINAR**
4. Probablemente sí → **TERMINAR**
5. Definitivamente Sí → → **TERMINAR**

P15. De la siguiente lista de clasificación de productos orgánicos, ¿Cuál le interesa comprar?
(Puede circular más de una alternativa)

1. Alimentos
2. Cosmética
3. Limpieza
4. Lácteos
5. Snacks
6. Vestimenta
7. Aromaterapia
8. Otra (especifique): _____

P16. Aproximadamente, ¿Cuánto podría gastar por cada compra?
(Circule solo una alternativa)

1. Más de S/. 100
2. De S/. 81 a S/. 100
3. De S/. 61 a S/. 80
4. De S/. 40 a 60
5. Menos de S/. 40

P17. ¿Con que frecuencia compraría estos productos?
(Circule la frecuencia para cada aroma de aceite esencial)

1. Semanal
2. Quincenal
3. Mensual
4. Bimensual
5. Trimestral
6. Semestral
7. Más tiempo
8. En eventos importantes

P18. ¿Qué valor agregado le gustaría recibir por la adquisición de estos productos orgánicos y/o naturales?
(Circule solo una alternativa)

1. Descuentos
2. Promociones
3. Talleres online
4. Otro (especifique): _____

P19. ¿Qué medio de comunicación prefiere para recibir noticias sobre estos productos? (Circule solo una alternativa)

1. Portales de internet
2. Redes sociales
3. Diarios escritos
4. Radio
5. Afiches
6. Otro (especifique): _____

P20. ¿A través de que medio o lugar prefiere comprar estos productos?
(Circule solo una alternativa)

1. Aplicación móvil
2. Página web
3. Tienda física

P21. ¿Qué nombre sería el ideal para este nuevo producto de aceites esenciales?
(Circule solo una alternativa)

1. Ecom ZonaBio (salud)
2. Easy Shop Bio
3. Zona Verde Ecom

FILTRO NSE

Con la finalidad de agrupar sus respuestas con las de otras personas de similares características a las de usted, nos gustaría que responda a las siguientes preguntas referentes al jefe de hogar:

JEFE DE HOGAR: Aquella persona, hombre o mujer, de 15 a más, que aporta más económicamente en casa o toma las decisiones financieras de la familia, y vive en el hogar. HOGAR: conjunto de personas que, habitando en la misma vivienda, preparan y consumen sus alimentos en común.

N1. ¿Cuál es el último año o grado de estudios y nivel que aprobó el jefe de hogar? (ACLARAR "COMPLETA O INCOMP.)

Sin educación/ Educación Inicial	0	Superior No Univ. Incompleta	2	Superior Univ. Completa	4
Primaria incompleta o completa/ Secundaria incompleta o completa	1	Superior No Univ. Completa/ Superior Univ. Incompleta	3	Post-Grado Universitario	6

N2. ¿Cuál de estos bienes tiene en su hogar que esté funcionando?

	NO	SI
Computadora o laptop en funcionamiento	0	1
Lavadora en funcionamiento	0	1
Horno microondas en funcionamiento	0	1
Refrigeradora/ Congeladora en funcionamiento	0	1
Total de bienes		

	Puntaje
0 bienes	0
1 bien	2
2 bienes	4
3 bienes	5
4 bienes	7

N3. ¿Cuál de los siguientes bienes o servicios tiene en su hogar que esté funcionando?

	NO	SI
Auto o camioneta para uso particular **(NO TAXI NI A TIEMPO PARCIAL)**	0	5
Conexión a internet (Banda ancha, móvil, Wi-Fi) **(TIENE QUE PAGAR MENSUAL)**	0	6
Servicio doméstico pagado (Sea por horas ó días) **(DEBE REALIZA UN PAGO NO PROPINAS)**	0	5
SUMAR PUNTAJES		

N4. ¿Cuál es el material predominante en los pisos de su vivienda? (CONSIDERAR ÁREA CONSTRUIDA. RESP. ÚNICA)

Tierra / Otro material (arena y tablones sin pulir)	0	Laminado tipo madera, láminas asfálticas o similares	6
Cemento sin pulir o pulido / Madera (entablados)/ tapizón	2	Parquet o madera pulida y similares; porcelanato, alfombra, mármol	7
Losetas / terrazos, mayólicas, cerámicos, vinílicos, mosaico o similares	4		

N5. ¿A qué sistema de prestaciones de salud está afiliado el jefe de hogar? (SI TIENE MÁS DE UNO ANOTAR EL DE MAYOR PUNTAJE)

No está afiliado a ningún seguro/ Seguro Integral de Salud (SIS)	0	Seguro Salud FFAA/ Policiales	3	Otro seguro de salud (especificar):	
ESSALUD	2	Entidad prestadora de salud (EPS)/ Seguro privado de salud	6		

N6. ¿Cuál es el material predominante en las paredes exteriores de su vivienda?

Estera	0	Piedra o sillar con cal o cemento	2
Madera/ Piedra con barro/ Quincha (caña con barro)/ Tapia/ Adobe	1	Ladrillo o bloque de cemento	4

N7. El baño o servicio higiénico que tiene en su hogar está conectado a:

No tiene	0	Baño compartido fuera de la vivienda. ***(Ejm.: quintas, corrolones, cuartos con baño compartido, etc.)***	2
Río, acequia o canal/ Pozo ciego o negro/letrina/ Pozo séptico	1	Baño dentro de la vivienda	4

N1	
N2	
N3	
N4	
N5	
N6	
N7	
Total	

9 puntos o menos	NSE E	9
De 10 a 15 puntos	NSE D	7
De 16 a 22 puntos	NSE C2	6
De 23 a 26 puntos	NSE C1	5

De 27 a 32 puntos	NSE B2	4
De 33 a 39 puntos	NSE B1	3
De 40 a 45 puntos	NSE A2	2
46 puntos a más	NSE A1	1

ANEXO B: Algoritmos de recomendación

MODELO DESARROLLADO EN GOOGLE COLAB

Integrando Google Colab y Google Drive e instalando librerías de Surprise

+ Code + Texto

SISTEMA RECOMENDADOR BASADO EN FILTROS COLABORATIVOS

▾ Configuración

```
[1] # Integrando Google Colab y Google Drive

    from google.colab import drive
    drive.mount('/content/drive')
```

```
Mounted at /content/drive
```

El sistema recomendador será desarrollado usando la librería **Surprise** (Librería para sistemas recomendadores) el cual corre bajo Python. De esta librería vamos a evaluar varios tipos de algoritmos basados en KNN).

```
[2] !pip install ansicolors
```

```
Collecting ansicolors
  Downloading https://files.pythonhosted.org/packages/[illegible]/ansicolors-1.1.8-py2.py3-none-any.whl
Installing collected packages: ansicolors
Successfully installed ansicolors-1.1.8
```

```
[3] !pip install scikit-surprise
```

```
Collecting scikit-surprise
  Downloading https://files.pythonhosted.org/packages/[illegible]/scikit-surprise-1.1.1.tar.gz (11.8MB)
     | 11.8MB 340kB/s
Requirement already satisfied: joblib>=0.11 in /usr/local/lib/python3.6/dist-packages (from scikit-surprise) (0.16.0)
Requirement already satisfied: numpy>=1.11.2 in /usr/local/lib/python3.6/dist-packages (from scikit-surprise) (1.18.5)
Requirement already satisfied: scipy>=1.0.0 in /usr/local/lib/python3.6/dist-packages (from scikit-surprise) (1.4.1)
Requirement already satisfied: six>=1.10.0 in /usr/local/lib/python3.6/dist-packages (from scikit-surprise) (1.15.0)
Building wheels for collected packages: scikit-surprise
  Building wheel for scikit-surprise (setup.py) ... done
  Created wheel for scikit-surprise: filename=scikit_surprise-1.1.1-cp36-cp36m-linux_x86_64.whl size=[illegible] sha256=[illegible]
  Stored in directory: /root/.cache/pip/wheels/[illegible]
Successfully built scikit-surprise
Installing collected packages: scikit-surprise
Successfully installed scikit-surprise-1.1.1
```

```
[4] from __future__ import print_function
    from colors import color, red, blue
```

```
[5] import numpy as np
    import pandas as pd
    from scipy import spatial

    import warnings
    warnings.filterwarnings('ignore')

    from IPython.core.display import display, HTML
    display(HTML("<style>.container { width:100% !important; }</style>"))

    import seaborn as sns
    import matplotlib.pyplot as plt
    %matplotlib inline
```

```
[6] pd.set_option('display.max_rows', 500)
    pd.set_option('display.max_columns', 500)
    pd.set_option('display.width', 200)
```

```
from surprise.model_selection import cross_validate

def obtener_nombre_modelo(model):
    return str(model).split('.')[-1].split(' ')[0].replace("'>", "")
def cv_multiples_modelos(data, models_dict, cv):
    results = pd.DataFrame()
    for model_name, model in models_dict.items():
        print('\n---> CV for %s...' % model_name)
        cv_results = cross_validate(model, data,['mae'], cv=cv)
        tmp = pd.DataFrame(cv_results).mean()
        tmp['model'] = model_name
        results = results.append(tmp, ignore_index=True)
    return results
def generar_modelos_dict(models, sim_names, user_based):
    models_dict = {}
    for sim_name in sim_names:
        sim_dict = {
            'name': sim_name,
            'user_based': user_based
        }
        for model in models:
            model_name = obtener_nombre_modelo(model) + ' ' + sim_name
            models_dict[model_name] = model(sim_options=sim_dict)
    return models_dict

# Recomendaciones para cada usuario a partir de las predicciones

from collections import defaultdict

def get_top_n(predictions, n=10):

    top_n = defaultdict(list)
    for uid, iid, true_r, est, _ in predictions:
        top_n[uid].append((iid, est))

    for uid, user_ratings in top_n.items():
        user_ratings.sort(key=lambda x: x[1], reverse=True)
        top_n[uid] = user_ratings[:n]

    return top_n

def k_from_details(details):
    try:
        return details['actual_k']
    except KeyError:
        return 1000
```

Conceptos de algoritmos de recomendación

- Estrategias

El objetivo de un **sistema de recomendaciones** es estimar las preferencias de un usuario con respecto a determinados items a partir de experiencias pasadas..

Sistemas de Recomendaciones están divididos en 3 categorías: **content based systems**, **collaborative filtering systems**, y **hybrid systems**(que usa una combinación de los otros 2).

Content base usa una serie de atributos de un ítem para recomendar otros items de propiedades similares.

Collaborative filtering el modelo es desarrollado a partir del comportamiento pasado del usuario (items previamente comprados o selecionados y/o ratings dados a esos items), así como también a partir de decisiones similares hechas por otros usuarios. Luego este modelo es usado para estimar items o (ratings de items) que podrían ser de interés del usuario.

Hybrid combina los dos anteriores.

Tipos diferentes de algoritmos dentro de la familia de filtros colaborativos: **memory based** and **model based**

Memory Based
Bajo esta categoría, los algoritmos aplican básicamente técnicas de estadísticas al dataset para las estimaciones de los ratings.
KNN es el algoritmo elegido para nuestro sistema.

Filtros colaborativos y etapas comprendidas

Etapas Comprendidas en Filtros Colaborativos (Memory Based)

Para desarrollar un sistema que automáticamente pueda recomendar items a usuarios teniendo en cuenta las preferencias de los otros usuarios, el primer paso es encontrar **usuarios similares o items similares**. El segundo paso es **estimar los ratings** de los items que aún no han obtenido rating alguno por parte de un usuario. El tercer paso es medir la precisión o 'accuracy' de los ratings calculados.

PASO 1: Encontrar usuarios o items similares

La técnica usada para encontrar **usuarios similares** basados en los ratings que ellos otorgan, se denomina **user-based** o user-user (Filtro Colaborativo basado en el usuario).

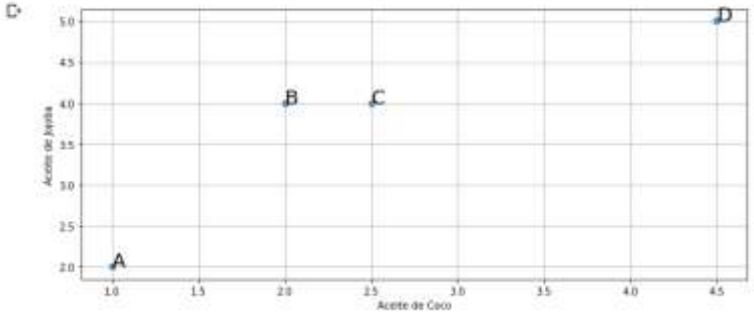
La técnica usada para encontrar **items similares** basados en los ratings que ellos reciben, se denomina **item-based** o item-item (Filtro Colaborativo basado en el item)

Para entender mejor el concepto de similitud, veamos con un ejemplo gráfico la similitud de usuarios.

La siguiente gráfica muestra 4 usuarios A, B, C y D, quienes han otorgado ratings a 2 productos.

```
a = [1, 2]
b = [2, 4]
c = [2.5, 4]
d = [4.5, 5]

labels = ["A", "B", "C", "D"]
data = np.append([a], [b], axis=0)
data = np.append(data, [c], axis=0)
data = np.append(data, [d], axis=0)
```

```
fig = plt.figure(figsize=(12,5))
ax = fig.add_subplot(111)
ax.scatter(data[:, 0], data[:, 1])
ax.set_xlabel("Aceite de Coco")
ax.set_ylabel("Aceite de Jojoba")
for i, word in enumerate(data):
    ax.annotate(labels[i], xy=(data[i, 0]+0.001, data[i, 1]), size=20)
plt.grid(True)
plt.show()
```



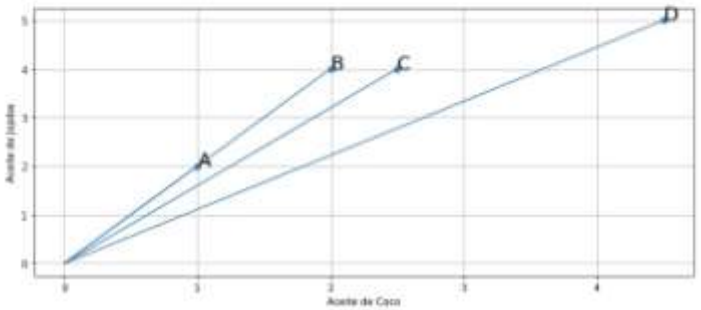
Una manera sencilla de estimar la similitud entre 2 usuarios o items es hallando simplemente la distancia que hay entre ellos mediante la fórmula de Euclides.

```
[14] print("C-A:",spatial.distance.euclidean(c, a))
     print("C-B:", spatial.distance.euclidean(c, b))
     print("C-D:", spatial.distance.euclidean(c, d))
```

```
C-A: 2.5
C-B: 0.5
C-D: 2.236067977499979
```

Cosine similarity (similitud del coseno): es una medida de similitud basada en el coseno del ángulo de 2 vectores.

```
[15] fig = plt.figure(figsize=(12,5))
     ax = fig.add_subplot(111)
     ax.scatter(data[:, 0], data[:, 1])
     ax.set_xlabel("Aceite de Coco")
     ax.set_ylabel("Aceite de Jojoba")
     for i, word in enumerate(data):
         ax.plot(np.array([0,data[i][0]]), np.array([0, data[i][1]]), color='steelblue')
         ax.annotate(labels[i], xy=(data[i, 0]+0.001, data[i, 1]), size=20)
     plt.grid(True)
     plt.show()
```




```
[16] print("C-A:", spatial.distance.cosine(c,a))
     print("C-B:", spatial.distance.cosine(c,b))
     print("C-D:", spatial.distance.cosine(c,d))
     print("A-B:", spatial.distance.cosine(a,b))
```

```
C-A: 0.00630452740604788
C-B: 0.00450452740604788
C-D: 0.011137225946083022
A-B: 0.0
```

You could say C is closer to D in terms of distance. But looking at the lower angle between the vectors of C and A gives a lower cosine distance value.

PASO 2: Predicción de Ratings

User-Based Vs Item-Based

Ambos enfoques son matemáticamente casi similares pero la diferencia se encuentra en el cálculo del rating en cada enfoque. Bajo el enfoque item-based, la fórmula de la estimación del rating es la siguiente:

$$\hat{r}_{ui} = \frac{\sum_{j \in N_u^k(i)} \text{sim}(i,j) \cdot r_{uj}}{\sum_{j \in N_u^k(i)} \text{sim}(i,j)}$$

2.1 Predicción de Ratings Bajo la Técnica de User-based:

For a user U, with a set of similar users determined based on rating vectors consisting of given item ratings, the rating for an item I, which hasn't been rated, is found by picking out N users from the similarity list who have rated the item I and calculating the rating based on these N ratings.

Español

El rating para un item **I** (que aún no tiene rating) se halla seleccionando N usuarios de la lista de usuarios similares que han otorgado rating al item **I**. Luego se calcula el rating basado en esos N ratings.

2.2 Predicción de Ratings Bajo la Técnica de Item-based:

For an item I, with a set of similar items determined based on rating vectors consisting of received user ratings, the rating by a user U, who hasn't rated it, is found by picking out N items from the similarity list that have been rated by U and calculating the rating based on these N ratings.

Español

El rating para un item **I** (que aún no tiene rating) se halla seleccionando N items de la lista de items similares que han obtenido rating del usuario **U**. Luego se calcula el rating basado en esos N ratings.

Matriz de Ratings

El objetivo de nuestro sistema recomendador es llenar la Matriz de Ratings, es decir predecir aquellos ratings que aún no se conocen.

	i_1	i_2	i_3	i_4	i_5
u_1	5		4	1	
u_2		3		3	
u_3		2	4	4	1
u_4	4	4	5		
u_5	2	4		5	2

PASO 3: Evaluación del Modelo

Como métrica de evaluación del modelo vamos a realizar el cómputo del Mean Absolute Error (MAE), que viene a ser el promedio de los valores absolutos de todas las diferencias entre el rating conocido y el rating previsto o pronosticado (predicted rating
A menor sea el MAE, mejor será el grado de precisión que tendrá el Sistema Recomendador para predecir los ratings.

Fórmula del MAE:

$$\text{MAE} = \frac{1}{|\hat{R}|} \sum_{\hat{r}_{ui} \in \hat{R}} |r_{ui} - \hat{r}_{ui}|$$

Importando los datos de los archivos productos y ratings

▾ Dataset

```
[17]  import os
      products_filename = os.path.join('/content/drive/My Drive/proyecto1/productos.csv')
      ratings_filename = os.path.join('/content/drive/My Drive/proyecto1/ratings.csv')
      df_productos = pd.read_csv(os.path.join(products_filename),
      usecols=['item_id', 'title'],
      dtype={'item_id': 'int32', 'title': 'str'})

      df_ratings = pd.read_csv(os.path.join(ratings_filename),
      usecols=['user_id', 'item_id', 'rating'],
      dtype={'user_id': 'int32', 'item_id': 'int32', 'rating': 'float32'})

      row_count1, column_count1 = df_ratings.shape
      row_count2, column_count2 = df_productos.shape
```

```
[18] print("Total de productos: ", row_count2)
     print()
     df_productos.style.hide_index()
     df_productos.head()
```

```
Total de productos:  1682
```

item_id	title
1	ACAI O HUASAI 300 GR GRANULADO AVANTARI
2	ACEITE DE COCO ORGANICO DE INDIA 1 LITRO AVANTARI
3	ACEITE DE COCO ORGÁNICO DE INDIA 500ML AVANTARI
4	ACEITE DE COPAIBA 30 ML - AVANTARI
5	ACEITE DE MAGNESIO 300 ML AVANTARI

```
[20]  print("Total de registros en la tabla 'ratings': ", row_count1)
      print()
      df_ratings.head()
```

```
Total de registros en la tabla 'ratings':  100003
```

	user_id	item_id	rating
0	0	50	5.0
1	0	172	5.0
2	0	133	1.0
3	196	242	3.0
4	186	302	3.0

```
[21] print("Número de productos: " + str(len(df_ratings.item_id.unique())))
```

```
Número de productos: 1682
```

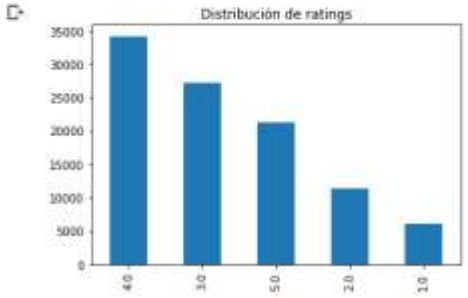
```
[22] print("Número de usuarios: " + str(len(df_ratings.user_id.unique())))
```

```
Número de usuarios: 944
```

Grafica mostrando la distribución de los ratings

Distribución de ratings entre los productos

```
[23] df_ratings.rating.value_counts().plot(kind="bar")
     plt.title("Distribución de ratings")
     plt.show()
```



Tipos de algoritmos KNN. Métricas de similitud y parámetros

Tipos de Algoritmos KNN, Métricas de Similitud y Parámetros

Antes de pasar a la etapa de entrenamiento y selección del modelo, veamos algunos tipos de algoritmos KNN y algunos de sus parámetros.

ALGORITMOS KNN

K-Nearest Neighbours algorithm calculates the distances between users or items and finds the closest ones = the most similar ones. Surprise package offers several variations of the model.

KNNBasic A basic collaborative filtering algorithm.
KNNWithMeans A basic collaborative filtering algorithm, taking into account the mean ratings of each user.
KNNWithZScore A basic collaborative filtering algorithm, taking into account the z-score normalization of each user.
KNNBaseline A basic collaborative filtering algorithm taking into account a baseline rating.

MÉTRICAS DE SIMILITUD

Entre las **métricas de similitud** disponibles para el cómputo de similitud entre usuarios o ítems tenemos MSD similarity, Cosine similarity o Pearson correlation coefficients, las cuales se basan simplemente en operaciones matemáticas.

`cosine`	Compute the cosine similarity between all pairs of users (or items).
`msd`	Compute the Mean Squared Difference similarity between all pairs of u
`pearson`	Compute the Pearson correlation coefficient between all pairs of users
`pearson_baseline`	Compute the (shrunk) Pearson correlation coefficient between all pairs

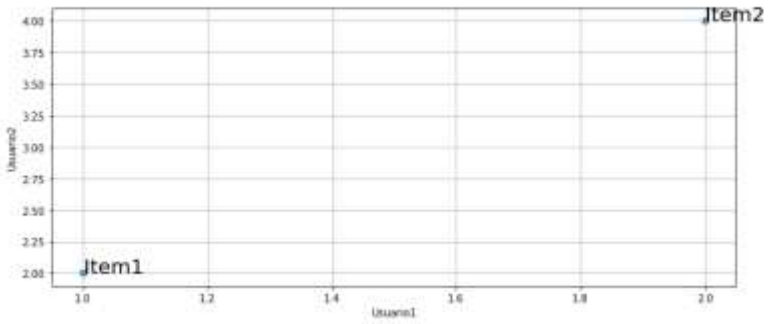
Cálculo del MSD entre 2 ítems

Cómo se calcula la MSD entre 2 items bajo la técnica "item-based"?

Para entender mejor, primero grafiquemos los ratings de dos usuarios otorgados a los items.

```
[24] a = [1, 2]
     b = [2, 4]
     #c = [2.5, 4]
     #d = [4.5, 5]

     labels = ["Item1", "Item2", "Item3", "Item4"]
     data = np.append([a], [b], axis=0)
     #data = np.append(data, [c], axis=0)
     #data = np.append(data, [d], axis=0)
```

```
fig = plt.figure(figsize=(12,5))
ax = fig.add_subplot(111)
ax.scatter(data[:, 0], data[:, 1])
ax.set_xlabel("Usuario1")
ax.set_ylabel("Usuario2")
for i, word in enumerate(data):
    ax.annotate(labels[i], xy=(data[i, 0]+0.001, data[i, 1]), size=20)
plt.grid(True)
plt.show()
```



En la gráfica mostrada arriba, cada punto representa un item y son graficados en relación a los ratings que ellos obtuvieron de dos usuarios: Usuario1 y Usuario2.

Y se lee por ejemplo: Item1 obtuvo 2 de rating del Usuario2; así como también item1 obtuvo 1 de rating del Usuario1.

MSD
Mean Square Distance
Distancia Cuadrática Media

Realiza el cómputo del **Mean Squared Distance similarity** entre todos los pares de **items**.

Solo **items** comunes son tomados en cuenta (En la grádica de arriba, Item1 es común para ambos usuarios; de la misma manera Item2 es común a ambos usuarios).

La '**Mean Squared Distance**' se define como:

$$\text{msd}(i,j) = \frac{1}{|U_{ij}|} \cdot \sum_{u \in U_{ij}} (r_{ui} - r_{uj})^2$$

Donde:
rui: el rating del usuario u al ítem i.
ruj: el rating del usuario u al ítem j.
Uij: el conjunto de todos los usuarios que han otorgado rating a ambos ítems i y j.

Tomando los valores de los ratings de nuestra gráfica mostrada arriba y reemplazándolos en la fórmula del msd(i,j), tenemos lo siguiente:

msd(i, j) = [(1 - 2)^2 + (2 - 4)^2] / 2 = **2.5**

Para calcular la similitud usando distancias se recurre a la función msd_sim(i, j) que retorna el valor de la similitud cuyo rango varía entre 0 y 1. A menor sea el **msd(i, j)** (distancia cuadrática media, mayor será la similitud o en otras palabras mayor será el valor del **msd_sim(i, j)**(Se acercará al valor de 1)

El **MSD-similarity** es definido como:

$$\text{msd\_sim}(i,j) = \frac{1}{\text{msd}(i,j) + 1}$$

Nota: el término **+1** solo es para evitar dividir entre 0.

Entonces remplazando el valor hallado, finalmente el **msd_sim(i, j)** sería:

msd_sim(i, j) = 1 / (2.5 + 1) = **0.2857**

Estimación o predicción del rating

Cómo se estima el rating de un ítem?

El rating estimado del usuario u al ítem i se calcula por la siguiente fórmula:

$$\hat{r}_{ui} = \frac{\sum_{j \in N_u^k(i)} \text{sim}(i,j) \cdot r_{uj}}{\sum_{j \in N_u^k(i)} \text{sim}(i,j)}$$

Notación:
sim(i, j): nos es otra cosa más que el msd_sim(i, j) expresado de otra manera. Se lee como la similitud entre el ítem i y el ítem j.
ruj: el rating real del usuario u al ítem j

En nuestro ejemplo, si quisiéramos estimar el rating del usuario1 al ítem2, al remplazar valores nuestra fórmula quedaría igual a:

(0.2857 * 2) / (0.2857) = 2

Nota: en este ejemplo solo estamos tomando un ítem similar... pero en la práctica encontraremos muchos ítems similares... es decir en el numerador de la expresión de arriba tendríamos una sumatoria de varios términos

Parámetros en los algoritmos KNN

PARÁMETROS

KNNBaseline(k=40, min_k=1, sim_options={})

1) **k**(int) – The (max) number of neighbors to take into for aggregation (see this note). Default is 40.

2) **min_k** (int) – The minimum number of neighbors to take into account for aggregation. Default is 1.

3) **sim_options** (dict) – A dictionary of options for the similarity measure. Options:

3.1) '**name**': The name of the similarity to use, as defined in the similarities module. Default is 'MSD'.
3.2) '**user_based**': Whether similarities will be computed between users or between items. This has a huge impact on the performance of a prediction algorithm. Default is True.
3.3) **min_support**:
For the **user-based approach**, is the minimum number of common items needed between users to consider them for similarity (when 'user_based' is 'True').
For the **item-based approach**, this corresponds to the minimum number of common users for two items to consider them for similarity (when 'user_based' is 'False')

Entrenamiento y evaluación MAE

Entrenamiento y Evaluación (split 75-25)

Split y Entrenamiento

```
[20] #Split
     from surprise.model_selection import train_test_split
     from surprise import Reader, Dataset

     reader = Reader(rating_scale=(1, 5))
     data = Dataset.load_from_df(df_ratings[['user_id', 'item_id', 'rating']], reader)

     trainset, testset = train_test_split(data, test_size=0.25)
```

```
[27] # Entrenamiento

     from surprise import KNNBasic, KNNWithMeans, KNNWithZScore, KNNBaseline, accuracy

     # To use user-based cosine similarity
     sim_options = {
         "name": "cosine",
         "user_based": True,  # Compute  similarities between users
     }
     model1 = KNNWithMeans(sim_options=sim_options)
     model1.fit(trainset)
```

```
Computing the cosine similarity matrix...
Done computing similarity matrix.
<surprise.prediction_algorithms.knns.KNNWithMeans at 0x7f932838da58>
```

Evaluación - MAE

```
predictions = model1.test(testset)
accuracy.mae(predictions)
```

```
MAE:  0.7603
0.7603468513346374
```

A continuación tenemos un ejemplo de cómo un usuario en particular 'U' calificaría al producto 'I' acorde al modelo:

```
userPrediction = model1.predict(186, 302, r_ui=3, verbose = True)
```

```
user: 186       item: 302       r_ui = 3.00   est = 3.94   {'actual_k': 40, 'was_impossible': False}
```

Entrenamiento usando K-Fold Cross Validation

Training Set y Testing Set

Antes de entrenar nuestro modelo necesitamos crear un training set y un testing set para el entrenamiento y evaluación respectivamente. También se puede dividir la data en subconjuntos (folds) donde parte de la data será usado para training y otra parte para testing.

Using only one pair of training and testing data is usually not enough. When you split the original dataset into training and testing data, you should create more than one pair to allow for multiple observations with variations in the training in testing data.

Algorithms should be cross-validated using multiple folds. By using different pairs, you'll see different results given by your recommender.

K-Fold Cross Validation

Método de la validación cruzada de K iteraciones mediante el cual se dividen los datos de entrada en K subconjuntos de datos. En nuestro caso vamos a dividir el dataset en 3 subconjuntos y vamos a entrenar o ajustar el modelo en 2 de los subconjuntos y, a continuación, evaluar el modelo en el subconjunto que no se ha utilizado para el entrenamiento. Este proceso se repite 3 veces, con un subconjunto diferente que se asigna para la evaluación o testing.

Selección del modelo

▾ Selección del Modelo

Selección del algoritmo KNN

Vamos a comparar primero diferentes tipos de algoritmos KNN, dejando como fijos el parámetro User-based = True y el parámetro de similitud = cosine. Esto con la finalidad de ver qué tipo de algoritmo KNN arroja menor error y así seleccionarlo y evaluarlo con diferentes parámetros.

Para efectos de training y testing, el dataset que vamos alimentar al modelo lo vamos a dividir en 3 subconjuntos iguales y realizaremos una validación cruzada de 3 iteracciones (K-Fold Cross-Validation)

```
models1 = generar_modelos_dict (
    models = [KNNBasic, KNNWithMeans, KNNWithZScore, KNNBaseline],
    sim_names = ['cosine'],
    user_based = True
)
results1 = cv_multiples_modelos (
    data=data,
    models_dict=models1,
    cv=3
)
models1 = None
display(results1)
```

```
---> CV for KNNBasic cosine...
Computing the cosine similarity matrix...
Done computing similarity matrix.
Computing the cosine similarity matrix...
Done computing similarity matrix.
Computing the cosine similarity matrix...
Done computing similarity matrix.

---> CV for KNNWithMeans cosine...
Computing the cosine similarity matrix...
Done computing similarity matrix.
```



```
---> CV for KNNWithZScore cosine...
Computing the cosine similarity matrix...
Done computing similarity matrix.
Computing the cosine similarity matrix...
Done computing similarity matrix.
Computing the cosine similarity matrix...
Done computing similarity matrix.

---> CV for KNNBaseline cosine...
Estimating biases using als...
Computing the cosine similarity matrix...
Done computing similarity matrix.
Estimating biases using als...
Computing the cosine similarity matrix...
Done computing similarity matrix.
Estimating biases using als...
Computing the cosine similarity matrix...
Done computing similarity matrix.
```

	fit_time	model	test_mae	test_time
0	0.802199	KNNBasic cosine	0.808326	5.134612
1	0.875009	KNNWithMeans cosine	0.759685	5.498317
2	0.899647	KNNWithZScore cosine	0.756181	5.938996
3	1.018789	KNNBaseline cosine	0.742883	6.346094

La comparación de esos modelos muestra que el algoritmo KNNBaseline tiene el mejor performance en términos de MAE.

El modelo KNNBaseline será probado más adelante con diferentes métricas de similitud y parámetros , así como tambien será probado usando las 2 técnicas User-based y Item-based.

User-based o Item-based

Selección de User-Based/Item-Based

User-Based

```
# USER-BASED

models2 = generar_modelos_dict([KNNBaseline], ['cosine', 'msd', 'pearson'], True)
results2 = cv_multiples_modelos(
           data=data, models_dict=models2, cv=3)
models2 = None
display(results2)
```

```
---> CV for KNNBaseline cosine...
Estimating biases using als...
Computing the cosine similarity matrix...
Done computing similarity matrix.
Estimating biases using als...
Computing the cosine similarity matrix...
Done computing similarity matrix.
Estimating biases using als...
Computing the cosine similarity matrix...
Done computing similarity matrix.

---> CV for KNNBaseline msd...
Estimating biases using als...
Computing the msd similarity matrix...
Done computing similarity matrix.
Estimating biases using als...
Computing the msd similarity matrix...
Done computing similarity matrix.
Estimating biases using als...
Computing the msd similarity matrix...
Done computing similarity matrix.
```

```
---> CV for KNNBaseline pearson...
Estimating biases using als...
Computing the pearson similarity matrix...
Done computing similarity matrix.
Estimating biases using als...
Computing the pearson similarity matrix...
Done computing similarity matrix.
Estimating biases using als...
Computing the pearson similarity matrix...
Done computing similarity matrix.
```

	fit_time	model	test_mae	test_time
0	1.018206	KNNBaseline cosine	0.741232	6.272307
1	0.454999	KNNBaseline msd	0.737265	6.273476
2	1.311676	KNNBaseline pearson	0.738294	6.176679

Queda claro que el **Mean Square Distance similarity** es el mejor en términos de MAE.

Item-Based

```
#ITEM-BASED

models3 = generar_modelos_dict([KNNBaseline], ['cosine', 'msd', 'pearson'], False)
results3 = cv_multiples_modelos(data=data, models_dict=models3, cv=3)
models3 = None
display(results3)
```

```
---> CV for KNNBaseline cosine...
Estimating biases using als...
Computing the cosine similarity matrix...
Done computing similarity matrix.
Estimating biases using als...
Computing the cosine similarity matrix...
Done computing similarity matrix.
Estimating biases using als...
Computing the cosine similarity matrix...
Done computing similarity matrix.

---> CV for KNNBaseline msd...
Estimating biases using als...
Computing the msd similarity matrix...
Done computing similarity matrix.


---> CV for KNNBaseline pearson...
Estimating biases using als...
Computing the pearson similarity matrix...
Done computing similarity matrix.
Estimating biases using als...
Computing the pearson similarity matrix...
Done computing similarity matrix.
Estimating biases using als...
Computing the pearson similarity matrix...
Done computing similarity matrix.
```

	fit_time	model	test_mae	test_time
0	1.609704	KNNBaseline cosine	0.748450	7.068021
1	0.590186	KNNBaseline msd	0.738953	7.112078
2	2.075053	KNNBaseline pearson	0.745611	6.909182

Aquí también observamos que el **Mean Square Distance** nos da el mejor resultado en términos de MAE.

User-Based vs Item-Based

Tanto User-Based y Item-Based arrojan un MAE similar y de aproximadamente 0.73. Cualquiera de los 2 técnicas usadas en nuestro sistema, tendría buen rendimiento.

Vamos a seleccionar el modelo: **Item-Based, KNN Baseline, MSD**

Afinando el modelo con diferentes parámetros

▾ Tuning: Afinando los Parámetros del Modelo

```
from surprise.model_selection import GridSearchCV
param_grid = {'k': [40,45,50],
              'min_k': [1,3,5],
              'sim_options': {'name': ['msd'],
                              'min_support': [1, 5],
                              'user_based': [False]}
              }
gs = GridSearchCV(KNNBaseline, param_grid, measures=['mae'], cv=3)

gs.fit(data)
```

```
Estimating biases using als...
Computing the msd similarity matrix...
Done computing similarity matrix.
Estimating biases using als...
Computing the msd similarity matrix...
Done computing similarity matrix.
Estimating biases using als...
Computing the msd similarity matrix...
Done computing similarity matrix.
```

```
[34]  print(gs.best_score['mae'])
      print(gs.best_params['mae'])
```

```
0.7352136330862677
{'k': 40, 'min_k': 5, 'sim_options': {'name': 'msd', 'min_support': 5, 'user_based': False}}
```

Los mejores parámetros

Como se observa, el algoritmo KNN Baseline basado en el item (Item-based) arroja el menor MAE si tenemos en cuenta los siguientes parámetros:

Métrica de similitud: Mean Square Distance (MSD)

'minimum support' = 5

k=40

min_k=5

Analizando el MAE

Análisis de los resultados del modelo de Filtros Colaborativos

Para la evaluación de nuestro modelo estamos utilizando el Mean Square Error (MAE), el cual mide la diferencia (como valor absoluto) entre la estimación del algoritmo y el valor real del rating. El cómputo se realiza sobre el subconjunto de datos apartados para el proceso de evaluación usando la siguiente fórmula:

$$\mathrm{MAE} = \frac{1}{|\hat{R}|} \sum_{\hat{r}_{ui} \in \hat{R}} |r_{ui} - \hat{r}_{ui}|$$

```
[35]  #Split
      from surprise.model_selection import train_test_split
      from surprise import Reader, Dataset

      reader = Reader(rating_scale=(1, 5))
      data = Dataset.load_from_df(df_ratings[['user_id', 'item_id', 'rating']], reader)

      trainset, testset = train_test_split(data, test_size=0.25)
```

```
[36]  # Entrenamiento

      from surprise import KNNBasic, KNNWithMeans, KNNWithZScore, KNNBaseline, accuracy

      # To use user-based cosine similarity
      sim_options = {
          "name": "msd",
          "user_based": False,  # Compute  similarities between users
          "min_support": 5,

      }
      model1 = KNNBaseline(k=40, min_k=5, sim_options=sim_options)
      model1.fit(trainset)
```

```
Estimating biases using als...
Computing the msd similarity matrix...
Done computing similarity matrix.
<surprise.prediction_algorithms.knns.KNNBaseline at 0x7f9328446278>
```

```
predictions = model1.test(testset)
accuracy.mae(predictions)
```

```
MAE:  0.7369
0.7368698121442393
```

El MAE arrojado por nuestro modelo es de aproximadamente 0.73, el cual consideramos un valor moderado y bueno si tenemos en cuenta que los valores del rating van del 1 al 5.

Ahora veamos el error absoluto de 5 muestras aleatorias en la siguiente tabla Rating Actual Vs. Rating Estimado

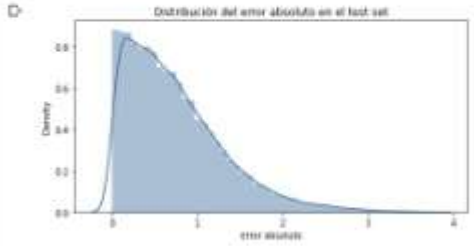
```
df_pred = pd.DataFrame(predictions, columns=['user_id', 'item_id', 'rating actual', 'rating estimado', 'details'])
#df_pred['rating estimado redondeado'] = df_pred['rating estimado'].round()
df_pred['error absoluto'] = abs(df_pred['rating estimado'] - df_pred['rating actual'])
df_pred.drop(['details'], axis=1, inplace=True)

df_versus = df_pred.sample(5)
estilo = df_versus.style.set_properties(**{'max-width': '-1', 'text-align': 'left', 'background-color': '#F9F8F8', 'color': '#7A7A7A'}).hide_index()
estilo.set_caption('Rating Actual Vs Rating Estimado')

estilo.set_table_styles([dict(selector='th', props=[('text-align', 'left'), ('color', '#7F7F7F'), ('background-color', '#E8EAFF')]),
                         dict(selector='caption', props=[('caption-side', 'top'), ('caption-alignment', 'left')])])
```

Rating Actual Vs Rating Estimado

user_id	item_id	rating actual	rating estimado	error absoluto
656	304	3.000000	2.886588	0.886588
878	137	3.000000	3.459403	0.459403
380	596	2.000000	3.674074	1.674074
13	40	2.000000	2.623057	0.623057
393	962	3.000000	2.914553	0.085447

Ahora veamos cómo está distribuido el error absoluto en el subconjunto de datos apartados para la evaluación (test set)

```
df_pred_err = df_pred.groupby('rating actual')['error absoluto'].mean().reset_index()
fig, (ax1) = plt.subplots(nrows=1, ncols=1, figsize=(6, 4))
sns.distplot(df_pred['error absoluto'], color='#2f6594', ax=ax1)
ax1.set_title('Distribución del error absoluto en el test set')
plt.show()
```



Finalmente podemos concluir de la gráfica que la mayoría de errores es pequeña: entre 0 y 1. A la derecha observamos una larga cola que nos indica que la minoría de errores fluctúa entre 1 y 4.

Entrenamiento del modelo final

▾ Modelo Final

Entrenamiento del modelo con los parámetros óptimos usando todo el dataset.

```
# Entrenamiento del modelo usando todo el dataset.

trainset = data.build_full_trainset()

sim_options = {
    "name": "msd",
    "user_based": False,  # realiza el computo de similitudes entre usuarios
    "min_support": 5,
}
modeloFinal = KNNBaseline(k=40, min_k=5, sim_options=sim_options)
modeloFinal.fit(trainset)
```

```
Estimating biases using als...
Computing the msd similarity matrix...
Done computing similarity matrix.
<surprise.prediction_algorithms.knns.KNNBaseline at 0x7f9328466f98>
```

Predicción de Ratings (Matriz de Ratings)

Predicción de ratings para todo par usuario-ítem que aún no se conoce.

	i_1	i_2	i_3	i_4	i_5
u_1	5		4	1	
u_2		3		3	
u_3		2	4	4	1
u_4	4	4	5		
u_5	2	4		5	2

```
# Predicción de ratings para todo par (usuario, producto) que no están en el dataset
testset = trainset.build_anti_testset()
predictions = modeloFinal.test(testset)
```

▾ 5 Mejores Recomendaciones

El reporte mostrado es el Top-5 de items con los ratings más altos que se han estimado para cada usuario en el dataset. Nosotros primero entrenamos el modelo optimizado usando todo el dataset, y luego estimamos todos los ratings para todo par user/item que no está en el dataset. Y finalmente seleccionamos las 5 mejores estimaciones por cada usuario.

```
# Obtener los N mejores productos recomendados para cada usuario
top_n = get_top_n(predictions, n=5)

# Muestra los N mejores productos recomendados para cada usuario
for uid, user_ratings in top_n.items():
    print(blue(uid), [(iid, round(est, 2)) for (iid, est) in user_ratings])
    #print(uid, [iid for (iid, _) in user_ratings])
```

```
82 [(408, 4.4), (603, 4.28), (169, 4.24), (488, 4.2), (187, 4.14)]
286 [(1062, 4.69), (318, 4.56), (1243, 4.54), (285, 4.5), (64, 4.49)]
200 [(64, 5), (603, 4.85), (488, 4.85), (12, 4.84), (408, 4.83)]
210 [(408, 4.95), (169, 4.95), (318, 4.93), (64, 4.88), (12, 4.79)]
224 [(64, 4.85), (966, 3.97), (50, 3.96), (12, 3.91), (272, 3.9)]
303 [(114, 4.83), (178, 4.75), (169, 4.75), (1449, 4.75), (272, 4.7)]
122 [(408, 4.67), (285, 4.62), (512, 4.6), (160, 4.50), (178, 4.56)]
194 [(603, 4.04), (114, 4.03), (408, 3.97), (169, 3.93), (313, 3.92)]
291 [(483, 5), (318, 5), (408, 4.99), (169, 4.95), (513, 4.93)]
234 [(408, 4.24), (114, 4.21), (169, 4.19), (272, 4.12), (302, 4.02)]
119 [(483, 4.84), (318, 4.83), (408, 4.75), (520, 4.74), (923, 4.71)]
167 [(64, 4.51), (483, 4.47), (488, 4.43), (50, 4.42), (12, 4.39)]
299 [(64, 4.25), (272, 4.23), (357, 4.23), (178, 4.21), (223, 4.15)]
308 [(114, 4.34), (694, 4.25), (1368, 4.21), (272, 4.19), (316, 4.19)]
95 [(169, 4.53), (408, 4.48), (318, 4.48), (12, 4.43), (114, 4.42)]
38 [(300, 5), (301, 4.99), (1164, 4.98), (1024, 4.93), (359, 4.85)]
102 [(169, 3.62), (408, 3.56), (64, 3.44), (318, 3.44), (114, 3.42)]
63 [(483, 4.17), (169, 4.09), (318, 4.04), (127, 4.04), (64, 4.04)]
160 [(12, 4.72), (318, 4.7), (98, 4.68), (64, 4.67), (178, 4.66)]
50 [(483, 4.37), (318, 4.36), (64, 4.3), (50, 4.23), (603, 4.23)]
301 [(1062, 4.37), (169, 4.37), (603, 4.31), (488, 4.29), (114, 4.28)]
225 [(483, 5), (408, 5), (318, 5), (169, 5), (12, 4.97)]
290 [(313, 4.13), (935, 4.03), (169, 4.03), (96, 4.01), (12, 4.01)]
97 [(285, 4.58), (64, 4.51), (12, 4.5), (134, 4.48), (483, 4.44)]
157 [(408, 4.83), (64, 4.82), (318, 4.82), (169, 4.79), (483, 4.75)]
101 [(1459, 3.15), (1269, 3.14), (1473, 3.05), (1004, 3.02), (1271, 3.0)]
278 [(408, 4.98), (169, 4.93), (318, 4.89), (483, 4.88), (64, 4.87)]
270 [(134, 4.84), (483, 4.83), (114, 4.82), (480, 4.75), (511, 4.75)]
7 [(313, 4.77), (963, 4.77), (251, 4.73), (408, 4.65), (59, 4.6)]
10 [(408, 4.06), (318, 4.04), (169, 4.77), (114, 4.75), (427, 4.60)]
```

Ejemplo de lo que se recomendaría a un cliente que ingresa a nuestra plataforma

▾ Interface

Bienvenido a Ecom ZonaBio

Fecha: 2020 / 10 / 7

Login:

usuario: 828

Productos Recomendados

Código	Producto
12	ACEITE ESENCIAL DE OREGANO 15 ML AVANTARI
98	HARINA DE ALFALFA 250 GR AVANTARI
169	QUINUA NEGRA GOURMET 300 GR AVANTARI
272	AE Abeto de Douglas (3 ml)
603	Touch Protein frutos rojos(540 gr)