Análisis de Riesgo Crediticio

Presentado por **Data & ML Innovators**

Alejandra Cruz R.

Brusly Patiño S.

Juan **García C.**

Marzo 2025

Bootcamp Xperience

El Desafío del Riesgo Crediticio

Entorno Competitivo

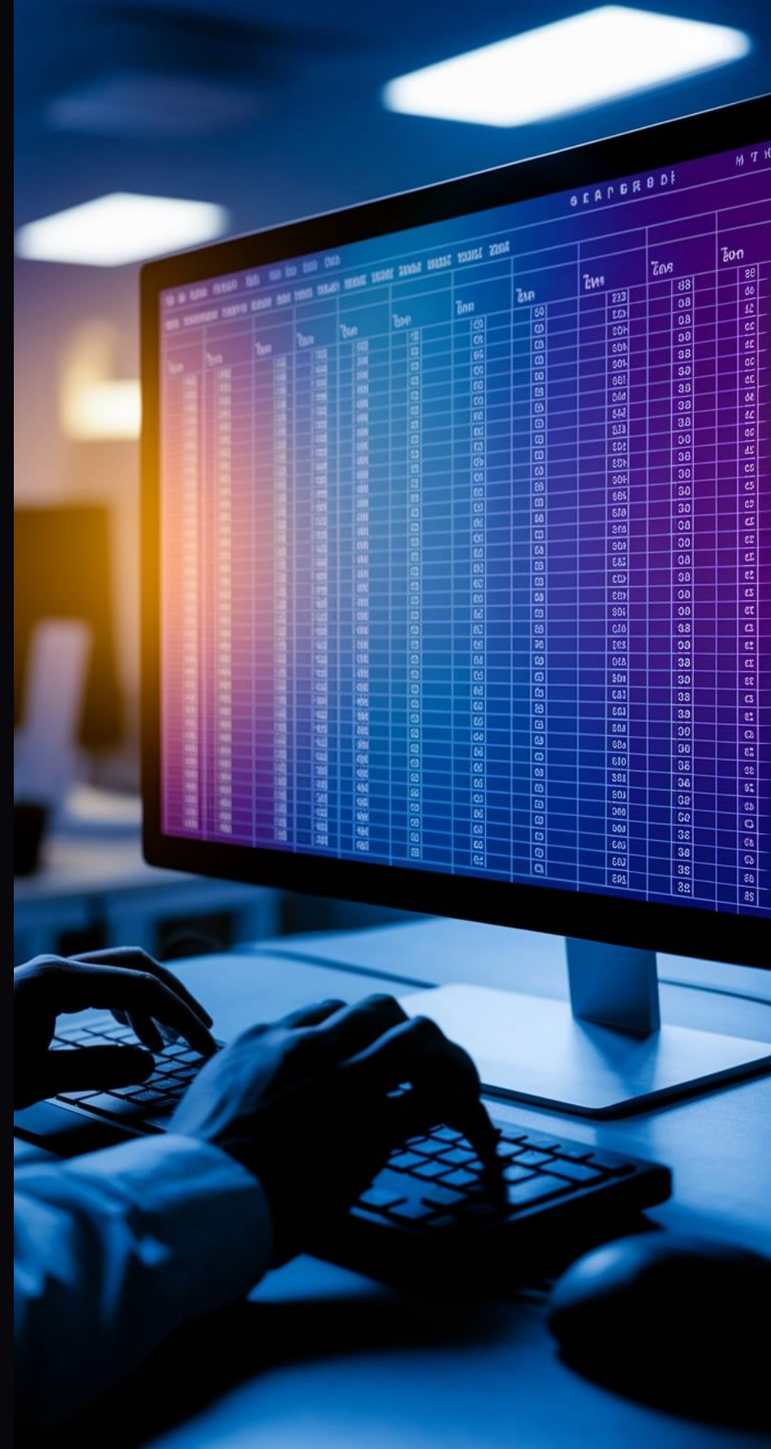
Decisiones precisas en un mercado regulado.

Confianza y Estabilidad

Una prioridad estratégica clave.

Gestión del Riesgo

Predecir y gestionar para mitigar riesgos.



Identificación de Clientes aptos

Enfoque Preciso

Identificar clientes aptos.

Riesgos

Disminuir los riesgos

Herramientas Confiables

Evaluar la solvencia.

Preguntas Clave del Negocio

¿Cuál es la probabilidad de incumplimiento crediticio de un cliente en función de sus características financieras y personales?

¿Qué variables tienen el mayor impacto en la predicción del riesgo crediticio?

¿Cómo se puede mejorar la precisión del modelo de predicción de riesgo crediticio utilizando técnicas de machine learning?

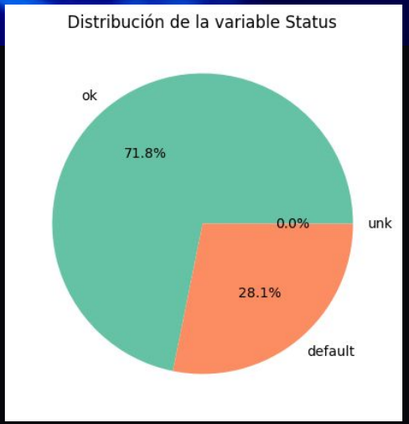
Etapas de desarrollo del proyecto

1. Preprocesamiento de datos
2. Exploración de datos
3. Construcción de Modelos
4. Evaluación y Selección de Modelos

Preprocesamiento de datos

```
1 data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4455 entries, 0 to 4454
Data columns (total 14 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   status     4455 non-null   object
 1   seniority  4455 non-null   int64
 2   home       4455 non-null   object
 3   time       4455 non-null   int64
 4   age        4455 non-null   int64
 5   marital    4455 non-null   object
 6   records    4455 non-null   object
 7   job        4455 non-null   object
 8   expenses   4455 non-null   int64
 9   income     4455 non-null   int64
 10  assets     4455 non-null   int64
 11  debt       4455 non-null   int64
 12  amount     4455 non-null   int64
 13  price      4455 non-null   int64
dtypes: int64(9), object(5)
memory usage: 487.4+ KB
```



Status → Default

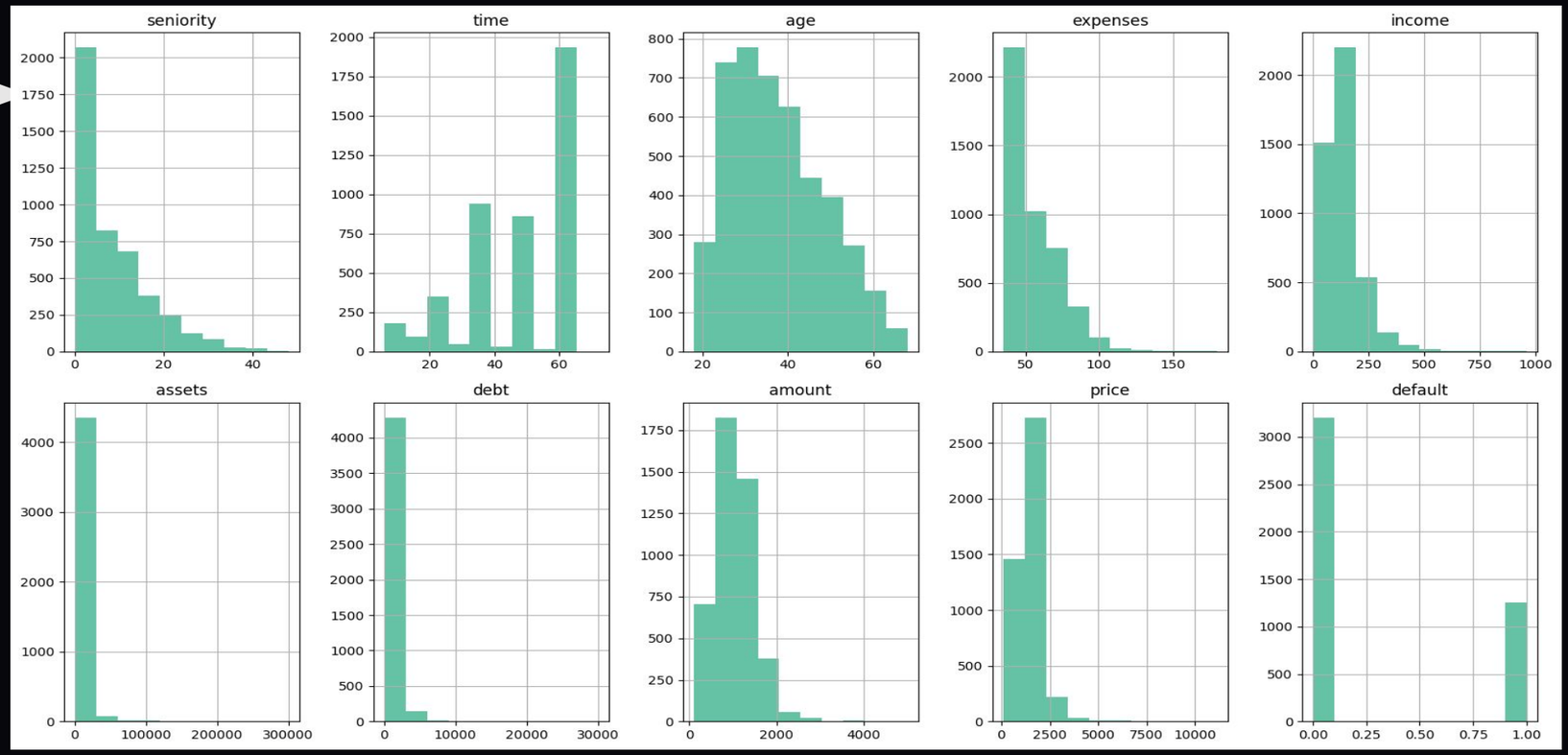
Identificar datos anómalos → Income, Assets, Debt

Descriptivos numéricos → Truncamiento de valores atípicos

Descriptivos categóricos → Agrupamiento de categorías

Exploración de datos

Datos Numéricos



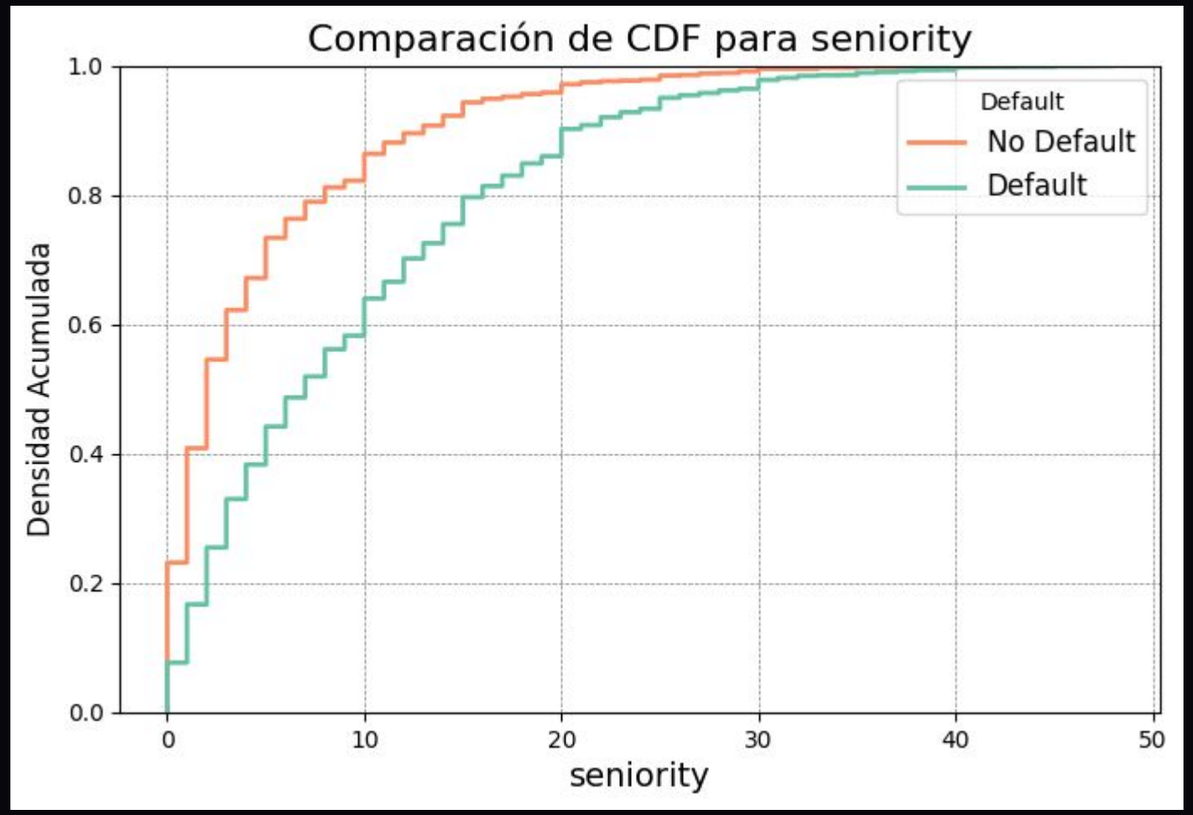
Histograma de las variables numéricas

Exploración de datos

Datos Numéricos

Los gráficos de distribución acumulada (CDF) para cada columna numérica, comparan cómo se distribuyen los valores entre las categorías de default (por ejemplo, si un cliente incumplió o no un pago). Esto es útil para identificar diferencias en las distribuciones de las variables numéricas entre estos dos grupos

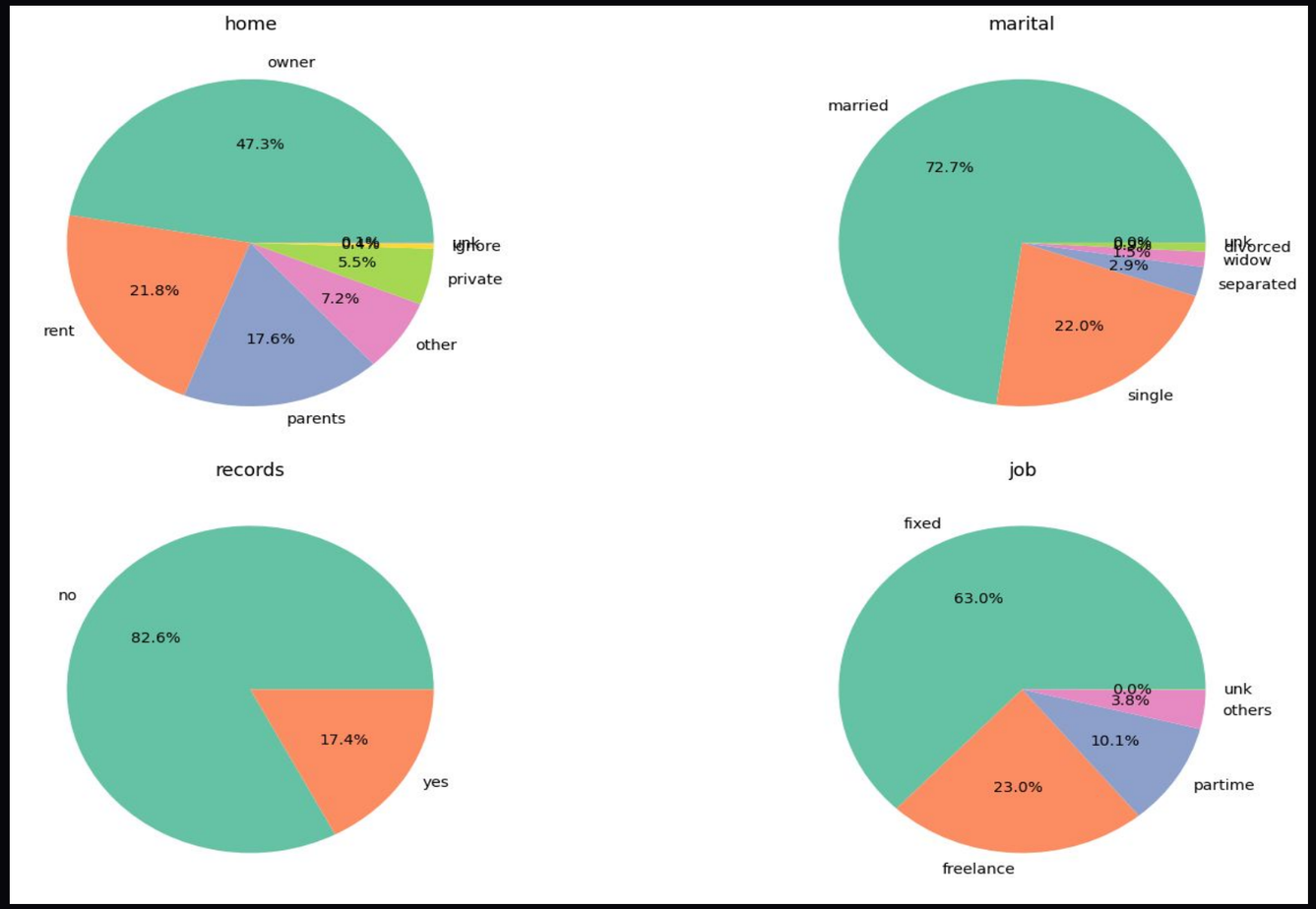
Para la variable seniority se tiene que el 80 % de los clientes buenos tienen aproximadamente 8 años de senioriyu. Y el 80 % de los clientes malos, tienen un seniority de 15 años.



Los gráficos de distribución acumulada (CDF) para cada columna numérica,

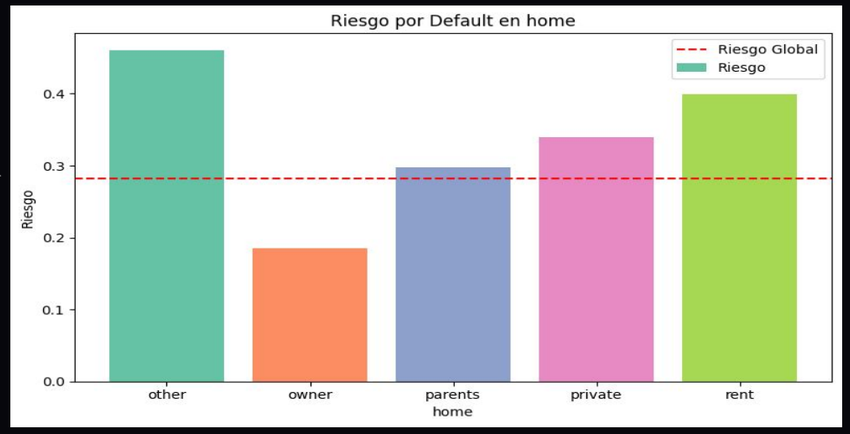
Exploración de datos

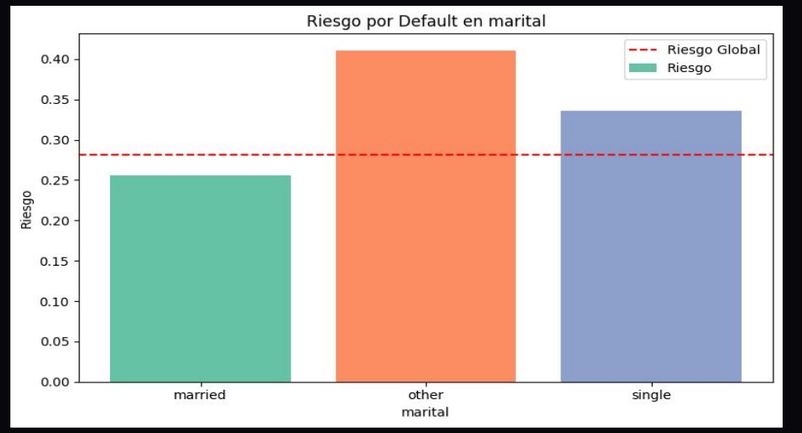
Datos Cualitativos



Exploración de datos

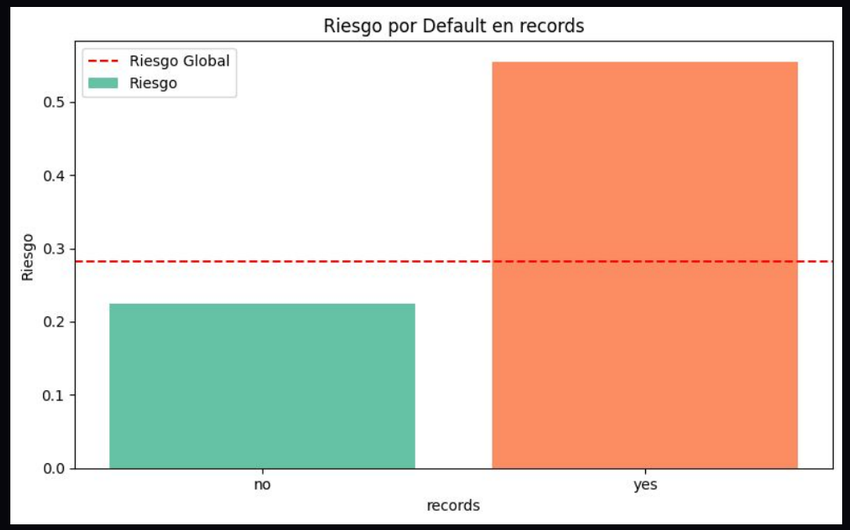
Datos Cualitativos

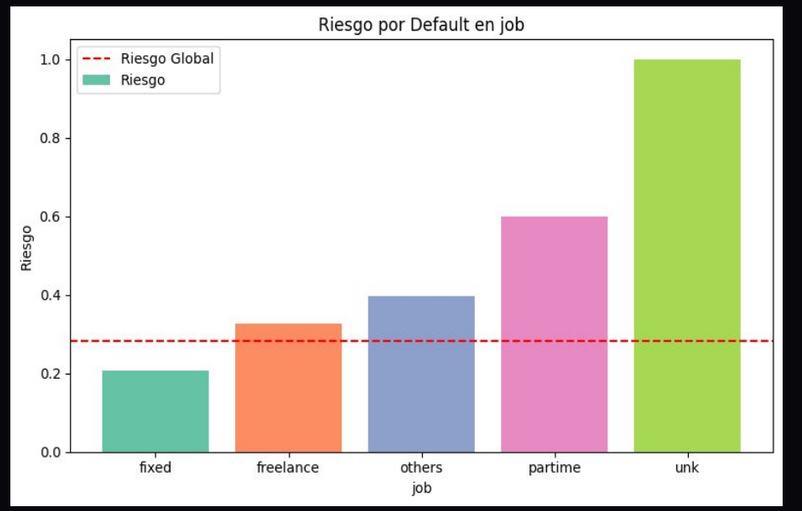




Observaciones:

1. En la categoría home, el mayor riesgo lo tienen la personas con valores other, private y rent
2. En la categoría marital, el mayor riesgo está en los que tienen como valor other y single
3. En la categoría records, tienen un mayor riesgo los que tienen valor yes
4. En la categoría job, tienen un mayor riesgo los que tienen unk, part time y other job

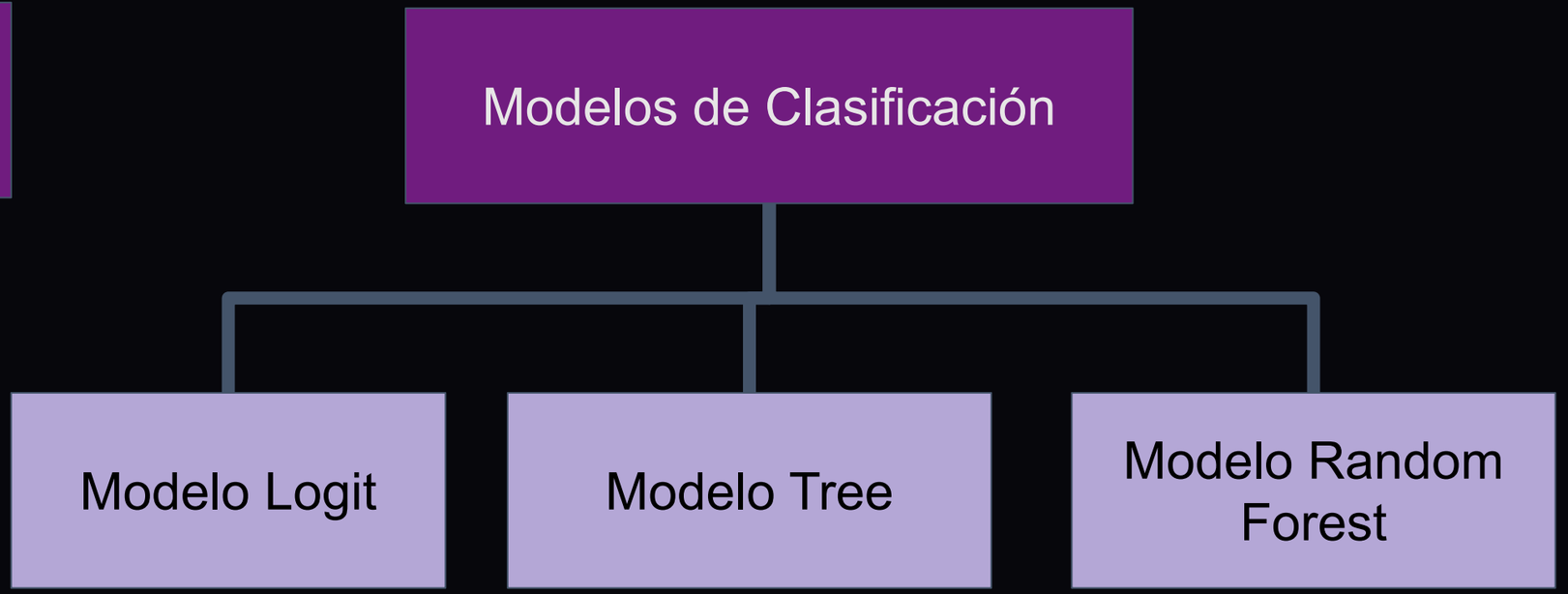




Construcción de Modelos

Partición de Datos

```
X_train,
X_test,
y_train,
y_test
```



Evaluación y Selección de Modelos

Modelo Logit

	precision	recall	f1-score	support
0	0.87	0.69	0.77	945
1	0.50	0.74	0.60	392
accuracy			0.71	1337
macro avg	0.68	0.72	0.68	1337
weighted avg	0.76	0.71	0.72	1337

AUC-ROC: 0.7734

Modelo Tree

	precision	recall	f1-score	support
0	0.87	0.72	0.79	945
1	0.52	0.74	0.61	392
accuracy			0.72	1337
macro avg	0.70	0.73	0.70	1337
weighted avg	0.77	0.72	0.73	1337

AUC-ROC: 0.7842

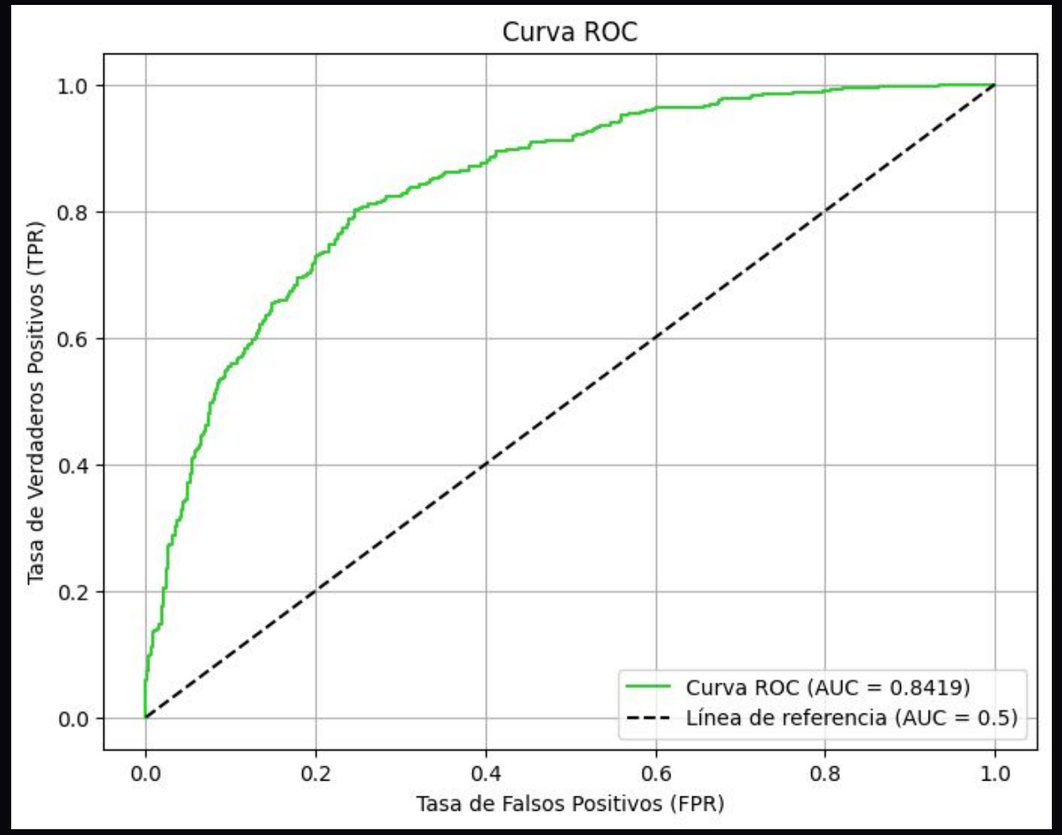
Modelo Random Forest

	precision	recall	f1-score	support
0	0.90	0.75	0.82	945
1	0.58	0.80	0.67	392
accuracy			0.77	1337
macro avg	0.74	0.78	0.75	1337
weighted avg	0.81	0.77	0.78	1337

AUC-ROC: 0.8419

Precisión del Modelo

Random Forest Más preciso que otros modelos.



col_0	0	1	All
default			
0	713	232	945
1	78	314	392
All	791	546	1337

	precision	recall	f1-score	support
0	0.90	0.75	0.82	945
1	0.58	0.80	0.67	392
accuracy			0.77	1337
macro avg	0.74	0.78	0.75	1337
weighted avg	0.81	0.77	0.78	1337

Conclusiones

El modelo Random Forest es el mejor entre los tres, ya que tiene:

La mayor precisión, recall y F1-Score para ambas clases.

La mayor exactitud (accuracy).

Además, el Random Forest tiene un mejor equilibrio entre precisión y recall, especialmente para la clase 1 (incumplimiento), lo cual es crucial en problemas de clasificación con clases desbalanceadas.

El modelo Random Forest es el mejor en términos de AUC-ROC, con un valor de 0.8418, lo que indica una buena capacidad para distinguir entre las clases.

El modelo Tree tiene un AUC-ROC ligeramente superior al modelo Logit, pero ambos están en el rango de moderadamente bueno.

Si el objetivo es maximizar la capacidad de discriminación del modelo, el Random Forest es la mejor opción.

Luego de esta comparación de los resultados,** se usará el modelo Random Forest** para la implementación final, ya que tiene el mayor AUC-ROC y, por lo tanto, la mejor capacidad para distinguir entre las clases.