Trabajo Final de Introducción a la ciencias de datos

Materia: Introducción a la Ciencias de Datos
Profesor: Claudio Pavón

Integrantes:

- Fernandez Julian
- Juárez María Ailén
- Simeón Villegas Andrés
- Viani Fiorella

Informe de datos: Marketing Bancario

1. Introducción

En este trabajo, como grupo nos propusimos analizar un conjunto de datos de campañas de marketing bancario con el objetivo de predecir si un cliente aceptará una oferta de depósito a plazo. Creemos que este tipo de análisis tiene una gran utilidad práctica, ya que puede ayudar a los bancos a mejorar la eficiencia de sus campañas y reducir costos operativos.

A lo largo del trabajo, cada integrante del grupo participó en distintas etapas del proyecto: exploración de datos, tratamiento de valores nulos, selección y transformación de variables, entrenamiento de modelos y análisis de resultados. Utilizamos herramientas como pandas, seaborn, matplotlib y scikit-learn, y nos enfocamos principalmente en dos modelos: regresión logística y random forest.

En el informe explicamos nuestras decisiones, mostramos los resultados obtenidos y reflexionamos sobre las ventajas y limitaciones de cada técnica aplicada.

---

2. Datos

Al analizar el conjunto de datos previsto, observamos que contiene múltiples variables que reflejan características personales y comportamientos de los clientes. Como grupo, nos enfocamos en comprender el significado de cada una de ellas para interpretar correctamente su impacto en la decisión de aceptar una oferta bancaria.

Entre las más relevantes destacamos la edad, la profesión, el estado civil, el nivel educativo, el saldo bancario y la duración de la llamada. También prestamos atención a variables relacionadas con campañas anteriores, como la cantidad de veces que fue contactado el cliente y el resultado de esos contactos. Este entendimiento inicial fue clave para definir cómo preparar los datos y qué modelos utilizar posteriormente.

A continuación, se muestra un fragmento del dataset utilizado:

	age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome	depos
0	59.0	admin.	married	secondary	no	2343	yes	no	unknown	5	may	1042.0	1.0	-1	0	unknown	yes
1	56.0	admin.	married	secondary	no	45	no	no	unknown	5	may	NaN	1.0	-1	0	unknown	yes
2	41.0	technician	married	secondary	no	1270	yes	no	unknown	5	may	1389.0	1.0	-1	0	unknown	yes
3	55.0	services	married	secondary	no	2476	yes	no	unknown	5	may	579.0	1.0	-1	0	unknown	yes
4	54.0	admin.	married	tertiary	no	184	no	no	unknown	5	may	673.0	2.0	-1	0	unknown	yes
5	42.0	management	single	tertiary	no	0	yes	yes	unknown	5	may	562.0	2.0	-1	0	unknown	yes
6	56.0	management	married	tertiary	no	830	yes	yes	unknown	6	may	1201.0	1.0	-1	0	unknown	yes
7	60.0	retired	divorced	secondary	no	545	yes	no	unknown	6	may	1030.0	1.0	-1	0	unknown	yes
8	37.0	technician	married	secondary	no	1	yes	no	unknown	6	may	608.0	1.0	-1	0	unknown	yes
9	28.0	services	single	secondary	no	5090	yes	no	unknown	6	may	1297.0	3.0	-1	0	unknown	yes
10	38.0	admin.	single	secondary	no	100	yes	no	unknown	7	may	786.0	1.0	-1	0	unknown	yes
11	30.0	blue-collar	married	secondary	no	309	yes	no	unknown	7	may	1574.0	2.0	-1	0	unknown	yes
12	NaN	management	married	tertiary	no	199	yes	yes	unknown	7	may	1689.0	4.0	-1	0	unknown	yes
13	46.0	blue-collar	single	tertiary	no	460	yes	no	unknown	7	may	NaN	2.0	-1	0	unknown	yes
14	31.0	technician	single	tertiary	no	703	yes	no	unknown	8	may	943.0	2.0	-1	0	unknown	yes

---

3. Preparación de datos

En esta etapa, nos distribuimos tareas para tratar adecuadamente los datos. Primero, identificamos valores faltantes en variables clave como la edad, la duración de la llamada y la cantidad de contactos. Decidimos imputarlos con la mediana, por ser una medida robusta frente a valores extremos.

Luego, transformamos algunas variables para facilitar su uso en los modelos. Por ejemplo, agrupamos profesiones en categorías generales, codificamos variables categóricas con One-Hot Encoding y normalizamos algunas columnas numéricas. También convertimos ciertas variables a booleanos para simplificar su interpretación.

Todo este trabajo en equipo nos permitió construir una base de datos limpia y bien estructurada, lista para el entrenamiento de modelos.

---

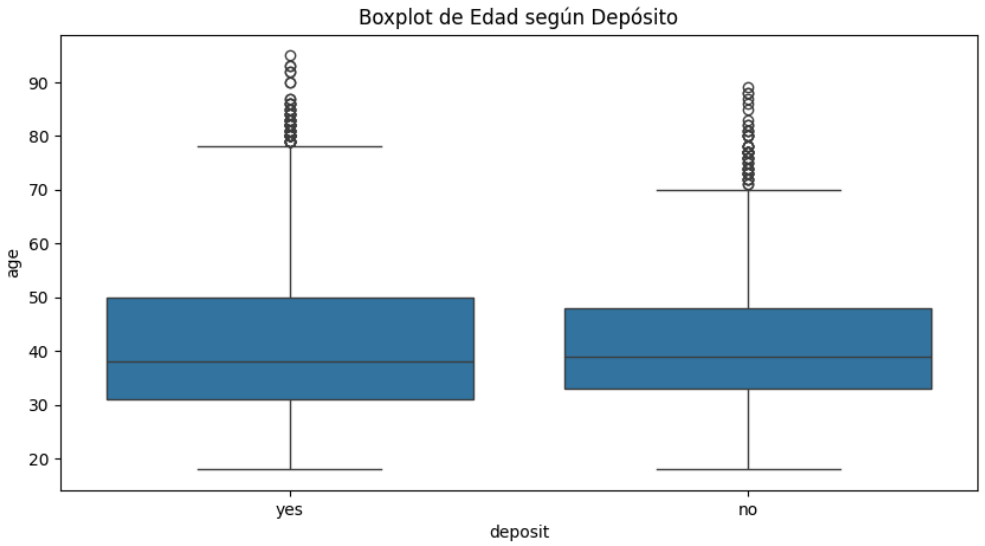
4. Análisis exploratorio

Como grupo realizamos distintos gráficos para visualizar mejor la información. Un integrante se encargó de los histogramas, otro de los boxplots, y así fuimos colaborando y compartiendo nuestras observaciones.

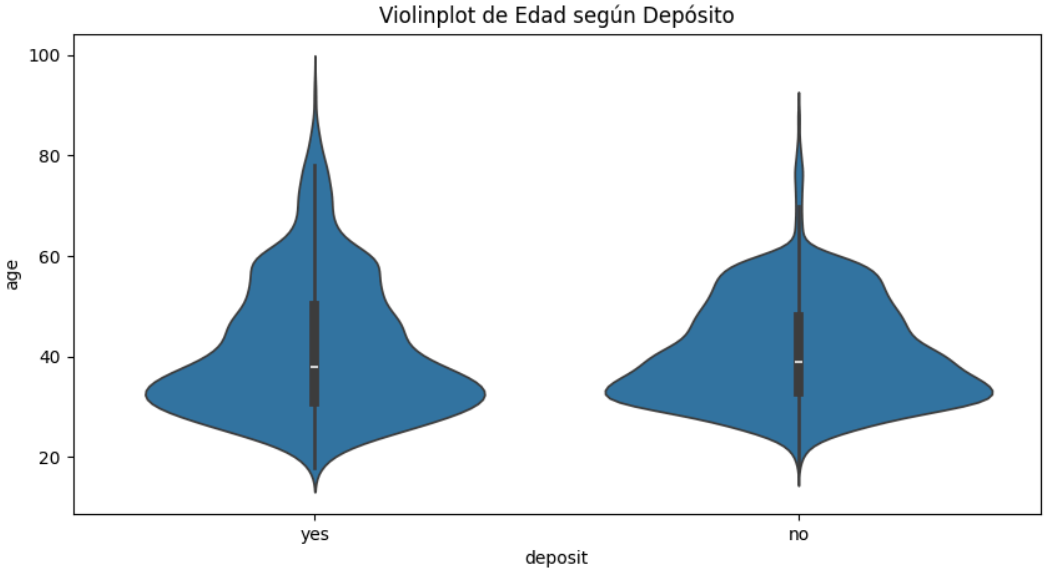
Detectamos que la edad y la duración de las llamadas influían bastante en la decisión del cliente. Las llamadas más largas estaban asociadas a un mayor número de aceptaciones del depósito. También descubrimos que había una alta concentración de clientes entre 25 y 45 años, lo que puede ser útil para focalizar campañas futuras.

A continuación, presentamos algunos de los gráficos generados:

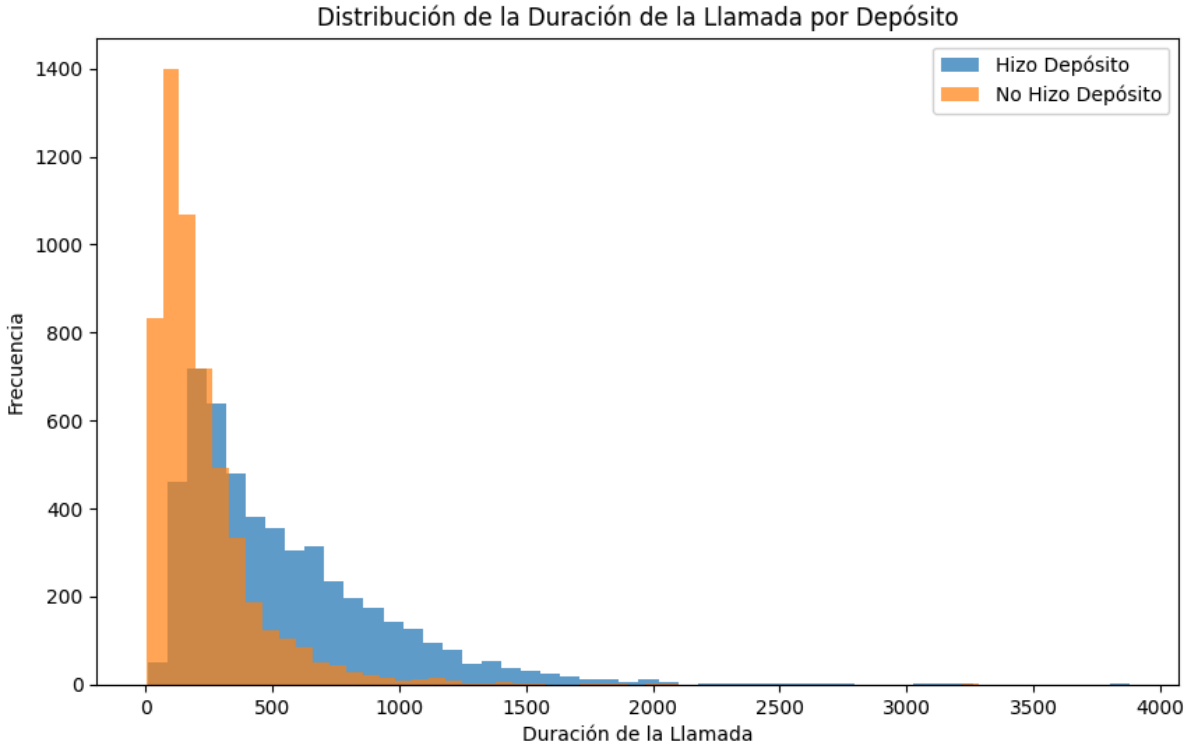
- **Boxplot de Edad según Depósito**
  Este gráfico muestra cómo varía la edad de los clientes según hayan aceptado o no el depósito. Se observan medianas similares, pero una ligera tendencia hacia edades más jóvenes en los que aceptaron.



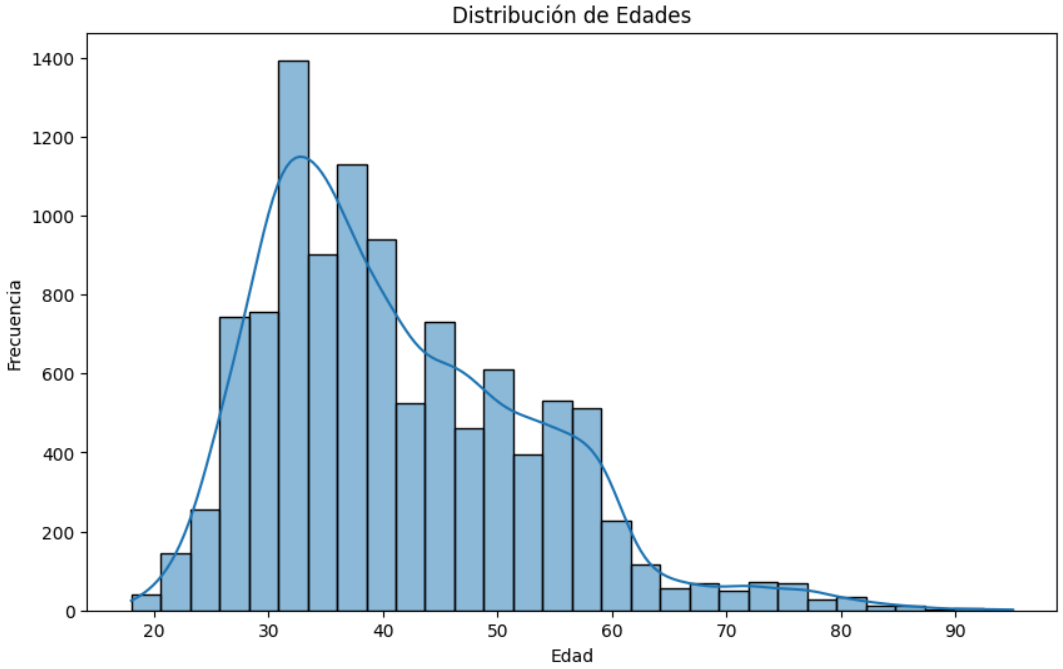
- **Violinplot de Edad según Depósito**
  Complementa al boxplot mostrando la distribución de la edad con más detalle, incluyendo la densidad.



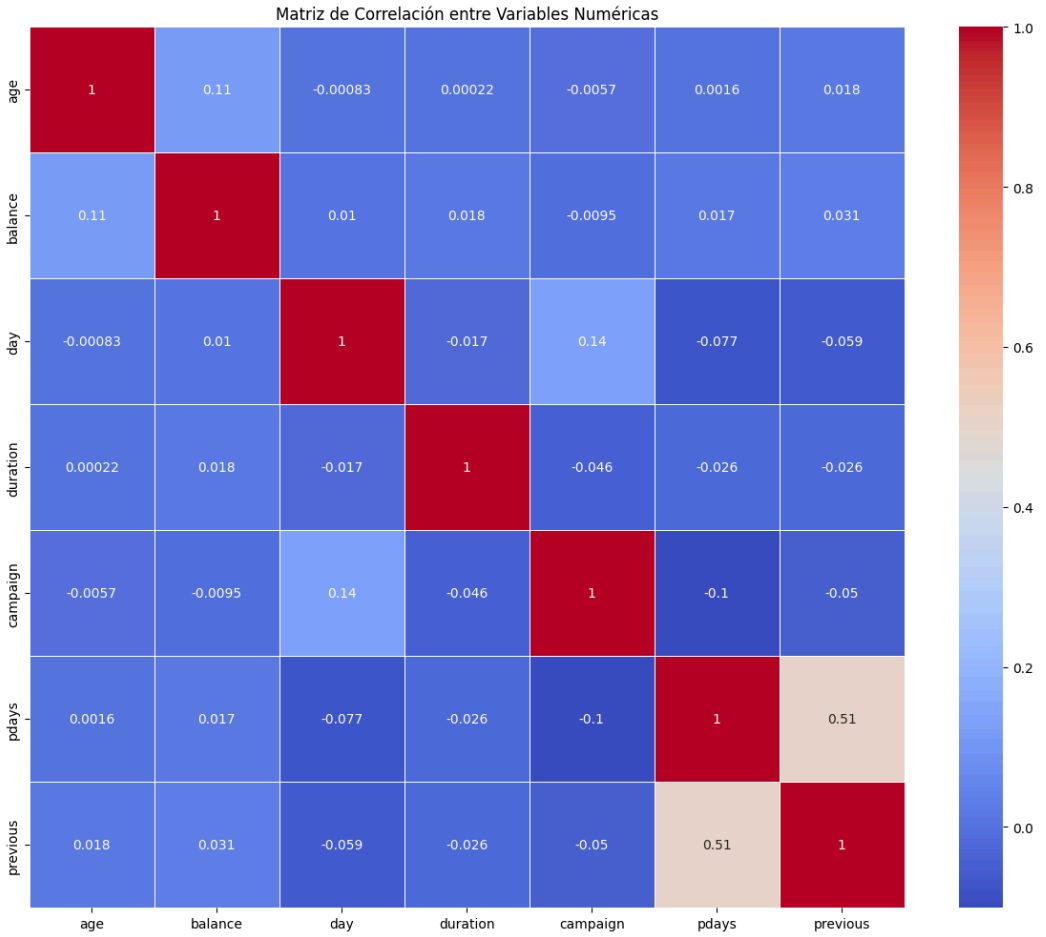
- **Histograma de la Duración de la Llamada por Depósito**
  Se aprecia que las llamadas más extensas están asociadas a una mayor probabilidad de aceptación.



- **Distribución de Edades con KDE**
  Este gráfico nos permitió ver la forma general de la distribución de edades en toda la base, ayudando a identificar rangos predominantes.



- **Matriz de Correlación entre Variables Numéricas**
  Aunque no se encontraron correlaciones muy fuertes, se destacó cierta asociación entre duración de llamada y la variable objetivo.



El análisis de correlación no arrojó relaciones muy fuertes, pero sí confirmó la importancia de algunas variables numéricas como balance, duration y campaign.

---

5. Modelado y resultados

Para esta parte del trabajo, nos dividimos en dos subgrupos. Uno se encargó de implementar la regresión logística y el otro el modelo de random forest.

La regresión logística fue fácil de interpretar y rápida de entrenar, pero su precisión fue menor (77.65%). En cambio, el random forest obtuvo una

precisión del 84.83%, mostrando una mejor capacidad para predecir correctamente los casos positivos (cuando el cliente acepta).

A partir de los resultados, todos coincidimos en que random forest es más robusto para este tipo de datos y que puede manejar mejor las variables no lineales. Las matrices de confusión y los reportes de clasificación nos ayudaron a comparar los modelos y a entender mejor su rendimiento.

La matriz de confusión es una herramienta fundamental para evaluar el rendimiento de los modelos de clasificación, ya que permite analizar la cantidad de aciertos y errores que comete el modelo al clasificar los datos. En este caso, se aplicó tanto a la regresión logística como al bosque aleatorio.

Regresión Logística

En este modelo, la matriz de confusión permitió observar cómo se desempeñó el algoritmo al predecir si un cliente contrataría un depósito a plazo fijo. Se tienen cuatro posibles salidas:

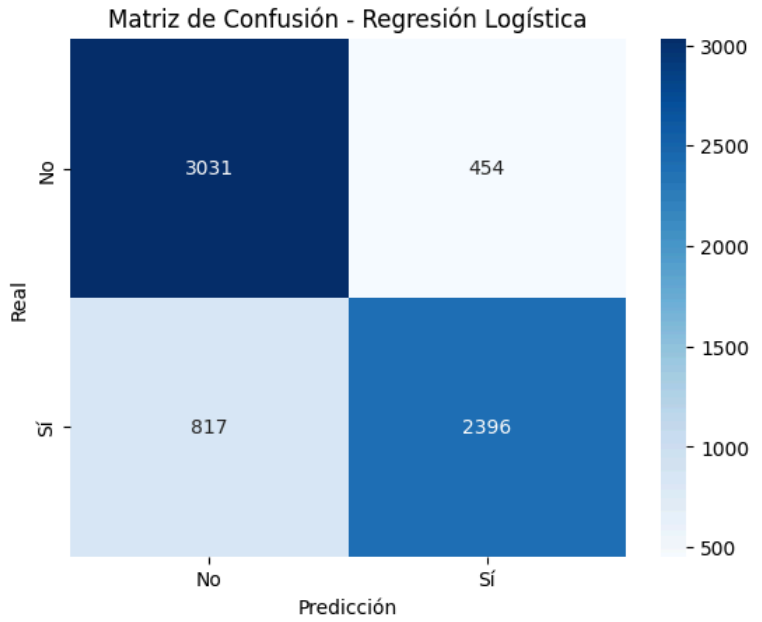
Verdaderos positivos (VP): clientes que efectivamente contrataron el depósito y fueron correctamente clasificados como tal.

Falsos positivos (FP): clientes que no contrataron el depósito, pero el modelo predijo que sí lo harían.

Verdaderos negativos (VN): clientes que no contrataron el depósito y fueron correctamente clasificados.

Falsos negativos (FN): clientes que sí contrataron el depósito, pero el modelo no lo predijo.

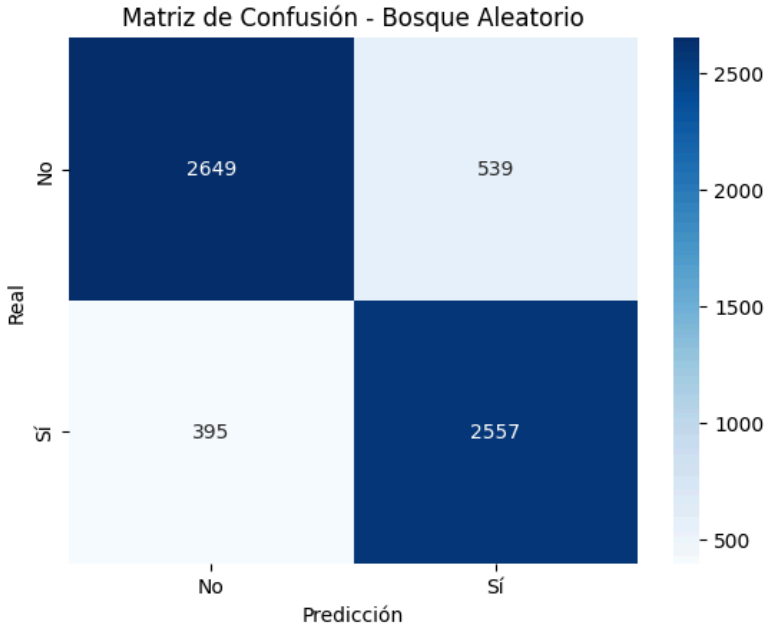
Este análisis permitió observar que la regresión logística tuvo un buen desempeño general, aunque presentó algunas limitaciones a la hora de detectar ciertos casos positivos, lo cual puede representar un problema si se busca maximizar la captación de potenciales clientes.



Bosque Aleatorio

El modelo de bosque aleatorio, al ser más complejo y basado en la combinación de múltiples árboles de decisión, mostró una matriz de confusión con mejor capacidad de clasificación en comparación con la regresión logística. Se observó un mayor número de verdaderos positivos y una reducción en los falsos negativos, lo que indica una mejor habilidad del modelo para detectar correctamente a los clientes interesados.

Este resultado sugiere que el bosque aleatorio podría ser una mejor opción para predecir con mayor precisión el comportamiento de los clientes, especialmente si el objetivo principal es minimizar los casos en los que se pierde la oportunidad de captar a alguien interesado en contratar un producto del banco.



---

6. Conclusión Final

Como grupo, este proyecto nos permitió aplicar de forma concreta muchos de los conceptos que fuimos aprendiendo durante las clases. No solo trabajamos con un conjunto de datos real, sino que también atravesamos todas las etapas del proceso de ciencia de datos: desde la exploración inicial, pasando por el preprocesamiento y análisis visual, hasta la construcción y evaluación de modelos predictivos.

Una de las cosas más valiosas que aprendimos fue la importancia de entender bien los datos antes de aplicar cualquier modelo. Nos dimos cuenta de que un modelo, por muy avanzado que sea, no puede dar buenos resultados si no se trabaja correctamente con la calidad de los datos. También valoramos la utilidad de visualizar los datos para encontrar patrones que no son obvios a simple vista.

Además, pudimos comprobar que distintos modelos tienen diferentes niveles de rendimiento y que no siempre el más complejo es el más adecuado. En nuestro caso, el modelo de Random Forest superó a la regresión logística, pero ambos aportaron información útil.

Por último, destacamos el trabajo en equipo. Nos organizamos bien, lo hicimos entre todos juntos e íbamos discutiendo las ideas, lo que nos permitió complementar nuestras habilidades y lograr un resultado más completo. Este trabajo no solo fue un ejercicio técnico, sino también una experiencia de colaboración que nos preparó mejor para futuros desafíos.

---

7. Consideraciones adicionales

Durante el trabajo surgieron muchas ideas para seguir explorando. Algunos del grupo propusieron usar más variables (como tipo de cuenta o nivel de ingresos), mientras que otros pensaron en aplicar modelos más avanzados como Gradient Boosting o redes neuronales.

Coincidimos en que este trabajo es solo un punto de partida y que se pueden hacer muchos más experimentos para mejorar la precisión y aplicabilidad del

modelo. También sería interesante trabajar con conjuntos de datos más grandes y variados para validar lo que descubrimos.

Referencia: https://github.com/XERYVEL/Trabajo_Final_Cs_Datos