UNIVERSIDAD PERUANA UNIÓN

FACULTAD DE INGENIERÍA Y ARQUITECTURA

Escuela Profesional de Ingeniería de Sistemas

Enfoques y algoritmos de aprendizaje automático para la clasificación de productos en comercio electrónico: Una revisión sistemática de la literatura

Trabajo de Investigación para obtener el Grado Académico de
Bachiller en Ingeniería de Sistemas

Por:
Harold Enrique Cotacallapa Mamani

Asesor:
Mg. Nemias Saboya Ríos

Lima, diciembre de 2020

DECLARACIÓN JURADA DE AUTORÍA DE TRABAJO DE INVESTIGACIÓN

Mg. Nemias Saboya Ríos, de la Facultad de Ingeniería y Arquitectura, Escuela Profesional de Ingeniería de Sistemas, de la Universidad Peruana Unión.
DECLARO:

Que el presente informe de investigación titulado: "**Enfoques y algoritmos de aprendizaje automático para la clasificación de productos en comercio electrónico: Una revisión sistemática de la literatura**" constituye la memoria que presenta el estudiante **Harold Enrique Cotacallapa Mamani** para aspirar Grado de Bachiller en Ingeniería de Sistemas, cuyo trabajo de investigación ha sido realizada en la Universidad Peruana Unión bajo mi dirección.

Las opiniones y declaraciones en este informe son de entera responsabilidad del autor, sin comprometer a la institución.

Y estando de acuerdo, firmo la presente constancia en Lima, 21 de diciembre del año 2020.

Asesor
Mg. Nemias Saboya Ríos.

ACTA DE SUSTENTACIÓN DE TRABAJO DE INVESTIGACIÓN

En Lima, Ñaña, Villa Unión, a……..los.....21.....día(s) del mes de…..diciembre.....del año 2020.... siendo las.....08:40.....horas, se reunieron los miembros del jurado en la Universidad Peruana Unión campus Lima, bajo la dirección del (de la) presidente(a): ..... Dra. Erika Inés Acuña Salinas....., el (la) secretario(a): ..... Ing. Diana Lidia Sanchez Torpoco..... y los demás miembros:.… Ing. Jenson Daniel Chambi Aguilar …..y el (la) asesor(a): ….. Mg. Nemias Saboya Rios.… con el propósito de administrar el acto académico de sustentación del trabajo de investigación titulado: "Enfoques y algoritmos de aprendizaje automático para la clasificación de productos en comercio electrónico: Una revisión sistemática de la literatura"……..

…….....….de los (las) egresados (as): a)…….....Harold Enrique Cotacallapa Mamani …….....……b)……..……

…… conducente a la obtención del grado académico de Bachiller en ……

……..……Ingeniería de Sistemas……..
(Denominación del Grado Académico de Bachiller)

El Presidente inició el acto académico de sustentación invitando …al… candidato(a)/s hacer uso del tiempo determinado para su exposición. Concluida la exposición, el Presidente invitó a los demás miembros del jurado a efectuar las preguntas, y aclaraciones pertinentes, las cuales fueron absueltas por …el… candidato(a)/s. Luego, se produjo un receso para las deliberaciones y la emisión del dictamen del jurado.

Posteriormente, el jurado procedió a dejar constancia escrita sobre la evaluación en la presente acta, con el dictamen siguiente:

Candidato/a (a): …… Harold Enrique Cotacallapa Mamani …….

CALIFICACIÓN	**ESCALAS**			**Mérito**
	Vigesimal	**Literal**	**Cualitativa**	
Aprobado	**20**	**A+**	**Con nominación de Excelente**	Excelencia

Candidato/a (b): ……..

CALIFICACIÓN	**ESCALAS**			**Mérito**
	Vigesimal	**Literal**	**Cualitativa**	

(*) Ver parte posterior

Finalmente, el Presidente del jurado invitó …al… candidato(a)/s a ponerse de pie, para recibir la evaluación final y concluir el acto académico de sustentación procediéndose a registrar las firmas respectivas.

Presidente
Dra. Erika Inés Acuña Salinas

Secretario
Ing. Diana Lidia Sanchez Torpoco

Asesor
Mg. Nemias Saboya Rios

Miembro

Miembro
Ing. Jenson Daniel Chambi Aguilar

Candidato/a (a)
Harold Enrique Cotacallapa Mamani

Candidato/a (b)

Enfoques y algoritmos de aprendizaje automático para la clasificación de productos en comercio electrónico: Una revisión sistemática de la literatura

Harold Cotacallapa[1]

[1] Universidad Peruana Unión
Lima, Perú
haroldcotacallapa@upeu.edu.pe

Resumen. Con el rápido crecimiento del comercio electrónico en los últimos 3 años, y tras su reciente aceleración, debido a la enfermedad por coronavirus (COVID-19), surge la necesidad de ofrecer un servicio de comercio electrónico óptimo que maximice las utilidades del negocio. Un servicio óptimo implica un motor de búsqueda y recomendación altamente eficaces, por tanto, depende significativamente de una excelente clasificación de productos, lo cual aún es un desafío para la ciencia y la industria, ya que, implica una clasificación jerárquica múltiple en tiempo real para un inmenso volumen de productos con descripciones no estructuradas y una larga lista de subcategorías con pocos datos. No obstante, el reciente avance de la Inteligencia Artificial, generó una amplia gama de algoritmos que abordan estos problemas. Por consiguiente, el presente artículo desarrolla una Revisión Sistemática de la Literatura (RSL) con el objetivo de identificar los algoritmos de aprendizaje automático, sus métricas de evaluación y los enfoques usados para la clasificación de productos en comercio electrónico. Esta RSL sigue una secuencia de pasos definidos en la guía de Kitchenham. Al finalizar este estudio, se concluye que los algoritmos más usados son K-Means, SVM, y Naive Bayes cuando el objeto de estudio es el texto, y la red neuronal convolucional jerárquica cuando se trabaja con imágenes; además, más del 70% de los artículos usan solamente los atributos textuales del producto y la mayoría de artículos cuyo enfoque es el aprendizaje supervisado usan el "accuracy" como única métrica para validación del modelo.

Palabras claves: clasificación, comercio electrónico, algoritmos, revisión sistemática, aprendizaje automático

1 Introducción

Tras el exitoso proyecto del inglés Tim Berners-Lee, más conocido como la World Wide Web y la celeridad del acceso a Internet alrededor del mundo, el gran salto del comercio electrónico solamente fue cuestión de tiempo [1], así, a finales del siglo XX, eBay y Amazon, disponibilizaron sus primeros portales exclusivos para esta actividad económica, lo que significó una revolución en las compras digitales. Desde entonces, este modelo de negocio ha tenido un crecimiento constante; aunque según eMarketer[2], el crecimiento del mercado minorista global se ha desacelerado considerablemente en comparación con los cinco años anteriores, no obstante, las ventas de comercio electrónico en todo el mundo superaron los $ 3,5 billones de dólares en el 2019, un aumento de aproximadamente el 18% con respecto al año anterior y se espera que el comercio electrónico casi se duplique para 2023 a más de $ 6.5 mil millones, asimismo, en 2019, la participación del comercio electrónico en las ventas minoristas globales totales fue del 14.1% y los analistas esperan que aumente un 2% anual hasta 2023.

No obstante, la enfermedad por coronavirus COVID-19 en el presente año, ha generado diversas medidas de prevención a nivel mundial, las cuales han tenido un fuerte impacto sobre el comercio electrónico global. A nivel mundial las plataformas minoristas han experimentado un aumento de tráfico global sin precedentes entre enero de 2019 y junio de 2020, superando incluso los picos de tráfico de la temporada navideña, en general, los sitios web minoristas generaron casi 22 mil millones de visitas en junio de 2020, frente a los 16.07 mil millones de visitas globales en enero de 2020 [3].

En particular, los principales mercados de comercio electrónico, Estados Unidos y China, experimentaron una fuerte demanda de compras en línea, según el reporte "2020 Digital Economy Index" de Adobe Analytics y una encuesta sobre el comportamiento de compra en línea de los chinos realizada por Rakuten Insight en mayo de 2020 [4], [5]. De la misma manera, en América Latina, tras el brote de COVID-19, los pedidos minoristas en línea se dispararon en México, Colombia y Perú a fines de marzo de 2020, mientras que en Brasil, el mayor aumento en pedidos minoristas electrónicos se registró a fines de febrero, donde los pedidos de e-retail crecieron un 35 por ciento en comparación con el mismo período del año anterior [6].

Ante la descripción del escenario actual de la venta global al por menor, la participación del comercio electrónico y el impacto que está produciendo la pandemia mundial en esta actividad económica, nasce la necesidad de fortalecer los servicios ofrecidos mediante una plataforma de comercio electrónico, con el fin de maximizar las ganancias. Por tanto, para alcanzar este objetivo es fundamental tener en cuenta el motor de búsqueda y recomendación [7], los cuales, a su vez, dependen significativamente de una óptima clasificación de productos [8].

Ahora bien, la clasificación de productos, es un campo de estudio ya recorrido por distintos investigadores, sin embargo, hasta la actualidad, las empresas aún buscan un algoritmo que se adapte a su propio contexto, principalmente por el idioma, y además, que resuelva diversos problemas que continúan siendo investigados, tales como, el orden jerárquico de los productos, que, a diferencia de una clasificación tradicional, esta posee niveles de clasificación; el voluminoso número de productos y categorías, cuyo crecimiento es diario y veloz; la información no estándar contenida en la descripción del producto; la limitada cantidad de datos para una clase determinada y por último la necesidad de una clasificación en tiempo real [8]; estos problemas, sin duda, podrían ser resueltos a través de una clasificación manual y con un número adecuado de trabajadores, no obstante, esta solución es ineficiente, no escalable e imprecisa, ya que, está sujeta al rendimiento y sesgo de los trabajadores, por ello, la clasificación manual queda limitada [9]. En efecto, cuanto mejor sea la clasificación, más información puede generar una buena categoría de productos [10].

Cabe resaltar que, el problema de la clasificación de productos de comercio electrónico no es reciente, antes del comienzo del siglo XXI ya se conocía, por ello en 1998 se creó el United Nations Standard Products and Services Code (UNSPSC) [10], el cual es definido como una taxonomía de productos y servicios para su uso en comercio electrónico, dicho estándar presenta una clasificación jerárquica con cinco niveles, aunque en la práctica el quinto nivel apenas se usa [11].

No obstante, recientes avances de la Inteligencia Artificial permitieron crear tecnología de clasificación basada en aprendizaje automático, mediante el uso de la información del producto, tanto gráfica como textual, brindada por los consumidores o agentes de comercio electrónico, además, han demostrado promisorios avances en la automatización de la clasificación de productos en comercio electrónico. Algunas compañías reconocidas globalmente por su presencia en el comercio electrónico, han desarrollado algoritmos de clasificación para sus productos o han disponibilizado un dataset para que puedan ser construidos por terceros, como es el caso de la compañía

japonesa más grande, Rakuten Ichiba. Cevahir y Murakami [12] propusieron un modelo de clasificación jerárquica, donde usaron dos modelos neuronales diferentes, “belief nets” y “deep autoencoders”, para un dataset de 150 millones de productos con 5 niveles, consiguiendo un 81% de precisión. Shen et. al [13] propuso también una clasificación jerárquica, que se descompone en tareas para el “coarse level” y “fine level”, y utilizó un algoritmo gráfico para descubrir automáticamente grupos de clases muy similares como categoría de producto en lugar de depender de una jerarquía definida por humanos; el modelo fue entrenado con 83 millones de productos de eBay. Por último y no mucho menos, Vandic et. al [14] propuso un marco de clasificación jerárquica de productos (HPC), el cual utiliza múltiples nodos de clasificación, los cuales son construidos a partir de una “receta de clasificación”, estas permiten clasificadores flexibles que se ajusten a las características específicas de la taxonomía; basado en 3000 descripciones de productos de Amazon.com, este marco alcanza una precisión general del 76.80%.

Con todo, los algoritmos de aprendizaje automático, aún enfrentan desafíos importantes para conseguir una clasificación eficaz y eficiente, tales como, la clasificación jerárquica múltiple [15] la cual difiere de una clasificación tradicional [16], una distribución con “cola” larga, es decir, pocos datos en muchas categorías de productos [17]; y el pre procesamiento para los datos no estructurados de la información textual del producto [18]; a esto se agrega la versatilidad del sistema de clasificación para detectar nuevas categorías de productos.

En consecuencia, dado que el conocimiento sobre nuevos algoritmos y modelos para abordar este problema se encuentra esparcido en la literatura científica, y debido a la necesidad de tener un estado de arte actualizado respecto a los algoritmos de aprendizaje automático empleados para la clasificación de productos de comercio electrónico, el presente estudio tiene como objetivo realizar revisión sistemática de la literatura e identificar los algoritmos de aprendizaje automático, sus métricas de evaluación y los enfoques usados para la clasificación de productos en comercio electrónico. Finalmente, este artículo se encuentra organizado de la siguiente manera: la sección 2 presenta el marco conceptual, la sección 3 describe el protocolo de la revisión sistemática de la literatura, la sección 4 presenta los resultados de la revisión y por último la sección 5 presenta las conclusiones y el trabajo futuro referente al objeto de estudio.

2 Marco conceptual

2.1 Comercio electrónico

Es de conocimiento común que el comercio electrónico es un modelo de negocio que ha sobrepasado las fronteras físicas y las limitaciones de tiempo y espacio que implica una adquisición de un producto o servicio de modo offline. No en tanto, es pertinente definir algunos conceptos relacionados a este tema y saber qué implica hablar sobre comercio electrónico.

Como sugiere el término, el comercio electrónico se refiere a diversas actividades comerciales en línea centradas en el intercambio de productos básicos por medios electrónicos, Internet en particular, por parte de empresas, fábricas, empresas industriales y consumidores [19]. Por otro lado, Turban et. al define comercio electrónico como el uso de Internet y otras redes (por ejemplo, intranets) para comprar, vender, transportar o intercambiar datos, bienes o servicios [20]. En términos más simples, la ISO define el comercio electrónico como: es el término general para el intercambio de información entre empresas y entre clientes y empresas [19]. No obstante, muchos confunden el término de e-business con e-commerce por lo cual es necesario saber que, e-business se refiere a una definición más amplia de e-commerce, no solo la compra y venta de bienes

y servicios, sino también la realización de todo tipo de negocios en línea, como el servicio a los clientes, la colaboración con socios comerciales, la entrega de aprendizaje electrónico y la realización de transacciones electrónicas dentro de las organizaciones [20].

Ahora bien, todo modelo de negocio, posee un mercado, el cual permite la interacción entre la demanda y la oferta; en términos de comercio electrónico, este espacio se denomina e-marketplace, también conocido como electronic market o mercado electrónico [20]. La clasificación del comercio electrónico ayuda a comprender mejor este campo diversificado, en general, vender y comprar electrónicamente puede ser de empresa a consumidor (B2C) o de empresa a empresa (B2B) [19], [20].

Por otro lado, el desarrollo del comercio electrónico, se basa fundamentalmente en el desarrollo de muchas otras ciencias y tecnologías relacionadas, entre ellas, las ciencias de las matemáticas, la informática, las comunicaciones y la gestión tienen una gran influencia en el desarrollo del entorno blando del comercio electrónico, no en tanto, como su propio nombre lo define, comercio electrónico, resaltamos la importancia de la informática. La historia del comercio electrónico avanza conforme al desarrollo de la tecnología, tras la creación de la primera computadora, la comunicación exitosa entre dos computadores, y posteriormente la creación del internet, trajo consigo las primeras transacciones electrónicas de dinero entre entidades financieras, luego apareció el intercambio electrónico de datos (EDI, por sus siglas en inglés). Más tarde, a principios de la década de 1990, la aparición de la World Wide Web, marcó un hito importante en el desarrollo del comercio electrónico, que permitía a las empresas tener presencia en Internet tanto con texto como con fotos, fue en estos años que se acuñó el término de comercio electrónico o e-commerce, desde entonces las plataformas de comercio electrónico se extendieron por todo el mundo [20].

En la actualidad, la tecnología continúa avanzando y su desarrollo ha permitido que las plataformas de comercio electrónico puedan crecer en un contexto seguro, ágil y cada día atraer a nuevos consumidores, no obstante, cabe mencionar que en el siglo XXI los rápidos avances en inteligencia artificial y sus distintas áreas de aplicación han repercutido también en la automatización inteligente de ciertas actividades en las plataformas de comercio electrónico, tales como el motor de búsqueda, motor de recomendación y por su puesto la clasificación de los productos.

2.1.1 Clasificación de productos de comercio electrónico

La clasificación de productos se refiere a un sistema de categorías en el que se coloca un grupo de productos; una plataforma de comercio electrónico, usualmente aprovecha la taxonomía del producto para ordenarlo jerárquicamente, el nivel más alto es denominado nivel superficial o coarse-grain y las categorías más bajas son usualmente más específicas, y a dicho nivel se denomina nivel profundo o fine-grain; también recibe el nombre de hojas, haciendo referencia a la estructura de un árbol. Sin embargo, comparando entre ambos niveles, la clasificación de productos a nivel profundo se mantiene mucho más desafiadora que una clasificación a nivel superficial. [15]

Si bien es cierto, esta tarea de clasificación puede ser hecha manualmente, sin embargo, una clasificación automática de productos puede reducir el trabajo manual, tiempo, costo y es escalable, dando a los clientes una mejor experiencia al buscar o subir nuevos productos. [21] Y cuando se habla de clasificación automática, precisión y gran cantidad de datos, abordamos los enfoques computarizados, donde el aprendizaje automático es una vía natural para que un algoritmo informático aprenda de un conjunto de datos mientras optimiza continuamente la operación de clasificación para reducir el

error, el tiempo y el costo [22]. El aprendizaje supervisado y el aprendizaje no supervisado son grupos de algoritmos de aprendizaje automático que se usan según el problema y el conjunto de datos, donde uno necesita datos etiquetados, y el otro no necesita, pues su objetivo es reconocer padrones dentro de un conjunto de datos.

Asimismo, cuando se propone una clasificación automática de productos, nos referimos a un problema de clasificación jerárquica, lo cual trae consigo algunos desafíos, según [23] el primer desafío es que el conjunto de datos suele presentar una gran cantidad de categorías con datos que son extremadamente escasos y con una distribución sesgada de cola larga; segundo, una taxonomía jerárquica impone restricciones a la activación de las etiquetas, es decir, si una etiqueta secundaria está activa, entonces es necesario que una etiqueta principal esté activa; tercero, para un uso práctico, la predicción debe suceder en tiempo real, idealmente en pocos milisegundos. Esta lista, no está completa si no mencionamos un cuarto desafío, que son las descripciones de texto ruidosas de los artículos de acuerdo a [24], ya que, según el tipo de comercio electrónico, sucede que el vendedor ingresa la información del producto según su criterio y muchos de ellos pueden tener su propia taxonomía con la cual guardan la información del producto, como es descrito en [7].

Uno de los primeros enfoques en esta área de investigación fue GoldenBullet, un sistema que aplicó una combinación de recuperación de información (Information Retrieval) y técnicas de aprendizaje automático (Machine Learning) para clasificar los productos de una plataforma de comercio electrónico según una estructura taxonómica pre definida como la de UNSPSC [25]

2.2 Aprendizaje automático

2.2.1 Algoritmos de aprendizaje supervisado de clasificación

El aprendizaje supervisado, tiene como objetivo usar un algoritmo para aprender la función óptima que mejor representa la relación entre la entrada y la variable objetivo, se denomina “supervisado” porque el algoritmo tiene conocimiento previo de cuáles deberían ser los valores de salida [26]. Dependiendo de si la variable objetivo es una variable categórica o continua, una tarea de aprendizaje supervisada es un problema de clasificación o regresión, respectivamente, algunos algoritmos más reconocidos son: Regresión logística, Support Vector Machine, Decision Tree y Naive Bayes [27], entre otros algoritmos más recientes como las redes neuronales y las redes neuronales convolucionales.

2.2.1.1 Árboles de decisión

Algoritmo de aprendizaje supervisado, diseñado para resolver tanto problemas de clasificación como de regresión dividiendo continuamente los datos según un determinado parámetro. Las decisiones están en las “hojas” y los datos se dividen en los nodos. Según [27] Este algoritmo es adecuado para problemas de regresión y clasificación, facilidad de interpretación, facilidad de manejo de valores categóricos y cuantitativos, capaz de completar los valores faltantes en los atributos con el valor más probable; no obstante, puede ser inestable, puede ser difícil controlar el tamaño del árbol, puede ser propenso a errores de muestreo y proporcionar una solución óptima localmente, no una solución global y óptima.

2.2.1.2 Redes neuronales artificiales

Las Redes Neuronales Artificiales (ANN) es un modelo de estructura neural biológica que recibe una entrada a través de un axón en un cuerpo celular y envía una salida a la siguiente neurona a través de una sinapsis. Este proceso involucra a muchas neuronas que están conectadas en paralelo y tienen la capacidad de aprender de un error para mejorar a sí mismas ajustando el peso de cada neurona. Un modelo de redes neuronales tiene tres tipos principales de capas. La primera capa es una capa de entrada para recibir datos y enviarlos a la siguiente capa. La segunda capa consta de capas ocultas que se encargan de calcular y mejorar los nodos. El rendimiento y la precisión de un modelo dependen de las características de esta capa, como el número de capas y nodos ocultos. Después de que los datos fueron procesados por las capas ocultas, la capa de salida determina la respuesta usando una función activada para un problema específico [22].

2.2.1.3 Redes neuronales convolucionales

Las redes neuronales convolucionales (CNN) son consideradas como una evolución de las primeras redes neuronales artificiales [26]. CNN es ampliamente utilizado con datos de imágenes. CNN sigue los supuestos básicos de Neutral Networks (NN) ya que incluye la función de pérdida y la capa totalmente conectada. Sin embargo, la mayor diferencia de CNN es que no todas las neuronas tienen conectividad completa, excepto la última capa completamente conectada. Esto se debe a que la conectividad total en todos los parámetros es un desperdicio y puede conducir a un sobreajuste. Además, CNN tiene una composición de capas diferente en su arquitectura, ya que se compone de una capa convolucional para generar mapas de características, una capa de agrupación para reducir la dimensionalidad de los mapas de características y una capa completamente conectada para comprender y clasificar las características extraídas [28].

2.2.1.4 Máquina de soporte vectorial

Máquina de soporte vectorial (SVM) es uno de los potentes algoritmos que pueden clasificar una amplia gama de tipos de datos [29]. SVM selecciona y utiliza instancias representativas adecuadas de un conjunto de entrenamiento como un vector de soporte. SVM construye hiperplanos entre vectores de soporte de cada clase, que luego pueden usarse básicamente para la clasificación lineal. Para hacer un modelo como un clasificador no lineal, se aplica una función de núcleo. Según [22] la función Kernel asigna datos de entrada en un espacio de características de dimensión inferior a un espacio de características de dimensión superior. Algunas funciones comunes del kernel incluyen la función de base polinómica y radial (RBF).

Tanto SVM como CNN tienen sus propios defectos para el problema considerado de clasificación de productos de comercio electrónico. SVM es más adecuado para problemas de clasificación binaria, pero el problema considerado es un problema de varias clases con una gran cantidad de categorías de productos. CNN generalmente requiere una gran cantidad de datos de entrenamiento, pero el problema considerado generalmente no tiene una gran cantidad de muestras para cada categoría [8].

2.2.1.5 Naïve Bayes

Esta técnica se basa en el teorema de Bayes con fuertes supuestos de independencia entre las características. Por lo general, se usa para la clasificación de texto al calcular las probabilidades de ocurrencia de elementos o probabilidades posteriores

dado que una ocurrencia y la ocurrencia anterior son independientes. Luego se elige la ocurrencia con mayor probabilidad. Este modelo se usa para predecir la frecuencia de las ocurrencias de corpus con la suposición de que la longitud del documento está relacionada con la etiqueta de acuerdo con el teorema de Bayes. Por lo tanto, cada documento representa una bolsa de palabras. Las palabras se cuentan para que se pueda calcular la probabilidad de cada etiqueta. [22]

2.2.2 Algoritmos de aprendizaje no supervisado de clasificación

Según [30],el modelo de aprendizaje no supervisado se utiliza para descubrir grupos de instancias similares dentro de los datos. Tiene la capacidad de aprender y agrupar información basada en diferentes similitudes entre las instancias y proporciona una posible solución para el problema de agrupamiento. A diferencia del aprendizaje supervisado, en una tarea de aprendizaje no supervisada, no hay un valor objetivo o una clase conocida. Existen dos aplicaciones principales del aprendizaje no supervisado útiles para la investigación de los trastornos cerebrales: la agrupación y la reducción de la dimensionalidad [31].

2.2.2.1 K-Means

La técnica de agrupación de K-means es el proceso para agrupar los elementos de un conjunto de datos en grupos de k-means. El proceso de agrupamiento se realiza calculando la distancia euclidiana entre los elementos y el centroide, donde el centroide es el elemento con densidad uniforme [32]. La simplicidad y la alta velocidad para ejecutar grandes conjuntos de datos son las principales fortalezas de la agrupación de K-Means. Sin embargo, generalmente es una tarea difícil seleccionar características óptimas y útiles para reducir una gran cantidad de características, que a menudo son ruidosas y redundantes para datos de alta dimensión [30].

2.2.2.2 Reducción de la dimensionalidad

La reducción de la dimensionalidad es útil en situaciones en las que el número de características es sustancialmente mayor que el número de observaciones. En tales casos, reducir el número de características puede ser beneficioso para aliviar las demandas computacionales, eliminar información redundante o irrelevante y reducir el riesgo de sobreajuste. Se pueden utilizar varias técnicas para realizar la reducción de la dimensionalidad, como el análisis de componentes principales (PCA), el análisis de componentes independientes o los autoencoders [31]. PCA es una técnica estadística clásica para transformar los atributos del conjunto de datos en un nuevo conjunto de atributos no correlacionados llamados componentes principales. Esta técnica es la más usada pues al realizar la reducción consigue mantener la mayor variabilidad posible del conjunto de datos [30], asegurando la integridad de los datos.

2.3 Algoritmos ensamblados

Muchos investigadores han investigado la técnica de combinar las predicciones de múltiples clasificadores para lograr una mayor precisión, un concepto que generalmente se conoce como ensemble learning. El método de bagging es un algoritmo de clasificación basado en votación que permite la combinación de múltiples clasificadores "débiles" para producir un modelo de clasificación más preciso y "fuerte". El nombre se deriva de las palabras Boostrap aggregating [33].

3 Método de la revisión sistemática de la literatura

La metodología de investigación desarrollada en este artículo para reunir y evaluar toda la información y evidencia disponible acerca del tópico de estudio, consiste en una serie de procedimientos definidos en la guía de Kitechenham, dicha metodología, también conocida como la Revisión Sistemática de la Literatura (RSL) tiene su origen en las ciencias de la salud y ciencias sociales, sin embargo, ha sido adaptada para reflejar los problemas específicos de la investigación en ingeniería de software. [34]

La Revisión Sistemática de la Literatura (RSL) es una metodología científica que puede ser usada para integrar la investigación empírica en el área de Ingeniería de Software [35]. Además, esta metodología ha sido implementada por diversos investigadores en el área de Inteligencia Artificial, siendo la misma área del presente artículo [36]–[38], lo cual, refuerza la validez de uso de la metodología.

3.1 Necesidad de la revisión sistemática

La presente Revisión Sistemática de la Literatura surge con el objetivo de determinar los algoritmos de aprendizaje automático que existen para realizar una clasificación de productos de comercio electrónico, conocer cuáles son sus características y qué enfoques existen para abordar el problema de clasificación automática de productos en el área del comercio electrónico. Esta necesidad se sustenta en la complejidad de tomar una decisión respecto al algoritmo adecuado para el problema mencionado anteriormente, ya que existen diversos modelos y enfoques para abordar dicho problema. En cierto grado se seguirá la plantilla Goal, Question, Metric (GQM) para establecer el objetivo de la investigación, lo cual se describe en la Tabla 1

Tabla 1 : Elaboración del objetivo de la investigación

Campo	Valor
Objeto de estudio	Clasificación de productos de comercio electrónico automático
Propósito	Identificar
Foco	Algoritmos, arquitecturas, modelos, métodos, enfoques
Contexto	Ninguno para este caso

3.2 Preguntas para la revisión sistemática

Con el fin de delimitar el objetivo del estudio se plantearon las preguntas de investigación, lo cual sirve como un punto de partida para el estudio. En la Tabla 2 se describen las preguntas propuestas, relacionadas al tópico de estudio.

Tabla 2: Preguntas de Investigación

ID	Pregunta	Motivación
PI-1	¿Qué **algoritmos** existen para la clasificación de productos en comercio electrónico?	Determinar los algoritmos que son usados para la clasificación de productos en comercio electrónico

PI-2	¿Cuáles son los **enfoques** abordados para la clasificación de productos en comercio electrónico?	Identificar los diversos enfoques con los que se aborda el problema de clasificación en productos de comercio electrónico.
PI-3	¿Cuáles son las **métricas** usadas para la validación de los algoritmos?	Identificar las métricas de validación de los algoritmos propuestas para la clasificación de productos en comercio electrónico.

De la misma manera, se formularon las preguntas bibliométricas, con el fin de obtener una visualización de la evolución y tendencia que existe en el tema de investigación, estas preguntas son detalladas en la Tabla 3.

Tabla 3: Preguntas bibliométricas

ID	Pregunta	Motivación
PB-1	¿ Cuántos artículos referentes a la clasificación de productos de comercio electrónico se publicaron desde el 2014 ?	Determinar la frecuencia de artículos publicados para poder establecer la relevancia del tema en el tiempo.
PB-2	¿ Cuál es la cantidad de publicaciones por tipo de artículo ?	Identificar la cantidad de publicaciones referentes a la clasificación de productos de comercio electrónico según el tipo de artículo para identificar la concentración de los mismos
PB-3	¿ Cuáles son las publicaciones en las que se han encontrado estudios relacionados a la clasificación de productos de comercio electrónico ?	Identificar en qué dominio de aplicación se concentra la mayor cantidad de publicaciones sobre este tema.

3.3 Definición del protocolo de investigación

Basado en la guía para la elaboración de la RSL de Kitchenham, el protocolo de la revisión consiste en la especificación formal de los pasos a seguir durante la realización de la revisión sistemática. A continuación se presentan los pasos para la conducción de la RSL [34].

3.3.1 Definición de las cadenas de búsqueda

Para la formulación de la cadena de búsquedas, se eligió la estrategia PICO, la cual fue propuesta como método para la recolección de evidencias, según la Práctica Basada en Evidencias (PBE) [39]. PICO representa un acrónimo para población, intervención, comparación y resultados (outcomes), que son los elementos fundamentales para la búsqueda bibliográfica de evidencias. Por la naturaleza de la investigación, el tercer elemento no se aplica a este estudio, asimismo en las Tablas 4 y 5 se detallan los términos principales, los términos alternos, y para cada uno se detalla su justificación.

Con respecto a las fuentes de datos, las librerías digitales consideradas para el presente estudio fueron escogidas por su relevancia e importancia que tienen sus publicaciones y para la carrera al cual el artículo va dirigido:

- Science Direct (https://www.sciencedirect.com/)
- ACM (https://dl.acm.org//)
- SpringerLink (https://link.springer.com/)
- IEEE Xplore (https://ieeexplore.ieee.org/Xplore/home.jsp)

Tabla 4: Proceso de construcción de la cadena de búsqueda según PICO

Término principal	Términos alternos	Justificación
Población		
Algoritmos	Modelos	Se selecciona algoritmos por ser el foco de estudio y los términos alternos: modelos por ser cercanos al término principal
Enfoques	Métodos, técnicas	Se selecciona enfoques por ser el foco de estudio y los términos alternos: métodos, técnicas por ser cercanos al término principal
Intervención		
Aplicado a la clasificación de productos de comercio electrónico	Comercio electrónico, clasificación de productos	Se elige el término principal por ser el objeto de estudio y elemento de intervención en la búsqueda de algoritmos/métodos/técnicas.
Aplicado a algoritmos de clasificación	Aprendizaje de máquina, aprendizaje supervisado, aprendizaje no supervisado.	Se elige el término principal por ser el objeto de estudio y elemento de intervención en la búsqueda de algoritmos/métodos/técnicas
Resultados		
Rendimiento de los algoritmos de clasificación de productos electrónicos	Precisión	Se selecciona estos términos para conocer los resultados de la aplicación de los algoritmos
Propuestas y experiencias de la aplicación de algoritmos de aprendizaje supervisado para la clasificación de productos de comercio electrónico.	Propuestas, experiencias	Se selecciona estos términos para conocer la experiencia de los investigadores al aplicar el algoritmo

En la Tabla 5 se detalla la construcción de la base de las cadenas de búsquedas, la cual se formó en base a los datos descritos en la estrategia PICO.

Tabla 5: Términos en inglés y conectores lógicos a ser usados en la búsqueda

Concepto	Términos
Población	(algor*) AND (model* or genet* or decis* or regres* or pred* or convolut*) AND (meth* or tech*)
Intervención	(*commerce or "electronic commerce") AND (class*) AND (produc*) AND ("learning" or "deep" or "machine" or "supervised")
Comparación	No aplica
Resultado	("propo*" or "exper*" or "app*") AND (acc* or perform*)

3.3.2 Criterios de inclusión y exclusión

De acuerdo a los lineamientos de la guía de Kitchenham, luego de elaborar la cadena de búsqueda en las librerías digitales, los resultados obtenidos deben ser sometidos a un proceso de evaluación, aplicando criterios de selección, con el fin de identificar los estudios primarios que proveen evidencia directa respecto a las preguntas de investigación. Las Tablas 6 y 7 muestran los criterios de inclusión y exclusión, respectivamente.

Tabla 6: Criterios de Inclusión

ID	CRITERIO
C.I.1	Se consideran todos aquellos artículos provenientes de librerías digitales indexadas. (IEEEXplore , Science Direct, SpringerLink, ACM)
C.I.2	Se considerarán todos los artículos que se encuentren dentro del rango de temporalidad definido. (2014 - 2020)
C.I.3	Los artículos deben provenir del área de Comercio electrónico y Ciencias de la Computación
C.I.4	Se aceptarán artículos provenientes de revistas científicas (Journals) y conferencias (Proceedings)
C.I.5	Se aceptarán artículos que contengan estudios de algoritmos para la clasificación de productos de comercio electrónico.

Tabla 7: Criterios de Exclusión

ID	CRITERIO
C.E.1	Serán rechazados los artículos que se encuentren en un idioma diferente a inglés
C.E.2	Serán excluidos los artículos cuya posición de relevancia es mayor que 100 según la máquina de búsqueda de la librería digital.
C.E.3	Serán excluidos los artículos duplicados.
C.E.4	Serán excluidos los artículos cuyo título no tenga relación con el objeto de estudio
C.E.5	Serán rechazados los artículos de contenido similar, quedándose solo los que tengan el contenido más completo.

Procedimiento para la selección de estudios: Se realizaron cuatro pasos para la selección de los artículos en la Revisión Sistemática de la Literatura, los cuales son formados por distintos criterios de inclusión y exclusión según la Tabla 8:

- Paso 1: En primer lugar, se realizó una búsqueda solamente en las librerías digitales seleccionadas, filtrando los artículos por el criterio de temporalidad, y excluyendo los artículos que no se encuentren en idioma inglés.
- Paso 2: Se realizó un proceso de filtrado según el área del artículo, estos deben pertenecer a Ciencias de la computación y comercio electrónico, además que deben provenir de revistas científicas o conferencias. Se filtran los artículos por relevancia según el motor de búsqueda de la librería digital, siendo escogidos los 100 primeros.
- Paso 3: Se excluyen los artículos duplicados y a su vez, se revisan los artículos por la similitud del título al tema de investigación propuesto.
- Paso 4: Finalmente, se realiza un análisis más profundo a través de la lectura del resumen, donde se considera solo a los artículos similares y que aportan al tema de investigación.

Tabla 8: Aplicación de los criterios de selección

Procedimiento	Criterio de selección
Paso 1	CI.1, CI.2 CE.1
Paso 2	CI.3, CI.4, CE.2
Paso 3	CE.3, CE.4
Paso 4	CI.5, CE.5

3.4 Criterios de calidad

Cumpliendo con los lineamientos de Kitchenham, se procede con la evaluación de calidad de los estudios primarios, lo cual se relaciona a la medida con que el estudio minimiza el sesgo y maximiza la validez interna y externa [34]. Para llevar a cabo dicha evaluación de calidad se construye un esquema compuesta por una lista de criterios; cada criterio está acompañado por unos indicadores cualitativos, los cuales basándose en la escala de Rouhani [40] se transforman en cuantitativos: Si cumple (S) = 1, Cumple parcialmente (P) = 0.5 y No cumple (N) = 0. Los resultados se presentan según el esquema presentado en la tabla 9.

Tabla 9: Esquema de evaluación de calidad

Código	Criterio de evaluación	Escala de evaluación
C1	**¿Se ha documentado adecuadamente la implementación del algoritmo?**	S: El algoritmo propuesto ha sido documentado adecuadamente.
		P: El algoritmo propuesto ha sido documentado parcialmente.
		N: No se ha documentado el algoritmo propuesto.
C2	**¿La elección del algoritmo ha sido justificado claramente?**	S: El estudio justifica claramente la elección del algoritmo
		P: El estudio justifica parcialmente la elección del algoritmo
		N: No se justifica la elección del algoritmo
C3	**¿Han sido descritos los aportes al estudio para los círculos científicos, académicos o para la industria?**	S: Han sido descritos los aportes al estudio claramente.
		P: Han sido descritos los aportes al estudio parcialmente.
		N: No se han mencionado aportes
C4	**¿Los resultados ayudan a responder las preguntas de investigación planteadas?** .	S: Los resultados ayudaron a responder todas las preguntas de investigación.
		P: Los resultados ayudaron a responder algunas preguntas de investigación.
		N: Los resultados no ayudaron a responder las preguntas de investigación.
C5	**¿Se ha documentado las limitaciones del algoritmo de manera clara?**	S: Las limitaciones del algoritmo han sido documentadas totalmente.
		P: Las limitaciones del algoritmo han sido documentadas parcialmente.
		N: No se han documentado las limitaciones del algoritmo.

3.5 Extracción de datos

El objetivo de esta etapa es extraer con precisión, información relevante y relacionada a las preguntas de investigación, la cual se encuentra en los estudios primarios seleccionados. Para realizar la extracción de datos y con el fin de reducir la posibilidad de sesgo se diseña un formulario de extracción, según el modelo presentado en la tabla 10 [34]. Para el diseño del formulario de extracción de datos, se tomaron en cuenta las guías descritas por Kitchenham y Brereton [41].

Tabla 10: Formulario para la extracción de datos

Criterio	Detalle	Relevancia
Identificador		-
Fuente		PB-2
Título		PB-2
Autores		PB-2
Publicación		PB-3
Años de publicación		PB-1
Tipo de publicación		PB-2
Objeto de estudio		PI-2
Tipo de aprendizaje usado		PI-2
Tipo de Algoritmo		PI-1
Algoritmos		PI-1
Características del algoritmo		PI-1
Métricas de validación		PI-3

3.6 Síntesis de Datos

El proceso de síntesis de datos, también conocido como análisis de datos, implica recopilar y resumir los resultados de los estudios primarios a través de una narración usando técnicas cuantitativas, cualitativas, o ambas, [34]. Dicho de otra manera, se inspeccionan los datos extraídos de los estudios primarios en busca de similitudes, para definir cómo se podría encapsular los resultados [42]. En este estudio los, resultados del análisis de los datos se describen en la sección 4.

4 Resultados

En esta sección, siguiendo los lineamientos de la guía de Kitchenham [34], se describen a detalle los resultados de cada paso del protocolo de revisión propuesto en la sección 3, con el propósito final de responder a las preguntas de investigación y bibliométricas planteadas en la misma sección.

4.1 Resultados de la búsqueda

Según el protocolo de revisión definido en la sección 3, el primer paso consiste en la ejecución de la cadena de búsqueda base en las librerías digitales seleccionadas. La Tabla 11 registra la cadena de búsqueda utilizada para la respectiva librería digital, la fecha y número total de bibliografía resultante.

Las cadenas de búsqueda fueron adaptadas de la cadena de búsqueda base según las restricciones o comportamiento de los resultados en las librerías digitales; es decir, algunas librerías presentan restricciones en el uso de comodines y otras mostraban un excesivo número de resultados, por lo que, la cadena de búsqueda se adaptó a las condiciones particulares de cada librería digital. Se realizaron los ajustes de la siguiente manera:

- La base de datos Science Direct no reconoce el comodín * y una abreviación del operador NOT es –
- La base de datos IEE Xplore presenta una restricción del uso de comodines (*?), el cual se restringe a un máximo de 6 comodines en la cadena de búsqueda.
- La base de datos SpringerLink presentó mucho ruido en sus resultados, es decir, mostraba resultados no relevantes, por ello se retiraron algunos comodines, para reducir el número de resultados de la búsqueda.

Con el fin de tener un mayor control de los términos de búsqueda, se utilizó la interfaz de búsqueda avanzada en dos librerías digitales : ACM y Science Direct, mientras que en IEEE Xplores y SpringerLink, solo se usó la sintaxis de búsqueda avanzada, es decir operadores y comodines, pero no la interfaz de búsqueda avanzada de tales librerías digitales, porque la cadena de búsqueda propuesta presentó iguales o mejores resultados que a través de la interfaz de búsqueda avanzada de la respectiva librería digital. Ejecutada la búsqueda y algunos filtros, se exportaron los resultados según el siguiente detalle:

- ACM: En esta librería digital la búsqueda fue realizada en la colección “The ACM Guide to Computing Literature” que presenta un mayor número de artículos que “ACM Full-Text Collection”. La librería cuenta con la opción de exportar todos los resultados por página, con un máximo de 50 resultados en cada una, en formato Bibtex (.bib).
- ScienceDirect: La librería tiene la opción de exportar todos los resultados por cada página, con un máximo de 100 resultados, en formato Bibtex (.bib).
- IEEE Xplore: La librería cuenta con la opción de exportar 100 resultados por página en formato Bibtex (.bib).
- SpringerLink: La librería tiene la opción de exportar solamente a csv y un máximo de 1000 resultados por cada exportación.

Cabe resaltar que el total de resultados que se detalla en la tabla 2 se refiere a los resultados obtenidos de la ejecución de la cadena de búsqueda sin ningún tipo de filtro (fecha, tema, tipo de artículo, idioma), lo cual aplica para todas las librerías digitales descritas en la misma tabla.

Las referencias obtenidas en formato Bibtex (.bib) fueron importadas en Zotero (https://www.zotero.org) y posteriormente exportadas en formato CSV para su lectura en Excel; SpringerLink, fue excento del procesamiento en Zotero, ya que el formato del archivo de los resultados era CSV.

Tabla 11: Cadena de búsqueda por Base de datos

Base de Datos	Fecha	Total
ACM	**Julio 2020**	**293**
[[Abstract: "commerce"] OR [Abstract: "electronic commerce"] OR [Abstract: "ecommerce"]] AND [[Abstract: classif*] OR [Abstract: categor*]] AND [[Abstract: product?] OR [Abstract: item]] AND NOT [[Publication Title: sentiment*] OR [Publication Title: review*] OR [Publication Title: recommend*] OR [Publication Title: predict*]] AND [Publication Date: (01/01/2014 TO 12/31/2020)]		
SCIENCE DIRECT	**Julio 2020**	**48**
Title, abstract, keywords: (e-commerce OR "electronic commerce" OR ecommerce) AND (product OR item OR ecatalog) AND (classification OR categorization) Title: -sentiment –review 2014-2020		
IEEE Xplore	**Julio 2020**	**328**
(*commerce OR "electronic commerce") AND (classif* OR categor*) AND (product OR "online product" OR *catalog) AND NOT (sentiment* OR review OR retriev*) Filters Applied: Conferences Journals electronic commerce learning (artificial intelligence) pattern classification text analysis image classification support vector machines neural nets 2014 – 2020		
SpringerLink	**Julio 2020**	**2242**
'e-commerce AND product AND (classification OR categorization) AND NOT (review* OR sentiment* OR recommend*)' within English Remove this filter 2014 - 2020		

4.2 Selección de estudios primarios

Paso 1: Este paso se realiza en el motor de búsqueda de cada librería digital indexada (ACM, ScienceDirect, IEEE Xplore y SpringerLink). Una vez ejecutada la cadena de búsqueda final, se ejecuta el filtro de tiempo, el cual determina que la fecha de publicación de los artículos debe estar entre el rango de 2014 y 2020. Así también, se descarta cualquier artículo que no esté escrito en inglés.

Paso 2: Este paso se realiza en el motor de búsqueda de cada librería digital. El primer filtro que se ejecuta, es relacionado al tema de investigación, para ello, se filtra solo los artículos relacionados a Ciencias de Computación y comercio electrónico, para este filtro se usan las etiquetas temáticas que disponga cada librería digital, relacionadas a ambas áreas; en el caso de la librería digital ACM, no es posible realizar este filtro, sin embargo, por la naturaleza de la librería, se entiende que todos los artículos están relacionados a computación. El siguiente filtro que se usa es el tipo de artículo, el cual debe provenir de una conferencia (conference paper) o revista (journal article). Por último, ordenamos los resultados por su relevancia, según la librería digital, y exportamos los metadatos de los 100 primeros resultados en formato Bibtex, excepto en SpringerLink, cuyos datos son exportados directamente en CSV.

Paso 3: Una vez que se tienen todos los metadatos en formato Bibtex, son importados en Zotero, y luego exportados en formato CSV, teniendo todas las bibliografías en un solo archivo CSV, se pasa a ordenar alfabéticamente por el título del artículo, con el fin de visualizar artículos duplicados y retirarlos. Luego, se realiza una revisión rápida de los títulos y se eliminan los artículos cuyo título no presenta una relación con el tema de clasificación de productos de comercio electrónico.

Paso 4: Al finalizar el paso 3 se obtuvieron 68 artículos, los cuales son sometidos a una revisión preliminar de su contenido, y basados en los datos presentados en el resumen, se descartan aquellos que no presenten algoritmos para la clasificación de productos de comercio electrónico y entre aquellos que son artículos similares, se mantienen los que demuestran tener un contenido más holístico.

En la Tabla 12 se muestra los resultados del proceso de selección en cada paso, y en el apéndice A se listan los artículos que fueron seleccionados como estudios primarios.

Tabla 12: Resultados del proceso de selección de estudios primarios

Base de Datos	Artículos descubiertos	Paso 1	Paso 2	Paso 3	Paso 4
IEEE Xplore	328	154	122	25	**10**
ACM	293	95	52	17	**7**
Science Direct	48	22	8	4	**2**
SringerLink	2242	931	568	22	**4**
Total	**2911**	**1202**	**750**	**68**	**23**

4.3 Evaluación de calidad

Como resultante del proceso de selección de estudios primarios, se tuvo una lista de 23 artículos, de los cuales 11 pertenecen a IEEE Xplore, siendo la librería digital que proveyó la mayor cantidad de artículos relacionados al tema de la presente revisión. Sobre esta lista de 23 artículos, se aplicaron los criterios de evaluación de calidad establecidos en el protocolo de revisión descritos en la sección 3. Los resultados de la evaluación se muestran en la Tabla 13; a partir de los cuales se observa que el 43.5%, menos de la mitad, muestra un puntaje entre 3.5 y 4, mientras que el 56.5% ostenta un puntaje entre 4 y 5, lo que se puede considerar como un buen indicador de la calidad de los estudios primarios seleccionados para la presente RSL.

Tabla 13: Evaluación de la calidad de estudios

ID	C1	C2	C3	C4	C5	Total
1	0.5	1	0.5	1	1	**4**
2	1	1	1	1	0.5	**4.5**
3	1	1	1	1	0.5	**4.5**

4	1	1	1	1	0.5	**4.5**
5	1	1	1	1	1	**5**
6	1	1	1	1	1	**5**
7	1	1	0.5	1	0.5	**4**
8	1	1	1	1	1	**5**
9	1	1	0.5	1	0.5	**4**
10	1	1	1	1	1	**5**
11	1	1	1	0.5	0.5	**4**
12	1	1	1	1	0.5	**4.5**
13	1	1	0.5	1	0.5	**4**
14	1	1	0.5	1	0.5	**4**
15	1	1	1	1	1	**5**
16	1	1	1	1	0.5	**4.5**
17	1	1	1	1	0;5	**4**
18	1	1	1	1	0.5	**4.5**
19	1	1	1	1	0.5	**4.5**
20	0.5	1	0.5	1	0.5	**3.5**
21	0.5	1	0.5	1	0.5	**3.5**
22	1	1	0.5	1	0.5	**4**
23	0.5	1	1	1	1	**4.5**

4.4 Extracción de información relevante

Según Kitchenham, en su guía para la implementación de la RSL en el ámbito de Ingeniería de Software, menciona que los formularios para la extracción de datos deben ser diseñados con el objetivo de recolectar toda la información necesaria para resolver las preguntas de investigación, por tal razón, se diseñó el formulario descrito en la sección 3. En la Tabla 14 se observa una muestra unitaria de los formularios utilizados; estos fueron completados simultáneamente al proceso de lectura y análisis del estudio primario respectivo, cuyos datos, se registraron en el mismo idioma del artículo. En los campos donde no se halló información relevante, fueron llenados con las siglas NI (No se encontró información).

Tabla 14: Ejemplo de extracción de datos de un estudio primario

Criterio	**Detalle**	**Relevancia**
Identificador	8	-
Fuente	IEEE Xplore	PB-2

Título	A Deep Forest Method for Classifying E-Commerce Products by Using Title Information	PB-2
Autores	Dai, J.; Wang, T.; Wang, S.	PB-2
Publicación	International Conference on Computing, Networking and Communications (ICNC)	PB-3
Años de publicación	2020	PB-1
Tipo de publicación	Conference Paper	PB-2
Objeto de estudio	Product title	PI-2
Tipo de aprendizaje usado	Supervised Learning	PI-2
Tipo de Algoritmo	Ensemble	PI-1
Algoritmos	The experiment results show that the classification accuracy using **gcForest** (Deep Network) is 92.38%, which outperforms SVM with RBF kernel (86.88%), SVM with linear kernel (89.73%) and CNN (86.86%).	PI-1
Características del algoritmo	then utilize gcForest to train the classification model, which has 2 completely random forests and 2 random forests, each owning 500 decision trees.	PI-1
Métricas de validación	Accuracy - 92,38%	PI-3

4.5 Análisis bibliométrico

En esta sección se procede a describir el análisis bibliométrico, al cual se somete la presente RSL, basado en las preguntas bibliométricas descritas en la sección 3, que están relacionadas a algunos factores como tiempo, tipo de artículo y tema.

4.5.1 Pregunta bibliométrica 1

¿Cuántos artículos referentes a la clasificación de productos de comercio electrónico se publicaron desde el 2014?

A través del proceso de recolección de recursos bibliográficos en 4 librerías indexadas, los estudios primarios seleccionados fueron 23, los cuales se distribuyen en el tiempo según la Figura 1. A partir de esta figura podemos afirmar que, en los últimos 5 años, el interés por la creación de algoritmos de aprendizaje automático para la clasificación de productos de comercio electrónico se ha incrementado, alcanzando la cima en el año 2018, con 7 artículos publicados y la misma cantidad se mantuvo para el año 2019. En el año 2020, hasta el mes de julio se cuenta con 2 estudios primarios referentes al tema de la presente RSL

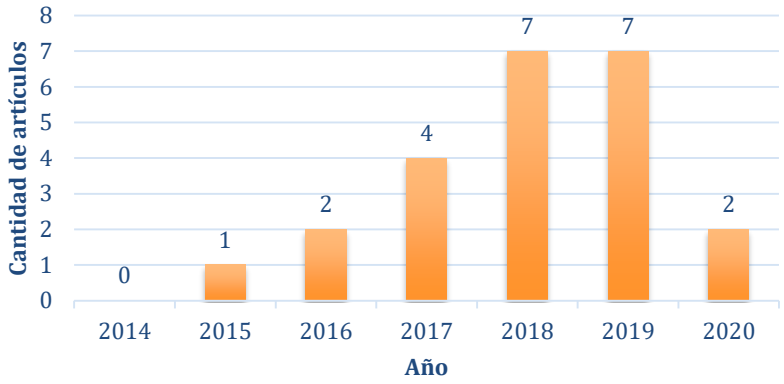

Figura 1 : Frecuencia de artículos por año

4.5.2 Pregunta bibliométrica 2

¿Cuál es la cantidad de publicaciones por tipo de artículo?

A continuación, se presenta un análisis de los 23 estudios primarios seleccionados según el tipo de artículo y la librería digital. En la figura 2 se observa que, de los 23 estudios primarios, el 22% son artículos de revistas, es decir fueron publicados en revistas, mientras que, el 78% fueron presentados en conferencias o workshops, y posteriormente publicados en el proceeding respectivo del evento.

Asimismo, en la figura 3, se observa que, los artículos de tipo "Conference Paper" provienen solamente de 3 librerías: ACM, IEEE Xplore y SpringerLink, dentro de las que la segunda librería mencionada presenta 10 artículos, siendo la cantidad máxima. De la misma manera, los artículos de tipo "Journal Article" provienen solamente de 2 librerías: ScienceDirect y ACM. Así también, destacamos la librería ACM por ser la única que presenta estudios primarios en ambos tipos.

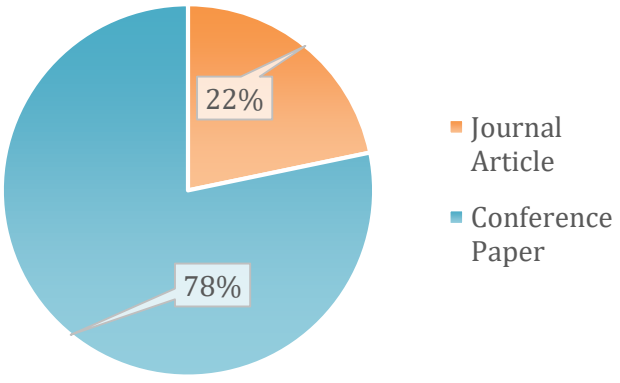

Figura 2 : Cantidad de publicaciones por tipo de artículo

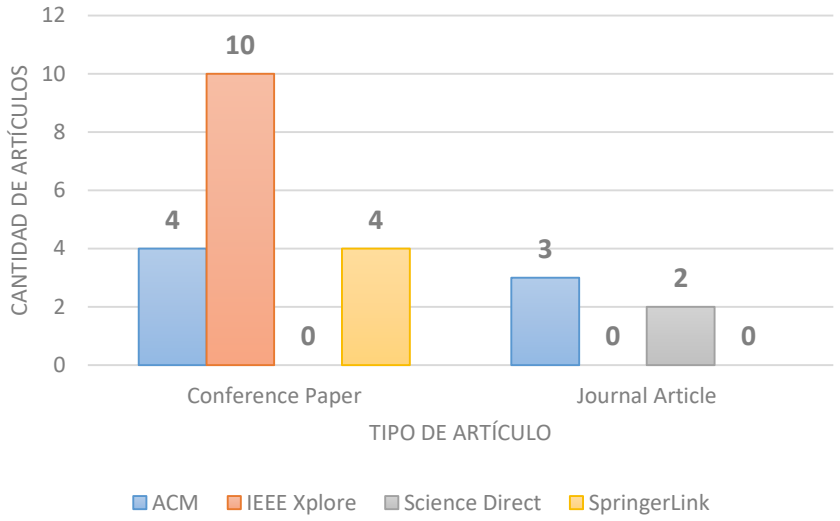

Figura 3: Cantidad de artículos por tipo y librería digital

4.5.3 Pregunta bibliométrica 3

¿Cuáles son las publicaciones en las que se han encontrado estudios relacionados a la clasificación de productos de comercio electrónico?

Cuando la pregunta bibliométrica se refiere a publicación, hacemos referencia a la revista científica o proceedings donde el artículo fue publicado. Como se observa en la Tabla 15, todas las publicaciones descritas tienen al menos un artículo, no obstante, se resalta la "ACM SIGKDD International Conference on Knowledge Discovery and Data Mining", ya que, es el único evento, aunque en diferentes tiempos, donde se presentaron 2 artículos relacionados al tema de investigación de la presente RSL.

Tabla 15: Lista de publicaciones

Publicación	Cantidad
2015 3rd International Conference on Artificial Intelligence, Modelling and Simulation (AIMS)	1
2016 IEEE International Conference on Big Data (Big Data)	1
2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL)	1
2017 IEEE International Conference on Information Reuse and Integration (IRI)	1
2017 Workshop of Computer Vision (WVC)	1
2018 4th International Conference on Electrical Engineering and Information Communication Technology (iCEEiCT)	1
2018 IEEE 30th International Conference on Tools with Artificial Intelligence (ICTAI)	1
2019 IEEE 23rd International Conference on Computer Supported Cooperative Work in Design (CSCWD)	1

2019 IEEE Conference on Big Data and Analytics (ICBDA)	1
2020 International Conference on Computing, Networking and Communications (ICNC)	1
ACM Trans. Manage. Inf. Syst.	1
ACM Transactions on Management Information Systems	1
Advances in Neural Networks - ISNN 2017	1
Big Data Analytics	1
Egyptian Informatics Journal	1
Electronic Commerce Research	1
Expert Systems with Applications	1
Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval	1
Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining	**1**
Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining	**1**
Proceedings of the 28th ACM International Conference on Information and Knowledge Management	1
Recent Advances in Information and Communication Technology 2018	1
Smart Computing and Informatics	1

4.6 Síntesis de datos

4.6.1 Pregunta de investigación 1

¿Qué algoritmos existen para la clasificación de productos de comercio electrónico?

Apoyándonos en los datos extraídos de cada artículo, los cuales fueron registrados según el formato del formulario de extracción de datos, se obtuvo una lista de algoritmos, los cuales fueron agrupados según el tipo de objeto de estudio del artículo, es decir, imagen o texto; a su vez, se evitó la duplicidad de los algoritmos mencionando el ID del artículo que lo implementa. Estos detalles se expresan en la Tabla 16 y 17.

A partir de los datos mostrados en la Tabla 16, se observa que cuando se trabaja con la información textual del producto de comercio electrónico, es decir, tanto el título y/o descripción del producto, la mayoría de los investigadores han optado principalmente por el uso de 3 algoritmos: K-means, Naive Bayes y Support Vector Machine, cada uno fue implementado en 2 artículos seleccionados. Es apropiado mencionar, que el algoritmo K-means, se usa para un aprendizaje no supervisado; además, el artículo con id (22) es particular, ya que el autor implementa el algoritmo Support Vector Machine (SVM) teniendo como datos tanto imagen como texto.

Tabla 16: Lista de algoritmos para texto

Algoritmos	Cantidad	Artículo ID
K-means	2	(10), (14)

Naive Bayes	2	(2), (20)
Support Vector Machine	2	(1), **(22)**
Artificial Neural network	1	(17)
AssocER - Associative Classifier for Entity Resolution	1	(13)
CNN & Bidirectional LSTM (Ensemble)	1	(1)
Convolutional Neural Networks	1	(16)
DeepCN: Multiple recurrent neural networks	1	(6)
GBTs : Gradient Boosted Trees	1	(16)
gcForest (Ensemble)	1	(8)
Hybrid Classifier : K-means - RepTree	1	(21)
Kim-CNN & Zhang-CNN (Ensemble)	1	(1)
K-nearest Neighbor	1	(23)
Logistic Regression	1	(4)
NPC : Neural Product Categorization	1	(5)
Ofsix LSTMs (Ensemble)	1	(1)
RNN & FFN (Ensemble)	1	(3)
SPC : Supervised Product classifier	1	(15)

A continuación, según la Tabla 17, se presenta una lista de 5 algoritmos usados cuando el tipo de objeto de estudio es la imagen del producto de comercio electrónico. En esta lista tenemos dos algoritmos que demuestran ser implementados en 2 artículos, los algoritmos son: Hierarchical Convolutional Neural Networks y Support Vector Machine. Cabe mencionar, que el artículo con id (22) implementa el algoritmo de SVM, sin embargo, el autor trabaja con datos textuales

Tabla 17: Lista de algoritmos para imágenes

Algoritmos	Cantidad	Artículo id
Hierarchical Convolutional Neural Networks	2	(11), (18)
Support Vector Machine	2	(12), **(22)**
Artificial Neural Network (Ensemble)	1	(19)
Deep neural network : ResNet50	1	(9)
MSURU : CNN based on ResNeXt	1	(7)

4.6.2 Pregunta de investigación 2

¿Cuáles son los enfoques abordados para la clasificación de productos de comercio electrónico?

A través de la figura 4, se observa que, de los 23 estudios primarios seleccionados, el 89% usó algoritmos de aprendizaje supervisado, esto quiere decir que, el data set usado por el autor tenía una clase definida, ya sea manualmente o por otro algoritmo. Además, un 7% de los estudios primarios usó algoritmos de aprendizaje no supervisado, lo cual significa que el algoritmo usado se encargó de agrupar automáticamente los productos de comercio electrónico según sus características, creando grupos y subgrupos de productos.

Asimismo, el 4% de los estudios primarios usó un algoritmo híbrido, combinando el aprendizaje supervisado y no supervisado, en este caso particular, no podemos afirmar que es un algoritmo semi-supervisado, pues las características del dataset no cumplen con estas características, al contrario, se tiene un dataset sin clases definidas, se aplica el algoritmo de clustering y dichos resultados son usados para clasificar nuevos productos, usando un algoritmo de aprendizaje supervisado, es por ello, que lo denominamos como híbrido [32].

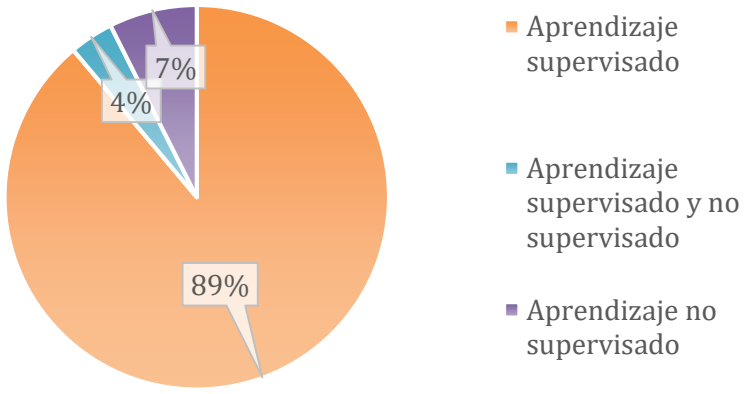

Figura 4: Enfoques según el tipo de aprendizaje

Por otro lado, los algoritmos usados por los estudios primarios, pueden ser divididos por el tipo de objeto de estudio, así, según se observa en la figura 5, el 74% de estos estudios realizaron la clasificación de productos de comercio electrónico usando la descripción textual del producto, mientras que el 22% de los estudios realizaron la clasificación usando la imagen del producto, y por último, el 4% de dichos artículos seleccionados, usaron tanto el texto como la imagen del producto para realizar una clasificación automática.

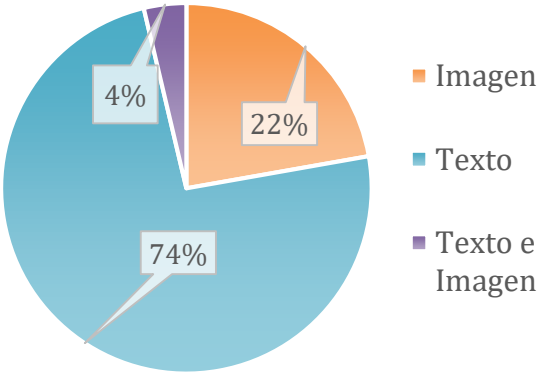

Figura 5: Enfoques según el objeto de estudio

Finalmente, se puede concluir que existen 2 tipos de enfoques para abordar el problema de clasificación de productos de comercio electrónico, ya sea según el conjunto de datos que se tiene, o el tipo de objeto de estudio escogido. Siendo que, el enfoque más utilizado es usar un algoritmo de aprendizaje supervisado para la clasificación de productos usando su información textual.

4.6.3 Pregunta de investigación 3

¿Cuáles son las métricas usadas para la validación de los algoritmos?

Las métricas de evaluación de los algoritmos, permite conocer la eficacia de los mismos, con el fin de tener la certeza de que proporciona confianza en los resultados producidos por el algoritmo. No obstante, ante la variedad de métricas y sus distintos usos según el tipo del algoritmo, esta pregunta busca reconocer las principales métricas de evaluación para algoritmos usados en el problema de clasificación de productos de comercio electrónico.

Según el detalle de métricas que presenta la Tabla 18 y 19, para el aprendizaje supervisado y no supervisado, respectivamente, se observa que, dentro de las métricas para algoritmos de aprendizaje supervisado, la más común es el nivel de Accuracy, y le sigue la versión con pesos de la precisión, recall y F1-score; la primera métrica, es usada en 9 estudios primarios, mientras que el conjunto de métricas mencionadas después son usadas, todas ellas en dos estudios primarios; por lo cual, al consider estas 4 métricas para evaluar el rendimiento de un algoritmo, es satisfactorio.

Tabla 18: Métricas de algoritmos de aprendizaje supervisado

Métricas	Cantidad	Artículo id
Accuracy	9	(2), (6), (8), (9), (11), (12), (15), (19), (23)
Weighted versions: precision, recall y F1 score.	2	(1), (3)
Accuracy, AUC	1	(17)
Accuracy, micro-AUROC and macro-AUROC	1	(22)
Accuracy, Precision, Recall, and F-measure.	1	(5)
AUROC : Area Under Receiver Operating Characteristic	1	(20)
F-measure	1	(4)
Loss and accuracy	1	(18)
Mean average precision (mAP) and ranking average precision (rankingAP)	1	(7)
Micro precision, F1	1	(16)
Micro-F1 and macro-F1	1	(13)

Por otro lado, las métricas usadas para algoritmos de aprendizaje no supervisado, según la RSL realizada sobre los 23 artículos, se detallan en la Tabla 19, donde las 3 métricas presentadas tienen un mismo peso, pues cada una es implementada en un solo estudio primario. No obstante, la segunda métrica mencionada, es interesante, pues en este estudio [43] se espera la validación de los usuarios para medir el rendimiento del algoritmo implementado, ya que así esperan medir el rendimiento de la clasificación del k-means, y saber si las nuevas categorías satisfacen a los usuarios.

Tabla 19: Métricas de algoritmos de aprendizaje no supervisado e híbrido

Métricas	Cantidad	Artículo id
Mean absolute error, root-mean-square error.	1	(21)
Satisfaction level of consumers	1	(14)
Silhouette score	1	(10)

4.7 Amenazas de la validez

En esta sección se discute las amenazas a la validez de la presente Revisión Sistemática de la Literatura. La selección de los estudios primarios, a pesar de estar sujetos a criterios de inclusión y exclusión definidos con claridad, puede haber sido afectada negativamente por la poca información presentada en los resúmenes de los artículos y en algunos casos, por la falta de claridad. Además, se provee toda la información necesaria para replicar el presente estudio. Todas las bases de datos y artículos se encuentran propiamente referenciados.

5 Conclusiones y trabajo futuro

En el presente estudio se presentan los resultados de una revisión sistemática sobre los algoritmos de aprendizaje automático para la clasificación de productos de comercio electrónico, a través del cual, se obtuvo un total de 23 estudios primarios, los cuales fueron sometidos a un proceso de selección y una evaluación de calidad, estos fueron extraídos de 4 base de datos indexadas y con alto impacto académico y científico. La revisión sistemática realizada en estos 23 artículos, ayudaron a responder las preguntas de investigación planteadas en la sección 3 y además las preguntas bibliométricas.

Como parte del análisis bibliométrico, se concluye que existe un reciente interés en el campo académico para resolver el problema de clasificación de productos de comercio electrónico, lo cual se fundamente en el número creciente de publicaciones durante los últimos 5 años, demostrando un interés continuo en los últimos 2 años ya que se presenta la misma cantidad de artículos publicados en ambos años. Asimismo, se comprobó una mayor cantidad de artículos publicados de tipo “Conference Paper” que artículos de tipo “Journal Article”. Además, no se ha detectado una publicación preferencial para el ámbito del problema presentado en este artículo, sin embargo, la ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, fue el evento donde se presentaron dos artículos de la presente revisión, aunque en diferentes tiempos.

Por otro lado, al finalizar la revisión sistemática, se consigue responder las 3 preguntas de investigación planteadas en el protocolo de revisión, así se concluye que, cuando se trabaja con la parte textual del producto se debe tomar en principal consideración los siguientes algoritmos: K-means, Naive Bayes y Support Vector Machine, cada uno se puede adoptar según las características particulares del problema. Asimismo, en caso de trabajar con las imágenes del producto, según la RSL se propone tomar en alta consideración los siguientes algoritmos: Hierarchical Convolutional Neural Networks y Support Vector Machine.

Del mismo modo, se concluye que existen 2 enfoques para abordar el problema de clasificación de productos de comercio electrónico, ya sea según el conjunto de datos que se tiene, o el tipo de objeto de estudio escogido. Siendo que, el enfoque más utilizado es usar un algoritmo de aprendizaje supervisado para la clasificación de productos usando su información textual.

Considerando la importancia de las métricas de evaluación de los algoritmos, se concluye que, para algoritmos de aprendizaje supervisado, según la presente revisión sistemática, se deben considerar el nivel de accuracy, y la versión con los pesos de precisión, recall y F1-score. De la misma manera, para algoritmos de aprendizaje no supervisado, las métricas usadas son el error absoluto medio, el error cuadrático medio y el silhouette score.

Este estudio busca resaltar la relevancia de conocer el estado de arte con respecto a la clasificación de productos en comercio electrónico, el cual facilita la elección de un algoritmo según el conjunto de datos obtenido por el investigador para su desarrollo y posterior implementación en el área industrial.

Para futuras investigaciones se propone realizar un estudio comparativo de los principales algoritmos que conforman el estado de arte actual, considerando indicadores de eficiencia y eficacia. Además, se plantea que, para el pre procesamiento de los datos, se construya un algoritmo que pueda ser reproducible en distintos idiomas.

Referencias

[1] T. J. Berners-Lee and R. Cailliau, "WorldWideWeb: Proposal for a HyperText project," 1990.

[2] eMarketer, "Global Ecommerce 2019," 2019. https://www.emarketer.com/content/global-ecommerce-2019 (accessed Jul. 19, 2020).

[3] Statista, "COVID-19 impact retail e-commerce site traffic 2020 | Statista," *Statista - The Statistics Portal*, 2020. https://www.statista.com/statistics/1112595/covid-19-impact-retail-e-commerce-site-traffic-global/ (accessed Oct. 21, 2020).

[4] Adobe, "2020 Digital Economy Index," *Adobe Analytics*, 2020. https://www.adobe.com/content/dam/www/us/en/experience-cloud/digital-insights/pdfs/adobe_analytics-digital-economy-index-2020.pdf (accessed Oct. 21, 2020).

[5] Statista, "China: coronavirus impact on future online purchase habits 2020 | Statista," *Statista - The Statistics Portal*, 2020. https://www.statista.com/statistics/1127034/china-coronavirus-impact-on-future-online-purchase-habits/ (accessed Oct. 21, 2020).

[6] Statista, "COVID-19: change in online retail orders LatAm | Statista," *Statista - The Statistics Portal*, 2020. https://www.statista.com/statistics/1142436/online-retail-orders-weekly-growth-latin-america-country/ (accessed Oct. 21, 2020).

[7] R. K. Khanuja, "Optimizing E-Commerce Product Classification Using Transfer Learning," 2019.

[8] J. Dai, T. Wang, and S. Wang, "A Deep Forest Method for Classifying E-Commerce Products by Using Title Information," in *2020 International Conference on Computing, Networking and Communications (ICNC)*, Feb. 2020, pp. 1–5, doi: 10.1109/ICNC47757.2020.9049751.

[9] D. Gao, W. Yang, H. Zhou, Y. Wei, Y. Hu, and H.-M. Wang, "Deep Hierarchical Classification for Category Prediction in E-commerce System," *ArXiv*, vol. abs/2005.0, 2020.

[10] "United Nations Standard Products and Services Code® (UNSPSC®)." https://www.unspsc.org/ (accessed Jul. 19, 2020).

[11] E. Schulten *et al.*, "The e-commerce product classification challenge," *IEEE Intell. Syst.*, vol. 16, no. 4, pp. 86–89, 2001.

[12] A. Cevahir and K. Murakami, "Large-scale Multi-class and Hierarchical Product Categorization for an E-commerce Giant," in *Proceedings of {COLING} 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, Dec. 2016, pp. 525–535, [Online]. Available: https://www.aclweb.org/anthology/C16-1051.

[13] D. Shen, J.-D. Ruvini, and B. Sarwar, "Large-Scale Item Categorization for e-Commerce," in *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, 2012, pp. 595–604, doi: 10.1145/2396761.2396838.

[14] D. Vandic, F. Frasincar, and U. Kaymak, "A Framework for Product Description Classification in E-Commerce," *J. Web Eng.*, vol. 17, no. 1–2, pp. 1–27, Mar. 2018.

[15] H. Chen, J. Zhao, and D. Yin, "Fine-Grained Product Categorization in E-Commerce," in *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, 2019, pp. 2349–2352, doi: 10.1145/3357384.3358170.

[16] A. Krishnan and A. Amarthaluri, "Large Scale Product Categorization using Structured and Unstructured Attributes," *CoRR*, vol. abs/1903.0, 2019, [Online]. Available: http://arxiv.org/abs/1903.04254.

[17] M. Li, L. Chen, T. Liu, and Y. Sun, "Short Text based Cooperative Classification for Multiple Platforms," in *2019 IEEE 23rd International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, May 2019, pp. 87–92, doi:

10.1109/CSCWD.2019.8791500.

[18] E. P. S. Castro, S. Chakravarty, E. Williamson, D. A. Pereira, and E. A. Fox, “Classifying Short Unstructured Data Using the Apache Spark Platform,” in *2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, Jun. 2017, pp. 1–10, doi: 10.1109/JCDL.2017.7991567.

[19] Z. Qin and Z. Qin, *Introduction to E-commerce*, vol. 2009. Springer, 2009.

[20] E. Turban, J. Whiteside, D. King, and J. Outland, *Introduction to electronic commerce and social commerce*. Springer, 2017.

[21] P. Wirojwatanakul and A. Wangperawong, “Multi-Label Product Categorization Using Multi-Modal Fusion Models,” *ArXiv*, vol. abs/1907.0, 2019.

[22] C. Chavaltada, K. Pasupa, and D. R. Hardoon, “A Comparative Study of Machine Learning Techniques for Automatic Product Categorisation,” in *Advances in Neural Networks - ISNN 2017*, 2017, pp. 10–17, doi: https://doi.org/10.1007/978-3-319-59072-1_2.

[23] V. Gupta, H. Karnick, A. Bansal, and P. Jhala, “Product Classification in E-Commerce using Distributional Semantics,” in *Proceedings of {COLING} 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, Dec. 2016, pp. 536–546, [Online]. Available: https://www.aclweb.org/anthology/C16-1052.

[24] D. Shen, J.-D. Ruvini, R. Mukherjee, and N. Sundaresan, “A Study of Smoothing Algorithms for Item Categorization on E-Commerce Sites,” in *Proceedings of the 2010 Ninth International Conference on Machine Learning and Applications*, 2010, pp. 23–28, doi: 10.1109/ICMLA.2010.11.

[25] L. Akritidis, A. Fevgas, and P. Bozanis, “Effective Products Categorization with Importance Scores and Morphological Analysis of the Titles,” in *2018 IEEE 30th International Conference on Tools with Artificial Intelligence (ICTAI)*, Nov. 2018, pp. 213–220, doi: 10.1109/ICTAI.2018.00041.

[26] M. G. Vieira and J. Moreira, “Classification of E-Commerce-Related Images Using Hierarchical Classification with Deep Neural Networks,” in *2017 Workshop of Computer Vision (WVC)*, Oct. 2017, pp. 114–119, doi: 10.1109/WVC.2017.00027.

[27] S. Ray, “A Quick Review of Machine Learning Algorithms,” in *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, 2019, pp. 35–39.

[28] Y. Seo and K. Shin, “Hierarchical convolutional neural networks for fashion image classification,” *Expert Syst. Appl.*, vol. 116, pp. 328–339, 2019, doi: https://doi.org/10.1016/j.eswa.2018.09.022.

[29] C. Chavaltada, K. Pasupa, and D. R. Hardoon, “Combining Multiple Features for Product Categorisation by Multiple Kernel Learning,” in *Recent Advances in Information and Communication Technology 2018*, 2019, pp. 3–12, doi: https://doi.org/10.1007/978-3-319-93692-5_1.

[30] N. M. N. Mathivanan, N. A. M. Ghani, and R. M. Janor, “Analysis of K-Means Clustering Algorithm: A Case Study Using Large Scale E-Commerce Products,” in *2019 IEEE Conference on Big Data and Analytics (ICBDA)*, Nov. 2019, pp. 1–4, doi: 10.1109/ICBDA47563.2019.8987140.

[31] S. Vieira, W. H. Lopez Pinaya, and A. Mechelli, “Chapter 1 - Introduction to machine learning,” in *Machine Learning*, A. Mechelli and S. Vieira, Eds. Academic Press, 2020, pp. 1–20.

[32] N. Midha and V. Singh, “Classification of E-commerce Products Using RepTree and K-means Hybrid Approach,” in *Big Data Analytics*, 2018, pp. 265–273, doi: https://doi.org/10.1007/978-981-10-6620-7_26.

[33] S. A. Oyewole and O. O. Olugbara, “Product image classification using Eigen Colour feature with ensemble machine learning,” *Egypt. Informatics J.*, vol. 19, no. 2, pp. 83–100, 2018, doi: https://doi.org/10.1016/j.eij.2017.10.002.

[34] B. Kitchenham and S. Charters, “Guidelines for performing systematic literature reviews in software engineering,” 2007.

[35] J. Biolchini, P. G. Mian, A. C. C. Natali, and G. H. Travassos, “Systematic review in software engineering,” *Syst. Eng. Comput. Sci. Dep. COPPE/UFRJ, Tech. Rep. ES*, vol. 679, no. 05, p. 45, 2005.

[36] S. S. Khanal, P. W. C. Prasad, A. Alsadoon, and A. Maag, “A systematic review: machine learning based recommendation systems for e-learning,” *Educ. Inf. Technol.*, vol. 25, no. 4, pp. 2635–2664, Jul. 2020, doi: 10.1007/s10639-019-10063-9.

[37] A. Hellas *et al.*, “Predicting Academic Performance: A Systematic Literature Review,” in

Proceedings Companion of the 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education, 2018, pp. 175–199, doi: 10.1145/3293881.3295783.

[38] I. Portugal, P. Alencar, and D. Cowan, "The use of machine learning algorithms in recommender systems: A systematic review," *Expert Syst. Appl.*, vol. 97, pp. 205–227, May 2018, doi: 10.1016/j.eswa.2017.12.020.

[39] C. M. da C. Santos, C. A. de M. Pimenta, and M. R. C. Nobre, "A estratégia PICO para a construção da pergunta de pesquisa e busca de evidências," *Rev. Lat. Am. Enfermagem*, vol. 15, no. 3, pp. 508–511, 2007.

[40] B. D. Rouhani, M. N. Mahrin, F. Nikpay, R. B. Ahmad, and P. Nikfard, "A systematic literature review on Enterprise Architecture Implementation Methodologies," *Inf. Softw. Technol.*, vol. 62, pp. 1–20, 2015, doi: https://doi.org/10.1016/j.infsof.2015.01.012.

[41] B. Kitchenham, O. P. Brereton, D. Budgen, M. Turner, J. Bailey, and S. Linkman, "Systematic literature reviews in software engineering--a systematic literature review," *Inf. Softw. Technol.*, vol. 51, no. 1, pp. 7–15, 2009.

[42] H. P. Breivold, I. Crnkovic, and M. Larsson, "A systematic review of software architecture evolution research," *Inf. Softw. Technol.*, vol. 54, no. 1, pp. 16–40, 2012, doi: https://doi.org/10.1016/j.infsof.2011.06.002.

[43] Y. Hsieh, S. Wu, L. Chen, and P. Yang, "Constructing Hierarchical Product Categories for E-Commerce by Word Embedding and Clustering," in *2017 IEEE International Conference on Information Reuse and Integration (IRI)*, Aug. 2017, pp. 397–402, doi: 10.1109/IRI.2017.81.

APÉNDICE

A. Artículos relacionados

Id	Biblioteca	Título	Autor	Año	Tipo de documento
1	ACM	A Dataset and Baselines for E-Commerce Product Categorization	Lin, Yiu-Chang; Das, Pradipto; Trotman, Andrew; Kallumadi, Surya	2019	Conference Paper
2	ACM	A Framework for Product Description Classification in E-Commerce	Vandic, Damir; Frasincar, Flavius; Kaymak, Uzay	2018	Article
3	ACM	E-Commerce Product Categorization via MachineTranslation	Tan, Liling; Li, Maggie Yundi; Kok, Stanley	2020	Article
4	ACM	Engineering Doc2vec for Automatic Classification of Product Descriptions on O2O Applications	Lee, Hana; Yoon, Young	2018	Article
5	ACM	Fine-Grained Product Categorization in E-Commerce	Chen, Hongshen; Zhao, Jiashu; Yin, Dawei	2019	Conference Paper
6	ACM	Large-Scale Item Categorization in e-Commerce Using	Ha, Jung-Woo; Pyo, Hyuna; Kim, Jeonghee	2016	Conference Paper

		Multiple Recurrent Neural Networks			
7	ACM	MSURU: Large Scale E-Commerce Image Classification with Weakly Supervised Search Data	Tang, Yina; Borisyuk, Fedor; Malreddy, Siddarth; Li, Yixuan; Liu, Yiqun; Kirshner, Sergey	2019	Conference Paper
8	IEEE Xplore	A Deep Forest Method for Classifying E-Commerce Products by Using Title Information	Dai, J.; Wang, T.; Wang, S.	2020	Conference Paper
9	IEEE Xplore	A Novel Idea of Classification of E-commerce Products Using Deep Convolutional Neural Network	Islam, C. S.; Alauddin, M.	2018	Conference Paper
10	IEEE Xplore	Analysis of K-Means Clustering Algorithm: A Case Study Using Large Scale E-Commerce Products	Mathivanan, N. M. N.; Ghani, N. A. M.; Janor, R. M.	2019	Conference Paper
11	IEEE Xplore	Classification of E-Commerce-Related Images Using Hierarchical Classification with Deep Neural Networks	G. Vieira, M.; Moreira, J.	2017	Conference Paper
12	IEEE Xplore	Classification of Product Images in Different Color Models with Customized Kernel for Support Vector Machine	Oyewole, S. A.; Olugbara, O. O.; Adetiba, E.; Nepal, T.	2015	Conference Paper
13	IEEE Xplore	Classifying Short Unstructured Data Using the Apache Spark Platform	Castro, E. P. S.; Chakravarty, S.; Williamson, E.; Pereira, D. A.; Fox, E. A.	2017	Conference Paper
14	IEEE Xplore	Constructing Hierarchical Product Categories for E-Commerce by Word Embedding and Clustering	Hsieh, Y.; Wu, S.; Chen, L.; Yang, P.	2017	Conference Paper

15	IEEE Xplore	Effective Products Categorization with Importance Scores and Morphological Analysis of the Titles	Akritidis, L.; Fevgas, A.; Bozanis, P.	2018	Conference Paper
16	IEEE Xplore	Large-scale taxonomy categorization for noisy product listings	Das, P.; Xia, Y.; Levine, A.; Di Fabbrizio, G.; Datta, A.	2016	Conference Paper
17	IEEE Xplore	Short Text based Cooperative Classification for Multiple Platforms	Li, M.; Chen, L.; Liu, T.; Sun, Y.	2019	Conference Paper
18	Science Direct	Hierarchical convolutional neural networks for fashion image classification	Seo, Yian; Shin, Kyung-shik	2019	Article
19	Science Direct	Product image classification using Eigen Colour feature with ensemble machine learning	Oyewole, S. A.; Olugbara, O. O.	2018	Article
20	SpringerLink	A Comparative Study of Machine Learning Techniques for Automatic Product Categorisation	Chavaltada, Chanawee; Pasupa, Kitsuchart; Hardoon, David R.	2017	Conference Paper
21	SpringerLink	Classification of E-commerce Products Using RepTree and K-means Hybrid Approach	Midha, Neha; Singh, Vikram	2018	Conference Paper
22	SpringerLink	Combining Multiple Features for Product Categorisation by Multiple Kernel Learning	Chavaltada, Chanawee; Pasupa, Kitsuchart; Hardoon, David R.	2019	Conference Paper
23	SpringerLink	Evaluating the Progressive Performance of Machine Learning Techniques on E-commerce Data	Cheekati, Bindu Madhuri; Padala, Sai Varun	2018	Conference Paper