# Aplicaciones de Machine Learning: Caso de Agrupacion de datod mediante NLP con técnicas de aprendizaje supervisado

Ariana Rosado

May 2024

**Abstract**

El estudio desarrollado busca identificar nuevas aplicaciones de conocimientos de aprendizaje automático en la ciencia e ingeniería de la computación para la industria. Esto se logra mediante el procesamiento de textos en lenguaje natural a través del aprendizaje supervisado, una rama de la inteligencia artificial, para clasificar y etiquetar los textos de forma binaria. Se aplicó en el contexto de algoritmos de optimización orientados a la investigación de operaciones. El método incluyó la minería de datos bibliográficos de más de 1000 registros en bases de datos relevantes como Scopus y Web of Science, siguiendo la metodología SEMMA para identificar patrones y predecir tendencias en aplicaciones de aprendizaje automático en la industria con procesamiento de lenguaje natural. Utilizando un algoritmo de aprendizaje supervisado y el lenguaje de programación Python, se optimizó un modelo de clasificación binaria, el cual mostró un 99

Palabras Clave : aprendizaje automático, algoritmos, procesamiento de lenguaje natural, tendencias de aprendizaje, industria 4.0, aprendizaje supervisado.

## 1 Introduction

El artículo aborda el tema de las aplicaciones de Machine Learning, específicamente en el contexto de la agrupación de datos mediante el Procesamiento del Lenguaje Natural (NLP, por sus siglas en inglés). El objetivo principal es identificar nuevas formas de aplicar el conocimiento de Machine Learning en la industria, particularmente a través del procesamiento de datos bibliográficos utilizando técnicas de aprendizaje supervisado.

Se describe un método que incluye la minería de datos bibliográficos de más de 1000 registros utilizando bases de datos relevantes como Scopus y Web of Science. La metodología SEMMA se emplea para identificar patrones en los datos y predecir tendencias en aplicaciones de Machine Learning en la industria con procesamiento de lenguaje natural. Se utiliza un algoritmo de aprendizaje

supervisado para la optimización del modelo de clasificación binaria utilizando Python como lenguaje de programación.

Los resultados de la investigación muestran un 99

El artículo también proporciona una revisión de la literatura sobre Machine Learning, destacando su evolución desde los años 50 y los diferentes tipos de aprendizaje, como supervisado, no supervisado, semi supervisado y activo. Se mencionan los pasos para entrenar un modelo en PNL, así como los conceptos básicos del Procesamiento del Lenguaje Natural y los tipos de algoritmos utilizados en este campo.

Finalmente, se describe la metodología aplicada en el estudio, que incluye la identificación del problema, la recolección y preparación de datos, el entrenamiento del modelo y la evaluación de su rendimiento. Se concluye que el modelo propuesto tiene un alto grado de precisión en la predicción, lo que sugiere su utilidad en la identificación de nuevas aplicaciones de Machine Learning en la industria.

# 2 REVISION DE LITERATURA

# 3 Machine Learning

El origen del aprendizaje automático (machine learning) se remonta a varias décadas atrás, con contribuciones significativas de diferentes investigadores en el campo de la inteligencia artificial y la estadística. No hay un único origen o autor definitivo, ya que el aprendizaje automático ha evolucionado a lo largo del tiempo gracias a numerosos avances teóricos y prácticos. Se menciona que las bases teóricas para la idea de una máquina universal desarrollada por Turing. Esta noción se considera un precursor crucial para el desarrollo de la computación y el aprendizaje automático.

# 4 Tipos de Aprendizajes

Existían 3 tipos de aprendizaje el Aprendizaje supervisado, Aprendizaje no Supervisado y Refuerzo según la naturaleza de los datos que son recibidos. En el transcurso del tiempo se ha ido evolucionando a un nuevo aprendizaje que es semi supervisados y el aprendizaje activo, aunque aun se ha encontrado que Bernard Schölkopf: ha realizado importantes investigaciones en el campo del aprendizaje se-mis supervisado junto con sus colaboradores. Su trabajo en métodos de aprendizaje basados en el núcleo ha tenido un impacto en la creación de algoritmos semi supervisados Para predecir un valor continuo o numérico, se utiliza la regresión. El objetivo de un algoritmo de regresión es descubrir una función que conecte las variables de entrada una variable de salida continua. Por ejemplo, la regresión se puede utilizar para predecir el precio de una casa en función de su ubicación, tamaño, número de habitaciones y otros factores.

$y = w\ x + b$ (1).

Donde w representa la pendiente (inclinación) de la línea recta y b es la intersección con el eje y. Cuando se intenta predecir una categoría o clase específica, se utiliza la clasificación [20]. Un algoritmo de clasificación tiene como objetivo asignar objetos de entrada en una clase o categoría predeterminada en función de

# 5 Pasos para entrenar un modelo NLP

Para ello existe una serie de pasos para entrenar un modelo, un proceso utilizado para realizar las predicciones con la aplicación de los algoritmos. Donde los datos recibidos y patrones son relacionados para entrenarlos, obteniendo nuevos datos, esta herramienta de uso matemática o algorítmica se puede aplicar para que la maquina aprenda y tomar decisiones a través de la información, para ello machine learning nos da pautas para entrenar un modelo con la finalidad que la maquina aprenda, una guía para aplicar en cualquier modelo supervisado que se menciona a continuación:

a) Tener claro el problema para correlacionar el objetivo a tratar.

b) Colección o recolección de datos, una parte que se podria decir es donde se lleva la mayor cantidad de tiempo.

c) Pre procesamiento y procesamiento de datos, con la ayuda de EDA.

d) Selección del modelo para la aplicación del algoritmo.

e) Evaluación del modelo para conocer el porcentaje de asertividad o predicción.

# 6 Procesamiento de Lenguaje Natural

Estos representan cada palabra mediante embeddings o vectores numéricos, que codifican su significado y función en las oraciones. Actores que integran dentro del procesamiento de Lenguaje Natural. Se conoce que existe una intervención de machine learning, Deep learning, procesamiento de lenguaje natural sin embargo en lo que se diferencia es la forma de interpretar el lenguaje natural, es decir el idioma lingüístico, por lo tanto, los actores que intervienen en NLP es:

a) NLU Natural language understanding

b) NLG Natural language generation

c) TM Tex mining

El TM es una minería de datos, descubre información asociada que se aplica en el texto que permite estructurar y transformar el texto. NLU encargada de entender el texto, una comprensión lingüística aplicado en las traducciones automática, las preguntas y respuestas, o las categorizaciones de textos sin olvidar el análisis de sentimiento. NLG encargada de la generación del lenguaje, ayudando a las creaciones de interfaces que realizan conversaciones. En la cual posee sus técnicas de procesamiento de lenguaje natural en la inteligencia artificial para texto

# 7 APLICANDO LA METODOLOGIA

La aplicación de herramientas de enseñanza esta contribuyendo significativamente, con gran uso en la medicina sin embargo ya se esta extendiendo la aplicación en otros campos, que en nuestro caso es identificar nuevas aplicaciones de conocimientos de machine learning en la enseñanza para la industria, identificación de los patrones y predicción de la tendencia en aplicaciones utilizando la metodología SEMMA.

Fase 1- Identificación del problema, esto permitió detectar la solución necesaria para alcanzar el objetivo e identificar cual sería la muestra para tomar, necesaria a pensar que modelo de aprendizaje automático se pudo optar, que sea útil en la identificación de criterios en la industria 4.0 para las líneas de investigación basado en machine learning con el procesamiento de texto (NLP).

Fase 2 – Implica la colección de datos, para la llamada Data set, en esta fase se recolectó aproximadamente mas de 1800 artículos científicos de alto impacto desde varias fuentes confiables en dos mecanismos:

a) El uso de los gestores bibliográficos.

b) Descargas de artículos científicos en revistas nacionales e internacionales, incluso de publicaciones de eventos científicos en scopus.

Fase 3 – Preparación de datos para su manipulación de datos antes de darle un entrenamiento a la maquina, debe ser pre procesadas, sin duda un trabajo de hormiga en la fase 2, que una vez terminada se logró recolectar aproximadamente 1800 registros que correspondía de artículos científicos, nos tomo 7 meses en solo recolectar sin embargo después de revisarla y verificar fue-ron 1064 registros que cumplían de ser producciones de alto impacto relacionadas a la industria de 4.0 con algoritmos de investigación de operaciones, algoritmos de optimización y sus aplicaciones, para esta fase se eliminaron lo que no cumplían con el parámetro de criterios bibliométricos, además de campos innecesarios que estaban incompletos. Una vez listo se procede a cargar los datos de la llamada Data set, utilizamos el lenguaje de Python con una herramienta científico de datos para el debido roceso de exploración, preprocesamiento y procesamiento.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
```

Figure 1: Librerías de Python a utilizar para manipulación de datos para la utilización de código y correr el proceso de exploración. 2023.

En la Fig. 1 nos muestra una parte de la data con el uso de comando Python, un lenguaje que ayuda a proporcionar una inspección directa e indirecta de datos, para la aplicación exploración de análisis de datos (EDA), el inicio de encontrar correlación de datos para aplicar el pre-procesamiento.

Fase 4 – Luego del pre-procesamiento y procesamiento se procede a modelar el algoritmo de clasificación binaria y red neuronal.Como primera instancia se

elimina las columnas no necesarias, luego revisamos para verificar el tratamiento de los datos nulos como se observa la figura 4. Al realizar el proceso quedaron 1058 registros y 22 columnas. Luego de ello se crea un nuevo archivo de datos procesados para la aplicación binaria con sus librerías de keras, con esto se entrena la variable a utilizar dividiendo los inputs y los outputs, eso facilita en el testeo, prediciendo con un algoritmo de regresión, completando el entrenamiento. La utilización de las fórmulas matemáticas es indispensable, se utilizará la de regresiones al modelo binario.

Y=b0+b1*X1 (2)

Y=b0+b1*X1+b2*X2.....+bn*Xn (3)

Con la intención de utilizar una vez detectado la variable independiente.

# 8 RESULTADOS

El desarrollo bibliométrico permitió identificar las nuevas aplicaciones de conocimientos de machine learning en la enseñanza que permite abordar el observatorio de creación de nuevas líneas de investigación para la industria obteniendo un nuevo conocimiento descubriendo cuando se realiza el análisis de contenidos, aplicado con Python con herramienta tecnológica, aplicando el tipo supervisado con el algoritmo de clasificación de procesamiento de lenguaje natural, dando el 99 del porcentaje de predicción que el conocimiento es “machine learning y modelos predictivos”, utilizando aplicaciones informáticas sin embargo realizando con la aplicación del algoritmo de redes neuronal nos vota el 69 del porcentaje, aplicando 16 capas de entrada, podemos concluir que a mayor capas mejores resultados.