Aplicaciones de Machine Learning: Caso de Agrupación de datos mediante NLP con técnicas de aprendizaje supervisado

Oliver Alexander Chiriboga Mero
Universidad Laica Eloy Alfaro de Manabí Extensión El Carmen
e1311270944@live.uleam.edu.ec

ABSTRACT

Se identifico aplicaciones para el conocimiento de machine Learning, se agruparon datos para ser procesados a través del aprendizaje máquina de tipo supervisado. El método incluyó minería de datos de más de 1000 registros bibliográficos publicados. Se procedió al uso de la metodología SEMMA para identificación los datos y predecir la tendencia en aplicaciones de Machine Learning, mediante el algoritmo de aprendizaje supervisado y el modelo de clasificación binaria junto a la herramienta de programación Python, se mostró una eficacia del 99 pociento en los resultados de la predicción.

KEYWORDS

Machine learning, algoritmos, procesamiento de lenguaje natural, tendencias de aprendizaje, industria 4.0, aprendizaje supervisado.

1 INTRODUCCIÓN

En la actualidad la tecnología está produciendo nuevos descubrimientos en diversas áreas. La inteligencia artificial en la computación y la industria está impulsando nuevas innovaciones. Abriendo caminos y diferentes métodos de estudios por medio de Machine Learning en el ámbito educativo. Contribuyendo significativamente en diversos países del mundo para un mejor aprendizaje en los campos de diversos conocimientos. En la educación se desarrolla diversos factores de la búsqueda continua del aprendizaje.

Analizando y comparando habilidades en diversos campos. En los empleadores se contribuye a una mejor efectividad mejorando las aplicaciones en las industrias. La docencia abre nuevos paradigmas por lo que se deben adaptar a los nuevos cambios, tomando decisiones para la gestión de la educación superior y la aplicación de nuevos algoritmos. Según el banco mundial esto ha llevado a transformar la industria, expandiéndose a varios países aportando altos y bajos ingresos a los medios.

Nuevos descubrimientos han surgido en la ingeniería, presentando procesos innovadores ante otros aprendizajes. Como lo es nuevas tendencias tecnológicas, nuevas disciplinas, bases de datos, las maquinas comprenden e interpretan el pensamiento humano, entre otros. El procesamiento natural es beneficioso debido a la extracción de información que se puede lograr de diversos documentos, obteniendo conocimientos que contribuyen a la orientación de la investigación de las aplicaciones como Machine Learning. Esto es posible dado que se analizaron 1800 datos de los cuales 1000 datos fueron válidos para la implementación de la metodología SEMMA.

2 REVISIÓN DE LITERATURA

Machine Learning

En la década de los 50 surgió el aprendizaje automático como una razón para que las maquinas piensen y razonen como los seres humanos. El origen del aprendizaje automático no está completamente definido debido a los constantes cambios en la evolución de los avances teóricos.

Tipos de Aprendizajes

Aprendizaje Supervisado, Aprendizaje no Supervisado, Refuerzo según la naturaleza de los datos que son recibidos, Aprendizaje semi supervisado, Algoritmo de regresión, Árboles de decisión, Bosques aleatorios, Redes neuronales, Algoritmos de clasificación.

Pasos para entrenar un modelo en NLP

a) Tener claro el problema para correlacionar el objetivo a tratar.
b) Recolección de datos, una parte que se podría decir es donde se lleva la mayor cantidad de tiempo.
c) Preprocesamiento 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.

Procesamiento de Lenguaje Natural

Representan cada palabra mediante vectores numéricos que se integran dentro del procesamiento del lenguaje natural para interpretar el idioma lingüístico. Actores que intervienen:
a) NLU Natural language understanding
b) NLG Natural language generation
c) TM Tex mining.
El TM es una minería de datos, que permite estructurar y transformar el texto. NLU encargada de entender el texto, aplicado en las traducciones automática, además de preguntas y respuestas. NLG encargada de las creaciones de interfaces que realizan conversaciones.

3 APLICANDO LA METODOLOGÍA

Se llevo 4 momentos considerados como fases basándose en la Metodología SEMMA (Sample Explore Modify Model Assess).

Fase 1- Identificación del problema, esto permitió detectar la solución necesaria para alcanzar el objetivo e identificar cual sería la muestra para tomar.

Fase 2 – Implico la colección de datos (Data set), en esta fase se recolectó 1800 artículos científicos en dos mecanismos:
a) El uso de los gestores bibliográficos.
b) Descargas de artículos científicos.
Ambos mecanismos pasaron por un proceso de validación. Esos datos fueron colocados en una matriz de metaanálisis con un tipo de archivo de .csv con la finalidad de automatizar el proceso de validación.

Fase 3 – Preparación de datos para su manipulación y darle un entrenamiento a la máquina (aproximadamente 1800 registros que correspondía de artículos científicos).

Fase 4 – Luego del preprocesamiento y procesamiento se procedió a modelar el algoritmo de clasificación binaria y red neuronal. Se elimino columnas innecesarias, luego los datos nulos. Al realizar el proceso quedaron 1058 registros y 22 columnas. Luego se entrenó la variable a utilizar dividiendo los inputs y los outputs, facilitando el testeo y predicción en el entrenamiento. La fórmula matemática que se usó:

$$Y = b_0 + b_1 X_1 \tag{1}$$

$$Y = b_0 + b_1 X_1 + b_2 X_2 + \cdots + b_n X_n \tag{2}$$

Con la intención de utilizar una vez detectado la variable independiente.
Fase 5- Se evaluó la selección de entradas, alcanzando el 99 porciento de predicción usando las variables de machine learning y modelos predictivos. Para visualización se aplicó la matriz confusión que ayudo al rendimiento del modelo de clasificación. Se aplicaron otras experimentaciones con otros modelos donde se apreció mejores resultados en los modelos de algoritmo binario y de redes neuronales. Con la ayuda del lenguaje de Python se pudo apreciar más resultados.

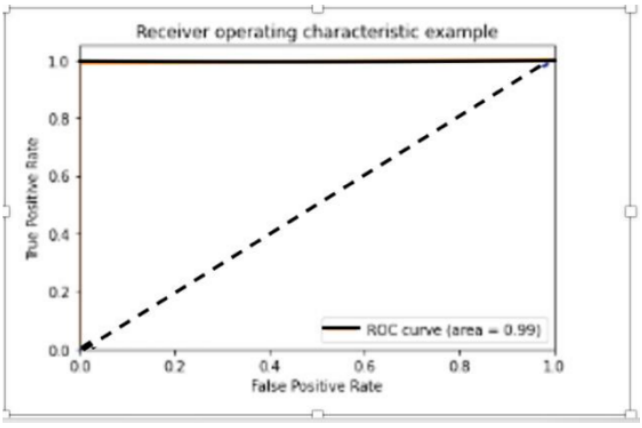

Figure 1: Resultado aplicando algoritmo de clasificación binaria. 2023

Descripción	Algoritmo de Clasificación Binaria	Algoritmo de red neuronal
accuracy_score:	0.99	0.69
precision_score:	1.00	1.00
recall_score:	0.98	1.00

Table 1: Tabla comparativa para visualizar la calidad de los algoritmos de clasificación binaria y una red neuronal.

4 RESULTADOS

El desarrollo bibliométrico permitió abordar nuevos conocimiento y descubrimientos en el análisis de contenidos. Por medio de Python y el algoritmo de clasificación de NLP se dio una predicción del 99 porciento, sin embargo, con el algoritmo de redes neuronal dio 69 porciento, aplicando 16 capas de entrada, concluyendo que a mayor capa mejores resultados.

ISBN: 978-1-950492-76-3
ISSN: 2771-6333

https://doi.org/10.54808/CICIC2024.01.339