|  |  |  |  |
| --- | --- | --- | --- |
| **Nombre:** | Tomás Tapia | Fecha: | 23-11 |

1. Explica qué es la normalización y desnormalización de datos y por qué es importante en el análisis de datos.

R: **La normalización** de datos es una estrategia de transformación de datos, para que estos puedan tener una escala común, de modo que estos se reescalan, para que los registros de diferentes magnitudes no afecten de forma desigual el análisis.

**La desnormalización** es el proceso inverso de la normalización, lo que implica retornar los datos a su escala original.

Una técnica muy utilizada para normalizar es la **normalización min-max**, donde esta convierte los valores a un rango entre 0 y 1.

Es de importancia normalizar los datos por varias razones:

Los algoritmos de ML como los algoritmos de clustering se benefician al poder realizar agrupaciones que podrían tener escalas muy diferentes. Por ejemplo agrupar valores asociados de un trabajador: sueldos (millones), cantidad de hijos, estatura, peso, ahorros en el banco,etc. Estos registros llevándolos a una escala común, podría mejorar la interpretación y en el caso de algoritmos de clasificación o regresión permitirían disminuir el sesgo, ya que algunos algoritmos pueden ser sensibles a diferencia de escala entre las distintas características

1. Imagina que has realizado un análisis de datos y has encontrado información valiosa. ¿Cómo presentarías tus hallazgos a un equipo no técnico?

Cuando se presentan hallazgos a un equipo no técnico, es crucial simplificar la información y enfocar la presentación de manera clara y comprensible.

A. Objetivo del análisis: Explicación de por que se realizó el análisis.

B. Explicar los hallazgos en términos simples: Descubrimos alzas de ventas en los meses de verano, de alrededor de un 34%, mientras que en los meses de julio, todos los años hay una caída importante en relación al resto de los meses del año’

C. Impacto en el negocio: explicar cómo los hallazgos pueden influir en el negocio, por ejemplo: ’Este hallazgo puede permitirnos anticiparnos y crear campañas para los distintos meses del año, pudiendo establecer cantidad de personal, recursos e infraestructura necesaria’

1. ¿Cómo manejas un conjunto de datos que falta varios valores?

R: En primer lugar, se debe estudiar la razón de la ausencia de estos, ya que podría implicar varias causas:

- Falla en la carga de datos

- Error de digitación

- Ausencia del valor

Algunas estrategias podrían ser las siguientes:

Crear una nueva categoría: En el caso de datos categóricos, donde hay datos faltantes, se puede indicar la ausencia de datos en una nueva etiqueta.

Eliminar filas con datos faltantes: En el caso que estos registros, fueran muy pocos en relación al total de datos, se podría tomar la decisión de borrar, pero podría implicar perder información valiosa contendida en otras características.

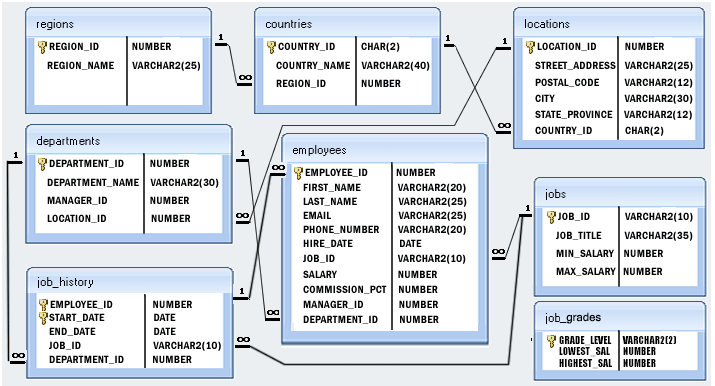
Imputar datos: Si fueran valores continuos, se pueden completar con media, mediana o moda, o en el caso de que fueran datos pertenecientes a datos temporales, se puede utilizar interpolación polinómica o lineal, convirtiendo esto en parte del ajuste, si fuera el caso del entrenamiento de un modelo de machine learning

1. ¿Cuáles son los métodos de selección de características utilizados para seleccionar las variables correctas?

**Correlación:** Eliminar las características altamente correlacionadas y a la vez las que están bajamente correlacionadas con la variable objetivo.

**Completitud de los datos:** En muchos casos, la información debe ser revisada para ver si esta tiene registros que sean útiles para un determinado análisis. Ejemplo: De un set de clientes, muchos no cuentan con la característica ’Edad’, y no contamos con otra columna que pueda ayudar a imputar esa información, podría ser causal de no contar con esa columna. En el caso de utilizar algún metodo de imputación de datos, se debería evaluar el seleccionar esta columna utilizando la técnica anterior.

1. ¿Qué es la reducción de dimensionalidad y cuáles son sus beneficios? Es un proceso útil para reducir la cantidad de características (dimensiones) en un conjunto de datos, manteniendo la mayor cantidad de información relevante, permitiendo un mejor manejo de los datos, sin perder calidad o precisión en los resultados. Algunos beneficios: Mejora en la eficiencia del algoritmo, eliminación de ruido (detección de características irrelevantes), Reducción de Sobre-ajuste.



1. Escribe un query SQL que retorne la cantidad de empleados que hay en cada departamento. Se requiere los departamentos que tienen más de 5 empleados. Debe retornar el nombre del departamento, el nombre de la ciudad y país del departamento y la cantidad de empleados.

SELECT

d.department\_name,

l.city,

c.country\_name,

COUNT(e.employee\_id) AS num\_employees

FROM

departments d

JOIN

locations l ON d.location\_id = l.location\_id

JOIN

countries c ON l.country\_id = c.country\_id

JOIN

employees e ON e.department\_id = d.department\_id

GROUP BY

d.department\_name, l.city, c.country\_name

HAVING

COUNT(e.employee\_id) > 5

ORDER BY

num\_employees DESC;

1. Escribe un query SQL para encontrar los managers. Devuelva la información completa de los managers.

---Suponiendo que los managers son aquellos empleados donde ---------manager\_id = 7.

SELECT

e.\*,

d.\*

FROM

employees e

JOIN

departments d ON e.department\_id = d.department\_id

WHERE

e.manager\_id = 7

ORDER BY

e.last\_name, e.first\_name;

1. Escribe un query SQL que calcule la experiencia de los managers. Debe retornar la siguiente información del manager: ID, Nombre, nombre del trabajo, nombre del departamento, nombre de la ciudad, nombre del país, fecha de contratación, salario y experiencia, indicando años, meses y días.

**Caso 1. Análisis de datos.**



El fraude es una amenaza constante en todas las escalas comerciales. Detectarlo requiere tecnología avanzada y análisis profundo de datos para identificar patrones anómalos y salvaguardar activos. Desde el robo de identidad hasta el fraude financiero, las organizaciones deben estar vigilantes. La detección implica una combinación de tecnología, experiencia y diligencia para mantener a raya a los estafadores y proteger la integridad de la empresa. Es un proceso continuo y dinámico en la lucha contra la actividad delictiva.

Tu misión es identificar actividades fraudulentas por medio del análisis de datos, cuentas con información de transacciones de tarjetas de crédito para realizar tu labor: [Conjunto de datos](https://drive.google.com/drive/folders/1F0nP_7Ttp-LS2HNKIlBW1f8OOHR8AVjg?usp=sharing)

El auditor Senior con el que trabajas te menciona actividades que son consideradas sospechosas:

1. Transacciones de montos altos durante las primeras horas de la mañana, es decir, entre las 6:00 a 8:00 a. m.
2. Algunos estafadores efectúan múltiples transacciones de bajo valor (generalmente menos de $2,00) utilizando tarjetas de crédito, lo cual suele pasar desapercibido para los titulares de estas.

Realiza un análisis descriptivo con los datos provisto para identificar actividades fraudulentas/anómalas.

* Crea un Diagrama Entidad Relación (DER) con base a los archivos CSV proporcionados. Usa la herramienta de diagramación de tu preferencia.
* Debes realizar esta tarea en un Notebook usando tu lenguaje de preferencia (Python, R).
* Entrega un video de no más de 15 minutos explicando los resultados y técnicas utilizadas.
* Todos los entregables se deben subir a Github y debes compartir el acceso al repositorio.

0.0

**Caso 2. Aprendizaje automático.**



El auditor Senior en conjunto con el equipo de investigación, ha construido un conjunto de datos de movimientos de tarjetas de créditos donde está marcada cuales transacciones han sido legítimas y cuáles han sido fraudulentas. Este set de datos es muy valioso porque ya contiene 29 variables que explican cuando la transacción es legítima y cuando es fraudulenta.

[Conjunto de datos](https://drive.google.com/file/d/1kQ3sHPp3-Vc16dQCE0ZUcWSojPAy-8K0/view?usp=sharing).

Tu misión es construir un modelo de aprendizaje automático que permita para nuevas transacciones identificar si es legítima o fraudulenta.

* Debes realizar esta tarea en un Notebook usando tu lenguaje de preferencia (Python, R).
* Entrega un video de no más de 25 minutos explicando los resultados y técnicas utilizadas. Enfoca la explicación en como logras medir el rendimiento de la detección de fraude a través de las Métricas de rendimiento.
* Todos los entregables se deben subir a Github y debes compartir el acceso al repositorio.