Logo, company name

Description automatically generated

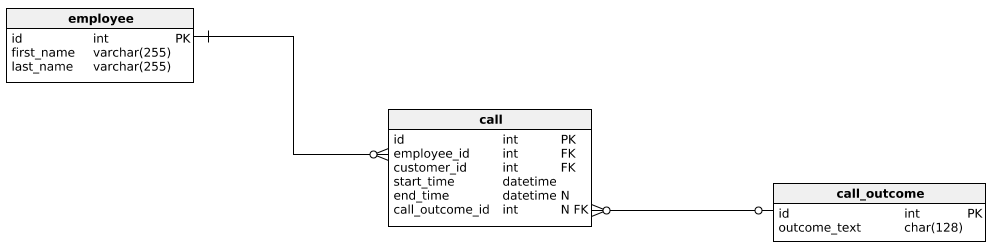
**Introducción a la Ciencia de Datos**

Parcial 12 de Octubre 2022

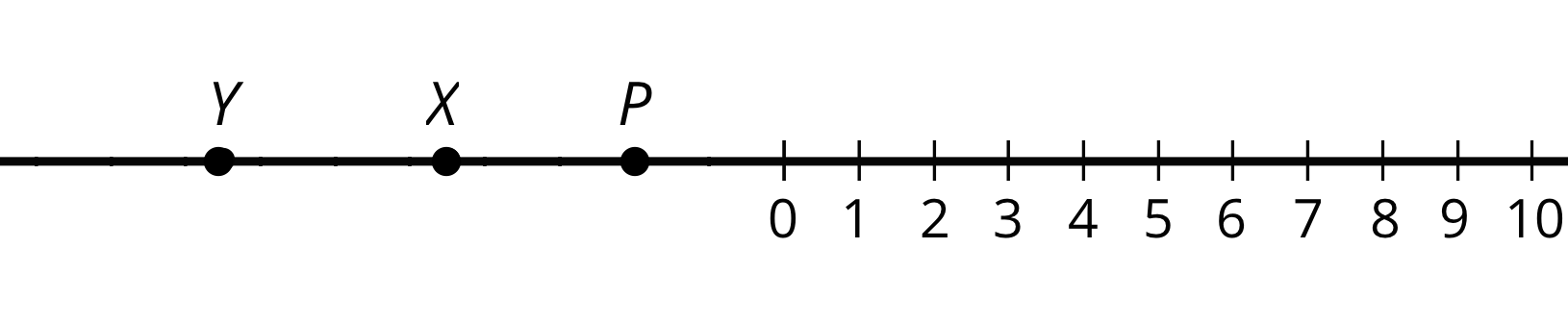
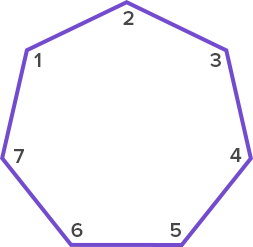
1. En clase hemos visto que el modelo de regresión logística es un modelo altamente interpretable. El mismo queda expresado por medio de cualquiera de las siguientes fórmulas (logit es la función inversa de la función logística):

Suponga que usted entrenó el modelo con 2 variables explicativas y obtuvo los coeficientes .

1. Interprete el modelo conociendo los coeficientes.
2. Interprete el modelo a partir de las posibilidades (odds):
3. Suponga que usted tiene un dataset con 10 columnas y quiere encontrar clusters de datos. De las 10 columnas, 9 son continuas y 1 es categórica. Las columnas continuas se encuentran en rangos de valores muy distintos y la columna categórica posee una altísima cardinalidad, aunque la frecuencia de los valores es alta solo para 3 categorías. Explique el pre procesamiento que llevaría a cabo.
4. Suponga que usted trabaja con una base de datos de un call center. Usted sospecha que el empleado Victor Lustig ha mentido sobre su performance y quiere verificarlo usando la base de datos. Necesitamos todas las llamadas con su hora de inicio y hora de finalización. Para cada llamada, queremos mostrar cuál fue el resultado, así como el nombre y apellido del empleado que realizó esa llamada. Realice la query.



1. Un feature engineering muy común para datos temporales es expresarlos mediante una función seno y coseno, es decir, “graficar” los datos en un círculo. ¿Por qué cree que esto es una buena idea en lugar de expresarlos numéricamente?

1. Qué problemas causan los outliers en el modelo de regresión lineal (supervised) y en k-Means (unsupervised)?

6) Una empresa internacional de jugos naturales sin azúcar quiere comenzar a comercializar sus productos en la cantina de la Universidad de Montevideo. Su estrategia de marketing se basa en información recolectada de diferentes fuentes que apunta a que los estudiantes universitarios consumen, en promedio, 3 litros de jugo natural a la semana.

Usted es recientemente contratado por esta empresa y duda si dicha premisa es correcta. Confeccione un experimento para comprobar dicha premisa definiendo

- Universo de estudio y muestra a realizar

- Proceso por el cual realizará el experimento

- Como comprobará estadísticamente la significancia de sus resultados

7) A continuación, se mencionan diferentes variables relativas a una persona. Determine la naturaleza de las siguientes variables en Nominales, Ordinales, Intervalo y Ratio

- Su nombre

- Su fecha de nacimiento en formato ISO 8601 (Ej: 2000 – 01– 15)

- Su edad

- Su género

- Su altura

- Su cédula de identidad

- La hora en más probable en la que se levanta los lunes

- La distancia que viaja hasta su trabajo

- La puntuación, del 1 al 5, que dio a la última película que vio.

- Su comida favorita

8) Se accede a información de consumos de la cantina de facultad. Usted sospecha que existen dos perfiles distinguibles de clientes en la cantina por lo que aplica un algoritmo de aprendizaje no supervizado k-means inicializándolo con 2 clusters. El algoritmo inicializa con ubicando 2 centrodies: A y B en las siguientes posiciones: A (2,100) y B (14,800). Determine a que cluster se asigna cada cliente en la primer iteración del algoritmo.

|  |  |  |
| --- | --- | --- |
| Frecuencia mensual compras en cantina | Ticket Promedio | Cluster |
| 15 | 500 |  |
| 6 | 200 |  |
| 4 | 240 |  |
| 8 | 0 |  |
| 12 | 1000 |  |

9) Que diferencias existen entre k-Means y DBSCAN? ¿En qué situaciones es preferible utilizar uno u otro?

10) Cuales son los componentes necesarios para una buena visualización? Justifique.