# 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.

        1.1.  COMPRENDER LOS USOS Y OBJETIVOS DEL ANALISIS NO SUPERVISADO.

Es una de las formas en que [Machine Learning](https://www.tibco.com/es/reference-center/what-is-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](https://www.tibco.com/reference-center/what-is-supervised-learning) 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](https://www.tibco.com/es/reference-center/what-is-a-data-scientist). 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.

        1.2.  IDENTIFICAR ALGORITMOS DE AGRUPACIÓN.

Existen diferentes tipos de algoritmos de agrupamiento que manejan todo tipo de datos únicos.

Basado en densidad

En el agrupamiento basado en densidad, los datos se agrupan por áreas de altas concentraciones de puntos de datos rodeadas por áreas de bajas concentraciones de puntos de datos. Básicamente, el algoritmo encuentra los lugares que son densos en puntos de datos y los llama grupos.

Lo bueno de esto es que los grupos pueden tener cualquier forma. No estás limitado a condiciones esperadas.

Los algoritmos de agrupamiento de este tipo no tienen en cuenta los valores atípicos en los grupos, por lo que se ignoran.

Basado en la distribución

Con un enfoque de agrupamiento basado en la distribución, se considera que todos los puntos de datos forman parte de un grupo según la probabilidad de que un punto pertenezca a un grupo determinado.‌‌

Funciona así: hay un punto central y, a medida que aumenta la distancia de un punto de datos desde el centro, la probabilidad de que forme parte de ese grupo disminuye.

‌‌

Si no estás seguro de cuál podría ser la distribución de tus datos, deberías considerar un tipo diferente de algoritmo.

Basado en Centroides

El agrupamiento basado en centroides es la que probablemente has escuchado más. Es algo sensible a los parámetros iniciales que le das, pero es rápida y eficiente.

Estos tipos de algoritmos separan puntos de datos en función de múltiples centroides en los datos. Cada punto de datos se asigna a un grupo en función de su distancia al cuadrado del centroide. Este es el tipo de agrupación más utilizado.

Basado en Jerarquías

El agrupamiento basado en jerarquías se utiliza normalmente en datos jerárquicos, como los que obtendrías de la base de datos de una empresa o de taxonomías. Construye un árbol de grupos para que todo esté organizado de arriba hacia abajo.

Esto es más restrictivo que los otros tipos de agrupamiento, pero es perfecto para tipos específicos de conjuntos de datos.

        1.3.  IDENTIFICAR ALGORITMOS DE REDUCCIÓN DE DIMENSIONALIDAD.

Normalmente, 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

En los problemas de [Machine Learning](https://aprendeia.com/que-es-lo-mas-importante-para-aprender-en-machine-learning/) y de la ciencia de datos, el objetivo principal sigue siendo encontrar las características más relevantes que juegan un papel dominante en la determinación e influencia de los resultados de la producción.

En la mayoría de los problemas de ciencia de datos, el conjunto de datos está sobrecargado con numerosas características que dan como resultado un ajuste excesivo y aumentan los enormes costes de formación, lo que hace que el proceso sea considerablemente lento. Los algoritmos desarrollados a lo largo del tiempo tienen como objetivo resolver algunos de los problemas básicos, incluyendo:

Reducir la dimensión del conjunto de datos de formación restableciendo la varianza y manteniendo intacta la información relevante.

Reducir el tiempo y el coste de formación.

Estructurar formas de visualización efectivas.

Comencemos con el por qué necesitamos realizar la reducción de la dimensionalidad antes de analizar los datos y llegar a algunas inferencias. A menudo es necesario visualizar el conjunto de datos, para tener una idea de ello. Pero, hoy en día, los conjuntos de datos contienen muchas variables aleatorias, también llamadas características, debido a las cuales se hace difícil visualizar el conjunto de datos. Aquí es donde nos encontramos con la reducción de la dimensionalidad.

Como su nombre lo indica, la reducción de la dimensionalidad es el proceso de reducir el número de variables aleatorias del conjunto de datos bajo consideración, mediante la obtención de un conjunto de variables principales.

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

La mayoría de los profesionales en el campo del Data Science coinciden en que, para la construcción de un modelo de data mining bueno, se debe invertir alrededor de un 75% del tiempo y esfuerzo en la fase de preprocesado de los datos —si aún no tienes claro qué es data mining (o minería de datos), [te recomiendo pasarte por mi anterior post](https://aukera.es/blog/data-science-que-es-y-que-no-es/) para resolver tus dudas—.

Una de las técnicas de preprocesado para modelos de aprendizaje supervisado es la reducción de la dimensionalidad, que no es más que la reducción del número de variables en una colección de datos. En este artículo vamos a explicar en qué consiste.

Beneficios de reducir la dimensionalidad

A menudo, el monstruo de Diógenes se nos aparece y nos dice que cuanta más información mejor, que da igual si no nos va a servir de mucho, que menos da una piedra y que mejor no tocar nada, por si nos cargamos algo.

Si no sabemos cómo hacerlo bien (y no tenemos intención de aprender) quizá sí es recomendable escuchar al monstruo Diógenes, pero en cualquier otro caso, deberíamos mandarlo a paseo y seguir leyendo este post.

Este data scientist no sabe lo que es la reducción de la dimensionalidad.

Las razones por las que nos interesa reducir la dimensionalidad son varias:

Porque interesa identificar y eliminar las variables irrelevantes.

Porque no siempre el mejor modelo es el que más variables tiene en cuenta.

Porque se mejora el rendimiento computacional, lo que se traduce en un ahorro en coste y tiempo.

Porque se reduce la complejidad, lo que lleva a facilitar la comprensión del modelo y sus resultados.

Porque el [KISS (Keep It Simple, Stupid!)](https://es.wikipedia.org/wiki/Principio_KISS) está de moda.

Técnicas, métodos y algoritmos para reducir la dimensionalidad

Vamos a ver algunas técnicas, métodos y algoritmos que se pueden aplicar para reducir la dimensionalidad de nuestro set de datos.

La técnica más sencilla y la que vamos a ver más a fondo en el post es la de selección de variables, que no es más que seleccionar el conjunto de variables óptimo. Otra técnica es la de análisis de componentes principales. Esta última es más complicada de entender, por lo que no profundizaremos tanto.

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

La reducción de dimensionalidad es otra de las grandes aplicaciones de los algoritmos no supervisados. El objetivo de este tipo de algoritmo es convertir un dataset de una cierta dimensión -digamos, 1000 características- en otro con una menor dimensión -digamos, 300-. Los motivos para desear realizar esta reducción son variados:

Puede reducir los recursos necesarios para pasar dichos datos por un algoritmo supervisado

Puede eliminar ruido presente en el dataset original

Los resultados pueden ser más fácilmente interpretables

Grosso modo, hay tres métodos para realizar reducción de dimensionalidad:

Selección de características, es decir, escoger un subconjunto de las características originales que, según cierto criterio, representen bien el conjunto de datos.

Derivación de características, método consistente en la creación de nuevas características a partir de las originales. Lógicamente, esto supone una reducción de dimensionalidad solo si creamos características que combinen dos o más de las originales y podemos sustituir las originales por las nuevas sin que se produzca una pérdida de información excesiva.

Conclusión

Como una conclusión a esto los temas de los algoritmos y de la métrica son uno de los temas que mas hay de que hablar cunado se tata de las base de datos y que son de mucha ayuda para nosotros y de eso tenemos que comprender.

Bibliografía

<https://blog.softtek.com/es/la-reduccion-de-dimensionalidad-en-el-machine-learning>

<https://aprendeia.com/reduccion-de-la-dimensionalidad-machine-learning/>

<https://aukera.es/blog/reduccion-dimensionalidad/>

<https://interactivechaos.com/es/manual/tutorial-de-machine-learning/reduccion-de-dimensionalidad>