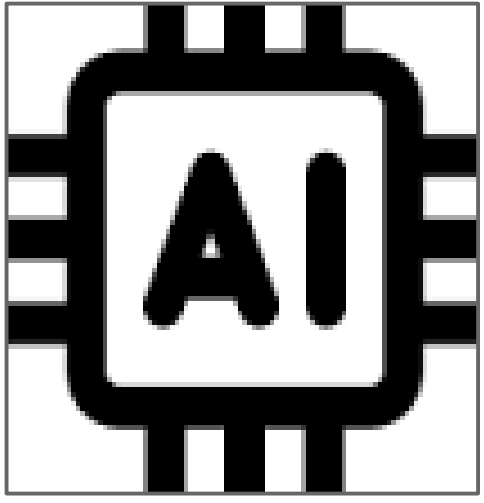
Introducción a la IA



Preprocesamiento de los Datos (Procesos ETL)

Datos e Información

Dato: Definiciones

Es el valor que resulta de una medición.

Es una representación simbólica (numérica, alfabética, espacial, etc.) de un atributo o variable cuantitativa o cualitativa.

Es la parte mínima de la información.

Datos e Información

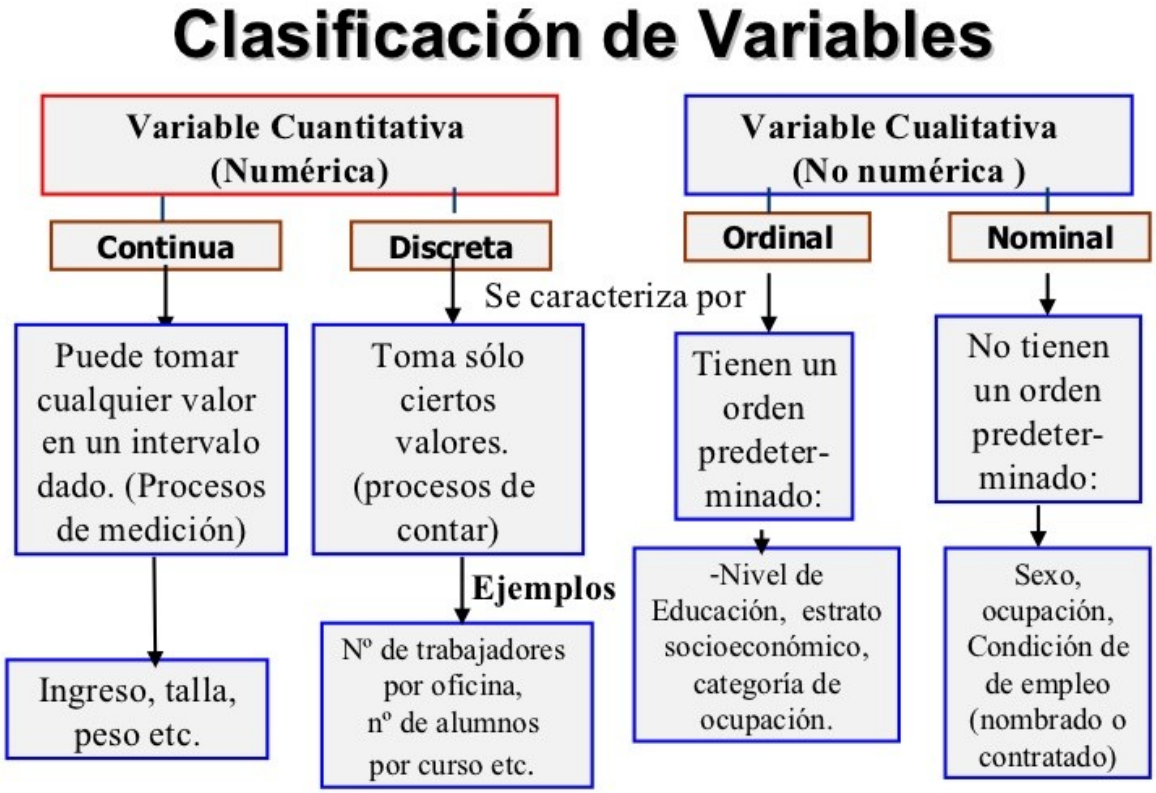
Variable

Definición: Es un concepto acerca de algún aspecto y/o magnitud de un elemento o unidad de análisis capaz de asumir diferentes cualidades y/o valores.

Las variables permiten presentar las características de los **atributos** de las unidades de análisis.

Datos e Información

Variables: Clasificación



Datos e Información

Los Datos y la información.

Los datos por sí solos no contienen información que sea humanamente relevante.

El análisis y procesamiento de un conjunto de datos puede sacar a la luz información o evidencias del fenómeno estudiado.

La **información** es un conjunto de datos **organizados y procesados**, que conforman un mensaje que cambia el estado actual del conocimiento del receptor del mensaje.

Procesos ETL

Rara vez los datos obtenidos en bruto se encuentran en el formato adecuado, sin errores, completos y con etiquetas y codificado adecuadamente. Se transforman los datos para darles un formato común.

Frecuentemente se los recopila en bancos de datos donde se consigue unificar toda la información.

Se debe encontrar y tratar datos faltantes, erróneos y anomalías. A esta etapa se la llama también Extracción-Transformación-Limpieza o etapa **ETL** (según sus siglas). Con esto se obtiene lo que se conoce como “vista minable” que es el conjunto de datos listo para aplicarle algoritmos de ML.

Esta tarea es una etapa esencial en el proceso de implementación de IA. De hecho en la práctica, muchas veces, esta tarea ocupa más tiempo que el minado de datos (De Jonge & Van Der Loo, 2013).

Procesos ETL: Tareas

Las principales tareas del proceso ETL son:

- Formateo de datos
- Detección de errores de carga o inconsistencias
- Detección de Datos Vacíos
- Detección de Outliers o valores anómalos
- Normalización de datos
- Equilibrio de Datasets
- Discretización de Variables

Procesos ETL: Tareas

Formateo de Datos:

Para el correcto tratamiento por parte de los algoritmos de ML es necesario que tengan el formato adecuado. Por ejemplo, los números deben tener formato numérico y no de string.

Procesos ETL: Tareas

Detección de Errores o Inconsistencias:

Con funciones sencillas es posible identificar muchos de ellos.

Por ejemplo, en una persona, se detecta un error si Edad (en años) > 150, o si Altura (en metros) > 2,50. Otros pueden requerir condiciones más complejas. Un ejemplo: un dato inconsistente: Embarazo='SI' y Sexo='Masculino'.

Procesos ETL: Tareas

Detección de Datos Vacíos:

Muchos algoritmos pueden soportar datos vacíos sin problema, pero otros no. Más allá de soportarlos o no, un dato vacío puede quitar mucha información valiosa y generar ruido. Es importante considerar como se los trata.

Y frente a ello, la sugerencia es consultar a los investigadores para obrar en consecuencia.

Procesos ETL: Tareas

Detección de Outliers o datos anómalos:

Dependiendo el tipo de estudio que se está llevando a cabo, el tratamiento de outliers puede ser crítico. Mientras que en algunas aplicaciones los valores anómalos generan ruido y es bueno eliminarlos, en otras, su detección es primordial, por ejemplo, en los sistemas de detección de fraudes.

Detectar outliers puede lograrse fácilmente usando funciones estadísticas.

Procesos ETL: Tareas

Normalización de Datos:

La reducción de escala para crear un rango más pequeño suele ser recomendado. En análisis multivariable, si no se normaliza las variables, pueden causar errores por el peso que la variable que trae consigo.

Procesos ETL: Tareas

Equilibrio del Dataset:

Si la variable objetivo tiene resultados 1 y 0, debería haber cantidades similares de registros con 1 que registros con 0.

Opciones:

- Eliminar los registros en exceso.
- Se duplican los registros en defecto

En general, la opción más recomendable es duplicar los registros que están en defecto, hasta estar en cierto equilibrio.

Procesos ETL: Tareas

Discretización de Datos:

Dado que la mayoría de los algoritmos de IA necesitan datos tipo numéricos, en caso de tener variables cualitativas nominales, pueden transformarselas en variables cuantitativas discretas.

Por ejemplo, si tenemos una variable “marca_auto” cuyo resultado es una lista de marcas, por ejemplo: [‘ford’, ‘vw’, ‘renault’, ‘nissan’, …] podemos presentar una variable por marca, como ser: “marca_auto_ford”, “marca_auto_vw”, … Y llenamos la columna “marca_auto_ford” con “1” en los registros donde la “marca_auto” = ‘ford’, por ejemplo.

Referencias

De jonge, Edwin; Van Der Loo, Mark. (2013). “An Introduction to Data Cleaning with R”. Statistics Netherlands, The Hague/Heerlen.

Jorge Kamlofsky, Vanesa Miana , Elio Prieto Gonzalez. “Uso de Técnicas de Inteligencia Artificial para el Análisis del Impacto de Ambientes Contaminantes en el Índice de Daño Genético Humano”. Revista RAIA (1998).