1) Carga y Exploración de Datos (EDA)

Se utilizó el dataset Bank Marketing, compuesto por 4.119 instancias de clientes de un banco.
El objetivo es predecir si el cliente se suscribirá a un depósito a plazo fijo (variable objetivo `y`)

En la variable objetivo (`y`), 3.668 clientes (89%) respondieron que sí y 451 (11%) que no, mostrando una leve desproporción hacia el grupo positivo.

Entre las variables predictoras, se destacan:

- Age: la mayoría de los clientes tiene entre 25 y 35 años, siendo este rango el más representativo del conjunto.

- Poutcome (resultado de la campaña anterior): la mayoría de los registros corresponden a `unknown` (3.231 “no” y 292 “sí”), lo que indica que la mayoría de los clientes no tenían un resultado previo conocido. Los valores de `success` (92 “no” y 50 “sí”) y failure (387 “no” y 67 “sí”) son menos frecuentes, pero muestran una relación directa entre haber tenido éxito en campañas anteriores y una mayor probabilidad de suscribirse nuevamente.

2) Preprocesamiento y División de Datos

Se utilizó el widget Select Columns para eliminar variables que no aportan información predictiva o que introducen sesgo, como:

- `duration`, `contact`, `day`, y `month`.

Luego, se aplicó Continuize para convertir las variables categóricas (`job`, `marital`, `education`, `housing`, `loan`, etc.) en formato numérico adecuado para los algoritmos de machine learning.

Finalmente, los datos fueron divididos mediante data sampler en un 80% para entrenamiento y 20% para prueba.

3) Modelado y Evaluación

Se entrenaron tres modelos de clasificación en el widget Test & Score:

- Regresión Logística

- Random Forest

- Naive Bayes

Los resultados de las métricas principales fueron los siguientes:

Modelo	AUC	F1-score	Accuracy
Logistic Regression	0.751	0.876	0.901
Random Forest	0.730	0.878	0.896
Naive Bayes	0.731	0.812	0.812

El modelo con mayor AUC y Accuracy fue la Regresión Logística, mientras que Random Forest obtuvo el mayor F1-score por una pequeña diferencia.

4) Matriz de Confusión (Regresión Logística)

Los resultados obtenidos en el widget Confusion Matrix para el modelo de Regresión Logística fueron:

- True Positives (TP): 90.8%
- True Negatives (TN): 69.2%
- False Positives (FP): 9.2%
- False Negatives (FN): 30.8%

Esto indica que el modelo tiene una alta capacidad para identificar correctamente los clientes que efectivamente se suscriben, aunque comete algunos errores en la predicción de los casos negativos.

5) Curva ROC y Conclusiones

La curva ROC muestra que la Regresión Logística presenta la mejor relación entre la Tasa de Verdaderos Positivos y la Tasa de Falsos Positivos . Su curva se ubica más cerca del vértice superior izquierdo del gráfico, lo que confirma su mejor desempeño general entre los tres modelos evaluados.

Conclusión:
El modelo de Regresión Logística es el más equilibrado para este problema de clasificación, logrando el mejor rendimiento en AUC y Accuracy, y un comportamiento estable frente a las demás métricas. Por lo tanto, se recomienda su uso como modelo final para predecir la suscripción de clientes a depósitos a plazo fijo.