ACTIVIDADES REPASO UNIDAD 1

Ejercicio 1: Lee las siguientes situaciones y elige cuál de las "V's" es la más relevante en cada caso. Justifica tu elección.

- Una empresa analiza millones de transacciones bancarias en tiempo real para detectar fraudes.

Veracidad: Tendriamos que detectar cuales de las transacciones son fiables

- Una empresa de seguros analiza tweets sobre desastres naturales para evaluar el riesgo de sus clientes en ciertas ubicaciones.

Viabilidad: Habría que seleccionar los datos cuidadosamente para saber cuales pueden rentabilizar la compañia

o

Valor: dependiendo de la evaluacion que hagan sobre el riesgo que existe

- Se está analizando una base de datos para limpiar y corregir datos de clientes duplicados.

Variedad: Por la cantidad de datos que hay, saber cuantos distintos son de sí o si hay duplicados

Veracidad

Ejercicio 2: Clasifica los siguientes modelos de negocio según la clasificación basada en Big Data (usuarios de datos, proveedores de datos, facilitadores de Big Data):

- Una empresa que vende datos agregados de consumo de energía a otras compañías.

Proveedor de datos

- Una tienda online que usa Big Data para mejorar las recomendaciones de productos.

Usuario de datos

- Una empresa que proporciona servicios de almacenamiento en la nube para el procesamiento de datos.

Facilitador de big data

Ejercicio 3: Dado un conjunto de algoritmos con las siguientes notaciones Big O, ordénalos de menor a mayor según su complejidad computacional:

- $O(n^2)$
- $O(n)$
- $O(2^n)$
- $O(\log n)$

Demuestra tu respuesta cuando n vale 100.

$O(n) < O(\log n) < O(n^2) < O(2^n)$

En n el valor de 100 es el mismo asignado

en n^2 al cuadrado, el valor es el mismo con una operacion sobre si mismo

en 2^n el valor de n, tiene que mas operaciones para llegar al resultado

en el log el calculo es 2

Ejercicio 4: Imagina que tienes un problema que requiere realizar un total de 1 billón de operaciones (1,000,000,000,000) para procesar un conjunto de datos en un solo ordenador. Si distribuimos estos datos de manera equitativa entre 1,000 máquinas:

- ¿Cuántas operaciones debe realizar cada máquina individualmente?

1 mil millon

- Si se necesitan 100 operaciones adicionales para combinar los resultados de cada máquina, ¿cuál es el número total de operaciones en el sistema distribuido?

100* 1000 = 100.000 ; 1.100.000 operaciones por cada maquina

Ejercicio 5: Clasifica los siguientes problemas en supervisado o no supervisado y especifica si son de regresión, clasificación o clustering.

- Predecir el precio de una casa.

supervisado, regresion

- Detectar si un correo electrónico es spam.

supervisado, clasificacion

- Agrupar a clientes en base a sus patrones de compra.

no supervisado, clustering

Ejercicio 6: Imagina que tienes un conjunto de datos que incluye una variable cualitativa que representa el país de origen de una persona. Diseña una codificación para esta variable en una escala numérica.

Codificacion frecuencial, para saber por cada pais el valor de cuantas veces se repite ese dato cualitativo por persona

Ejercicio 7: Lee cada situación y determina en qué fase del procesamiento de datos estás: Depuración, Criba o Valores Faltantes. Explica por qué has elegido esa fase e indica cómo procederías.

- Tienes un conjunto de datos donde la columna de edad contiene valores como "35 años", "40", "cincuenta" y "45".

Depuracion, Todavía los datos no se han clasificado según su naturaleza, es decir, tenemos datos conjuntos en vez de específicos por lo que nuestro sistema no será capaz de procesarlos.

- Encuentras registros duplicados en una base de datos de clientes, donde un mismo cliente aparece dos veces con ligeras variaciones en su nombre.

Criba, Una vez que tenemos esos datos, cuando eliminamos datos repetidos encontrariamos este caso

- En un *dataset* de ventas, encuentras valores que son atípicos, como un pedido de 10,000 unidades cuando el promedio es de 10 unidades.

Criba: Son valores fuera de lo común, por lo que no nos darian un modelo fiable.

- Algunas celdas en la columna de ingresos mensuales están vacías.

Valores faltantes: Basicamente faltan valores en nuestros datos.

- En un *dataset* de análisis de clima, observas que la temperatura está en Celsius en unas filas y en Fahrenheit en otras.

Depuración: Debemos tener los datos establecidos de una determinada forma para que lleguen a la criba.

- Tienes datos categóricos con respuestas como "Sí", "sí", "SI" y "si".

Depuración: No podemos tener datos con formas distintas antes de la criba

- Algunas filas en un *dataset* de encuestas sobre satisfacción al cliente no incluyen respuestas en la columna de "Nivel de Satisfacción".

Valores faltantes: faltan datos

- Encuentras un registro en una columna de precios con el valor "-100", que no es válido.

Depuración: Ha encontrado un valor no valido por lo que no puede proceder

- Un *dataset* de consumo energético contiene valores atípicos extremos en el consumo mensual de electricidad en algunos hogares.

Criba: Eliminación de esos valores atipicos para que no afecten a nuestro modelo

Depuracion:

- Un *dataset* de ingresos muestra valores extremadamente altos para algunos empleados con respecto al resto de la empresa.

Criba: Valores que no proceden a lo normal.

Depuración: