Revista Científica Zambos
ISSN: 3028-8843
Vol. 4 - Núm. 2 / Mayo – Agosto 2025

Research Article

Optimización de Criterios de Búsqueda avanzada para Nuevas Tendencias en la Académica mediante Machine Learning

Optimization of Advanced Search Criteria for New Trends in Academia through Machine Learning

Sangacha-Tapia, Lady [1]
https://orcid.org/0000-0002-5169-8918
lmst0004@red.ujaen.es
España, Jaén, Universidad de Jaén

González-Cañizalez, Yomar [2]
https://orcid.org/0000-0002-6348-866X
yomar.gonzalezc@ug.edu.ec
Ecuador, Guayaquil, Universidad de Guayaquil

Rivas-Herrera, John [3]
https://orcid.org/0009-0001-8640-0106
john.rivas.est@tecazuay.edu.ec
Ecuador, Cuenca, Instituto Superior Tecnológico del Azuay

Autor de correspondencia [1]

DOI / URL: https://doi.org/10.69484/rcz/v4/n2/114

Resumen: La creciente disponibilidad de grandes volúmenes de datos ha generado nuevos activos para diversas industrias. Esto plantea un desafío clave para la educación superior: medir, definir y estructurar dichos activos mediante la identificación de líneas de investigación relevantes. Estas líneas deben servir de guía en la formación de nuevos profesionales, atendiendo a la creciente demanda de competencias tecnológicas dentro de la Industria 4.0. El presente estudio tiene como objetivo establecer criterios sólidos que permitan identificar áreas de investigación aplicables a este contexto. Para lograrlo, se ha empleado el modelo Muestrear, Explorar, Modificar, Modelar y Evaluar, el cual abarca todas las etapas del proceso de minería de datos, desde la recopilación inicial hasta la evaluación final de los modelos. Mediante un análisis bibliométrico, basado en cuatro características clave, se identificaron campos de conocimiento esenciales para el desarrollo de líneas de investigación por el análisis de 1,300 artículos científicos de alto impacto. Como resultado, la aplicación del algoritmo Near Zero automatizó la clasificación de criterios de búsqueda. Este enfoque no solo facilita la identificación de áreas emergentes, sino que también abre nuevas oportunidades en sectores industriales diversos, relevancia de la académica para el avance tecnológico como parte de las transformaciones digitales.

Palabras clave: máquina; industria; líneas de investigación; Productores de bases de datos bibliográficas.

Recibido: 20/Mar/2025
Aceptado: 09/Abr2025
Publicado: 31/May/2025

Cita: Sangacha-Tapia, L., González-Cañizalez, Y., & Rivas-Herrera, J. (2025). Optimización de Criterios de Búsqueda avanzada para Nuevas Tendencias en la Académica mediante Machine Learning. *Revista Científica Zambos*, 4(2), 197-211. https://doi.org/10.69484/rcz/v4/n2/114

Ecuador, Santo Domingo, La Concordia Universidad Técnica Luis Vargas Torres de Esmeraldas – Sede Santo Domingo
Revista Científica Zambos (RCZ)
https://revistaczambos.utelvtsd.edu.ec

Abstract:

The growing availability of large volumes of data has generated new assets for various industries. This poses a key challenge for higher education: measuring, defining, and structuring these assets by identifying relevant lines of research. These lines should serve as a guide in the training of new professionals, responding to the growing demand for technological skills within Industry 4.0. The present study aims to establish solid criteria for identifying areas of research applicable to this context. To achieve this, the Sample, Explore, Modify, Model, and Evaluate model has been used, which covers all stages of the data mining process, from initial collection to final evaluation of the models. Through a bibliometric analysis based on four key characteristics, essential fields of knowledge for the development of lines of research were identified by analyzing 1,300 high-impact scientific articles. As a result, the application of the Near Zero algorithm automated the classification of search criteria. This approach not only facilitates the identification of emerging areas but also opens up new opportunities in diverse industrial sectors, highlighting the relevance of academia to technological advancement as part of digital transformations.

Keywords: machine; industry; lines of research; bibliographic database producers.

1. Introducción

La optimización de criterios de búsqueda avanzada para la identificación de nuevas tendencias en la investigación académica mediante machine learning, es una alternativa en los diferentes campos de conocimientos de la investigación científica por su continua evolución como parte de su transformación digital. Se ha descubrimiento los nuevos enfoques como tecnologias transformadas que surgen a un ritmo acelerado, lo que representa un desafío para científicos e investigadores que deben esforzarse por mantenerse actualizados (Sandoval-Almazán, 2011). Blanch et al. (2016) consideró que, durante la última década las organizaciones han experimentado un proceso de cambio con escasos o inexistentes precedentes en la historia debido al acelerado ritmo de crecimiento de la competencia a nivel global, surgen nuevos retos y a gran velocidad en el enfoque tecnologicos, todo para adaptarse a los cambios, el entorno cambiante y la identificación de nuevas líneas de investigación prometedoras, para el progreso del conocimiento y la innovación, para la definición de una nueva línea de investigación puede llegar a ser relativamente compleja por el gradiente que puede existir en función de su generalidad o especialización.

Como se mencionó anteriormente en el trabajo la identificación de nuevas líneas de investigación han sido dependiente en gran medida del juicio, experiencia y recomendaciones de investigadores individuales, empresarios educadores, si bien este enfoque es valioso para crear e identificar de las mismas también presentan limitaciones claras como la subjetividad y la variabilidad entre las consideraciones de

los expertos que pueden estar orientadas a sesgos y omisiones de características y hasta líneas de investigación completas, lo que vuelve más tedioso el proceso de identificación y creación, por lo que se propon la reduccción de manera significativa la capacidad de una identificación oportuna de nuevas líneas y áreas de investigación. En respuesta a estas múltiples limitaciones hemos propuesto un nuevo enfoque basado en el procesamiento de lenguaje natural, según (Augusto et al., 2009) El lenguaje natural (LN) es uno de los medios que usamos cotidianamente para mantener comunicación. Mientras que (PLN) según se entiende como la capacidad de una máquina para procesar la información comunicada, no solo las letras y sus sonidos.

En palabras de (Augusto et al., 2009) el procesamiento de lenguaje natural (PLN) consiste en la utilización de un lenguaje natural para comunicarnos con la computadora, debiendo ésta entender las oraciones que le sean proporcionadas, el uso de estos lenguajes naturales, facilita el desarrollo de programas que realicen tareas relacionadas con el lenguaje o bien, desarrollar modelos que ayuden a comprender los mecanismos humanos relacionados con el lenguaje. Por medio de esta investigación aporta la optimización de criterios de búsquedas avanzadas para minimizar de manera significativa la inversión de tiempo y esfuerzo requerida por parte de expertos en la creación de nuevas líneas de investigación.

2. Metodología

Se pretende en este articulo como propuesta la identificación de indicadores para la revisión periódica de líneas de investigación de forma automatizada como caso; en carreras de ingeniería industrial. Para ello se ha considero aplicar el modelo SEMMA, que permite abarcar todas las etapas del proceso de minería de datos. El enfoque del modelo incluye 3 etapas, desde la fase inicial de recolección de datos hasta la evaluación final de los modelos resultantes. A lo largo del proceso, se realizó un análisis bibliométrico, centrándose en cuatro características clave para identificar los campos de conocimiento esenciales que guiarían el desarrollo de nuevas líneas de investigación. Para este análisis, se utilizaron datos que sumaban más de 2500 registros de artículos científicos con alto impacto en sus respectivas áreas de estudio.

Con la selección de instrumentos de procesamiento la información que soporte Python, para el análisis descriptivo de los datos, aplicando técnicas de machine learning para mejorar la calidad de los datos, experimentación y evaluación de los diferentes modelos y finalmente la implementación del modelo de mejor desempeño basada en machine learning. El Machine Learning proporciona a la herramienta de procesamiento de información útil para enriquecer y transformar adecuadamente los materiales en aras a mejoras en los procesos productivos e industriales; puntualmente, el machine Learning aporta a las máquinas una capacidad para generar patones o describir relaciones, adaptarse a los cambios y resolver problemas sin programación previa. (López et al., 2024). Con la intención de potenciar nuestros

resultados se analizaron 4 características cruciales dentro de los criterios de búsqueda bibliográficos las cuales son la ocurrencia, coocurrencia, el impacto y los campos de conocimiento. En palabras de Estrella & Lastra-Bravo (2019), el análisis bibliométrico es útil para analizar la cantidad y la calidad de las publicaciones científico-técnicas, como las temáticas y áreas de estudio. En el estudio presentado por (Angarita-Becerra, 2014) los bibliométricos facilitan no solo conocer más acerca de un tema particular sino cualificar internamente el proceso científico, brindando información importante sobre la producción científica en cuanto tal, que puede abrir otras líneas de investigación mediante ideas de investigación que surgen de la misma reflexión inferida desde la cuantificación y reflexión acerca de lo que se produce.

Para alimentar nuestra data nos dimos a la tarea de aplicar metodología cualitativa en la revisión de cada uno de los artículos recolectados para el desarrollo de la arquitectura de nuestro modelo predictivo, en cuanto a los procesos de recopilación, debe ser de fuentes fiables basadas en artículos científicos publicados en revistas de alto impacto indexadas en bases de datos electrónicas las cuales constan en el sistema regional de información en línea para revistas científicas del Caribe España y Portugal (latindex) Scientific Electronic Library Online (Scielo), Red de revistas científicas de América Latina y el Caribe, España y Portugal (Redalyc) y Scopus. El fin de esta recopilación de datos tenía como objetivo el conocer nuevas tendencias creando un metadato en la cual se contengan datos como el tema, años de publicación, códigos ISBN/ISSN, el enlace de publicación, la sección de resumen y conclusión del articulo las cuales fueron indispensables en el proceso de etiquetación para la correcta identificación de los paradigmas y criterios asociados al área de interés. (Rojas et al., 2023) Menciono que la metodología SEMMA es más que un método de minería de datos, es un conjunto de herramientas funcionales que se enfocan en los aspectos de autodesarrollo de los modelos de minería. Para la justificación de la aplicación de la metodología SEMMA se preparó un cuadro comparativo encargado de justificar su importancia:

Tabla 1

Cuadro comparativo para la selección del proceso de minería de datos

Metodología	Descripción	Pasos	Modelo%
KDD	Identificar patrones implícitos en los grandes volúmenes de datos y convertirlos en conocimiento, filtrando y descartando aquellos hallazgos que no resulten útiles para los objetivos fijados	Selección Preprocesamiento Transforma-cion Minería de datos Interpretación	60%
SEMMA	Desarrollada por el SAS Institute. Va en consonancia con el sistema de trabajo de SAS Enterprise Miner, su software de minería de datos	Muestreo Exploración Modificación Modelado Evaluación	100%

CRISP-DM	Un avance con respecto a KDD y SEMMA en el campo de los procesos de minería de datos para el análisis en Big Data, en el sentido de que anticipa la repercusión de los hallazgos	Comprensión del negocio Comprensión del negocio Preparación de los datos Modelado Evaluación Despliegue	80%
CATALYST	Conocido como P3TQ significa Product place price time quality, se encuentra conformada por dos modelos que es el modelo de negocio y la de explotación de información	Los datos son para encontrar patrones Cada problema es una identificación de oportunidad Identificar donde aplicar la minería de datos	85%
FOLKSOMIA	La folksomia ha recopilado los datos necesarios para su uso y los ha publicado en una plataforma web para que puedan consultarlos, modificarlos o usarlos	Comparación para la extracción	50%
MAM	Es para uso de plataforma en multimedia, para técnicas convencionales	Transformación Aplicación de técnicas de minería con preprocesamiento Procesamiento de las imágenes Análisis de multimedia	20%

Nota: Las diferentes metodologías de minería de datos donde se indica por qué la selección el método SEMMA. Lady Sangacha (2025).

Variables identificadas:

Se utilizo la columna de Abstract como nuestra variable independiente tanto de nuestra dataset 1 como de la dataset 2.

Cumpliendo el proceso considerando las 4 características de gran relevancia en los criterios de búsqueda bibliográfica que son la ocurrencia, coocurrencia, el impacto y los campos de conocimiento.

A continuación, te doy a conocer el desarrollo de cada etapa a través del modelo SEMMA (Rojas et al., 2023).

3. Resultados

3.1. Etapa 1: Proceso de muestreo

Se fortaleció la recopilación de los productos científicos (papel, libro o articulo) de impacto, para ello se preseleccionó todo producto que provenga de una fuente confiable considerando las siguientes fuentes de base científica:

Tabla 2

Fuentes confiables usadas para el desarrollo del proyecto

Bases de datos científicas	Bibliotecas virtuales UG
https://scielo.org/	https://link.springer.com/
https://dialnet.unirioja.es/	elibro
https://www.redalyc.org/	https://www.sciencedirect.com/
https://www.sciencedirect.com/	
https://scholar.google.com/	

Nota: Listado corto de las bases de datos científicos. Lady Sangacha (2025).

Además de aplicar una rápida lectura se considera 3 pasos, el primero es en leer el titulo del documento hallado, debe estar relacionado al tema de la propuesta. El segundo es leer el resumen del documento, esto proporciona partes del tema en interés y finalmente leer las conclusiones ya que proporciona resultados de interés.

Para asegurar el impacto de los productos, según los expertos académicos y científicos es necesario asegurarse de que los productos de base científica cumplan con los criterios de preselección para su posterior adición a la metadata para luego realizar un meta análisis adecuado al momento de identificar el muestreo con sus variables como se da a conocer a continuación:

Tabla 3

Elementos necesarios del articulo para su registro en el dataset

Elementos considerados para la metadata
AUTOR PRINCIPAL
Apellido del primer autor del Paper/Libro/Tesis
CITAS APA 7
REFERENCIA APA 7
FECHA DE PUBLICACION
(DD/MM/AAAA)
AÑO DE PUBLICACION
ISBN/ISSN
TITULO DEL ARTICULO
NOMBRE DE LA REVISTA O FUENTE DONDE FIE PUBLICADO
URL
Resumen que trae el articulo
Descarga directa desde la fuente oficial (SI/NO)
Descarga indirecta por SCI-hub (SI/NO)

Idioma original del Paper/Libro/Tesis
Tipo de documento
Paper/Libro/Tesis
NOMBRE DEL ARCHIVO.pdf
Hallazgos del articulo usando palaras, indicadores p variables destacadas, Evites redacciones vacías o generalizaciones
Construya varios criterios o palabras claves que definan los hallazgos del articulo (Ej. Canonical Polyadic Descomposition (CPD), Trensor-Train Network (TT), etc).
Abstract traducido al español usando Deepl.com
Conclusiones que trae el artículo en español usando Deepl.com

Nota: Criterios para el cumplimiento de productos de calidad. Lady Sangacha (2025).

3.2. Etapa 2: Proceso de exploración

Una vez asegurado la preselección de los productos, es sometido a la meta análisis anteriormente mencionada visualizando los datos de la dataset para determinar las líneas de investigación de la Universidad de Guayaquil enfocando únicamente la línea de investigación de “operaciones, algoritmos de automatización y aplicación”, orientadas a la industria 4.0. La industria 4.0 y la manufactura inteligente son parte de una transformación, en la que las tecnologías de fabricación y de la información se han integrado para crear innovadores sistemas de manufactura, gestión y formas de hacer negocios, que permiten optimizar los procesos de fabricación. Se vio la necesidad de dividir en 2 la dataset, con el propósito de cargar la primera data con las palabras clave clasificadas en base a la variable de la columna de “Abstract traducido al español con Deepl.com” mientras que la segunda dataset fue cargada en base a los campos de educación y capacitación 2013 de la CINE

a) La primera dataset es etiquetada con la identificación de palabras clasificadas que se encuentren orientadas a la industria 4.0 en la automatización, control y supervisión de procesos se partió desde las áreas y líneas de investigación de la Universidad de Guayaquil, pero enfocados únicamente a la línea de investigación de “operaciones de automatización y aplicación”, orientadas a la industria 4.0 (Solano-Gutiérrez, 2024).

b) En la segunda dataset con los mismos criterios del paso 1 con una creación de columnas en los distintos campos de educación y capacitación 2013 de la CINE que constan en 1) Educación; 2) Artes y humanidades; 3) Ciencias sociales, periodismo e información; 4) Administración de empresas y derecho; 5) Ciencias naturales, matemáticas y estadística; 6) Tecnologías de la información y comunicación; 7) Ingeniería, industria y construcción; 8) Agricultura, silvicultura, pesca y veterinaria; 9) Salud y Bienestar ; 10) Servicios.

Como parte del proceso de exploración se identificó la necesidad de la aplicación del diccionario de datos puesto que esto facilitaría los procesos de modelamiento por lo que se pudo realizar a través de líneas de comando del lenguaje Python o de forma manual donde al final llamaremos preprocesamiento en el proceso de modelamiento

de la metodología SEMMA para luego ser procesado a través de la máquina virtual de google colab.

Además de la identificación de variables para alcanzar el objetivo de la propuesta se consideró la columna de “Abstract traducido a español usando Deepl.com” para su aplicación en la arquitectura del modelo de aprendizaje supervisado en procesamiento de lenguaje natural con la segunda dataset de áreas de conocimiento, en palabras de (Camacho et al., 2013) Un sistema de PLN se basa en el reconocimiento de conceptos en el texto y la comprensión de las relaciones entre esos conceptos. Procesamiento natural del lenguaje (PNL) es el campo de estudio que busca entender cómo funciona el lenguaje, su construcción, la generación de nuevo lenguaje, así como todas las tareas que tienen relación con el tratamiento del lenguaje. (Beltrán & Rodriguez, 2021). La etiquetación de la dataset ha sido 26 columnas clasificadas creadas detectadas como se da a conocer en la tabla a continuación y en la dataset 2 con 10 campos de conocimiento:

Tabla 4

Columnas identificando los elementos de coocurrencia e impacto de las palabras clasificadas en 26 columnas

Palabras clasificadas de impacto
Inteligencia Computacional, (2) Procesamiento de Datos, (3) Creación de algoritmos, (4) Inteligencia Artificial, (5) Búsquedas Automatizadas, (6) Machine Learning, (7) Industria 4.0, (8) Modelos Predictivos, (9) Internet de las Cosas, (10) Seguridad y Salud, (11) Control de Sistemas, (12) Data Mining, (13) Sistema Eléctrico, (14) Motor de Control, (15) Supervisado, (16) Aplicación Empresarial, (17) Utilización de Datos, (18) Mantenimiento de Sistema, (19) Indicadores de Rendimiento, (20) Diagnósticos Industriales, (21) (22) Predicción de Datos, (23) Big Data, (24) Robótica, (25) Cuántica, Aprendizaje Automático, (26) Deep Learning.

Nota: Los campos detallados se diferencian en función de los métodos, técnicas, herramientas e instrumentos.” (UNESCO, 2014.)

Tabla 5

Columnas identificando los campos de conocimiento de la CINE 2013

ÁREAS DE CONOCIMIENTO
(1) educación, (2) artes y humanidades, (3) ciencias sociales, periodismo e información, (4) administración de empresas y derecho, (5) ciencias naturales, matemáticas y estadística, (6) tecnologías de la información y comunicación, (7) INGENIERÍA INDUSTRIA y construcción, (8) agricultura, silvicultura, pesca y veterinaria, (9) salud y bienestar, (10) servicios

Nota: Autores (2025).

Los estudios bibliométricos pertenecen a un campo de investigación interdisciplinario que tiene el potencial de extenderse a casi todos los campos científicos. La metodología de la Bibliometría comprende componentes de la Matemática, las

Ciencias Sociales, las Ciencias Naturales, la Ingeniería, la Informática, la estadística, entre otras (Romaní et al., 2011).

3.3. Etapa 3: Proceso de modificación y modelado

En este paso se aplicaron los 5 pasos de la creación de una arquitectura de modelo algorítmico basado en machine learning que se mencionan a continuación:

Fase 1: La recopilación de datos, luego este proceso, la metadata será sometida a un proceso de meta análisis en donde se identifican claramente las variables independientes y dependientes, para su proceso de modelamiento en la arquitectura respectiva, en este proceso de verifica que los datos necesarios se encuentren dentro de la dataset 1 y 2.

Fase 2: La aplicación del preprocesamiento envuelve todos los procesos de data cleaning, data transformación, data reduction, para este caso se aplicó data cleaning y data transformation por medio de la ejecución de líneas de comando de lenguaje de programación llamado Python. Un lenguaje de programación es un lenguaje formal definido como un conjunto de elementos (componentes léxicos) organizados a través de constructores (reglas gramaticales) que permiten escribir un programa y que éste sea entendido por el computador y pueda ser trasladado a computadores similares para su funcionamiento en otros sistemas.(Augusto et al., 2009)

Esta preparación es crucial antes de darle un entrenamiento a la maquina los datos deben ser preprocesados, una vez terminado este proceso se logró actualizar un aproximado de 1300 instancias correspondientes a los artículos científicos que cumplían con ser producciones de alto impacto relacionadas a la industria 4.0 con algoritmos de operaciones, algoritmos de optimización y sus aplicaciones, para esta fase se eliminan los registros que incumplen con esta norma además de campos innecesarios y/o incompletos. La Industria 4.0 es la revolución más reciente de la industria, que se centra en gran medida en la interconectividad, los macrodatos, los sistemas ciberfísicos y la inteligencia artificial. centrada en la interconectividad, los macrodatos, los sistemas ciberfísicos y el ML, también se conoce como Internet Industrial de las Cosas (IIoT), que significa el proceso de fabricación inteligente, que fusiona las operaciones físicas con procesos inteligentes impulsados por la inteligencia artificial y la automatización. Una vez terminado el proceso se procede con la carga de los dataset 1 y 2, para continuar el proceso utilizamos el lenguaje Python con una herramienta científica de datos para el debido proceso de exploración, preprocesamiento y procesamiento.

Se presenta una sección de los datos utilizando comando de Python para facilitar la exploración directa e indirecta de los datos y llevar a cabo una exploración EDA (Exploración de Análisis de Datos)" El análisis permite recopilar y evaluar información relevante para identificar oportunidades y desafíos marcando el comienzo de la búsqueda de correlaciones en los datos y la aplicación de métodos de preprocesamiento, durante la exploración se identificaron aquellos valores vacíos

dentro de la matriz, los cuales se procedió a dar su correspondiente tratamiento, durante este proceso se tuvieron en cuenta las variables dependientes e independientes considerando el modelo algorítmico diseñado para la clasificación de texto y la columna que permitiría aplicarlo.

Fase 3: Una vez listas las datasets aplicadas se contempla la identificación de la muestra parta su proceso de modificación y modelado, es necesario haberse aplicado el preprocesamiento, posteriormente se realiza una última exploración así asegurando que todos los datos estén correctamente tratados y no existan cabos sueltos “La Visualización es generalmente utilizada para obtener un entendimiento preliminar de los datos al inicio del proceso de KDD, y con esto se logra refinar los objetivos y tareas definidas inicialmente en la fase de formulación del problema.

Fase 4: Aplicación del modelo de aprendizaje supervisado multietiqueta basado en machine learning, para concluir con este paso se realizaron varias pruebas de arquitectura hasta dar con la final.

Tabla 6

Resultados de la aplicación de modelos

Modelo	Precisión final del modelo	% de precisión final del modelo
MultinomialNB	0.47	47%
Logistic Regression	0.42	42%
Random Forest	0.40	40%
SVC	0.81	81%

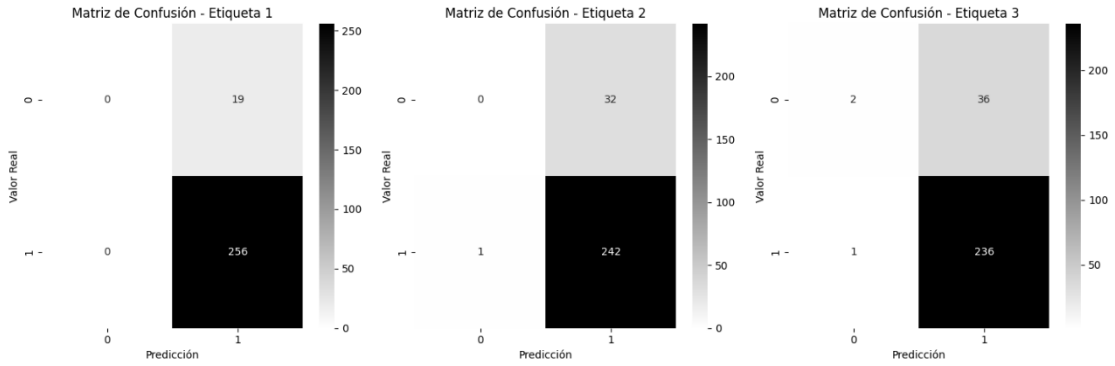
Nota: Se obtiene los resultados de los diferentes modelos basado en machine learning. Lady Sangacha. 2025

La clasificación multietiqueta es una variante de la clasificación tradicional de etiqueta única, en la que un objeto ya no se clasifica exclusivamente por una etiqueta. En su lugar, este aprendizaje pretende asignar a un objeto una o más clases de etiquetas de un conjunto predefinido de clases. (Bello et al., 2023)

Fase 5: La aplicación de las métricas de evaluación correspondientes al modelo final ya que “El proceso científico requiere de una serie de repeticiones o replicaciones que permitan una acumulación de observaciones para expresar un juicio de probabilidad de la existencia de dichas relaciones entre variables o conceptos.” para asegurar su calidad finalmente obteniéndose los siguientes resultados:

Figura 1

Matriz de confusión del modelo con la dataset 1



Nota: Resultados de los modelos. Lady Sangacha. 2025

Se procede a correlacionar con el campo de conocimiento de ingeniería, industria y construcción para proponer líneas de investigación obteniendo lo siguiente:

Tabla 8

Tabla De Correlación

Resultados Bibliométricos	Utilización De Datos	Predicción De Datos	Supervisado
Registros	27%	26%	25%
Campos De Conocimiento O Áreas De Conocimiento	Ingenieria, Industria Y Construccion. Tecnologias De La Informacion Y Comunicacion. Educacion. Salud Y Bienestar.	Ingenieria, Industria Y Construccion. Tecnologias De La Informacion Y Comunicacion. Ciencias Sociales, Periodismo E Informacion. Salud Y Bienestar.	Ingenieria, Industria Y Construccion. Tecnologias De La Informacion Y Comunicacion. Ciencias Sociales, Periodismo E Informacion. Salud Y Bienestar. Educacion.

Nota: Resultados correlaciones. Lady Sangacha (2025).

4. Discusión

El presente estudio utilizó motores de búsqueda como SciELO y ScienceDirect, entre otros, para la creación de un conjunto de datos compuesto por más de 2500 registros de las cuales se logró disernir a 1300 registros adecuados, permitiendo un análisis exhaustivo de la literatura disponible enfocando únicamente la línea de investigación de “operaciones, algoritmos de automatización y aplicación” aplicado a la Industria 4.0, en una primera adataset se etiquetaron 26 clasificaciones relacionadas la línea de investigación de “operaciones, algoritmos de automatización y aplicación” y 10

clasificaciones adicionales en un segundo dataset que abarcan los campos del conocimiento reconocidos por la CINE en 2013.

La investigación identificó cuatro criterios esenciales de búsqueda bibliométrica: coocurrencia, impacto, campos de conocimiento y visitas del artículo. Existe la bibliométria como tecnica cuantitativa en los respectivos del analisis para la investigación de la académica educativa superior. Estos criterios permitieron una evaluación detallada de las tendencias y la relevancia de los estudios en la disciplina. Los indicadores bibliométricos son datos numéricos calculados a partir de las características bibliográficas observadas en los documentos publicados en el mundo científico y académico, y el analisis para su producción (Flores-Fernández & Aguilera-Eguía, 2019), en palabras de (Arbeláez Gómez & Onrubia Goñi, 2014) la bibliometría o los estudios bibliométricos se utilizan para analizar la información relacionada con la producción científica.

Menciona Escorcia-Otalora (2008) que el uso en la bibliometría se buscan calcular los diferentes indicadores en base al tipo o diversidad de documento como es el caso de los artículos cientificos publicados en revistas de investigación los cuales pueden analizarse bibliométricamente. Estos estudios a partir de bases de datos, puede ser contable con la producción científica por distintos países y, además la identificación de grupos de trabajo, áreas de excelencia, en asociaciones de temáticas, la interdisciplinariedad, la disciplinas en emergentes, las prioridades en la ciencia, etc. (Alonso & Reyna, 2005)

Como lo indica (Espinosa-Castro et al., 2019) los indicadores bibliométricos es el uso de herramientas de mayor frecuencia para la medición del producto de la investigación científica, porque la documentación (independientemente del tipo de soporte) es el vehículo más prolífico y exitoso para la transferencia del conocimiento científico, conjuntamente con su por medio de conferencias y comunicaciones personales. Los resultados indican que las principales líneas de investigación emergentes se centran en 'Utilización de Datos', 'Predicción de Datos' y 'Supervisados'. El análisis bibliométrico reveló nuevas aplicaciones de machine learning en la enseñanza para la Industria 4.0, destacando la evolución del conocimiento y su impacto en la formación académica y profesional.

El análisis de contenidos, realizado con Python utilizando algoritmos de procesamiento de lenguaje natural, confirmó la efectividad del enfoque supervisado en la clasificación y comprensión de grandes volúmenes de datos textuales. Este método no solo facilitó el descubrimiento de nuevas aplicaciones, sino que también puede ser replicado en otras líneas de investigación, como el Internet de las Cosas (IoT).

5. Conclusiones

Este estudio ha realizado un análisis exhaustivo de la literatura científica relacionada con la aplicación de operaciones, algoritmos de automatización y machine learning en el contexto de la Industria 4.0. A través de técnicas bibliométricas y análisis de contenido recopilando más de 2500 registros por la metodologia SEMMA idoneo para el proceso, se ha logrado identificar las tendencias emergentes, los actores clave y las áreas de mayor interés en esta disciplina. Los resultados obtenidos evidencian un creciente interés en la utilización de datos para la toma de decisiones, la predicción de eventos y el desarrollo de modelos supervisados. La implementación de algoritmos de procesamiento de lenguaje natural ha sido fundamental para clasificar y comprender grandes volúmenes de texto científico por el modelo SVC, lo que ha facilitado la identificación de patrones y tendencias. Este enfoque no solo es aplicable a la Industria 4.0, sino que también puede ser extendido a otros campos de conocimiento.

Referencias Bibliográficas

Alonso Gamboa, J. O., & Reyna Espinosa, F. R. (2005). Compilación de datos bibliométricos regionales usando las bases de datos clase y periódica. *Revista Interamericana de Bibliotecología*, 28(1), 63-78. https://doi.org/10.17533/udea.rib.8596

Angarita Becerra, L. D. (2014). Estudio bibliométrico sobre uso de métodos y técnicas cualitativas en investigación publicada en bases de datos de uso común entre el 2011-2013. *Revista Iberoamericana de Psicología*, *7*(2), 67–76. https://doi.org/10.33881/2027-1786.rip.7207

Arbeláez Gómez, M. C., & Onrubia Goñi, J. (2016). Análisis bibliométrico y de contenido. Dos metodologías complementarias para el análisis de la revista colombiana Educación y Cultura. *Revista De Investigaciones · UCM*, *14*(23), 14–31. https://doi.org/10.22383/ri.v14i1.5

Augusto Cortez Vásquez, M., Hugo Vega Huerta, M., Jaime, L., & Quispe, P. (2009). *Procesamiento de lenguaje natural*. *Revista de Ingeniería de Sistemas e Informática*, 6(2), 45-54. https://sisbib.unmsm.edu.pe/bibvirtual/publicaciones/risi/2009_n2/v6n2/a06v6n2.pdf

Bello García, M., Bello Pérez, R., Nápoles, G., Vanhoof, K., García Lorenzo, M., & Aguilera Calzadilla, Y. (2023). Desarrollo de técnicas para el preprocesamiento y la predicción de problemas de clasificación multietiqueta. *Anales de la Academia de Ciencias de Cuba, 13*(3), e1344. https://revistaccuba.sld.cu/index.php/revacc/article/view/1344

Beltrán Beltrán, N. C., y Rodríguez Mojica, E. C. (2021). Procesamiento del lenguaje natural (PLN) - GPT-3.: Aplicación en la Ingeniería de Software . *Tecnología*

Investigación y Academia, *8*(1), 18–37. https://revistas.udistrital.edu.co/index.php/tia/article/view/17323

Blanch, J., Gil, F., Antino, M., & Rodríguez-Muñoz, A. (2016). Modelos De Liderazgo Positivo: Marco Teórico Y Líneas De Investigación. In *Papeles del Psicólogo / Psychologist Papers,* 37(3). https://www.papelesdelpsicologo.es/pdf/2772.pdf

Camacho, J., Moreno, S., Suarez-Obando, F., Puyana, J. C., & Gómez-Restrepo, C. (2013). El procesamiento de lenguaje natural y su relación con la investigación en salud mental. *Revista Colombiana de Psiquiatría, 42*(2), 227-233. https://www.redalyc.org/articulo.oa?id=80629187011

Escorcia, T. A. (2008). *Análisis bibliométrico como herramienta para el seguimiento de publicaciones científicas, tesis y trabajos de grado*. Recuperado de: http://hdl.handle.net/10554/8212

Espinosa-Castro, J.F., Hernández-Lalinde, J., Rodríguez, J. E., Chacín, M., & Bermúdez-Pirela, V. (2019). Indicadores bibliométricos para investigadores y revistas de impacto en el área de la salud. *AVFT – Archivos Venezolanos De Farmacología Y Terapéutica*, *38*(3). http://saber.ucv.ve/ojs/index.php/rev_aavft/article/view/16806

Estrella, N., & Lastra-Bravo, X. B. (2019). Análisis bibliométrico de los trabajos de titulación de ocho Universidades de Pichincha, Napo y Orellana (Ecuador). *Siembra*, *6*(1), 050–067. https://doi.org/10.29166/siembra.v6i1.1720

Flores-Fernández, C., & Aguilera-Eguía, R. (2019). Bibliometric indicators and their importance in clinical research. Why know them?, *Revista de la Sociedad Espanola del Dolor 26*(5). (315–316). https://doi.org/10.20986/resed.2018.3659/2018

López Telenchana, L. S. L., Serrano Torres, G. J., Quintana López, X. A., & Reina Haro, D. M. (2024). Machine Learning in Industry 4.0: a systematic review. *Salud, Ciencia y Tecnología*, *4*, 1068. https://doi.org/10.56294/saludcyt20241068

Muñoz-Estrada, G. K., Chumpitaz Caycho, H. E., Barja-Ore, J., Valverde-Espinoza, N., Verde-Vargas, L., & Mayta-Tovalino, F. (2022). Bibliometric analysis of the world scientific production on the flipped classroom in medical education. *Educacion Medica*, *23*(5). https://doi.org/10.1016/j.edumed.2022.100758

Rojas, G. C., Carreño, S. C., Ovalle, C., & Chávez, E. H. R. (2023). Intelligent predictive model applying Data Mining strategies for a credit evaluation of a commercial company. *Proceedings of the LACCEI International Multi-Conference for Engineering, Education and Technology*, *2023-July*. https://doi.org/10.18687/laccei2023.1.1.1148

Romaní, F., Huamaní, C., & González-Alcaide, G. (2011). Estudios bibliométricos como línea de investigación en las ciencias biomédicas: Una aproximación para el pregrado. *CIMEL Ciencia e Investigación Médica Estudiantil Latinoamericana, 16*(1), 52-62. http://www.redalyc.org/articulo.oa?id=71723602008

Sandoval-Almazán, Rodrigo. (2011). Mentes en peligro: El daño de internet en nuestro cerebro. *Convergencia*, *18*(56), 241-248. Recuperado en 31 de mayo de 2025, de http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S1405-14352011000200010&lng=es&tlng=es.

Solano-Gutiérrez, G. A. (2024). La Tecnología en la Educación a Distancia: Revisión de Progresos y Obstáculos a Superar. *Revista Científica Zambos*, *3*(2), 48-73. https://doi.org/10.69484/rcz/v3/n2/17

Unesco (2014), La Clasificación Internacional Normalizada de la Educación (CINE) forma parte de la familia internacional de Clasificaciones Manual que acompaña la Clasificación Internacional Normalizada de la Educación 2011 Campos de educación y capacitación 2013 de la CINE (IsCEd-F 2013). https://doi.org/10.15220/978-92-9189-157-3-sp