# UNIVERSIDAD TECNOLÓGICA DE CAMPECHE

## INGENIERÍA EN DESARROLLO Y GESTIÓN DE SOFTWARE

**ACTIVIDAD:**

INVESTIGACION

## ASIGNATURA:

EXTRACCION DE CONOCIMIENTOS DE BASE DE DATOS

**NOMBRE DEL ALUMNO**

CRISTIAN GONZÁLEZ JIMÉNEZ

## NOMBRE DEL DOCENTE:

MARÍA LOURDES CÁRDENAS MALDONADO

## GRADO Y GRUPO:

9no “B”

## PERIODO:

MAYO – AGOSTO 2023

**TEMA 1. ALGORITMOS DE APRENDIZAJE NO SUPERVISADO**.

COMPRENDER LOS USOS Y OBJETIVOS DEL ANALISIS NO SUPERVISADO.

El **aprendizaje no supervisado** es una de las formas en que Machine Learning (ML) "aprende" los datos. El aprendizaje no supervisado tiene datos sin etiquetar que el algoritmo tiene que intentar entender por sí mismo. El aprendizaje supervisado es en el que se etiquetan los conjuntos de datos para que haya una clave de respuestas con la que la máquina pueda medir su precisión. Si Machine Learning fuera un niño que aprendiera a andar en bicicleta, el aprendizaje supervisado es el padre que corre detrás de la bicicleta y la sostiene en posición vertical. El aprendizaje no supervisado consiste en entregar la bicicleta, darle palmaditas en la cabeza al niño y decirle "buena suerte".

El objetivo es simplemente dejar que la máquina aprenda sin ayuda o indicaciones de los los científicos de datos. En el camino, también deberá aprender a ajustar los resultados y agrupaciones cuando haya resultados más adecuados, permitiendo que la máquina comprenda los datos y los procese como mejor le parezca.

El aprendizaje no supervisado se utiliza para explorar datos desconocidos. Puede revelar patrones que podrían haberse pasado por alto o examinar grandes conjuntos de datos que serían demasiado para que los abordara una sola persona.

Tendremos que comprender el aprendizaje supervisado. Si una computadora estuviera aprendiendo a identificar frutas en un entorno de aprendizaje supervisado, se le darían imágenes de ejemplo de frutas etiquetadas, a esto se le llama datos de entrada. Por ejemplo, las etiquetas dirían que los plátanos son largos, curvos y amarillos, que las manzanas son redondas y rojas, mientras que una naranja es esférica, de aspecto ceroso y anaranjada.

Después de un tiempo conveniente, la máquina debería poder identificar con seguridad qué fruta es cuál, basándose en esos descriptores. Si se le presenta una manzana, por ejemplo, podría decir con seguridad que no es de color naranja, por lo tanto, no es una naranja, pero también que no es amarilla y larga, por lo tanto, no es un plátano. Entonces, será una manzana debido a que es redonda y roja.

IDENTIFICAR ALGORITMOS DE AGRUPACIÓN.

es una tarea de aprendizaje automático no supervisada. Es posible que también lo hayas escuchado como análisis de agrupación debido a la forma en que funciona este método.

El uso de un algoritmo de agrupamiento significa que le darás al algoritmo una gran cantidad de datos de entrada sin etiquetas y te permitirá encontrar cualquier agrupación en los datos que se pueda.

Esas agrupaciones se denominan agrupaciones. Una agrupación es un grupo de datos que son similares entre sí en función de su relación con los datos circundantes. El agrupamiento se utiliza para cosas tales como la ingeniería de características o el descubrimiento de patrones.

Cuando empiezas con datos de los que no sabes nada, el agrupamiento puede ser un buen punto de partida para obtener información.

**1.3.  IDENTIFICAR ALGORITMOS DE REDUCCIÓN DE DIMENSIONALIDAD.**

los algoritmos que se emplean en el Machine Learning son capaces de extraer información importante de conjuntos de datos que cuentan con muchas características, ya sean tablas con muchas columnas y filas o imágenes con millones de píxeles. Si a este hecho se le suman los grandes avances de computación en la nube, el resultado es que cada vez se pueden ejecutar modelos más grandes de aprendizaje automático con una gran potencia.

No obstante, cada característica que se agrega aumenta la complejidad del ejecutable, lo que hace que localizar la información gracias a estos potentes algoritmos también sea complicado. La solución es la reducción de dimensionalidad que consiste en emplear un conjunto de técnicas para eliminar características excesivas y no necesarias de los modelos de Machine Learning.

Asimismo, la reducción de dimensionalidad reduce de forma severa los costes del aprendizaje automático y permite la resolución de problemas complejos con modelos simples.

Esta técnica es especialmente útil en los modelos predictivos, ya que son conjuntos de datos que contienen un elevado número de características de entrada, y hace más complicada su función.

Por tanto, la técnica de reducción de dimensionalidad se define como una forma de convertir un conjunto de datos de dimensiones elevadas en un conjunto de datos de dimensiones menores, asegurando que la información que proporciona en similar en ambos casos. Como se ha mencionado, esta técnica se emplea a menudo en el aprendizaje automático para obtener un modelo predictivo más ajustado mientras se resuelven los problemas de regresión y clasificación que presentan los algoritmos.

Los datos de alta dimensión, el reconocimiento de voz, visualización de datos, reducción de ruido o el procesamiento de señales, entre otros, son los principales campos de aplicación de la reducción de dimensionalidad.

**TEMA 2.  METRICAS DE EVALUACIÓN DE MODELOS DE PROCESAMIENTO DE DATOS.**

**2.1.  IDENTIFICAR METRICAS DE EVALUACIÓN DE MODELOS DE AGRUPACIÓN Y REDUCCIÓN DE DIMENSIONALIDAD**

Métricas de evaluación de modelos

Cuando se habla de validar o evaluar un modelo de aprendizaje automático, es importante saber que las técnicas de validación empleadas no sólo ayudan a medir el rendimiento, sino que también contribuyen en gran medida a entender el modelo a un nivel más profundo. Esta es la razón por la que se dedica una cantidad significativa de tiempo al proceso de validación de resultados y evaluación del modelo mientras se construye un modelo de aprendizaje automático.

La validación de resultados es un paso muy crucial, ya que garantiza que nuestro modelo da buenos resultados no sólo en los datos de entrenamiento, sino, lo que es más importante, también en los datos reales o de prueba.

En el caso del aprendizaje supervisado, la evaluación se realiza sobre todo midiendo las métricas de rendimiento como la exactitud, la precisión, exhaustividad, el AUC, etc. en el conjunto de entrenamiento y en los conjuntos de espera. Estas métricas de rendimiento ayudan a decidir la viabilidad del modelo. A continuación, podemos ajustar los hiper parámetros y repetir el mismo proceso hasta conseguir el rendimiento deseado.

Sin embargo, en el caso del aprendizaje no supervisado, el proceso no es muy sencillo, ya que no tenemos la verdad de base (las etiquetas). En ausencia de etiquetas, es muy difícil identificar cómo se pueden validar los resultados.

Por lo tanto, es en general difícil evaluar la calidad de un algoritmo no supervisado debido a la ausencia de una métrica de bondad explícita como la que se utiliza en el aprendizaje supervisado.

Reducción de dimensionales

La reducción de dimensiones es un algoritmo usado como una etapa de pre proceso para condensar la información de un conjunto de variables y escoger un subconjunto de variables, de tal manera, que el espacio de características quede óptimamente reducido de acuerdo a un criterio de evaluación, cuyo fin es distinguir el subconjunto que representa mejor el espacio inicial de entrenamiento. Cada característica que se incluye en el análisis, puede llegar a incrementar el costo y el tiempo de proceso de los sistemas, por lo que hay una fuerte motivación para diseñar e implementar sistemas con pequeños conjuntos de características. Para ello existen técnicas que facilitan este pre proceso, sin embargo, muchas de esas técnicas pueden solamente manipular ciertos tipos de datos.

**2.2.  IDENTIFICAR EL PROCESO DE ENTRENAMIENTO Y EVALUACIÓN DE AGRUPACIÓN Y REDUCCIÓN DE DIMENSIONALIDAD.**

Los métodos no supervisados son algoritmos que basan su proceso de entrenamiento en un juego de datos sin etiquetas o clases previamente definidas. Es decir, a priori no se conoce ningún valor objetivo o de clase, ya sea categórico o numérico. El aprendizaje no supervisado está dedicado a las tareas de agrupamiento, también llamadas clustering o segmentación, donde su objetivo es encontrar grupos similares en el conjunto de datos.

El análisis de agrupación consiste en reunir elementos a partir de las características que comparten, con el propósito de realizar una evaluación profunda comparando los resultados obtenidos con la sección a la que pertenece cada elemento. 

La reducción de dimensionalidad puede parecer que no tenga sentido, ya que cuando se presenten demasiadas funciones, también se necesitara un modelo más complejo, con más datos de entrenamiento y más potencia de cálculo para para entrenar al modelo de manera adecuada.

La reducción de dimensionalidad se encarga de identificar y eliminar las características que disminuyen el rendimiento del modelo de aprendizaje automático. Además, son varias las técnicas de dimensionalidad que se verán a continuación, y cada una de ellas es útil para determinadas situaciones.

**2.3.  IDENTIFICAR EL PROCESO DE OPTIMIZACIÓN DE MODELOS DE AGRUPACIÓN Y REDUCCIÓN DE DIMENSIONALIDAD.**

Los modelos de agrupación en clústeres se centran en la identificación de grupos de registros similares y en el etiquetado de registros según el grupo al que pertenecen. Esto se lleva a cabo sin la ventaja de disponer de conocimientos previos sobre los grupos y sus características. De hecho, puede que ni siquiera sepa exactamente cuántos grupos va a buscar. Esto es lo que diferencia a los modelos de agrupación en clústeres de otras técnicas de aprendizaje de máquinas: no hay campo objetivo o de salida predefinidos para el modelo que se va a predecir. A menudo se hace referencia a estos modelos como modelos de **aprendizaje no supervisado**, ya que no hay ningún estándar externo con el que juzgar el rendimiento de la clasificación del modelo. No hay respuestas correctas o incorrectas para estos modelos. Su valor viene determinado por su capacidad de capturar agrupaciones interesantes en los datos y proporcionar descripciones útiles de dichas agrupaciones.

Los métodos de reducción de la dimensionalidad más comunes y conocidos son los que aplican transformaciones lineales, como:

**PCA (Análisis de Componentes Principales):** Popularmente utilizado para la reducción de la dimensionalidad en datos continuos, PCA rota y proyecta los datos a lo largo de la dirección de la varianza creciente. Las características con la máxima varianza son los componentes principales.

**Análisis factorial:** una técnica que se utiliza para reducir un gran número de variables a un número menor de factores. Los valores de los datos observados se expresan como funciones de una serie de posibles causas para encontrar cuáles son las más importantes.

**LDA (Análisis Discriminante Lineal):** proyecta los datos de forma que se maximice la separabilidad de las clases. Los ejemplos de la misma clase se colocan muy juntos por la proyección. Los ejemplos de clases diferentes se colocan lejos por la proyección.

**Análisis de componentes independientes:** transforma el conjunto de datos en columnas de componentes independientes. La separación ciega de fuentes y el "problema del cóctel" son otros de sus nombres.

CONCLUSIONES

Bibliografía

Betancourt, J. C. (22 de Agosto de 2023). *freecodecamp*. Obtenido de freecodecamp: https://www.freecodecamp.org/espanol/news/8-algoritmos-de-agrupacion-en-clusteres-en-el-aprendizaje-automatico-que-todos-los-cientificos-de-datos-deben-conocer/

Rueda, J. F. (22 de Agosto de 2023). *Health Data Miner*. Obtenido de Health Data Miner: https://healthdataminer.com/data-mining/aprendizaje-supervisado-y-no-supervisado/#:~:text=Los%20m%C3%A9todos%20no%20supervisados%20(unsupervised,ya%20sea%20categ%C3%B3rico%20o%20num%C3%A9rico.

Softtek. (22 de Agosto de 2023). *Softtek*. Obtenido de Softtek: https://blog.softtek.com/es/la-reduccion-de-dimensionalidad-en-el-machine-learning

Tique, F. G. (22 de Agosto de 2023). *Search Medium*. Obtenido de Search Medium: https://medium.com/analyst-ryma/aprendizaje-no-supervisado-reducci%C3%B3n-de-dimensiones-67d3aa2385f3