# Presentación de resultados: prueba Científico de datos junior

Analista de lenguaje

Alejandro Montenegro Taborda

05 de febrero de 2024

# Propósitos

Objetivos y alcances de la presentación

# Propósitos

- Presentar la propuesta de algoritmo clasificatorio para la prueba de Científico de datos.
  - Mostrar la etapa de diseño desde una perspectiva de Desarrollador.
  - Evidenciar las etapas del proceso de diseño e implementación de algoritmos.
- Provocar en el público reacciones, comentarios y/o sugerencias que puedan aportar a la construcción y ejecución del trabajo realizado.


# Contextualización

Información sobre el problema planteado

# Caso Compañía A

- **Problema:** Ventas efectivas bajas
- **Interés:** Incrementar y optimizar el número de ventas efectivas
- **Propuesta:** Creación de un modelo de predicción de clientes con mayores probabilidades de pago

**Estado actual e impresión de resultados.**

# Conjunto de datos

- **Client_ID:** El Id del Cliente
- **Name:** Nombre del cliente
- **Age:** Edad del Cliente
- **Location:** Estado donde vive el Cliente
- **Income:** Ingresos del Cliente
- **TAX:** Impuestos que paga el cliente
- **previous sales_#:** Cantidad de compras que ha realizado en el pasado
- **Type_of_Products:** Tipo de producto que ha comprado
- **Contact_Channel:** Canal por el que se ha contactado al cliente
- **Contact_hour:** Hora de contacto
- **Num_Contacts:** # de intentos que se han realizado para el contacto
- **Satisfaction_Score:** Medida de satisfacción (CSAT) 1 a 5 siendo 5 muy satisfecho
- **Sales:** Si la venta fue efectiva o no (1 Si, 0 No)

*Pero antes…*

*La arquitectura es agnóstica a la tecnología*

# Etapa de diseño

Diseño de la arquitectura

## Aspectos tenidos en cuenta

- **POO y Entornos virtuales:** Diseño pensado en algunas oportunidades de mejora observadas.
- **Infraestructura en la nube:** Optimización pensada en infraestructura de la nube (cobro por procesos).
- **PEP8 y seguridad:** Implementación usando buenas prácticas (internacionales).

*Ahora sí*

*¡Empecemos!*

# Procesos llevados a cabo

Etapa 1

Etapa 2

Etapa 3

Etapa 4

Etapa 5

# Descripción general

**Limpieza de datos**

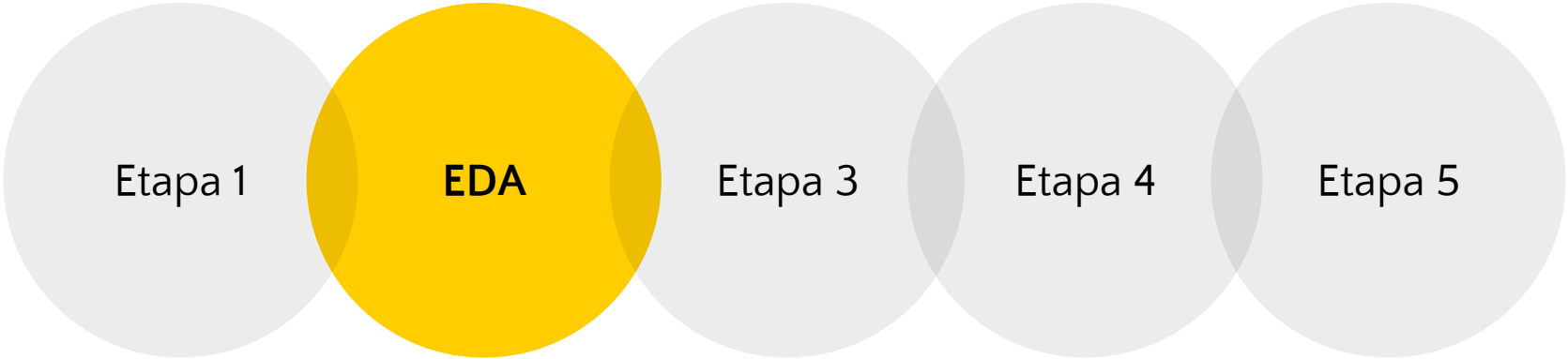
Etapa 2

Etapa 3

Etapa 4

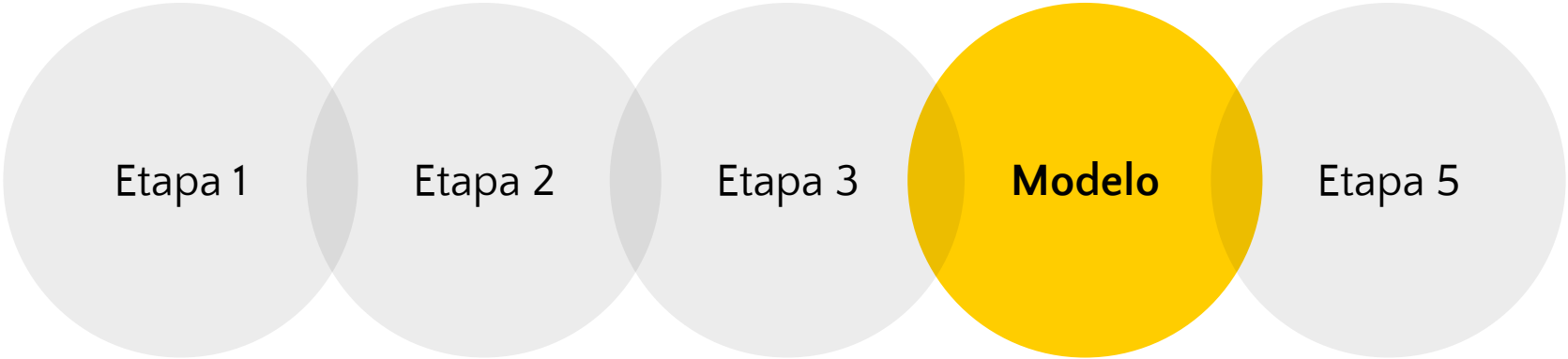
Etapa 5

# Descripción general

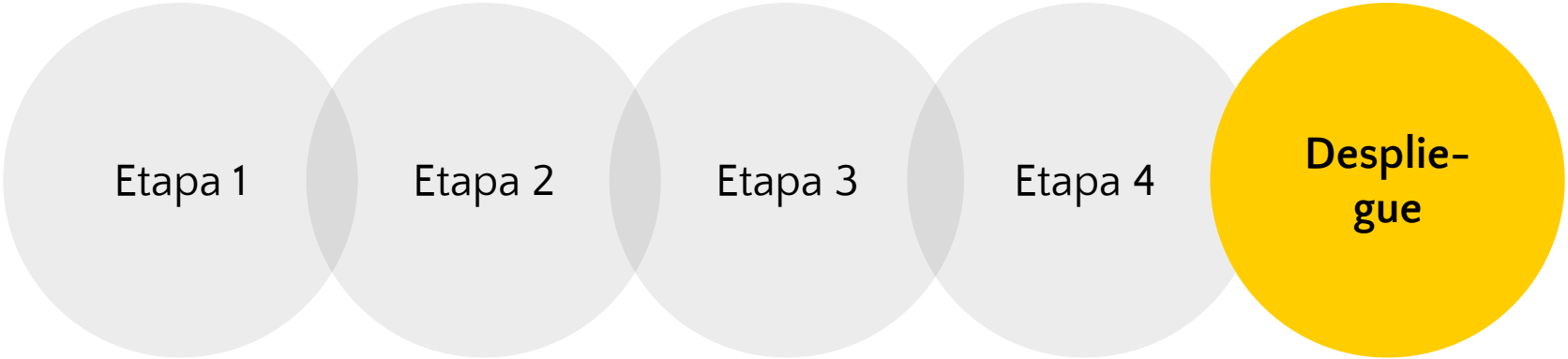


# Descripción general

# Descripción general



# Descripción general



# Limpieza de datos

Homogeinización de datos y capas de seguridad

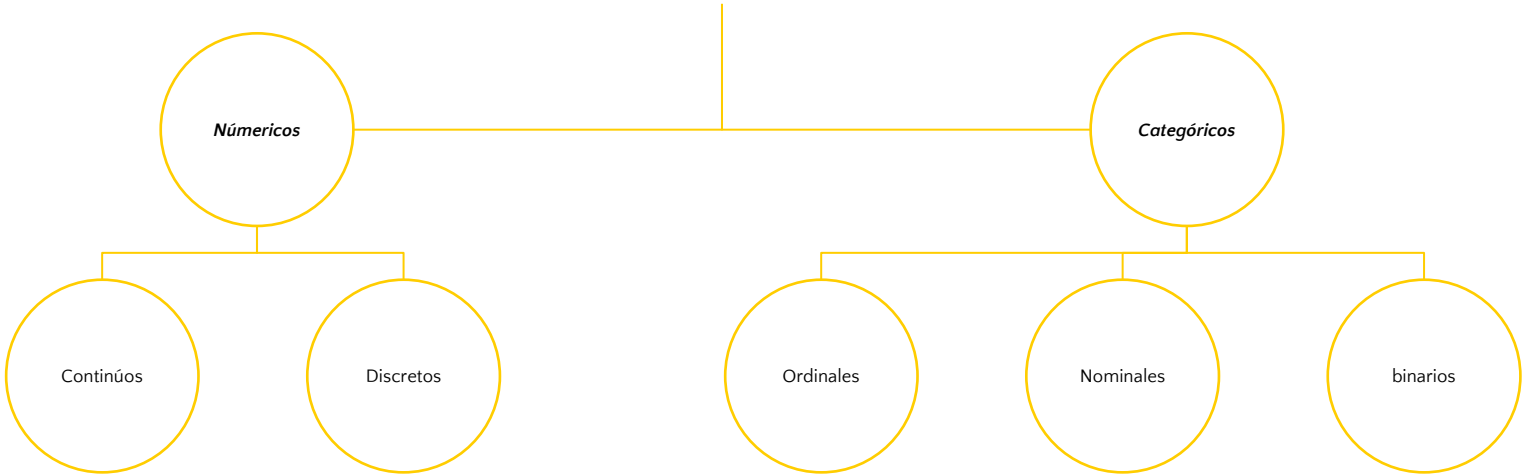
# Limpieza de datos

- Nombramientos (homogenización de columnas)
  - Camel Case
  - Snake Case
  - Pascal Case
  - Kebab Case
- Espacios vacios
- Caracteres especiales
- Protección de datos sensibles
  - Id del cliente
  - Nombre
- Ciberseguridad UUID
  - Encriptación de identidad
  - Desencriptación del usuario

# EDA

Analisis exploratorio de datos

# Tipos de datos




- *Income*
- TAX

- *Age*
- *Previous_sales*
- *Num_Contacts*

- *Client_ID*
- *Contact_hour*
- *Satisfaction_score*
- *Sales*

- *Name*
- *Location*
- *Type_products*
- *Contact_channel*

| | Age | Income | TAX | Previous_sales | Num_contacts | Satisfaction_score | Sales* |
|---|---|---|---|---|---|---|---|
| count | 1000 | 996 | 996 | 1000 | 1000 | 1000 | 1000 |
| mean | 41.191 | 63100.329317 | 9465.049398 | 2.052 | 3.492 | 2.558 | 0.269 |
| std | 12.259234 | 21638.692537 | 3245.803881 | 2.635952 | 2.399102 | 1.331318 | 0.443662 |
| min | 18 | 5000 | 750 | -7 | 1 | 1 | 0 |
| 25% | 31 | 46277.75 | 6941.662500 | 0 | 2 | 1 | 0 |
| 50% | 40.5 | 62770.5 | 9415.575000 | 1 | 3 | 2 | 0 |
| 75% | 51 | 80618 | 12092.7 | 4 | 5 | 3 | 1 |
| max | 125 | 165355 | 24803.25 | 25 | 33 | 5 | 1 |

*Análisis sin sentido debido a que la variable es del tipo Categórica.

# Dataviz

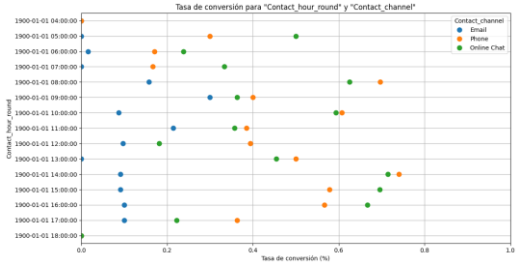
Visualización de datos en el EDA

# Visualización EDA

- Análisis por categoría


- Análisis univariado

- Análisis bivariado



*Una vez se han identificado las oportunidades de mejora, así como las fortalezas de la compañía, se realizarán transformaciones a los datos en pos de explotar dichas características*

# Preprocesamiento

Transformación de datos guiada por el EDA

# Innovación y explotación

- **Agrupar** horas de contacto
- **Determinar** horas pico (Big Data)
- Creación de variable binaria (enriquecimiento de datos)
- **Imputación** de datos
  - Imp. Simple
  - Imp. Hot-Deck (IA)*
- **Codificación de variables** categóricas

*Se explayará al final de la presentación.

# Modelo

Creación del modelo

# AutoML y Modelo

**Hiperparametrización**

Preparación para hiperparametrización de 19 posibles modelos.

**Comparación**

Comparación de resultados de clasificación (16 modelos).

**Selección del modelo**

Selección del modelo *Random Forest a* juicio humano.

**Entrenamiento iterativo**

Entrenamiento iterativo (10) usando los datos de entrenamiento.

**Resultados**

Resultados finales del modelo: Precisión del 91 % en la clasificación de venta efectiva en clientes.

**Distribución del modelo**

Guardado del modelo para su posterior distribución a los demás miembros del equipo en caso de necesitarse.

# Despliegue

Distribución del modelo

# Ambiente productivo

1. Lectura de datos y modelo.
2. Preprocesamiento de datos (datos a predecir).
3. Predicción en datos finales.
4. Simulación de exportación de datos a base de datos SQL.

# Comentarios

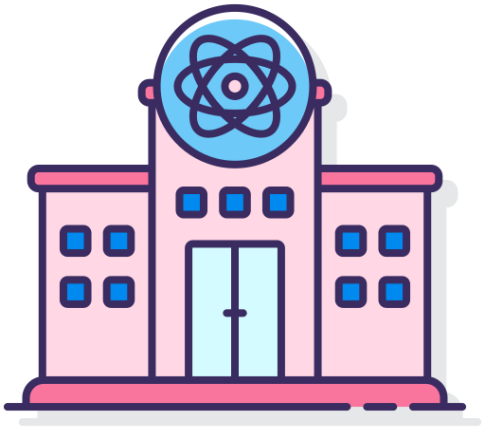
Comentarios finales sobre el proceso

# ¿El modelo puede mejorar?

Seguramente sí.

# ¿Cómo?

Trato de datos atípicos, aplicar imputación Hot-Deck e implementar arquitecturas diferentes.



# 91 %

De precisión en la clasificación de ventas efectivas.

# ¡Gracias!


***Alejandro Montenegro Taborda***

Analista de lenguaje

alejandro.montenegrotaborda@teleperformance.com