

SPRINT 4

Python

Grupo Sherlock

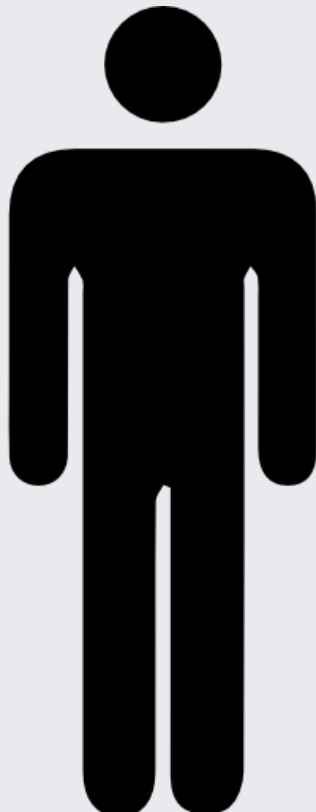




Lo que intentaremos responder:

**¿El cliente se suscribirá
a un depósito a plazo fijo?**

Perfil del cliente ideal



Ocupación

**Estudiantes y
Retirados**

Educación

**Alto nivel
académico**

Préstamos

**Sin préstamo
personal previo**

Resultado en campañas anteriores

**Se suscribió en campañas
anteriores al plazo fijo**

¿Cómo llegamos al resultado?

01

Análisis de Datos

Comunicar los datos de manera precisa y detectar relaciones.

02

Tratamiento del Dataset

Transformación y limpieza del dataset.

03

Evaluación de modelos predictivos

Crear, entrenar y evaluar modelos de Machine Learning.

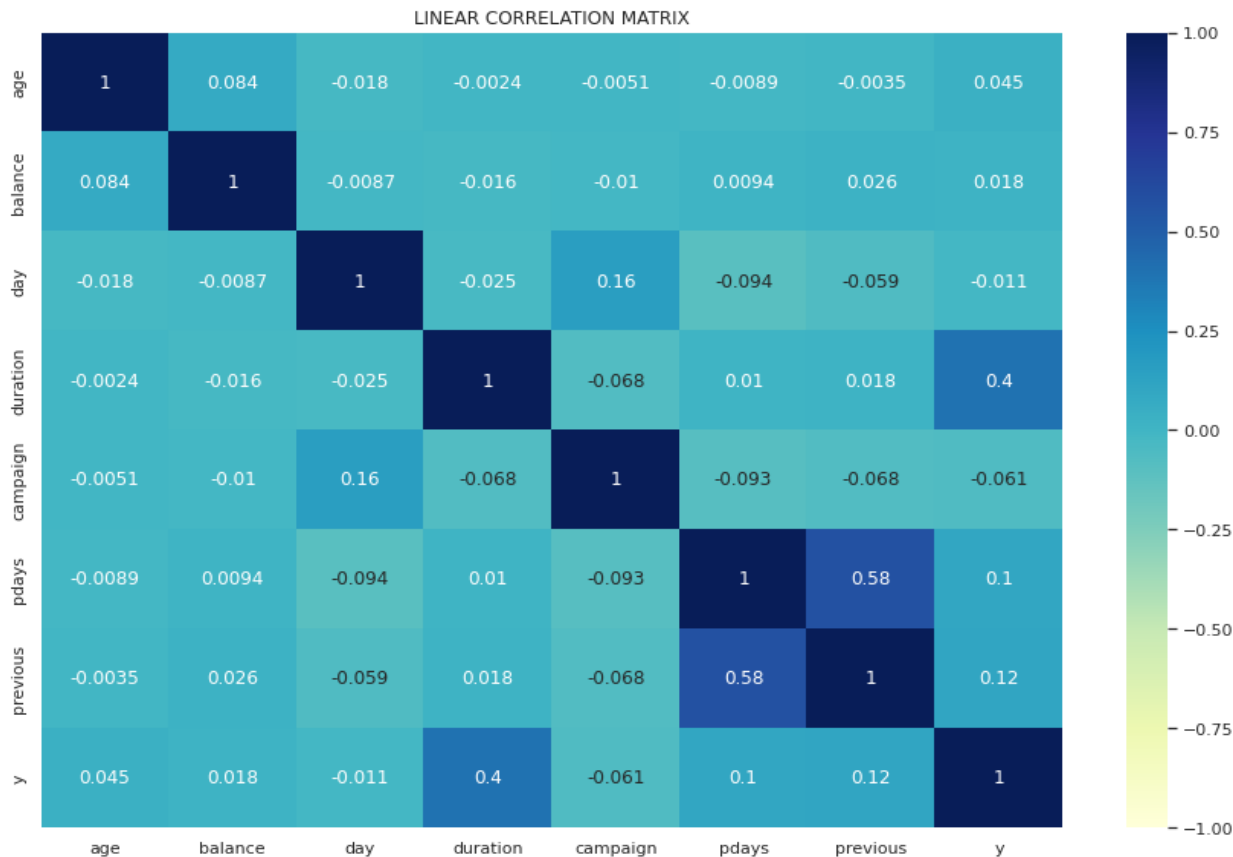
01

Analisis de Datos

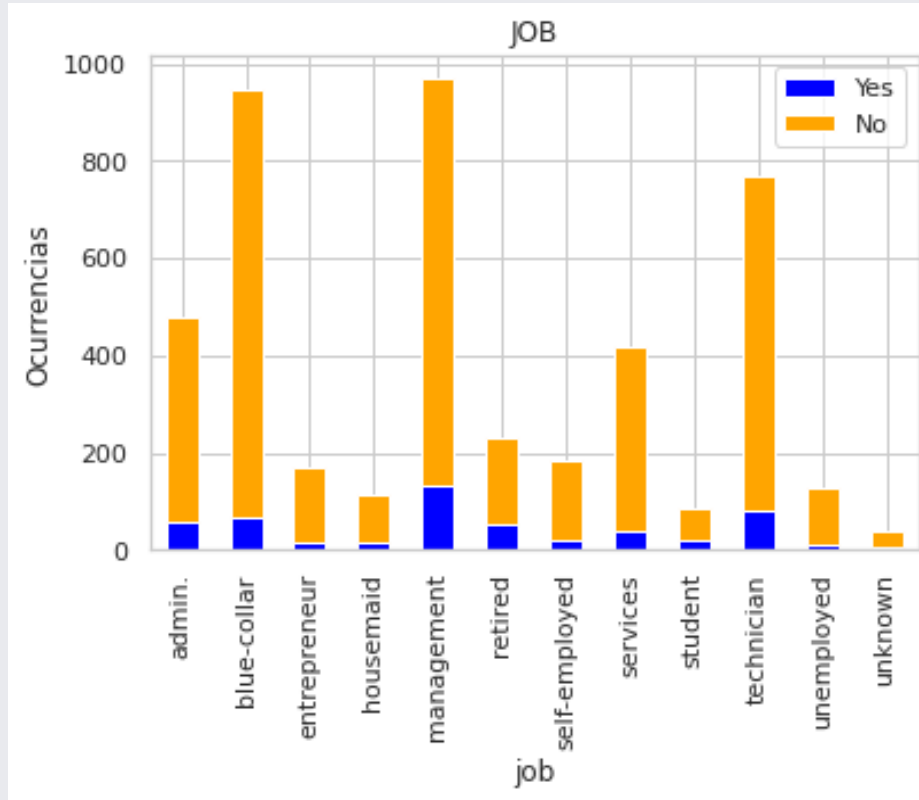
Analizar los datos de manera precisa, detectar relaciones y visualizarlos.

Matriz de Correlación

Las variables con mayor relación son "Previous" y "Pdays".



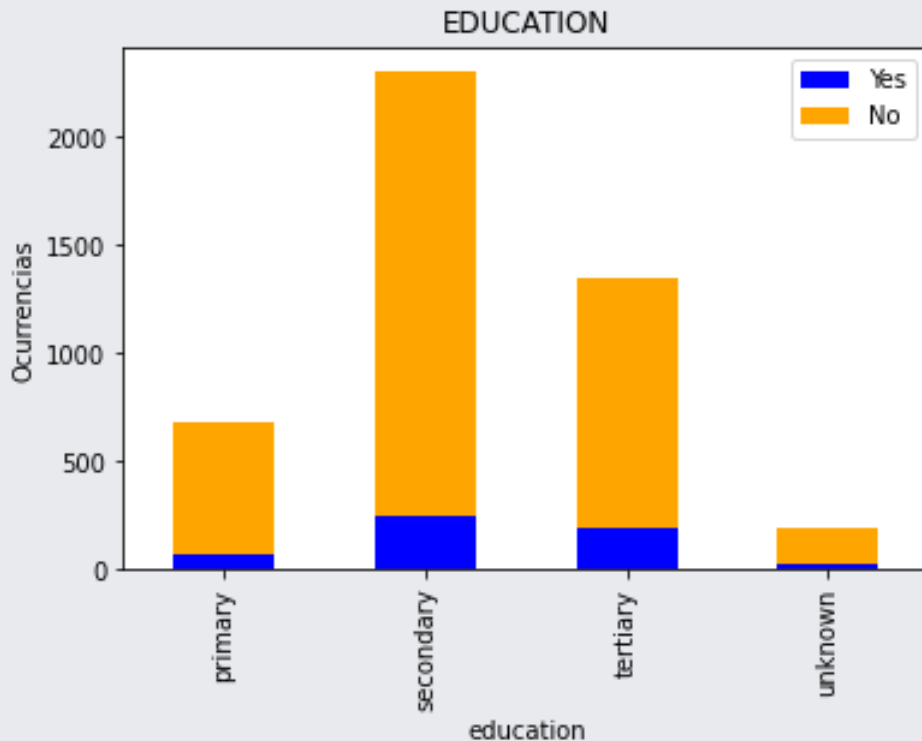
Occupation: Ocupación/Trabajo de los encuestados



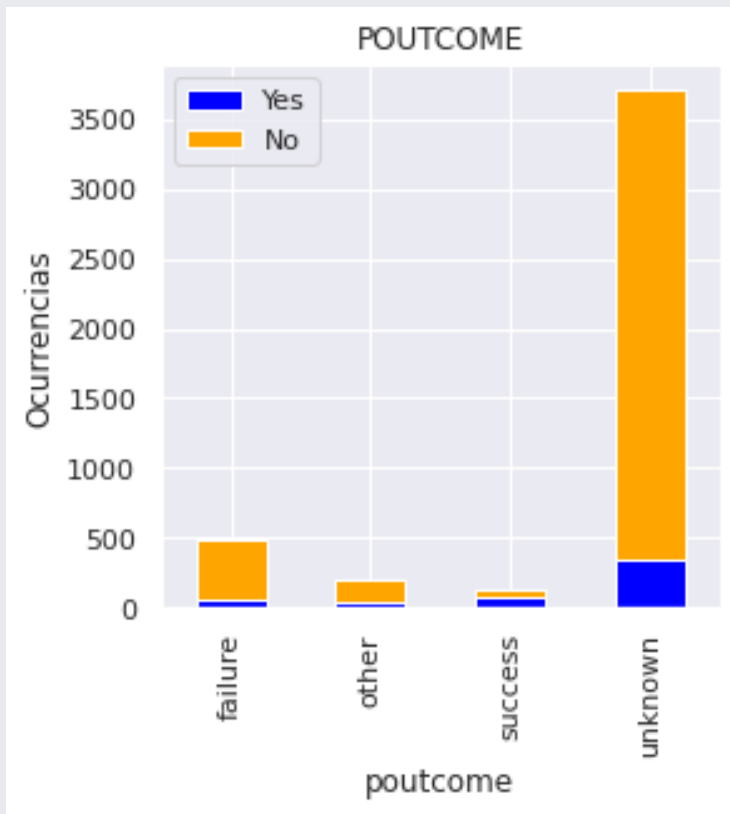
Mayor probabilidad de aceptación del crédito entre los estudiantes y los retirados.

Education: Educación alcanzada de los encuestados

Hay cierta tendencia de que a mayor nivel académico alcanzado, mayor probabilidad de suscripción.



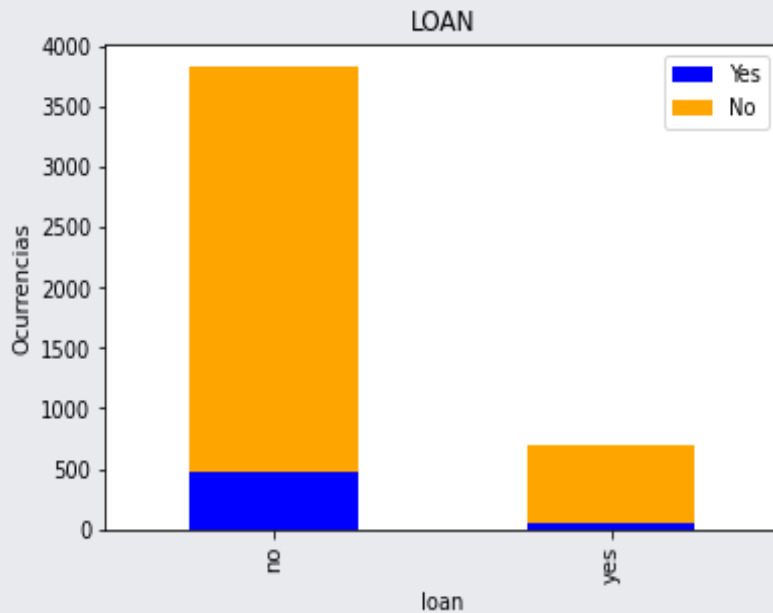
Poutcome: Resultado de la campaña de marketing anterior



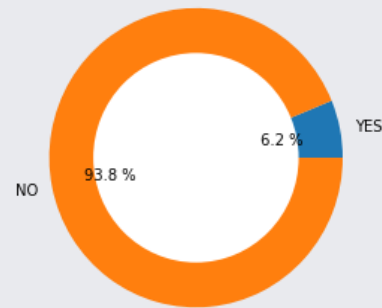
"Success" es el mas influyente en la suscripción de la campaña actual.

Mayor cantidad de ocurrencias de respuesta en la barra "unknown" de resultados de campañas anteriores.

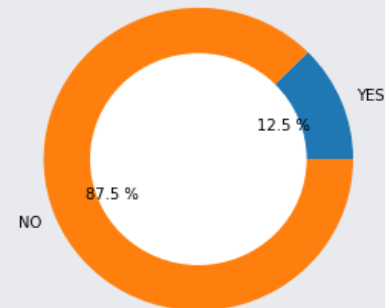
Loan: ¿El encuestado cuenta con un préstamo personal?



Si el encuestado tiene préstamo personal, hay tendencia a la NO suscripción,

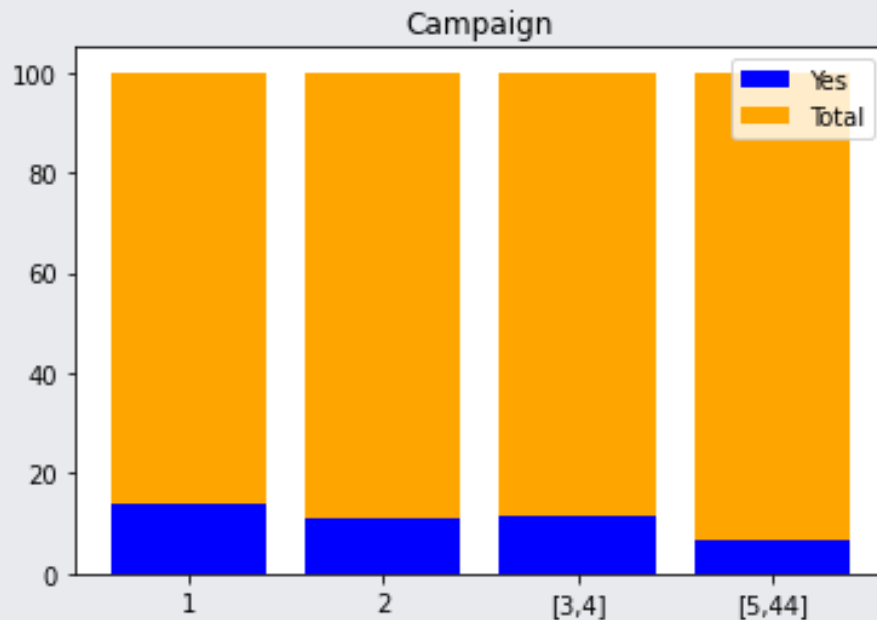


LOAN



NO LOAN

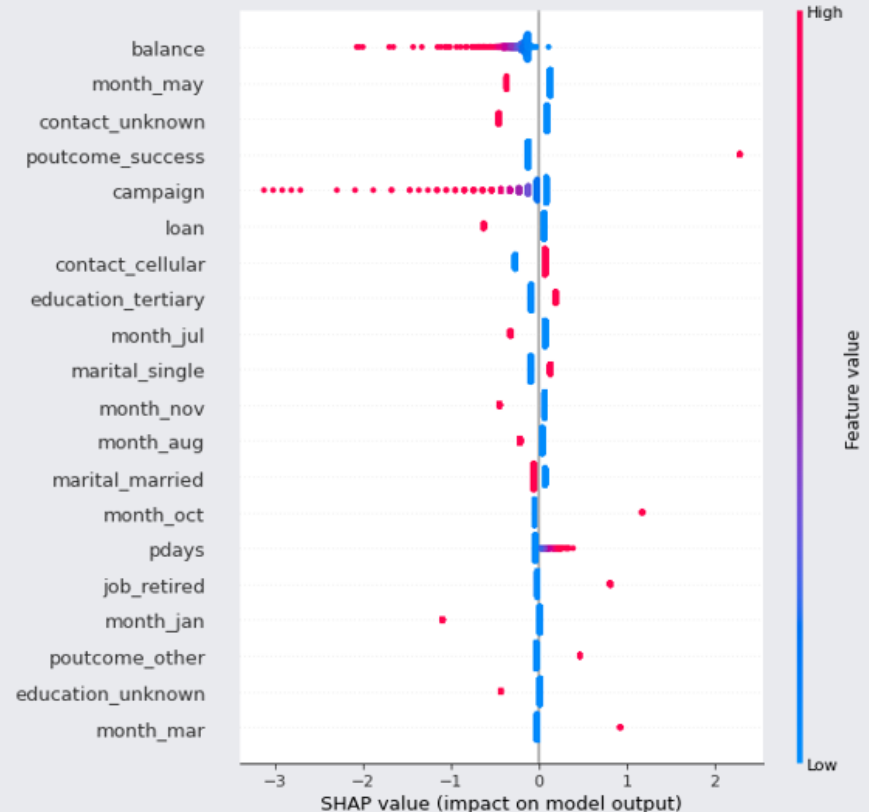
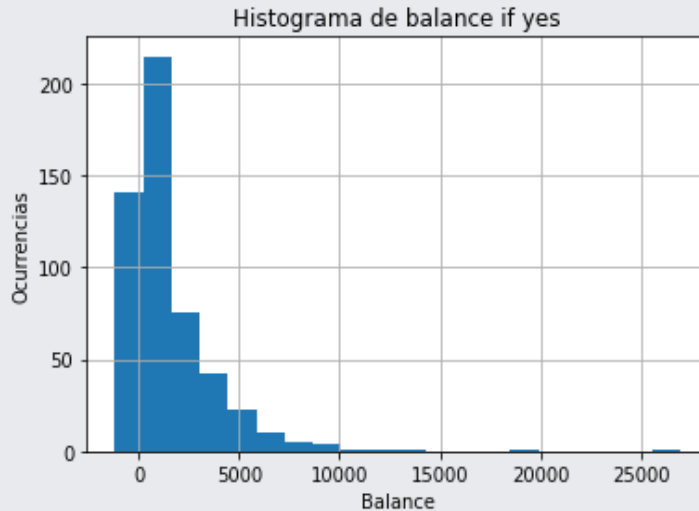
Campaign: Cantidad de contactos en la campaña actual



Mayor porcentaje de suscripciones a menor cantidad de contactos

¿Por qué no utilizamos el balance como variable predictora?

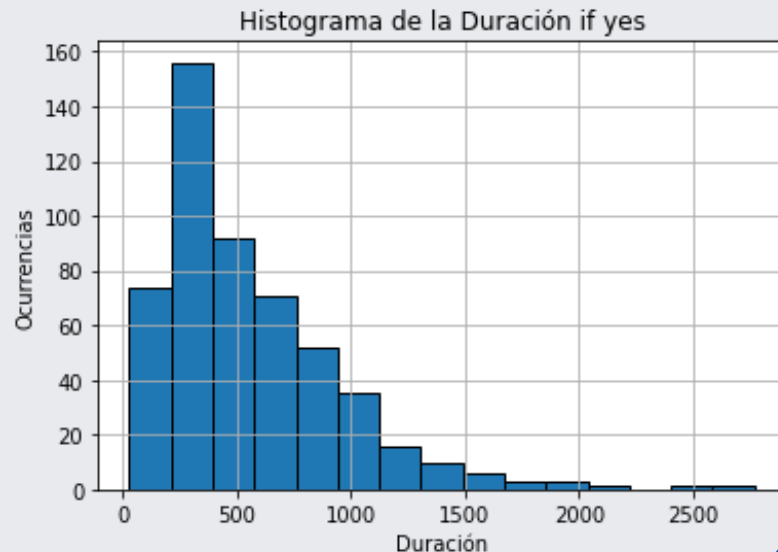
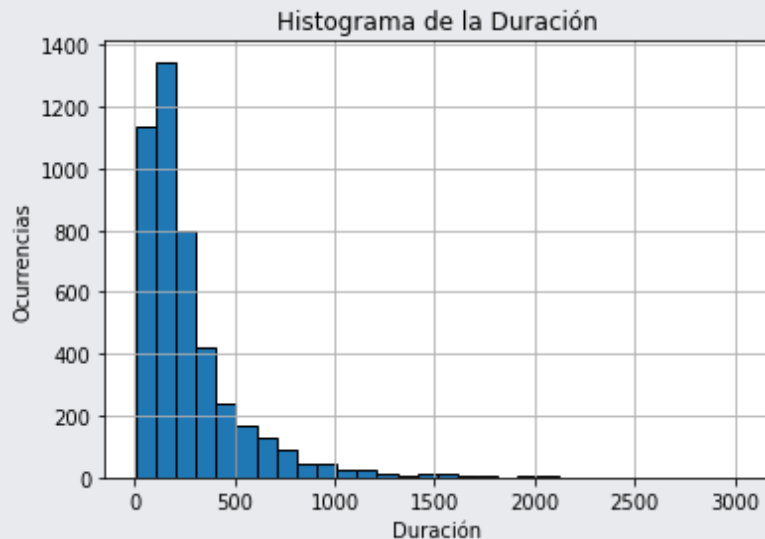
Balance: Saldo medio anual de los encuestados (€)



El 75%
de los encuestados tienen un saldo
medio inferior a los € 2160

¿Por qué no utilizamos la duración como variable predictora?

Duration: Tiempo de último contacto con el cliente (seg).



Marcado cambio en la forma del histograma cuando vemos la duración únicamente de los éxitos

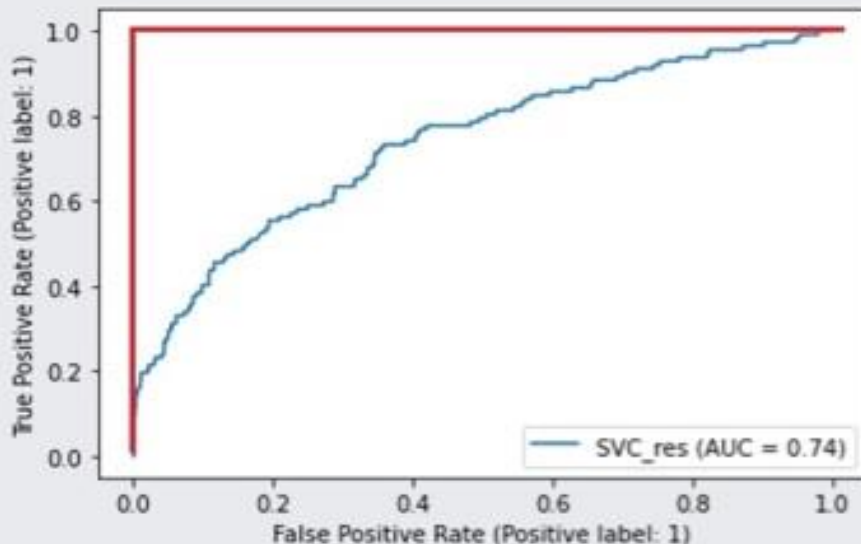
La duración del último contacto fue mayor si hubo un éxito, lo cual era de esperarse.

Modelo Predictivo Elegido:

SVM

El modelo

SVM: Support Vector Machine

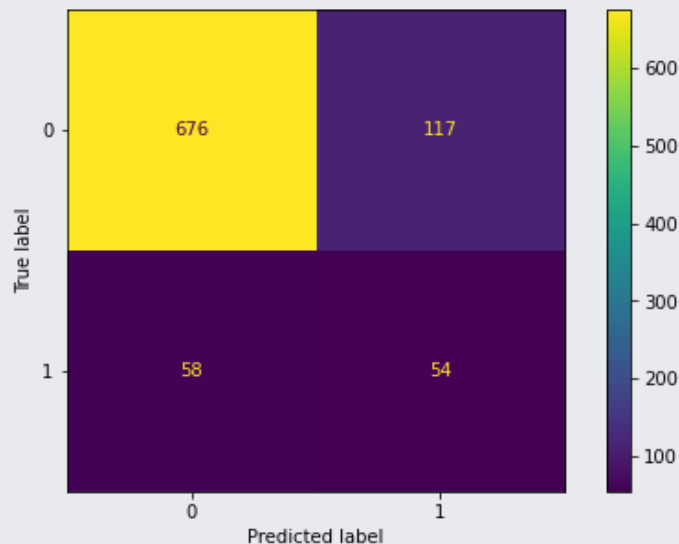


74% de probabilidad de que el modelo pueda distinguir entre clase positiva y clase negativa.

Curva ROC

- Mide qué tan bien se clasifican las predicciones, en lugar de sus valores absolutos.
- Mide la calidad de predicciones del modelo.

Métricas



VN: verdadero negativo. 676 de 793 negativos

FN: falso negativo. 58 de 112 positivos

VN: verdadero negativo. 676 de 793 negativos

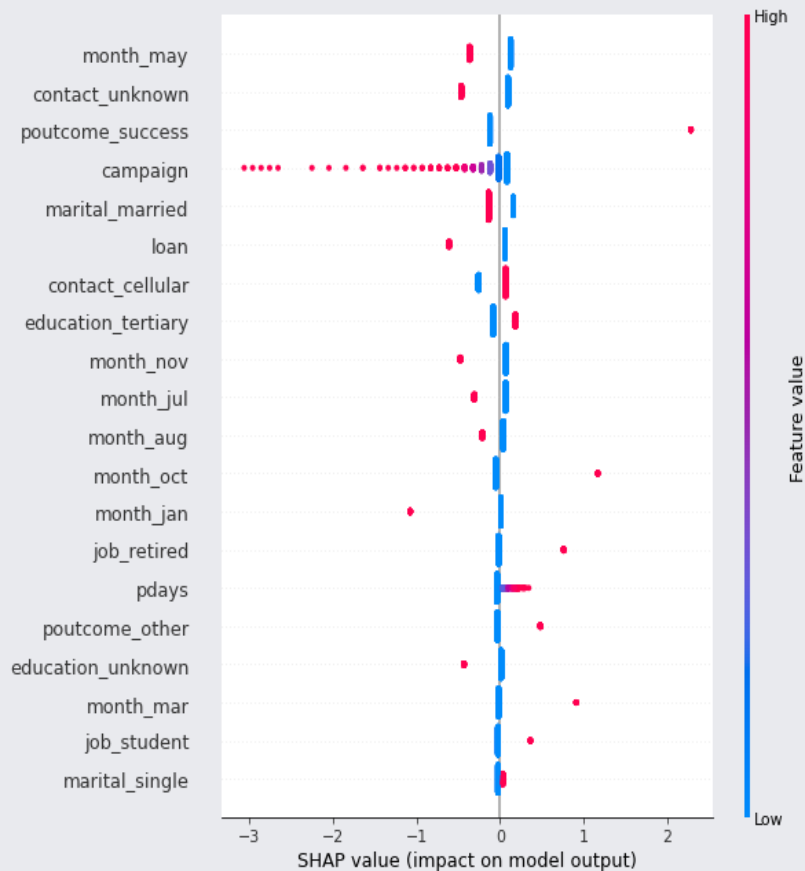
FP: falso positivo. 117 de 763 negativos

Accuracy	Recall 0	Recall 1	Precision 0	Precision 1	F1 score 0	F1 Score 1
0,81	0,85	0,48	0,92	0,32	0,89	0,38

SHAP Values

Variables descartadas:

Duration
Balance
Age
Day
Previous
Default
Housing



02

Tratamiento del Dataset

Transformación de los datos, estandarización
y limpieza de los datos.

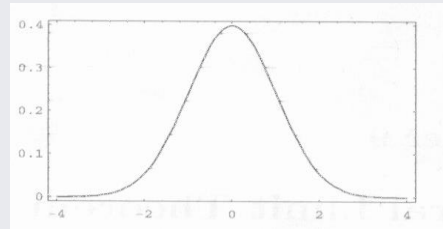
Normalizado

- Traduce los datos a una escala representativa de cada variable en el que los datos toman valores entre 0 y 1.
- Esto previene que haya errores por las distintas escalas entre las variables.

Estandarizado

$$Z_i = \frac{(X_i - \mu)}{\sigma}$$

- Ajusta y escala a los datos para que tengan la forma de una distribución normal estandar, con una media 0 y un desvio estandar de 1.



Con resampling

Sin resampling



Resampling de datos - ImbLearn

Crea registros similares a los que se tienen de forma de aumentar la cantidad y tener una cantidad representativa a la hora de entrenar el modelo lo cual aumenta la performance.

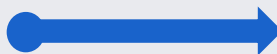
A través de...

Librería ImbLearn en Python

Dummy Variables

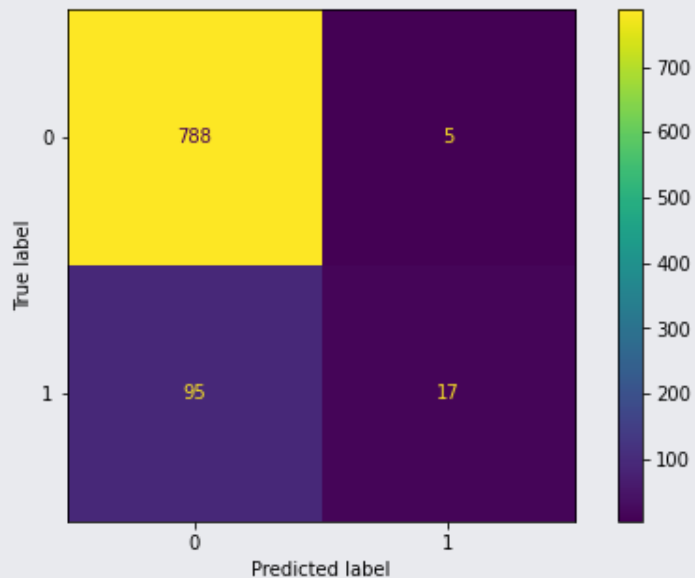
Variables categóricas que se re-expresan en variables numéricas.

Sueldo	Edad	Estado Civil
50.000	25	Soltero
200.000	45	Casado
100.000	40	Soltero

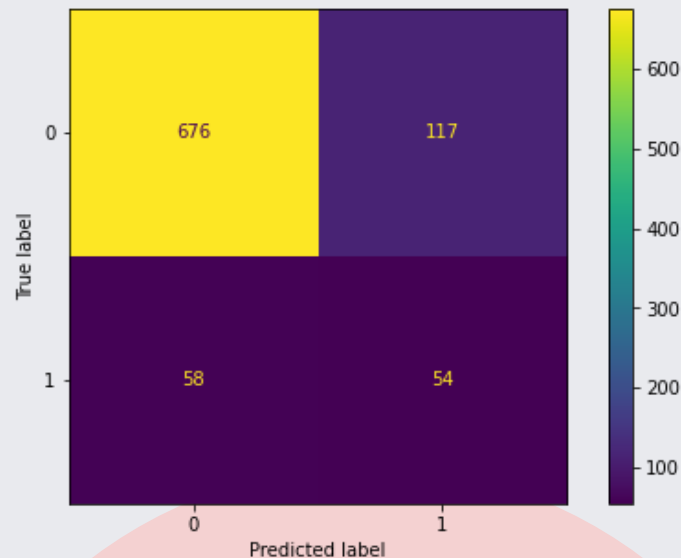


Sueldo	Edad	Soltero	Casado
50,000	25	1	0
200,000	45	0	1
100,000	40	1	0

Matriz de confusión (SVM)



Trabajando sobre el mismo Test Set



Con remuestreo

03

Evaluación de modelos predictivos

Crear, entrenar y evaluar modelos de Machine Learning

Problema

Necesitamos determinar
el modelo mas preciso.

Solución

Entrenar los modelos,
analizando las métricas y
predicciones obtenidas en
cada caso.

Modelos estudiados



Decision Tree

Hace uso de la metodología de ramificación para ejemplificar todos los resultados posibles en función de ciertas condiciones.



Logistic Regression

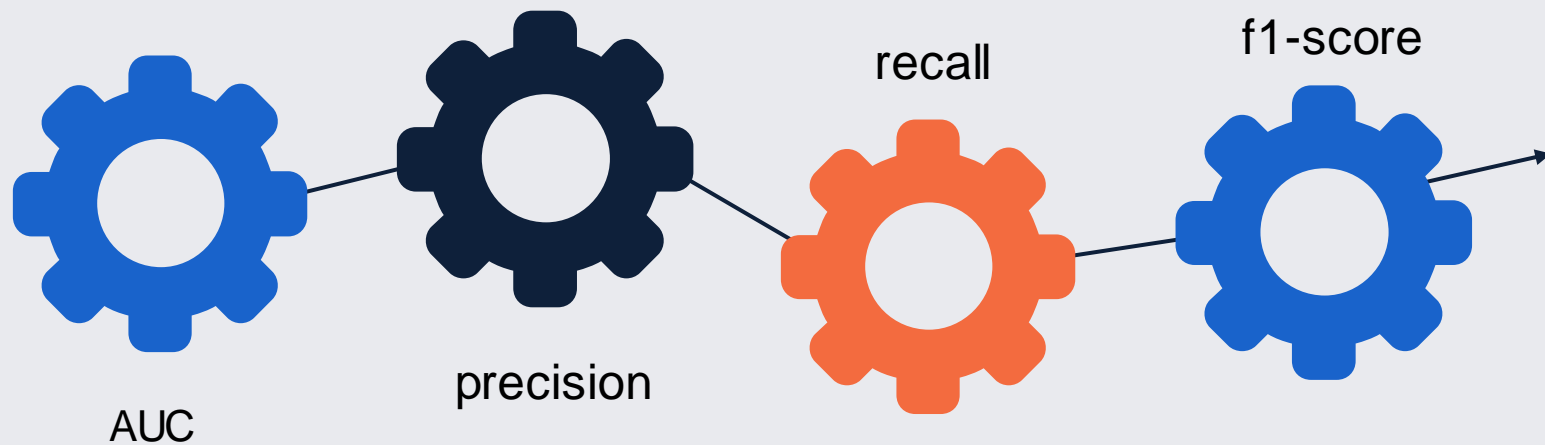
Utiliza una función llamada función Sigmoidal para mapear las predicciones y sus probabilidades.



SVM

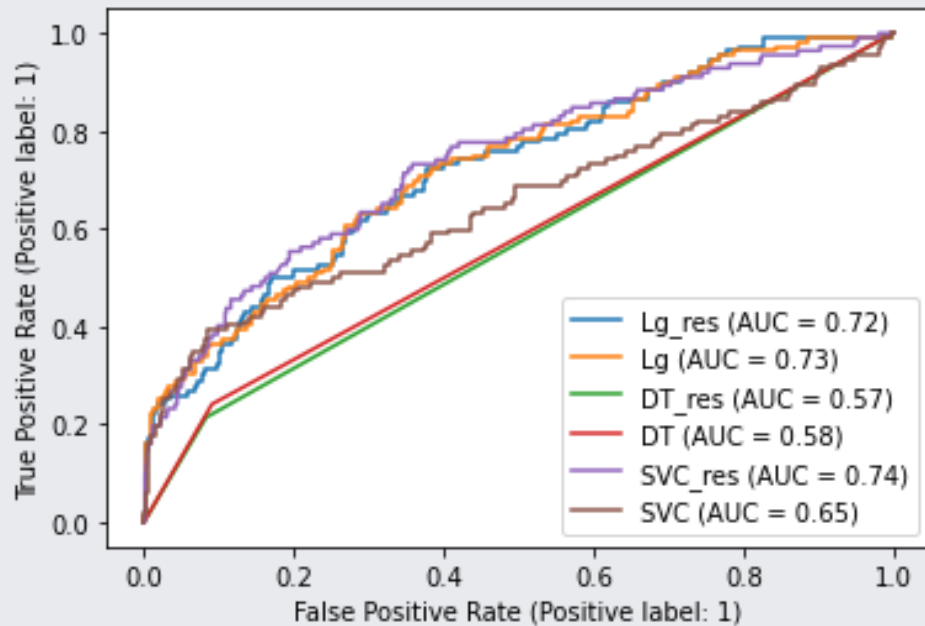
Algoritmo flexible usado para la clasificación y la regresión. Popular por su capacidad para manejar múltiples variables continuas y categóricas. El objetivo es dividir los conjuntos de datos en clases para encontrar un hiperplano que las separe.

¿Que métricas analizamos para elegir el mejor modelo?



¿Por qué no nos enfocamos en el valor de Accuracy?
(exactitud)

Comparación de Curvas ROC



Logistic Regression
Resampleado

(AUC = 0.72)

Logistic Regression

(AUC = 0.73)

DecisionTree
resampleado

(AUC = 0.57)

DecisionTreeClassifier

(AUC = 0.58)



SVM resampleado

(AUC = 0.74)

SVM

(AUC = 0.65)

Comparación de Métricas

MODELOS	Accuracy	Recall 0	Recall 1	Precision 0	Precision 1	F1 score 0	F1 score 1
Log Reg	0,89	1	0,13	0,89	0,94	0,94	0,23
Log Reg Resampleado 	0,79	0,84	0,46	0,92	0,29	0,88	0,35
Decision Tree	0,83	0,91	0,24	0,89	0,70	0,93	0,26
Decision Tree Resampleado	0,8	0,88	0,25	0,89	0,22	0,88	0,27
SVM	0,89	0,99	0,15	0,89	0,77	0,94	0,25
SVM Resampleado 	0,81	0,85	0,48	0,92	0,32	0,89	0,38



**Muchas
Gracias**