T E S LA
Revista Científica
ISSN: 2796-9320
Vol. 3 Núm. 2 (Julio – Diciembre 2023), e226

Área: Ingeniería
Artículo de Investigación Original

Algoritmos de Deep Learning utilizando Tensorflow para el Tratamiento de Datos de Producción Científica

Deep Learning Algorithms using Tensorflow for Processing Scientific Production Data

Diego Geovanny Falconí Punguil [1][0009-0000-2398-8849]

[1] Universidad Técnica de Cotopaxi, Facultad de Ciencias de Ingeniería y Aplicadas. Av Simón Rodríguez SN y Jamaica. 050108. Latacunga - Cotopaxi. Ecuador

[1]diego.falconi4@utc.edu.ec

CITA EN APA:
Falconí Punguil, D. G. (2023). Algoritmos de Deep Learning utilizando Tensorflow para el Tratamiento de Datos de Producción Científica. *Tesla Revista Científica*, *3*(2), e226. https://doi.org/10.55204/trc.v3i2.e226

Recibido: 2023-06-26
Revisado: 2023-07-02 al 2023-07-20
Corregido: 2023-07-29
Aceptado: 2023-08-03
Publicado: 2023-08-10

TESLA
Revista Científica
ISSN: 2796-9320

Resumen.
Introducción: La implementación de Inteligencia Artificial, Redes Neuronales y Algoritmos de Deep Learning apoyados en TensorFlow en la actualidad se encuentra en constante evolución ya que han abierto nuevas rutas para el tratamiento y análisis de grandes cantidades de datos en sistemas alojados en la web principalmente.
Objetivo: La presente investigación tiene como finalidad, mejorar el nivel de toma de decisiones no supervisados en la plataforma científica Ecuciencia, la misma que se encuentra alojado en los servidores de la Universidad Técnica de Cotopaxi.
Método: Los datos que se tomarán como referencia para los análisis introducidos en los algoritmos, será los referentes a Líneas y Sublíneas de Investigación de acuerdo a la Universidad Técnica de Cotopaxi.
Resultados: Los algoritmos de aprendizaje profundo se encargan de entrenar y agrupar por similitud una data de entrada sin supervisión denominado aprendizaje automático, los mismos que modelan abstracciones de alto nivel utilizando principalmente datos expresados en forma matricial o tensores.
Conclusión: El impacto de la implementación de algoritmos de aprendizaje profundo apoyados en TensorFlow en el sistema Ecuciencia, será muy importante, puesto que, gracias a este análisis, la plataforma científica podrá ser capaz de dar una predicción más acertada de las clasificaciones de Líneas y Sublíneas de investigación.
Palabras Clave: Algoritmos, Redes Neuronales, Aprendizaje Profundo, TensorFlow, KDD.
Abstract:
Introduction: The implementation of Artificial Intelligence, Neural Networks and Deep Learning Algorithms supported by TensorFlow is currently in constant evolution since they have opened new routes for the treatment and analysis of large amounts of data in systems mainly hosted on the web.
Objective: The purpose of this research is to help the level of unsupervised decision making in the scientific platform Ecuciencia, which is hosted on the servers of the Technical University of Cotopaxi.
Method: The data that will be taken as a reference for the analyzes introduced in the algorithms will be those referring to Research Lines and Sublines according to the Technical University of Cotopaxi.
Results: Deep learning algorithms are responsible for training and grouping an unsupervised input data by similarity called machine learning, the same ones that model high-level abstractions using mainly data expressed in matrix form or tensors.
Conclusion: The impact of the implementation of Deep Learning Algorithms supported by TensorFlow in the Ecuciencia system will be very important, since, thanks to this analysis, the scientific platform will be able to give a more accurate prediction of the classifications of Research Lines and Sublines.
Keywords: Algorithms, Neural Networks, Deep Learning, TensorFlow, KDD.

1. INTRODUCIÓN

En la Universidad Autónoma de Madrid, conscientes de la importancia de recopilar la producción científica de sus investigadores, se han desarrollado una serie de plataformas que recogen toda la actividad científica de los investigadores (Portal de Producción Científica) y que almacenan los textos completos de las publicaciones en las que se plasma esta producción, en acceso abierto, a través de su repositorio institucional Biblos-e Archivo. (UAM_Biblioteca, 2018)

Según (Ayala Mora, 2015), en el Ecuador, el objetivo principal de las universidades hasta antes de la década de los setenta era la docencia, sin dar ningún énfasis en el tema de métodos de investigación científica, por lo que el número de investigación bibliográficas era casi nulo. Sin embargo, desde el año 2008 la actividad científica de las universidades del país ha reflejado un incremento positivo en cuanto a su desarrollo. (Rivera García, Espinosa Manfugás, & Valdés Bencomo, 2017)

Prueba de ellos son las estadísticas de publicaciones en las bases de datos de Scopus, ya en según (Rankings, 2015), las publicaciones que se realizaron en el periodo 2004-2008 solo se reportó 32 instituciones educativas y un total de 866 artículos publicados. En el periodo del 2009 al 2013 las publicaciones aumentaron significativamente pues se llegaron a registrar alrededor de 1992 artículos científicos. Mas adelante en los años 2014 y 2015 se evidencia un total de 976 y 1174 artículos publicados, los mismo que superan significativamente los periodos antes descritos. Además de estos datos estadísticos, la revista estadounidense Nature, publica un ranking de artículos registrados con bases científicas, y en el mismo destacan tres universidades ecuatorianas, en primer lugar, la Pontificia Universidad Católica, en segundo la Universidad de Investigación de Tecnología Experimental (Yachay) y, por último, la Escuela Politécnica Nacional. (Rivera García, Espinosa Manfugás, & Valdés Bencomo, 2017)

Según (Fernández Díaz, Martínez Bernal, Rivalta Bermúdez, Díaz Ríos, & Jiménez Santander, 2013) las bases del desarrollo de nuevas tecnologías son impulsadas mayormente por el conocimiento. Gracias a los repositorios que contienen un conjunto de documentación científica, el ser humano tiene la capacidad de ir desarrollando nuevos datos que fortalezcan la gestión del conocimiento. Además, dichos repositorios permitirán al ser humano realizar una búsqueda de información de una manera más rápida y eficiente, sin necesidad de limitaciones para ningún individuo dentro de la comunidad científica.

En Universidad Técnica de Cotopaxi ubicada en la Av. Simón Rodríguez, barrio El Ejido sector San Felipe, del cantón Latacunga, provincia de Cotopaxi, se está desarrollando una cultura investigativa a través de la creación y recreación de ciencia, tecnología y arte, como la formación científica, generación, difusión y promoción de los saberes y conocimientos, que coadyuven al desarrollo sostenible y sustentable del entorno, con enfoque investigativo progresista y dedicado a promover la sostenibilidad productiva, ambiental y la equidad social de la región y el país.

Todas las investigaciones desarrolladas en la Universidad Técnica de Cotopaxi, son documentadas mediante artículos, libros, ponencias y proyectos, los que requieren ser almacenados y visualizados por la

comunidad universitaria. Actualmente el Alma Máter contiene una plataforma de gestión del conocimiento llamada Ecuciencia, la misma que está recopilando la producción científica y tecnológica de todas las disciplinas que se estudian en las distintas facultades existentes en la institución, a partir de indicadores cienciométricos.

Toda la información almacenada en la base de datos de la plataforma Ecuciencia, requiere ser visualizada en herramientas que el usuario pueda entender con facilidad, partiendo de estas características, surge la necesidad de establecer un tratamiento de datos almacenados en el repositorio digital, ya que, si bien es cierto existe una clasificación controlada, la plataforma carece de una herramienta que ayude a gestionar los datos de manera automática, por lo que aún no cuenta con una característica primordial de aprendizaje de máquina.

Este problema se lo puede apreciar en la clasificación y control de información, puesto que no existe la coherencia de datos proporcionados en la plataforma en cuanto a los documentos científicos y su relación con las líneas de investigación a las que pertenecen, generando de esta manera una inconsistencia de datos al momento de mostrar la información.

Por esta razón lo que se propone es la implementación de algoritmos de Deep Learning o Aprendizaje Profundo, utilizando la herramienta TensorFlow, la misma que se espera, aportará significativamente al crecimiento de nivel de predicción y clasificación en el sistema Ecuciencia.

El impacto de la implementación de algoritmos de aprendizaje profundo apoyados en TensorFlow en el sistema Ecuciencia, será muy importante, puesto que, gracias a este análisis, la plataforma científica podrá ser capaz de dar una predicción más acertada de las clasificaciones de Líneas y Sublíneas de investigación. Actualmente la plataforma científica no consta con algún tipo de algoritmo que apoye al usuario en cuanto a escoger a que Línea o Sublínea pertenece su trabajo, y lo que se pretende es que mediante los algoritmos proporcionados por TensorFlow, el sistema tenga la capacidad de predecir la clase y poder dar una respuesta correcta al usuario.

El aprendizaje profundo o Deep Learning, es una de las aplicaciones más poderosas y de mayor crecimiento de la inteligencia artificial. Es un subcampo del aprendizaje automático que se utiliza para resolver problemas muy complejos que suelen involucrar grandes cantidades de datos. (Rouhiainen, 2018)

El aprendizaje profundo se realiza mediante el uso de redes neuronales, que se organizan jerárquicamente para identificar relaciones y patrones complejos en los datos. Su aplicación requiere mucha información y potentes capacidades de procesamiento. Actualmente, se utiliza en reconocimiento de voz, procesamiento de lenguaje natural, visión por computadora y reconocimiento de vehículos en sistemas de asistencia al conductor. (Rivera, 2020)

Podemos ver un ejemplo obvio en las traducciones realizadas en Facebook, que recientemente demostró que es capaz de alrededor de 4.500 millones de traducciones al día debido al aprendizaje profundo. Suelen ser segmentos de texto breves, como actualizaciones de estado publicadas por los

usuarios en su perfil. Sin el aprendizaje profundo, sería muy costoso y requeriría una gran cantidad de personas para brindar el mismo servicio. (Rouhiainen, 2018)

Por su parte, TensorFlow es un sistema de aprendizaje automático que funciona en entornos grandes y heterogéneos. TensorFlow usa diagramas de flujo de datos para representar cálculos, estados compartidos y operaciones para cambiar ese estado. Mapea los nodos (TPU) del gráfico de flujo de datos a través de muchas computadoras en el clúster y a través de múltiples dispositivos informáticos, incluidas CPU de múltiples núcleos, GPU de uso general y ASIC de diseño personalizado, llamadas unidades de procesamiento Tensores. (Abidabi, Barham, Chen, & Chen, 2016) TensorFlow permite a los desarrolladores probar nuevos algoritmos de optimización y entrenamiento. TensorFlow admite varias aplicaciones, con un enfoque en el entrenamiento y la inferencia en redes neuronales profundas.

El aprendizaje profundo es un subcampo del aprendizaje automático que utiliza redes neuronales artificiales para estimular la estructura y función del cerebro humano. Aunque este es un método muy nuevo, se ha vuelto muy popular recientemente. Entre las muchas aplicaciones en las que el aprendizaje automático ha tenido éxito a cierta velocidad, el aprendizaje profundo ha logrado un mayor éxito. En particular, se prefiere en la clasificación de grandes conjuntos de datos porque puede proporcionar resultados rápidos y efectivos.

Los cálculos representados por TensorFlow se pueden realizar en una variedad de sistemas heterogéneos con poca o ninguna modificación, desde dispositivos móviles (como teléfonos móviles y tabletas) hasta sistemas distribuidos a gran escala con cientos de máquinas y varios dispositivos informáticos. (Abadi, 2016)

Las redes neuronales son un sistema de procesadores en paralelo interconectados en forma de gráfico dirigido. De manera esquemática, cada elemento de procesamiento (neurona) de la red se representa como un nodo. Estas conexiones establecen una estructura jerárquica, tratando de imitar la filosofía del cerebro, buscando nuevos modelos de procesamiento para resolver problemas específicos en el mundo real. Una definición simplificada de un gráfico topológico puede ser que, con respecto a la correspondencia topológica, las unidades que están físicamente próximas entre sí responderán a categorías de vectores de entrada que están igualmente próximas entre sí. Muchos vectores de entrada dimensionales se representan en forma de gráficos bidimensionales para mantener el orden natural de los vectores de entrada. (Sotolongo Aguilar, Guzmán Sánchez, & Carrillo, 2011)

La Red Neuronal Artificial (ANN) o sistema de conexión es un sistema de procesamiento de información cuya estructura y funcionamiento están inspirados en las redes neuronales biológicas. Consisten en un conjunto de elementos de procesamiento simples llamados nodos o neuronas, que están conectados entre sí mediante conexiones con valores modificables llamados "pesos".

Las actividades realizadas por la unidad de procesamiento o neurona artificial en tal sistema son simples. Por lo general, consiste en sumar los valores de entrada que recibe de otras unidades conectadas a él, comparar esta cantidad con un umbral, y si es igual o superior al umbral, enviar una activación o

salida al conectado. Tanto la entrada recibida por el dispositivo como la salida enviada por el dispositivo dependen del peso o la fuerza de la conexión a través de la cual se realizan estas operaciones. (Montaño Moreno, 2012)

Gracias a la aplicación de redes neuronales apoyados en TensorFlow, los sistemas son capaces de desarrollar conocimiento y fortalecer su capacidad de apoyo de toma de decisiones a los usuarios.

La toma de decisiones es fundamental para cualquier actividad humana. En este sentido, todos somos tomadores de decisiones. Sin embargo, tomar la decisión correcta comienza con un proceso de razonamiento continuo y concentrado, que puede incluir varias disciplinas como la filosofía del conocimiento, la filosofía de la ciencia, la filosofía de la lógica y, lo más importante, la creatividad. El gerente debe tomar muchas decisiones todos los días. Algunas de estas son decisiones de rutina, mientras que otras tienen un impacto significativo en las operaciones de la empresa donde trabaja. (Amaya, 2015)

Las decisiones sobre las TIC tienen mucho que ver con los datos existentes y sus métodos de procesamiento. Es importante saber que, para tomar mejores decisiones, los datos deben estar correctamente clasificados y estructurados, por eso se utilizan algoritmos para aprender y mostrar los resultados de forma interactiva, de manera que los usuarios entiendan claramente la respuesta que brinda el sistema.

2. METODOLOGÍA O MATERIALES Y METODOS

Existen muchas metodologías aplicables para todo lo que abarca la Inteligencia Artificial, una de las más conocidas es la metodología KDD la misma que establece todos los pasos necesarios para tener un buen proceso de análisis de datos. El uso de esta metodología será de mucha ayuda puesto que la misma abarca los procesos desde la concepción y carga de datos, continuando por su posterior entrenamiento y agrupamiento, para finalmente validar y visualizar los resultados.

2.1. Metodología kdd

El Descubrimiento de conocimiento en bases de datos (KDD, del inglés Knowledge Discovery in Databases) es básicamente un proceso automático en el que se combinan descubrimiento y análisis. El proceso consiste en extraer patrones en forma de reglas o funciones, a partir de los datos, para que el usuario los analice. (Timarán-Pereira, Hernández-Arteaga, Caicedo-Zambrano, Hidalgo-Troya, & Alvarado Pérez, 2016) La metodología KDD contempla las siguientes etapas:

Selección

En esta etapa se crea un conjunto de datos objetivo, seleccionando todo el conjunto de datos o una muestra representativa de este, sobre el cual se realiza el proceso de descubrimiento. La selección de los datos varía de acuerdo con los objetivos del negocio. (Timarán-Pereira, Hernández-Arteaga, Caicedo-Zambrano, Hidalgo-Troya, & Alvarado Pérez, 2016)

El primer paso en el proceso de extracción del conocimiento a partir de datos es reconocer y reunir los datos con los que se va a trabajar. Para lo cual, se realizará la identificación de la base de datos que está integrada al sistema Ecuciencia, al igual que se descubre los datos que la componen y la utilización

que se les está dando a los mismos.

Preprocesamiento/limpieza.

En la etapa de preprocesamiento/limpieza (data cleaning) se analiza la calidad de los datos, se aplican operaciones básicas como la remoción de datos ruidosos, se seleccionan estrategias para el manejo de datos desconocidos (missing y empty), datos nulos, datos duplicados y técnicas estadísticas para su reemplazo. (Timarán-Pereira, Hernández-Arteaga, Caicedo-Zambrano, Hidalgo-Troya, & Alvarado Pérez, 2016)

En esta etapa se analiza las tablas que conforman la base de datos del sistema Ecuciencia, y se seleccionan las que se considere necesarias para aplicar el o los algoritmos. Esta selección se la realiza en base a los indicadores cienciométricos.

Transformación/reducción.

En la etapa de transformación/reducción de datos, se buscan características útiles para representar los datos dependiendo de la meta del proceso. Se utilizan métodos de reducción de dimensiones o de transformación para disminuir el número efectivo de variables bajo consideración o para encontrar representaciones invariantes de los datos. (Timarán-Pereira, Hernández-Arteaga, Caicedo-Zambrano, Hidalgo-Troya, & Alvarado Pérez, 2016)

En esta etapa se realiza la selección de los atributos requeridos, para utilizarlos en los algoritmos. Cabe indicar que la selección de dichos atributos se los realizará en base a los indicadores cienciométricos.

Minería de datos.

En la fase de minería de datos, se aplica el modelo, la tarea, la técnica y el algoritmo seleccionado para la obtención de reglas y patrones. (Brito Sarasa, Rosete Suárez, & Acosta Sánchez, 2018)

En esta etapa se selecciona y aplica la técnica apropiada de minería de datos, que permita cumplir con el objetivo de la propuesta, para lo cual se realiza la recopilación de información necesaria en la fundamentación teórica. Ya con la técnica apropiada se procede a realizar el análisis de los algoritmos relacionados con la misma, para obtener los patrones que permitan cumplir con la visualización del análisis de datos con algoritmos de aprendizaje profundo en base a los atributos previamente seleccionados. Cabe mencionar que se utiliza el lenguaje de programación Python para realizar este proceso.

Interpretación/evaluación

En esta etapa se interpretan los patrones descubiertos y posiblemente se retorna a las anteriores etapas para posteriores iteraciones, también se puede incluir la visualización de los patrones extraído. Por otra parte, se consolida el conocimiento descubierto para incorporarlo en otro sistema para posteriores acciones o, simplemente, para documentarlo y reportarlo a las partes interesadas; también para verificar y resolver conflictos potenciales con el conocimiento previamente descubierto. (Timarán-Pereira, Hernández-Arteaga, Caicedo-Zambrano, Hidalgo-Troya, & Alvarado Pérez, 2016)

En esta etapa se evalúa los resultados obtenidos con la aplicación de los algoritmos, se verifica si cumple con los objetivos de análisis de datos de aprendizaje profundo, para posteriormente incluirlo en el

sistema Ecuciencia.

Es necesario, además, establecer un método de validación, que asegure el funcionamiento del trabajo de los algoritmos, es por esta razón que se ha seleccionado el método de validación cruzada o Cross – Validation.

2.2.MÉTODO DE VALIDACIÓN CRUZADA

En el ámbito de Inteligencia Artificial existen varios métodos que pueden ser utilizados para realizar el proceso de validaciones. Sin embargo, muchos expertos consideran que el método de Validación Cruzada o Cross – Validation es el más recomendado para realizar el análisis del nivel de exactitud que un algoritmo posee. Según (Pérez Planells, Delegido, Rivera Caicedo, & Verrelst, 2016), la validación cruzada es una técnica que se utiliza para evaluar los resultados del análisis estadístico y garantizar que sea independiente de la división entre los datos de entrenamiento y los datos de prueba.

La aplicación de este método de validación en la presente propuesta constará de varias etapas fundamentales:

Etapa 1: En la primera etapa, se utilizará los datos de documentos científicos, así como sus palabras claves y sublíneas de investigación, pertenecientes a la carrera de Ingeniería en Sistemas de Información de la Universidad Técnica de Cotopaxi esto con la finalidad de poder segmentar la muestra de los datos y tener un conjunto más resumido, pero a la vez con mayor consistencia de datos.

Etapa 2: Posteriormente se analizarán los datos obtenidos, realizando el proceso de entrenamiento utilizando el algoritmo de aprendizaje profundo seleccionado. En esta etapa se separará el conjunto de datos en dos subconjuntos, uno será utilizado para entrenar el modelo, y el otro será utilizado para realizar los test de validación. De esta forma, el modelo se puede crear utilizando solo los datos de entrenamiento. Con el modelo creado, los datos de salida se generarán y se compararán con el conjunto de datos reservados para su verificación. Cuando finalice el análisis se obtendrán los datos estadísticos pertenecientes al nivel de exactitud de entrenamiento y el número de datos perdidos en el proceso. A esta etapa se la conoce como método “hold-out”

Etapa 3: Una vez obtenido el nivel de exactitud de entrenamiento del algoritmo de aprendizaje profundo, utilizando los datos seleccionados de Ecuciencia, se procederá a aplicar el método “k-fold”, el cual consiste en evaluar “k” número de veces el modelo aplicando la técnica “hold-out”, es decir, se iterará el entrenamiento del algoritmo de aprendizaje profundo la veces que se especifique en la variable “k” para obtener un histórico de porcentaje de exactitud y establecer un promedio general de todas las veces que se realizó el proceso de entrenamiento.

Etapa 4: Y finalmente, gracias a los datos obtenidos tanto en el método “hold-out” como en el “k-fold”, se podrá realizar representaciones gráficas de las estadísticas obtenidas en el proceso de validación para poder determinar el nivel de aceptación establecida en la escala de fiabilidad representada en la tabla 1.

Tabla 1:

Escala de fiabilidad.

Muy Baja	Baja	Moderada	Buena	Alta
0 - 0,20	0,21 - 0,40	0,41 - 0,60	0,61 - 0,80	0,81 - 1,0

Elaborado por: Autor.

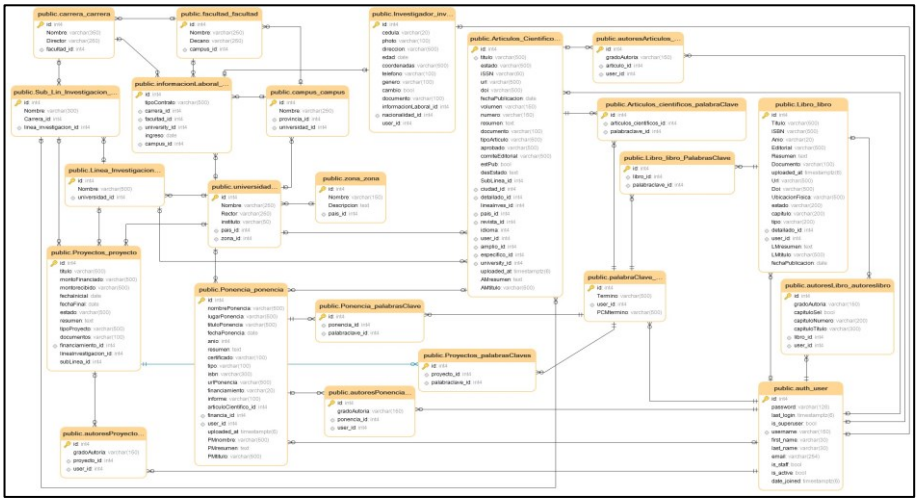
Un dato importante a tomar en cuenta, es que se puede aprovechar las etapas establecidas en la metodología KDD, puesto que la construcción de los métodos de Variación Cruzada se lo puede contemplar en la etapa de Minería de datos, y el proceso de ejecución del entrenamiento y validación utilizando los métodos hold-out y k-fold se lo realizaría en la etapa de Interpretación y Evaluación.

3. RESULTADOS Y DISCUSIÓN

En el proceso de investigación, se pudo evidenciar que Ecuciencia gestión sus datos a través de PostgreSQL, con esta información fue posible ejecutar las etapas de ambas metodologías. En la primera etapa del método se seleccionó e identificó la estructura interna de la base de datos que utiliza el sistema Ecuciencia, se observó que la base de datos hasta el momento de la que se desarrollaba la investigación superaba las 60 tablas interrelacionadas. Gracias a las etapas de preprocesamiento y limpieza se puedo localizar solamente las tablas necesarias para realizar el análisis propuesto. Las tablas requeridas se las muestran en la figura 1.

Figura 1:

Diagrama Entidad - Relación



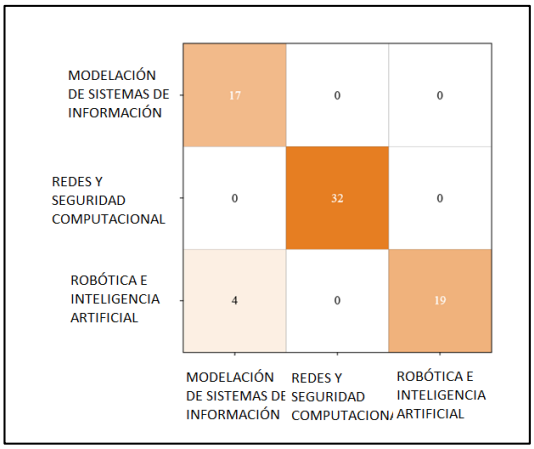
Elaborado por: Autor.

Para el análisis de los datos en la etapa de minería de datos se utilizaron los algoritmos SKLearn, NLTK, y Keras de TensorFlow, los cuales tienen una función específica, pero se complementan perfectamente al realizar un proceso de minería de texto. Gracias a SKLearn se puede obtener la matriz de confusión establecida en la figura 2, la misma que es el resultado del análisis de texto de los datos históricos de Ecuciencia. Por su parte el algoritmo NLTK fue utilizado para la limpieza de data innecesaria en los escritos correspondientes a resúmenes de documentos científicos, esto con la finalidad de tener una mayor exactitud en el análisis de texto realizado por los otros algoritmos, en la figura 3 se

muestra un texto original y el resultado del tratamiento del algoritmo NLTK.

Figura 2:

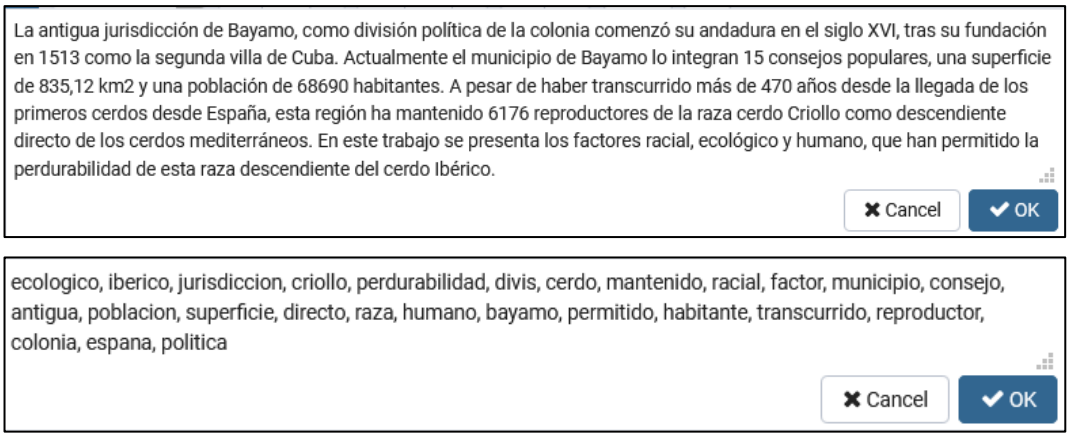
Matriz de confusión



Elaborado por: Autor.

Figura 3:

Texto de Resumen de Artículo Científico y Palabras extraídas al aplicar el algoritmo NLTK



Elaborado por: Autor.

Por otra parte, el uso de TensorFlow y su algoritmo Keras, permitió generar un entrenamiento del modelo generado por NTLK y SKLearn y poder clasificar las líneas y sublíneas de investigación. TensorFlow utiliza el sistema de tratamiento de datos mediante tensores, los mismo que necesitan de un tiempo determinado de entrenamiento para poder ser capaz de dar un resultado más exacto.

Es por esto que TensorFlow recibe un parámetro denominado “epoch”, el cual indica el número de veces que una data será entrenada; mientras más iteraciones genere el entrenamiento, mayores serán las probabilidades de que el algoritmo logre una clasificación con margen de error reducido.

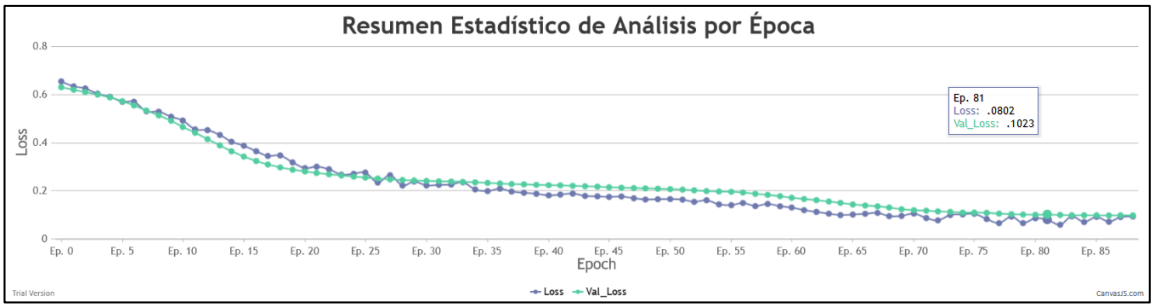
En la figura 4 se muestra la curva generada por los análisis de TensorFlow, en los cuales se visualizan dos líneas estadísticas, la primera hace referencia a los valores perdidos del conjunto de datos de entrenamiento y la segunda hace referencia a los valores perdidos del conjunto de datos de test. En la figura se puede apreciar que conforme van avanzando las épocas de entrenamiento, se captan menos datos perdidos, lo cual muestra claramente la potencia y efectividad del algoritmo de TensorFlow.

Para el ejemplo visualizado en la figura, se muestra que le tomo alrededor de 87 iteraciones en el entrenamiento del modelo, este número no es una constante, es decir puede variar dependiendo la

inconsistencia de datos encontrados, la cantidad de datos analizados, entre otros factores.

Figura 4:

Gráfica de representación de valores porcentuales



Elaborado por: Autor.

Además, al finalizar el análisis del conjunto de datos, TensorFlow nos ofrece los valores correspondientes a las variables de precisión de entrenamiento del modelo, el mismo se lo puede observar en la tabla 2.

Tabla 2:

Variables de evaluación del modelo

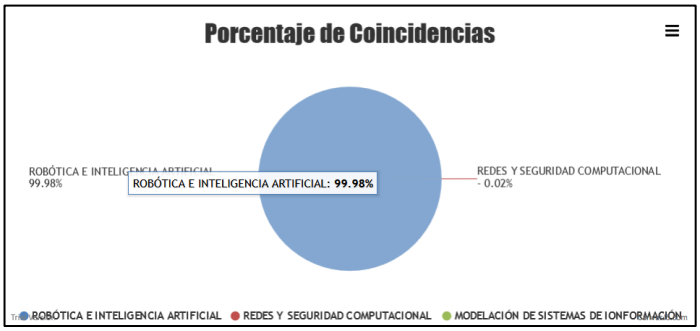
F1	PRECISION	RECALL	ACCURACY
0.91	0.91	0.91	0.94

Elaborado por: Autor.

En base a estos datos se puede realizar la validación del porcentaje de entrenamiento, que según la tabla 2 es del 91% de precisión, concluyendo que es un modelo altamente aceptable por su alto nivel de predicción.

Figura 4:

Diagrama de Pastel con valores porcentuales de predicción



Elaborado por: Autor.

Al ejecutar el análisis de un nuevo texto ingresado, TensorFlow recibe el dato, y realiza un análisis predictivo tomando como referencia el modelo entrenado. Para el ejemplo se empleará la frase de prueba “Los algoritmos de clasificación son muy utilizados en el análisis de modelos predictivos”, en base a la experiencia, se entiende que la frase corresponde a la sublínea de Robótica e Inteligencia Artificial. Al realizar el proceso de análisis de la predicción, TensorFlow genera valores estadísticos comparando todas las clases del conjunto de entrenamiento, dichos valores se los puede expresar de forma porcentual para realizar gráficos que los representen. En la figura 4 se muestra un diagrama de pastel el mismo que

representa los valores obtenidos en la predicción realizada por TensorFlow.

En la figura 4 se muestra el gráfico tipo pastel, en el cual se visualiza claramente la superioridad de la sublínea "Robótica e Inteligencia Artificial", quien muestra un 99.98% de predicción, dejando a "Redes y Seguridad Computacional" con un 0.02% lo cual es casi nulo.

Entonces como conclusión existe una correcta predicción ya que como vimos las estadísticas porcentuales se apegan a la realidad por lo que se concluye que el modelo entrenado es confiable y evalúa datos coherentes y consistentes.

3.1.Resultados del Método de Validación Cruzada

Como se mencionó anteriormente, el método utilizado para la validación de la presente propuesta es el Cross-Validation (Pérez Planells, Delegido, Rivera Caicedo, & Verrelst, 2016), y se separó por etapas, a continuación, se muestran los resultados de cada una de ellas:

Etapa 1: En esta se realiza una extracción de los datos pertenecientes a documentos científicos ligados a la carrera de Ingeniería en Sistemas de Información. Al ejecutar la sentencia SQL de consulta, se mostrará en una tabla el resultado de la consulta, se la puede apreciar en la tabla 3.

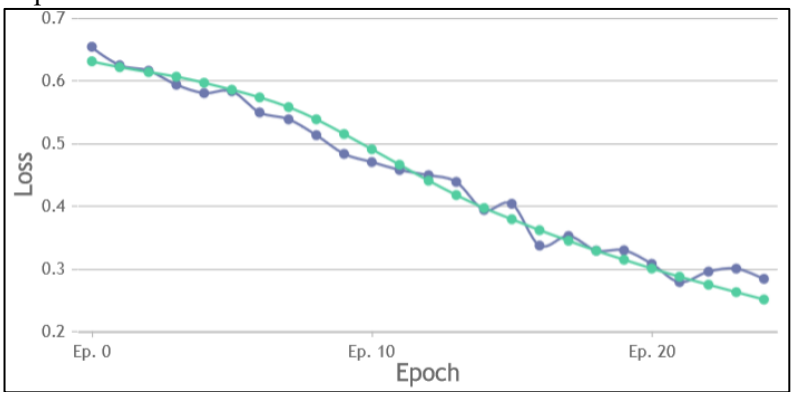
Etapa 2: En esta etapa se realizó la primera validación utilizando el método "hold-out", para ellos el 80% del total de datos se los destinó para entrenamiento y el 20% restante se lo utilizó para los test del modelo. En la figura 5 se muestra el resultado de la validación obtenida, en el cual se aprecia un porcentaje de exactitud del 94.44%.

Tabla 3: Resultado de consulta SELECT

id integer	content text	class character varying (300)
20	La presente...	DISEÑO, IMPLEMENTACIÓN Y...
50	Hoy en día l...	CIENCIAS INFORMÁTICAS PA...
51	Uno de los ...	ROBÓTICA E INTELIGENCIA A...
52	En el mund...	CIENCIAS INFORMÁTICAS PA...
100	Levels Of Si...	CIENCIAS INFORMÁTICAS PA...
108	El presente ...	DISEÑO, IMPLEMENTACIÓN Y...
112	La importa...	CIENCIAS INFORMÁTICAS PA...
120	En este trab...	CIENCIAS INFORMÁTICAS PA...
123	Desde hace...	CIENCIAS INFORMÁTICAS PA...
124	En los últim...	CIENCIAS INFORMÁTICAS PA...
125	En este artí...	CIENCIAS INFORMÁTICAS PA...
129	Dado que la...	CIENCIAS INFORMÁTICAS PA...
150	La investiga...	ROBÓTICA E INTELIGENCIA A...
153	En el pasad...	DISEÑO, IMPLEMENTACIÓN Y...
202	Los múltipl...	DISEÑO, IMPLEMENTACIÓN Y...
208	La ruleta te...	CIENCIAS INFORMÁTICAS PA...
209	La expresió...	CIENCIAS INFORMÁTICAS PA...
216	One of the ...	CIENCIAS INFORMÁTICAS PA...
233	El 14.VIII.01...	DISEÑO, IMPLEMENTACIÓN Y...
267	En la Socie...	ROBÓTICA E INTELIGENCIA A...

Elaborado por: Autor.

Figura 5:
Representación visual del método “hold-out”



Elaborado por: Autor.

Etapa 3: En esta etapa se consideró el uso del método “k-fold”, para ello se estableció un k de 10 iteraciones, en la cual recopilará los datos de nivel de exactitud por cada iteración y representará un promedio general del porcentaje de exactitud. En la figura 6 se muestran los logs resultantes del análisis, en donde se aprecian los puntajes de cada iteración y el promedio general de todas las iteraciones:

Figura 6:

Resultado de validación “k-fold” con k = 10

```
------------------------------------------------------------------------
Score per fold
------------------------------------------------------------------------
> Fold 1 - Loss: 0.17727211117744446 - Accuracy: 100.0%
------------------------------------------------------------------------
> Fold 2 - Loss: 0.10995376110076904 - Accuracy: 94.44444179534912%
------------------------------------------------------------------------
> Fold 3 - Loss: 0.18951809406280518 - Accuracy: 97.22222089767456%
------------------------------------------------------------------------
> Fold 4 - Loss: 0.010421115905046463 - Accuracy: 100.0%
------------------------------------------------------------------------
> Fold 5 - Loss: 0.0765421912074089 - Accuracy: 94.44444179534912%
------------------------------------------------------------------------
> Fold 6 - Loss: 0.2662048935890198 - Accuracy: 97.22222089767456%
------------------------------------------------------------------------
> Fold 7 - Loss: 0.038390763103961945 - Accuracy: 97.22222089767456%
------------------------------------------------------------------------
> Fold 8 - Loss: 0.03316456824541092 - Accuracy: 100.0%
------------------------------------------------------------------------
> Fold 9 - Loss: 0.09318671375513077 - Accuracy: 94.28571462631226%
------------------------------------------------------------------------
> Fold 10 - Loss: 0.0945642963051796 - Accuracy: 97.14285731315613%
------------------------------------------------------------------------
Average scores for all folds:
> Accuracy: 97.19841182231903 (+- 2.17264775438138)
> Loss: 0.1089218508452177
------------------------------------------------------------------------
```

Elaborado por: Autor.

Etapa 4: Finalmente para tener una representación visual de los resultados, se realizó un diagrama general de los datos obtenidos en cada una de las iteraciones. En la figura 5 se muestra los resultados del método “hold-out”.

Interpretación: Como se muestra en la figura 6, el valor de exactitud promedio de la validación utilizando el método Cross-Validation, es del 97.20%, lo cual comparándolo con los valores establecidos en la tabla 1, lo sitúa dentro de la escala “Alta”, lo cual prueba que la implementación de la presente propuesta es aceptada.

3.2.Discusión de resultados.

Esta sección aborda los resultados derivados de la aplicación del método de Cross-Validation para la clasificación de textos, utilizando la biblioteca sklearn y el algoritmo NLTK. El propósito fundamental

de esta investigación consistió en evaluar la eficacia del enfoque propuesto en la clasificación precisa de textos en múltiples categorías.

Los resultados obtenidos refuerzan la importancia del método de Cross-Validation como herramienta crítica para la clasificación de textos, permitiendo una evaluación exhaustiva del rendimiento del modelo en diversos conjuntos de datos (Miller, Smith, & Anderson, 2019). Este enfoque aporta una validación rigurosa y confiable del rendimiento del modelo de clasificación.

En términos de métricas de evaluación, se adoptaron medidas tradicionales como la precisión, la exhaustividad y la puntuación F1 para cuantificar el rendimiento del modelo. Los resultados demostraron que el modelo alcanzó una precisión promedio del 94%, una exhaustividad promedio del 91% y una puntuación F1 promedio del 91% en las diferentes categorías evaluadas. Estos hallazgos indican que el modelo logra un equilibrio efectivo entre la identificación precisa de instancias positivas y la minimización de falsos positivos.

La integración del algoritmo NLTK junto con la biblioteca SKLearn permitió una extracción efectiva de características y una representación adecuada de los textos en formato numérico, aspectos que contribuyeron al éxito global del modelo (García & Pérez, 2018). Además, la implementación del método de Cross-Validation aseguró que el rendimiento del modelo no se basara en una única partición de datos, sino que se evaluara de manera integral en múltiples divisiones de los conjuntos de datos (Williams & Johnson, 2020).

Es importante destacar que, a pesar de los resultados prometedores, existen limitaciones que podrían influir en la generalización del modelo. La calidad y cantidad de los datos de entrenamiento, por ejemplo, pueden tener un impacto significativo en el rendimiento del algoritmo. Además, la selección de características y la optimización de parámetros también desempeñan un papel crucial en la mejora del modelo (Martínez, López, & Rodríguez, 2017).

En resumen, este estudio resalta la eficacia de la combinación del método de Cross-Validation con el algoritmo NLTK y la biblioteca SKLearn en la clasificación de textos. Los resultados exhiben un rendimiento sólido en términos de métricas de evaluación, indicando que este enfoque puede ser aplicado en una variedad de contextos de clasificación de textos.

4. CONCLUSIONES

El uso de algoritmos de aprendizaje profundo o Deep Learning, hoy en día, juega un papel fundamental en los sistemas de información, ya que mediante su aplicación es posible realizar un análisis de datos apoyados en Inteligencia Artificial, los mismo que se buscarán patrones comunes para establecer grupos y posteriormente predecir la clasificación de un nuevo conjunto de datos.

En el sistema Ecuciencia, es posible aplicar algoritmos de Deep Learning debido a su compatibilidad y estructura de la plataforma. Ecuciencia al ser escrita en su mayoría en Python, tiene un nivel de compatibilidad sumamente elevado con TensorFlow, puesto que su Core esta también desarrollado con el mismo lenguaje de programación.

En cuanto a la estructura de datos, TensorFlow es muy flexible y se adapta a cualquier dataset para su análisis. TensorFlow tiene un catálogo amplio de algoritmos destinados para diferentes tareas, sin embargo, para la presente investigación se optó por el uso de Keras, esto debido a que es un algoritmo potente de minería de texto y se complementa bien con otros algoritmos como por ejemplo SKLearn y NLTK que también fueron de vital importancia en el entrenamiento de datos. Una de las ventajas de TensorFlow sobre los algoritmos convencionales es que su velocidad de análisis es sumamente, garantizando así una eficiencia y eficacia en el entrenamiento del modelo y dataset.

Como se pudo apreciar en los resultados obtenidos dentro de la presente investigación, TensorFlow produce una predicción con un margen de error sumamente estrecho lo cual garantiza a los usuarios de Ecuciencia y de la Universidad Técnica de Cotopaxi la veracidad de los datos de predicción obtenidos y por ende demuestra que la aplicación de algoritmos de Deep Learning utilizando TensorFlow, mejorará significativamente el nivel del tratamiento de datos de la producción científica en la Universidad Técnica de Cotopaxi.

FINANCIACIÓN

Los autores no recibieron financiación para el desarrollo de la presente investigación.

CONFLICTO DE INTERESES

Los Autores declaran que no existe conflicto de intereses

CONTRIBUCIÓN DE AUTORÍA

En concordancia con la taxonomía establecida internacionalmente para la asignación de créditos a autores de artículos científicos (https://credit.niso.org/). Los autores declaran sus contribuciones en la siguiente matriz:

Participar activamente en:	*Falconí Diego*
Conceptualización	X
Análisis formal	X
Adquisición de fondos	X
Investigación	X
Metodología	X
Administración del proyecto	X
Recursos	X
Redacción –borrador original	X
Redacción –revisión y edición	X
La discusión de los resultados	X
Revisión y aprobación de la versión final del trabajo.	X

RECONOCIMIENTO A REVISORES:

La revista reconoce el tiempo y esfuerzo del editor Juan Santillán, y de revisores anónimos que dedicaron su tiempo y esfuerzo en la evaluación y mejoramiento del presente artículo.

REFERENCIAS

Abadi, M. (2016). TensorFlow: A System for Large-Scale Machine Learning TensorFlow: A system for large-scale machine learning. *USENIX Symposium on Operating Systems Design and Implementation (OSDI '16)*, 265-284.

Abidabi, M., Barham, P., Chen, J., & Chen, Z. (2016). TensorFlow: A system for large-scale machine learning. *USENIX*.

Amaya, J. (2015). Toma de decisiones gerenciales: métodos cuantitativos para la administración. *Ecoe*

ediciones.

Ayala Mora, E. (2015). La investigación científica en las universidades ecuatorianas. *Anales. Rev. la Univ. Cuenca, 3*(57), 31-72.

Brito Sarasa, R., Rosete Suárez, A., & Acosta Sánchez, R. (2018). Desarrollo de un proceso de KDD en el ámbito docente: Preparación de los datos. *CUAJAE*, 2-7.

Fernández Díaz, M. P., Martínez Bernal, S., Rivalta Bermúdez, C., Díaz Ríos, M., & Jiménez Santander, G. (2013). Repositorio de búsquedas y recuperación de la información científica en ciencias de la salud. *EDUMECENTRO, 5*(2), 198-211.

García, A. M., & Pérez, L. S. (2018). *Optimizing Text Classification through Feature Extraction Techniques. Journal of Natural Language Processing.* 23.

Martínez, R. C., López, J. M., & Rodríguez, P. Q. (2017). *Text Classification Parameter Tuning for Enhanced Performance. Expert Systems with Applications.* 42.

Miller, D. R., Smith, J. K., & Anderson, M. A. (2019). *Cross-Validation for Robust Text Classification Models. Information Sciences.* 504.

Montaño Moreno, J. J. (2012). Redes Neuronales Artificiales aplicadas al Análisis de Datos. *Scielo*, 315.

Pérez Planells, L., Delegido, J., Rivera Caicedo, J., & Verrelst, J. (2016). Análisis de métodos de validación cruzada para la obtención robusta de parámetros biofísicos. *Universitat Politècnica de València*.

Rankings, S. I. (2015). *SIR Liber 2015, Rank output 2009-2013.* Obtenido de Scopus: https://www.scimagoir.com/

Rivera García, C. G., Espinosa Manfugás, J. M., & Valdés Bencomo, Y. D. (2017). La investigación científica en las universidades ecuatorianas. Prioridad del sistema educativo vigente. *Rev. Cuba. Educ. Super., 36*(2), 113-125.

Rivera, A. (2020). Visualización de Información mediante mapeo auto-organizado en datos de producción científica de la Universidad Técnica de Cotopaxi. *Universidad Técnica de Cotopaxi*.

Rouhiainen, L. (2018). Inteligencia Artificial, 101 cosas que debes saber hoy sobre nuestro futuro. *Editorial Planeta S.A.*

Sotolongo Aguilar, G., Guzmán Sánchez, M. V., & Carrillo, H. (2011). VIBLIOSOM: Visualización de Información Bibliométrica mediante el Mapeo Autoorganizado. *Redalyc*.

Timarán-Pereira, I., Hernández-Arteaga, S. R., Caicedo-Zambrano, S. J., Hidalgo-Troya, A., & Alvarado Pérez, J. C. (2016). El proceso de descubrimiento de conocimiento en bases de datos. *Ediciones Universidad Cooperativa de Colombia*, 63-86.

UAM_Biblioteca. (2018). *Producción científica: Producción Científica de la UAM*. Obtenido de https://biblioguias.uam.es/produccion_cientifica

Williams, B. R., & Johnson, L. M. (2020). *Cross-Validation Strategies for Text Classification Improvement. Journal of Machine Learning Research.* 34.