Introducción al Aprendizaje Automático, ECyT UNSAM

Noviembre 2024

“¿El cliente se suscribirá a un plazo fijo?”

Grupo 2

Autores: **Marcos Achaval** - **Federico Menicillo** - **Lucas Golchtein**

Universidad Nacional de San Martín

Agenda

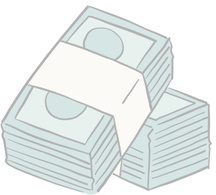
1. Objetivo de estudio
2. Presentación de los datos
3. Análisis Exploratorio de los Datos (EDA)
4. Modelado
5. Conclusión

Objetivo

Predecir si los clientes de un banco se van a suscribir a un plazo fijo

Responderemos a la pregunta de estudio a través de una serie de análisis de la muestra de datos y modelos predictivos.

Comencemos...



45211 Instancias

17 Atributos

Variable	Tipo	NaN
age	int	-
job	str	-
marital	str	-
education	str	-
default	str	-
balance	int	-
housing	str	-
loan	str	-
contact	str	-
day	int	-
month	str	-
duration	int	-
campaign	int	-
pdays	int	-
previous	int	-
poutcome	str	-
y	str	-

Fuente: https://archive.ics.uci.edu/dataset/222/bank+marketing

EDA

1. Eliminamos “**duration**”
2. Unificamos categorias (en lo posible)

contact

- **telephone** → **cellular**
- **cellular** → **cellular**
- **unknown** → **unknown**

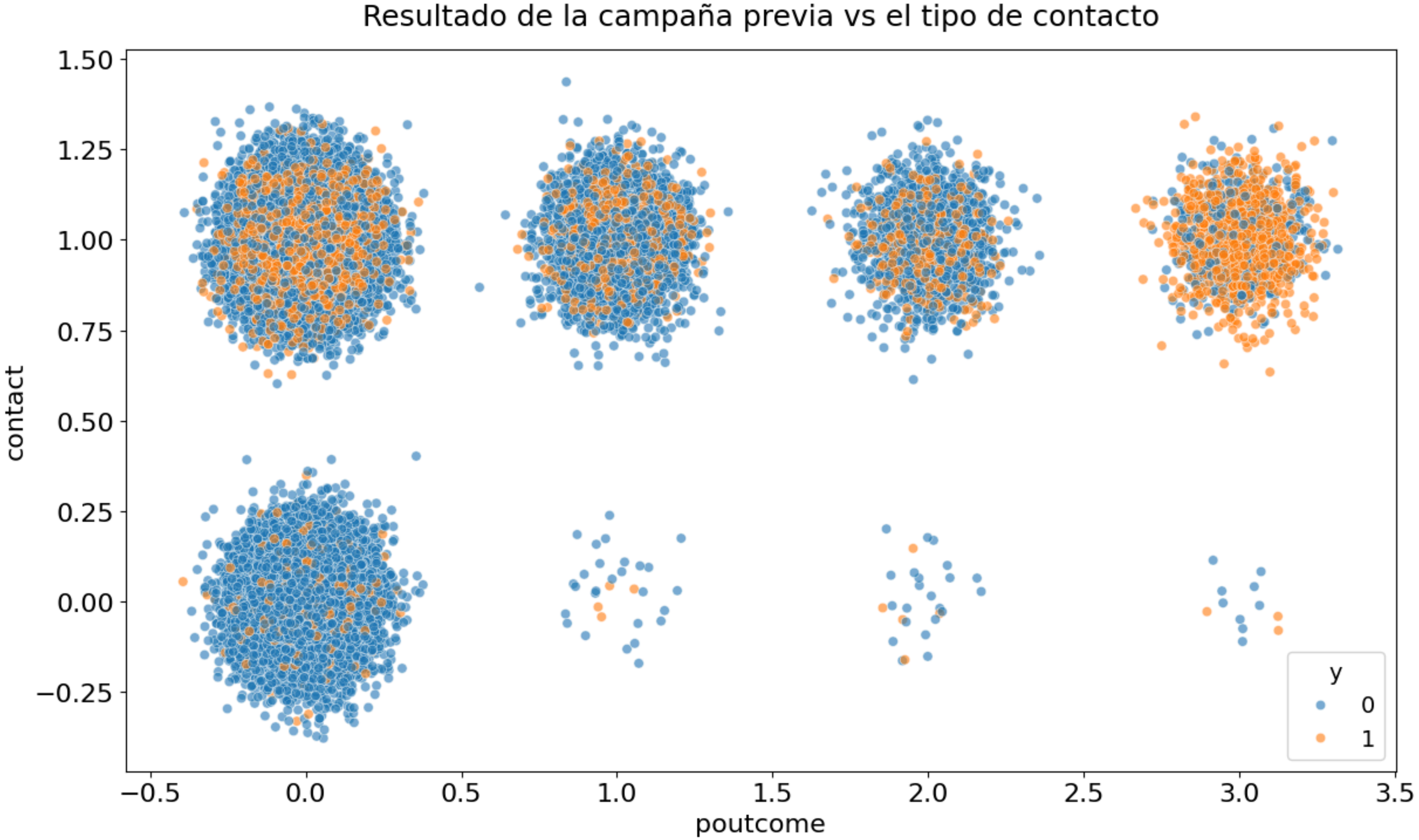
3. Convertimos variables “**str**” a “**int**”

default
housing
loan
y

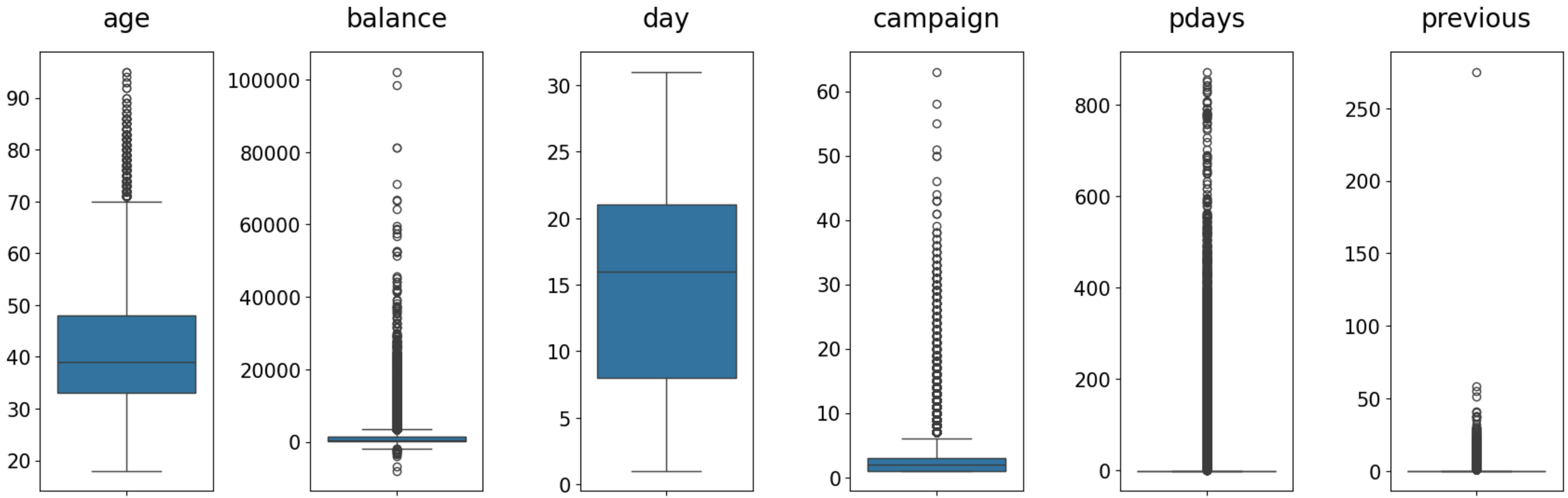
{“yes”: 1, “no”: 0}

education {“unknown”: 0, “primary”: 1, ...}

month {“jan”: 0, “feb”: 1, “mar”: 2, ...}


Resultado de la campaña previa vs el tipo de contacto
contact
poutcome
y
0
1
1.50
1.25
1.00
0.75
0.50
0.25
0.00
−0.25
−0.5
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5

Boxplots de las variables originalmente numéricas



Variables con mayor correlación con “*y*”

poutcome	1.00	0.26	0.00	0.71	0.49	0.26
contact	0.26	1.00	-0.21	0.25	0.15	0.15
housing	0.00	-0.21	1.00	0.12	0.04	-0.14
pdays	0.71	0.25	0.12	1.00	0.45	0.10
previous	0.49	0.15	0.04	0.45	1.00	0.09
y	0.26	0.15	-0.14	0.10	0.09	1.00
	poutcome	contact	housing	pdays	previous	y

Modelado

Introducción al modelado

Proporción de clases:

Clase 0	Clase 1
88% (39.922)	12% (5.289)

Train-Test Split (70/30)

Train:

Clase 0	Clase 1
88% (27.945)	12% (3.702)

Test:

Clase 0	Clase 1
88% (11.977)	12% (1.587)

Métrica de evaluación:

$$\textbf{F1-Score} = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

Benchmark:

Nuestro modelo benchmark, se trata de un **DummyClassifier**. La técnica del modelo es simple y eficiente, clasifica a partir de la proporción de clases existentes.

Al contar con el 88% de instancias totales correspondientes a la clase 0 y el 12% de instancias clasificadas con 1, este modelo predice aproximadamente 1 de cada 10 filas como positivas y 9 de cada 10 como negativas.

Evaluación del Benchmark

F1 Score	**0.12**

Modelo 1:

- **Modelo simple**
- **Árbol de Decisión**
- Variables predictoras: “**poutcome**” y “**housing**”

Evaluación del Modelo 1

F1 Score	**0.28**

Variables a utilizar en el modelo 2

age	1.00	0.00	0.10	-0.01	0.01	0.03
job	0.00	1.00	-0.03	-0.03	0.01	0.02
balance	0.10	-0.03	1.00	0.12	0.04	0.05
day	-0.01	-0.03	0.00	1.00	-0.07	-0.03
poutcome	0.01	0.01	0.04	-0.07	1.00	0.26
y	0.03	0.02	0.05	-0.03	0.26	1.00
	age	job	balance	day	poutcome	y

Modelo 2:

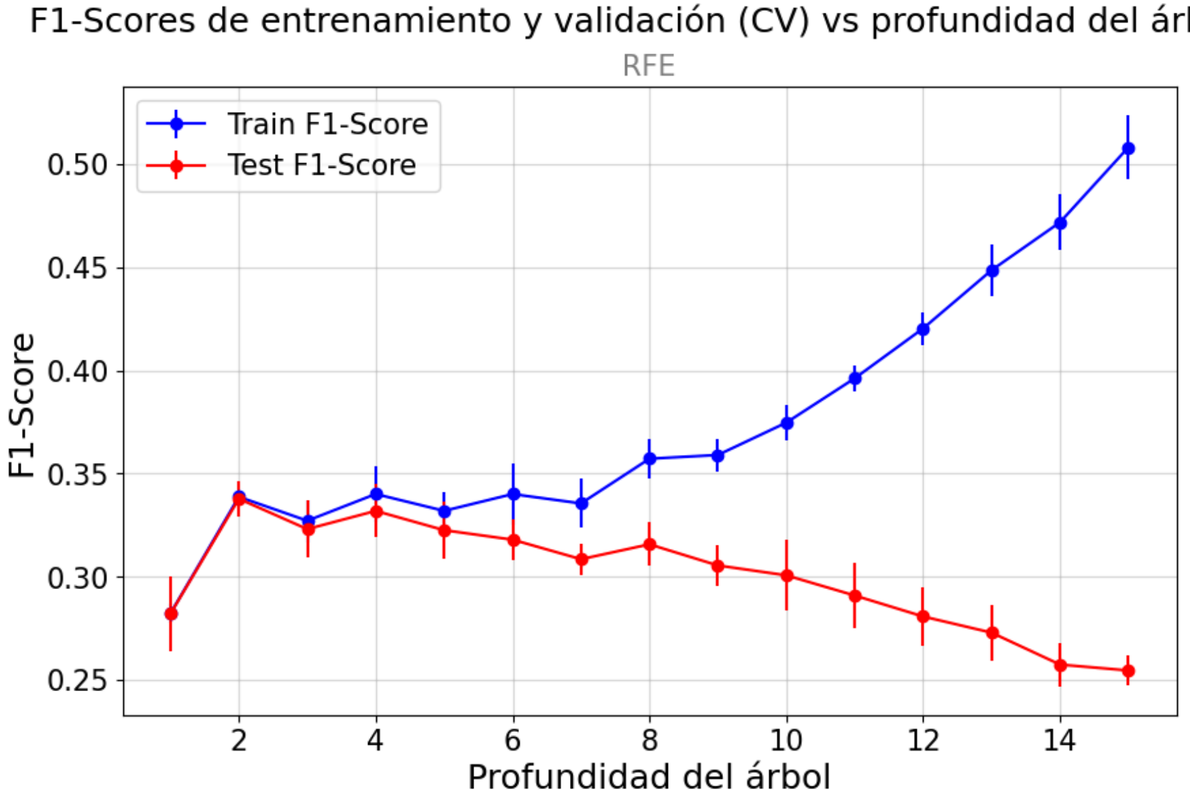
- **Modelo más complejo**
- **Árbol de Decisión**

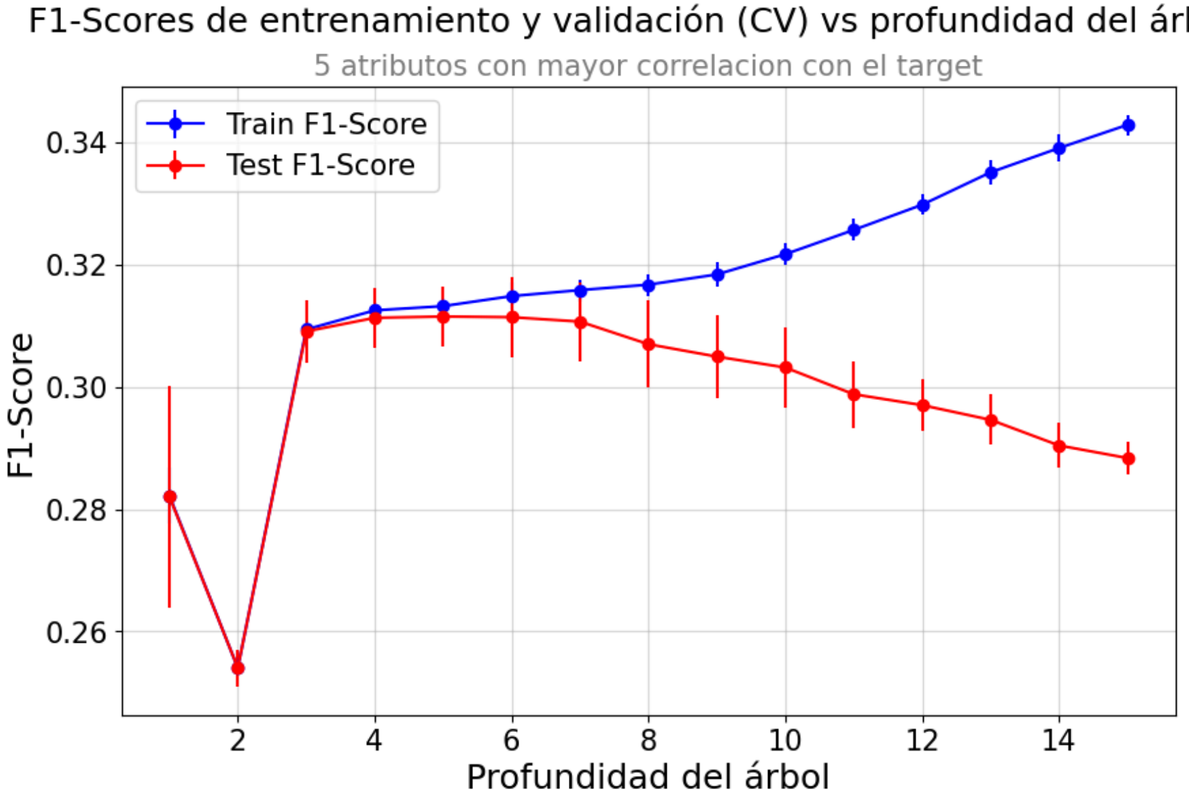
- Recursive Feature Elimination (RFE)
- Curva de complejidad
- Optimización de hiperparámetros
  - RandomizedSearchCV
  - GridSearchCV

Evaluación del Modelo 2

GridSearchCV

F1 Score (train)	**0.35 ± 0.00**
F1 Score (test)	**0.35 ± 0.02**


F1-Scores de entrenamiento y validación (CV) vs profundidad del árbol
RFE
Train F1-Score
Test F1-Score
F1-Score
Profundidad del árbol
0.50
0.45
0.40
0.35
0.30
0.25
2
4
6
8
10
12
14


F1-Scores de entrenamiento y validación (CV) vs profundidad del árbol
5 atributos con mayor correlacion con el target
Train F1-Score
Test F1-Score
F1-Score
Profundidad del árbol
0.34
0.32
0.30
0.28
0.26
2
4
6
8
10
12
14

Modelo 3:

- **Random Forest** } RandomizedSearchCV

Evaluación del Modelo 3

RandomizedSearchCV

F1 Score (train)	**0.37 ± 0.00**
F1 Score (test)	**0.36 ± 0.02**

Comparación de modelos:

1) Benchmark:

F1 Score	**0.12**

2) Modelo 1: Árbol simple

F1 Score	**0.28**

3) Modelo 2: Árbol complejo

GridSearchCV

F1 Score (train)	**0.35 ± 0.00**
F1 Score (test)	**0.35 ± 0.02**

4) Modelo 3: Random forest

RandomizedSearchCV

F1 Score (train)	**0.37 ± 0.00**
F1 Score (test)	**0.36 ± 0.02**

Conclusiones:

Modelo elegido:

Modelo 2 (Árbol complejo)

Evaluación en validación

F1 Score	**0.36**

Clientes que se suscriben:

- Son estudiantes.
- En la campaña anterior se suscribieron al plazo fijo.
- Tienen más de 60 años y no se suscribieron previamente.

Proximos pasos:

- Probar nuevas técnicas para balancear la muestra:
  - Oversampling
  - Undersampling
- Probar otros modelos para mejorar los resultados obtenidos en este proyecto:
  - XGboost
- Evaluación de modelos con el área bajo la curva de precisión-exahustividad (PR-AUC)

¡Muchas gracias por su atención!

Esperamos que les haya parecido interesante :)

Presentaron: **Marcos Achaval** - **Federico Menicillo** - **Lucas Golchtein**

¿Alguna pregunta?