

**TEAM DATA  
SCIENCE**

# **CREDIT SCORING PREDICTION**

**BOOTCAMP PROFE ALEJO**

# TEAM DATA SCIENCE



Presentaremos los resultados y las conclusiones de nuestro proyecto de análisis de riesgo crediticio para una institución financiera alemana. Como científicos de datos, hemos sido desafiados con la tarea de construir un modelo de Machine Learning que sea preciso y confiable para evaluar la probabilidad de obtener un buen cliente, aquel que demuestra ser un pagador confiable.



En un mundo donde la gestión de riesgos es esencial, nuestro objetivo ha sido proporcionar a la institución herramientas innovadoras y predictivas para tomar decisiones fundamentadas.



# OBJETIVOS

- 01** La importancia de reducir el riesgo crediticio es clave en el sector financiero. Nuestro cliente, una institución financiera alemana, ha reconocido la necesidad de adoptar enfoques innovadores para mejorar su capacidad de evaluar el riesgo crediticio de los clientes.
- 02** La misión principal es identificar y clasificar a los clientes en dos categorías: "Buen Cliente (0)" y "Mal Cliente (1)"
- 03** Este análisis permitirá a la institución tomar decisiones informadas y mitigar el riesgo de pérdidas crediticias.



# ETAPAS DEL PROYECTO

Planificación del Proyecto



Preprocesamiento de Datos



Análisis Exploratorio de Datos (EDA)



Desarrollo del Modelo



Optimización del Modelo



# PLANIFICACIÓN DEL PROYECTO

01

**Recopilación de Datos:** Identificación y adquisición de conjuntos de datos. Análisis inicial de variables.

02

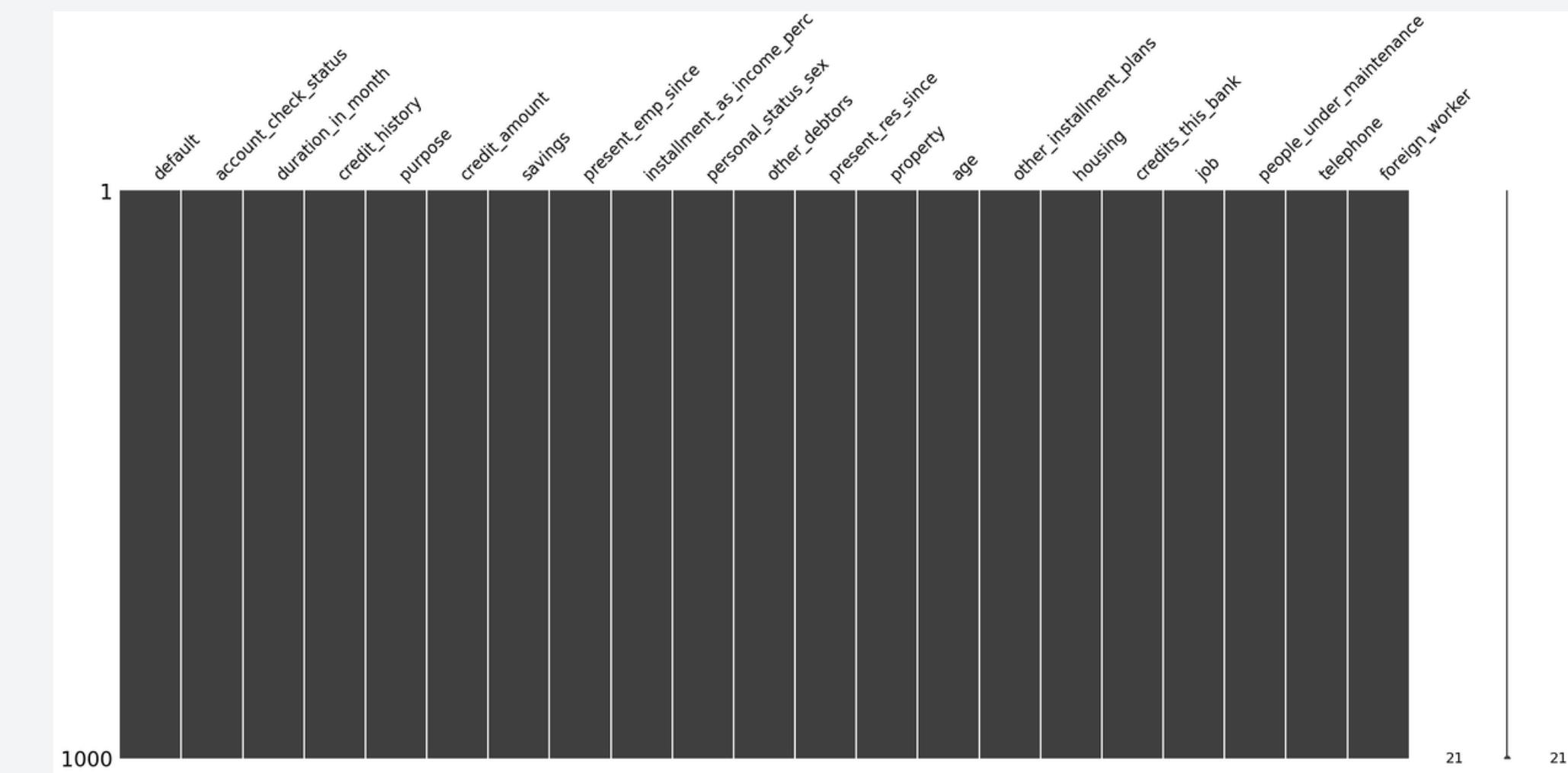
**Definición de Objetivos:** Establecimiento de metas y métricas clave para evaluar el rendimiento del modelo. Métodos de Trabajo en equipo.



# PREPROCESAMIENTO DE DATOS

01

**Eliminación de Duplicados:**  
Observamos que no hay presencia de elementos duplicados.



02

**Tratamiento de Nulos:**  
Validamos que no hay presencia de datos Nulos.  
Mostramos dos evaluaciones realizadas, La Matriz de Nulos y el Dendograma.



# PREPROCESAMIENTO DE DATOS

03

## Conversión de Variables

**Categóricas a Numéricas:** Para esto hicimos uso de la discretización de valores. De esta manera cada variable categórica tiene su correspondiente valor numérico discreto.

04

## Conversión de Variables numéricas

**en Discretas:** De la misma forma, notamos que se trabajaba con valores extensos en algunas variables, es así que decidimos colocar rangos y discretizar las variables.

05

**Feature Engineering:** Se realizó la creación de nuevas variables a partir de algunas ya existentes, como es el caso de la creación de la variable **Sexo y estado\_civil**.

	personal_status_sex	sexo	estado_civil
0		3	0
1		2	1
2		3	0
3		3	0
4		3	0
...	...	...	...
995		2	1
996		1	0
997		3	0
998		3	0
999		3	1

# PREPROCESAMIENTO DE DATOS

Dataframe Transformado

	default	account_check_status	credit_history	purpose	savings	present_emp_since	installment_as_income_perc	other_debtors	present_res_since	property		
0	0		1	5	5	1		4		1	4	1
1	1		2	3	5	5	3	2	1		2	1
2	0		4	5	8	5	2	2	1		3	1
3	0		1	3	4	5	2	2	3		4	2
4	1		1	4	1	5	3	3	1		4	4

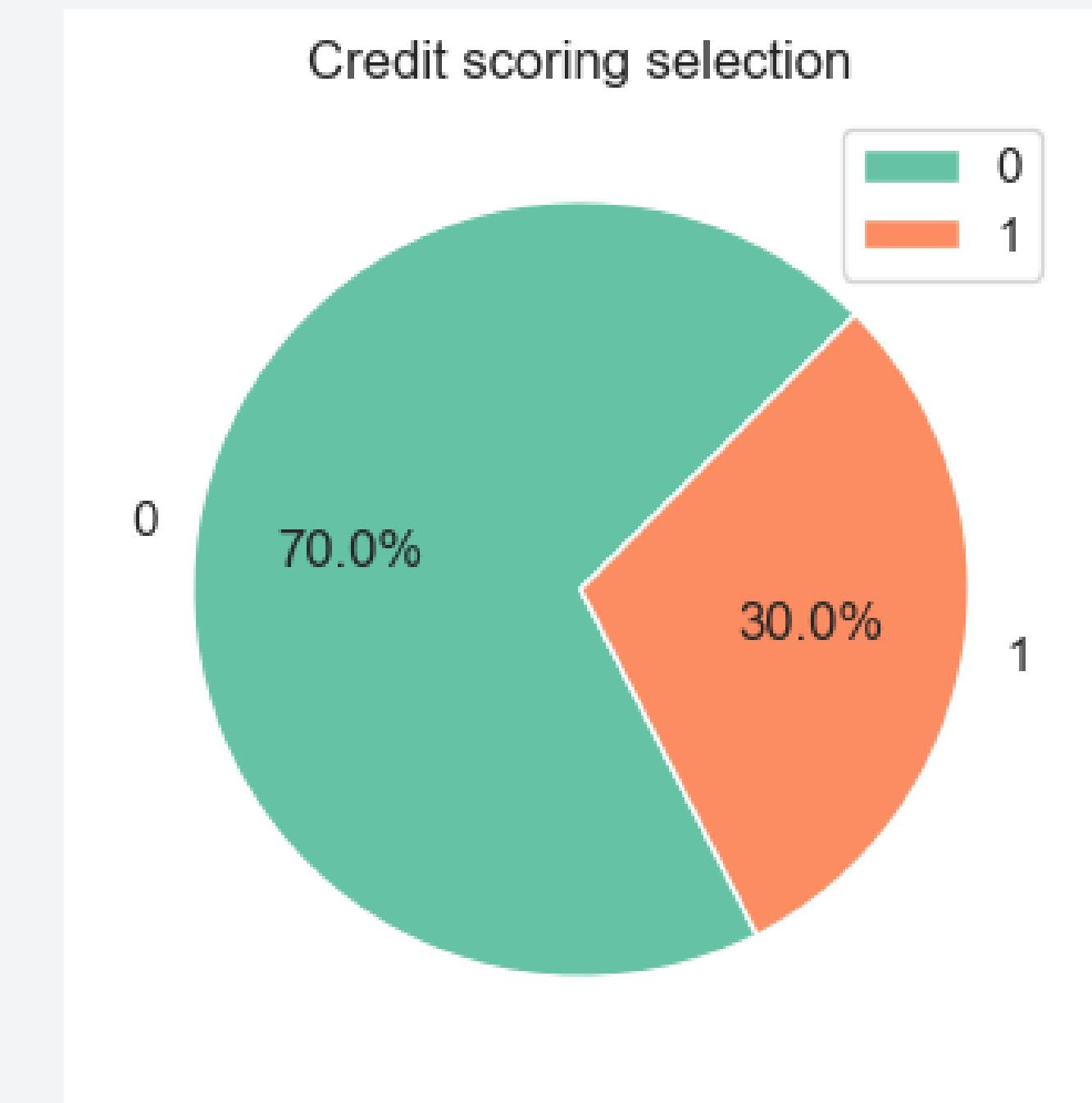
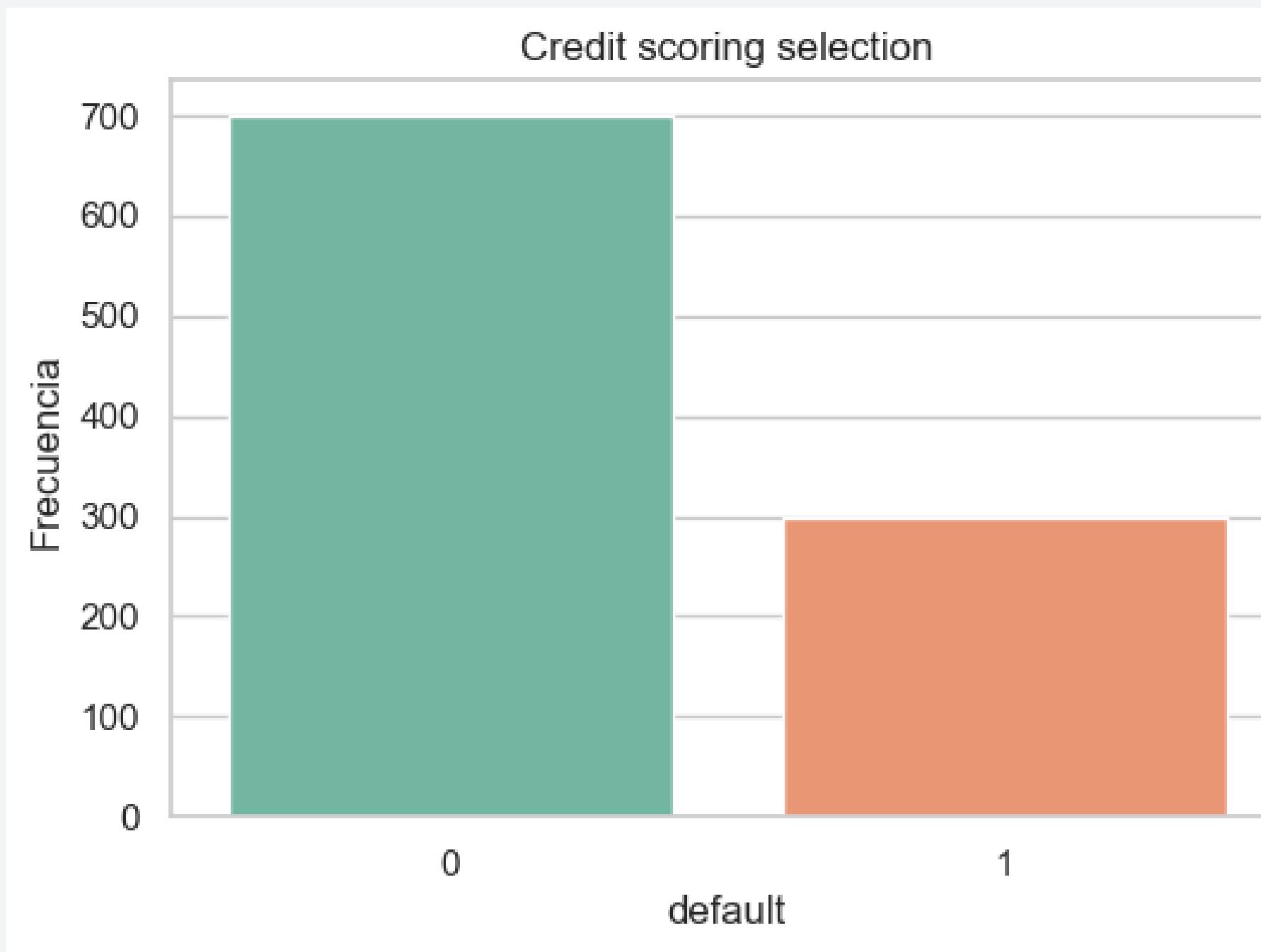


	default	account_check_status	credit_history	purpose	savings	present_emp_since	installment_as_income_perc	other_debtors	present_res	
count	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000
mean	0.300000	2.577000	3.54500	4.277000	3.895000	2.616000	2.973000	1.145000	2.8	
std	0.458487	1.257638	1.08312	2.739302	1.580023	1.208306	1.118715	0.477706	1.1	
min	0.000000	1.000000	1.00000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.0
25%	0.000000	1.000000	3.00000	2.000000	3.000000	1.000000	2.000000	1.000000	2.000000	2.0
50%	0.000000	2.000000	3.00000	4.000000	5.000000	3.000000	3.000000	3.000000	1.000000	3.0
75%	1.000000	4.000000	5.00000	5.000000	5.000000	3.000000	4.000000	1.000000	4.000000	4.0
max	1.000000	4.000000	5.00000	10.000000	5.000000	5.000000	5.000000	4.000000	3.000000	4.0

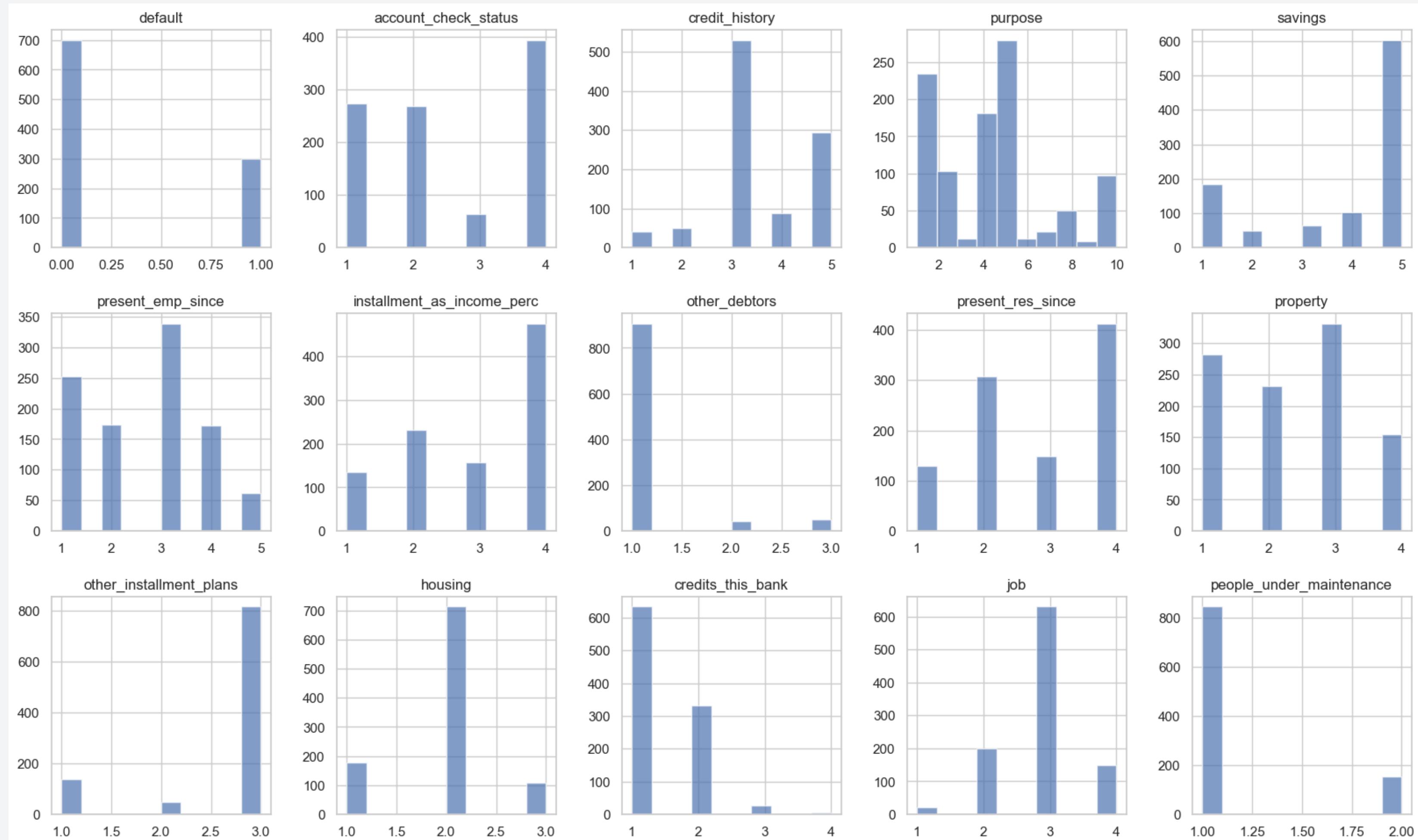


# ANÁLISIS EXPLORATORIO DE DATOS

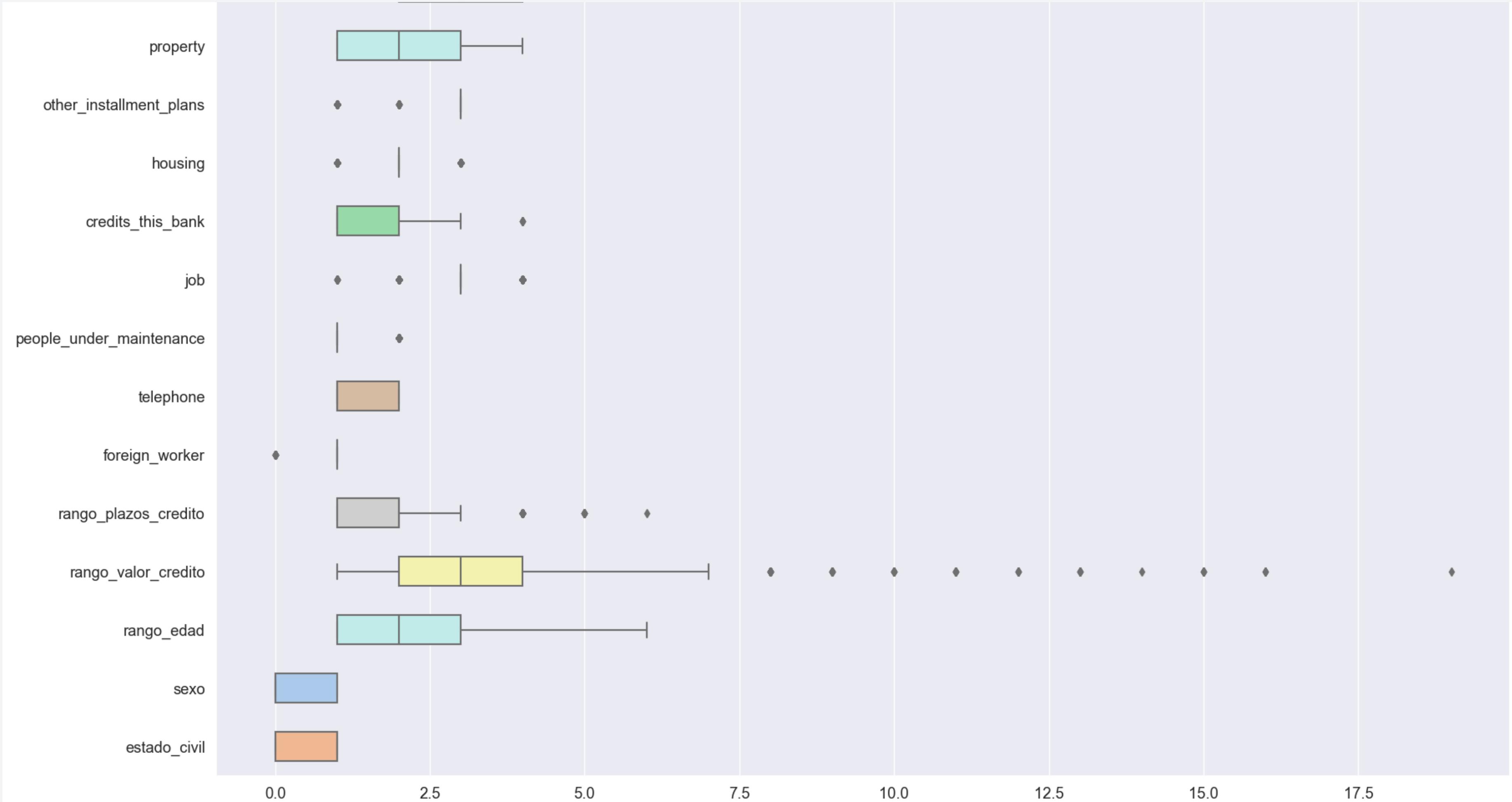
Variable Target: 'Default'



# Análisis Univariado

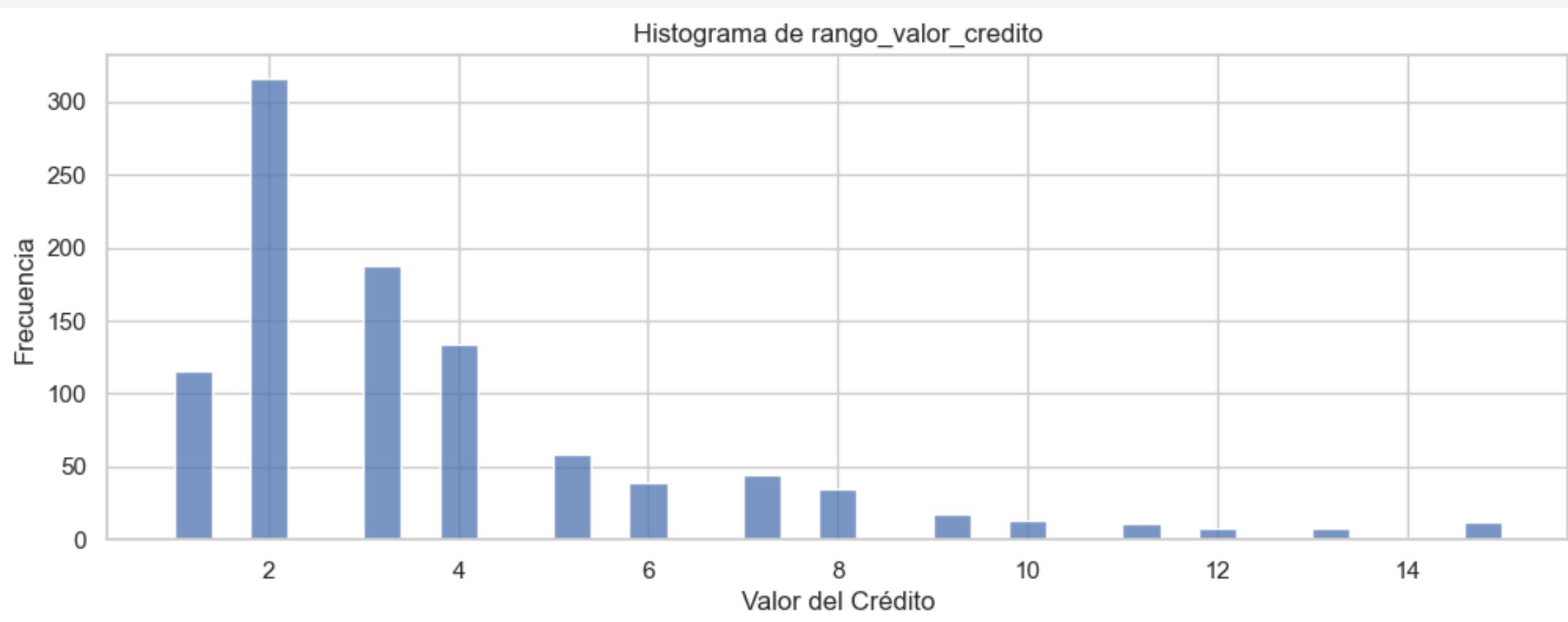
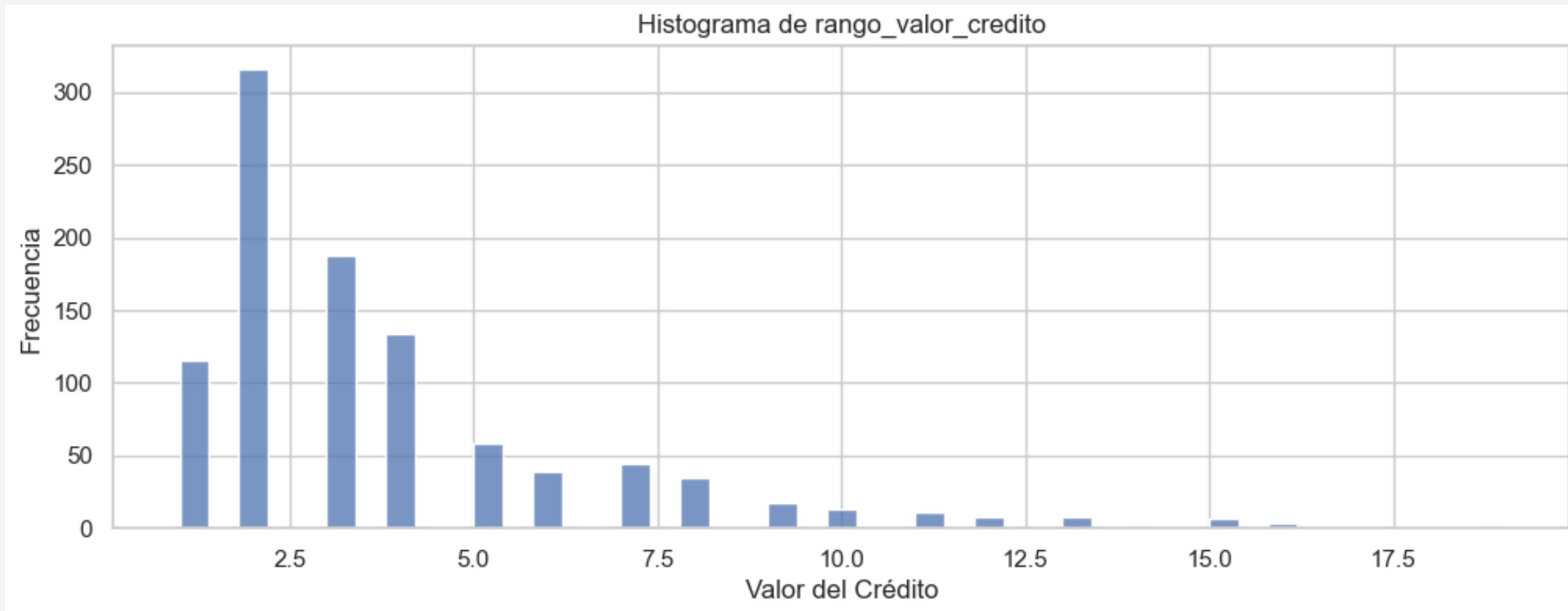


# Análisis Univariado



**Tratamiento de Atípicos:** Se realiza el análisis por percentiles. Truncamiento de percentiles. Observamos que nuestra variable que existen ciertas variables que poseen valores anómalos, pero también tenemos en cuenta que previamente realizamos una discretización. Por esta razón, la única que consideraremos para tratamiento es la variable '**rango\_valor\_credito**'

## Antes del Truncamiento



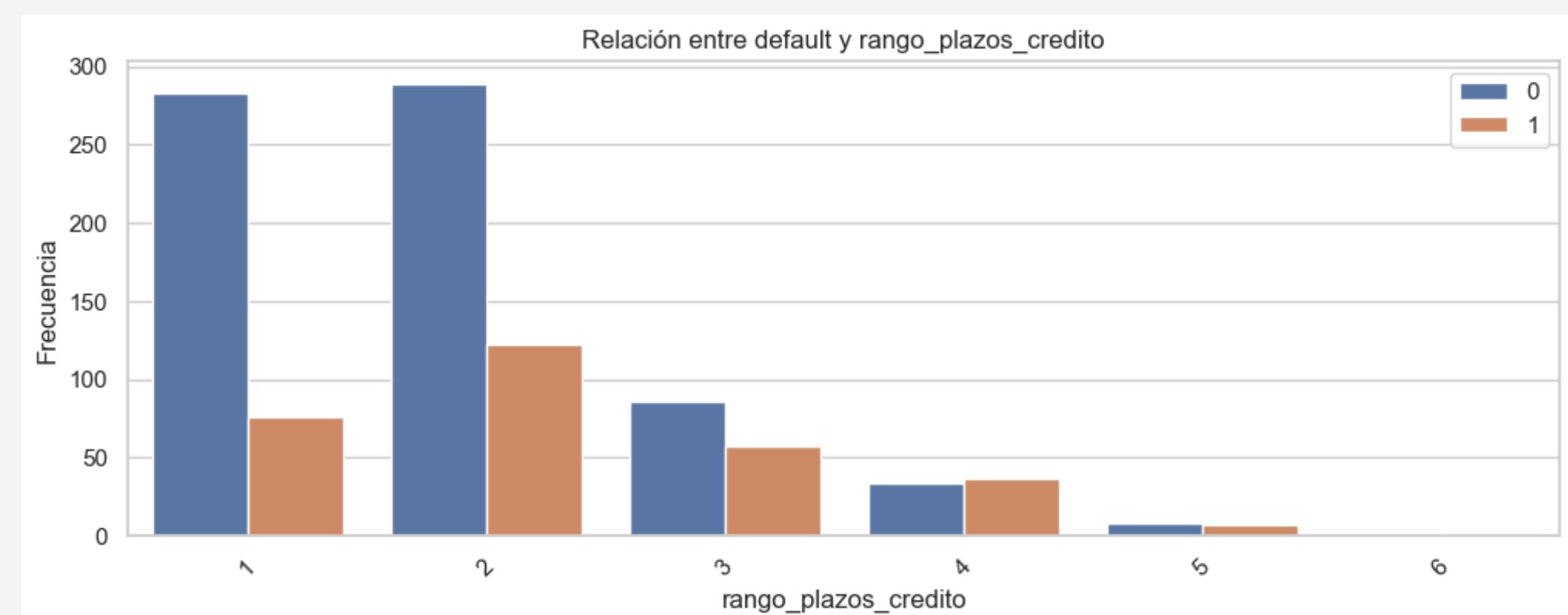
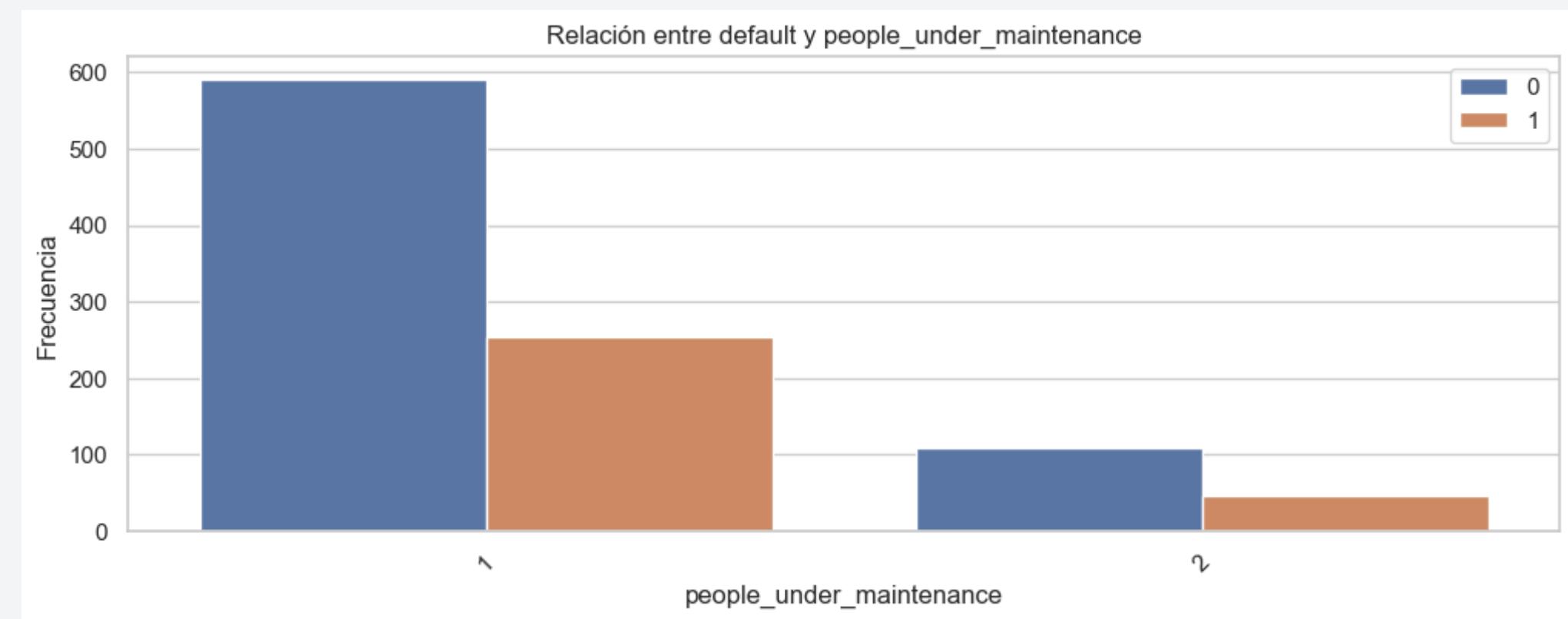
## Despues del Truncamiento

# ANÁLISIS EXPLORATORIO DE DATOS

## Análisis Bivariado

Después de realizar nuestro análisis Bivariado entre nuestras variables, podemos deducir que tenemos información para todo nuestro conjunto de variables.

Así también como en el análisis Univariado, notamos que hay categorías en las variables que se podrían considerar como atípicos. Pero tener en cuenta que ya habíamos discretizado previamente nuestras variables, el realizar demasiado ajuste podría no ser beneficioso para nuestro modelo. Decidimos continuar sin seguir modificando por el momento.



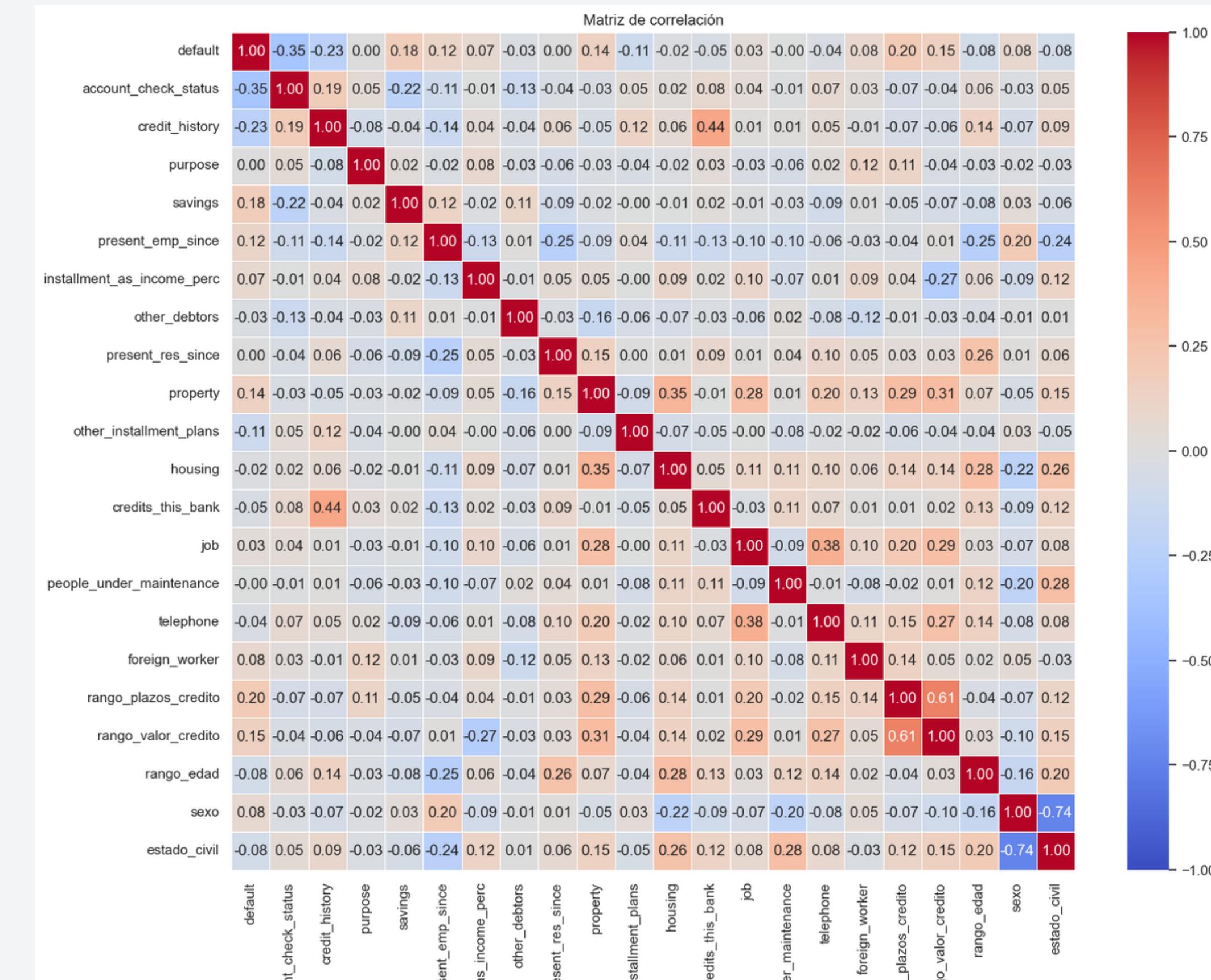
# ANÁLISIS EXPLORATORIO DE DATOS

## Análisis Multivariado

Observamos que muchas variables se correlacionan entre ellas, esto puede ser perjudicial para nuestro modelo, ya que puede ocasionar que los coeficientes de nuestro modelo no sean los mejores y por ende nos daría un modelo que no cumpla con la predicción deseada.

Por esta razón hacemos uso de un método de análisis para detectar multicolinealidad, estamos hablando del **VIF - Factor de Inflación de la Varianza-** El **VIF** mide la multicolinealidad entre las variables predictoras. Un **VIF mayor indica una mayor multicolinealidad**.

Matriz de Correlación



# ANÁLISIS EXPLORATORIO DE DATOS

Si bien es cierto que valores altos de VIF pueden indicar multicolinealidad, no necesariamente implica que debamos eliminar esas variables automáticamente.

Normalmente se considera que valores de VIF mayores de 10, son malos y generan multicolinealidad, pero también lo analizaremos con nuestra matriz de correlación y nuestro experto del negocio.

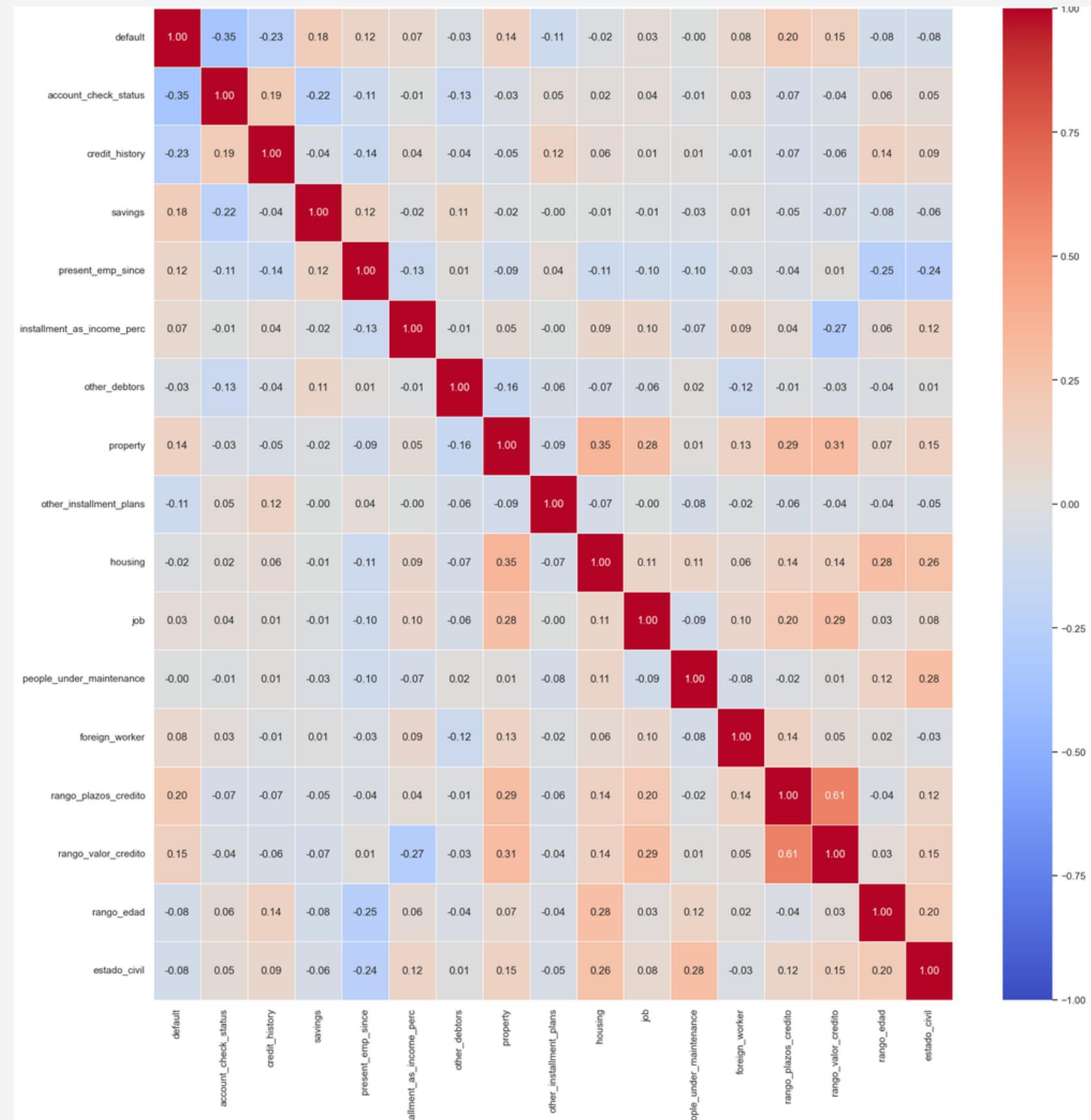
Ahora bien, también tenemos que notar que **algunas variables con VIF elevados son importantes para nuestros análisis**, según lo deducido previamente con nuestra matriz de correlación.

Descartamos ciertas variables, que consideramos no importantes para nuestro análisis. Estas variables elegidas son: **['purpose','credits\_this\_bank', 'sexo', 'telephone', 'present\_res\_since']**.

Variable	VIF
account_check_status	5.689848
credit_history	15.486267
purpose	3.646975
savings	7.446907
present_emp_since	6.071848
installment_as_income_perc	10.067531
other_debtors	6.535912
present_res_since	8.626305
property	8.464266
other_installment_plans	14.335426
housing	17.864891
credits_this_bank	8.986070
job	25.254947
people_under_maintenance	11.282341
telephone	11.659648
foreign_worker	25.506647
rango_plazos_credito	9.520950
rango_valor_credito	6.191227
rango_edad	5.512415
sexo	3.199740
estado_civil	5.477965

# ANÁLISIS EXPLORATORIO DE DATOS

Variable	VIF
account_check_status	5.572781
credit_history	12.132360
savings	7.327697
present_emp_since	5.911627
installment_as_income_perc	9.954497
other_debtors	6.464158
property	8.142708
other_installment_plans	14.015535
housing	17.739843
job	22.380021
people_under_maintenance	10.906730
foreign_worker	23.703003
rango_plazos_credito	9.278219
rango_valor_credito	5.985482
rango_edad	5.110627
estado_civil	2.803631



Observamos que ciertos valores de VIF se modificaron, bajaron en magnitud. Esto es bueno ya que nos indica que estas variables que descartamos no nos aportaban valor y nos generaban cierta multicolinealidad.

Notamos que ciertas variables mantienen una correlación media baja, pero consideramos que son importantes para el análisis de nuestros modelos.

Decidimos continuar con este nuevo conjunto de datos, los resultados se verán al momento de obtener las métricas de nuestros modelos generados.

# DESARROLLO DEL MODELO

01

**Separamos la data en entrenamiento y test:** Lo separamos considerando un 30% de data para prueba y el resto para entrenamiento.

```
pd.DataFrame(x_train_scaled).head()
```

	0	1	2	3	4	5	6	7	8	9	10	11
0	-0.438898	1.329526	0.712004	-1.312270	0.929207	-0.291554	0.622892	-2.298946	0.114720	0.131869	2.315953	0.196407
1	-1.236894	-0.498899	0.712004	0.306389	0.035739	-0.291554	-1.281145	0.475644	0.114720	0.131869	-0.431788	0.196407
2	1.157094	1.329526	-1.783129	0.306389	0.035739	-0.291554	0.622892	0.475644	0.114720	0.131869	-0.431788	0.196407
3	1.157094	1.329526	0.712004	0.306389	0.929207	-0.291554	-1.281145	0.475644	0.114720	0.131869	-0.431788	0.196407
4	-0.438898	-0.498899	0.088221	0.306389	-1.751197	-0.291554	1.574911	0.475644	2.026717	0.131869	2.315953	0.196407

02

**Estandarizamos los datos:** Hacemos uso de la estandarización por StandardScaler, que es el que mejor funciona con manejo de valores Atípicos..

```
pd.DataFrame(x_test_scaled).head()
```

	0	1	2	3	4	5	6	7	8	9	10	11
0	-1.236894	1.329526	0.088221	-1.312270	0.035739	4.193898	-1.281145	0.475644	0.114720	0.131869	2.315953	-5.091471
1	-0.438898	-0.498899	0.712004	-0.502941	0.035739	-0.291554	-1.281145	0.475644	-1.797278	-1.381387	-0.431788	0.196407
2	-0.438898	-0.498899	-1.783129	-1.312270	0.929207	-0.291554	1.574911	-2.298946	2.026717	0.131869	-0.431788	0.196407
3	-1.236894	-1.413111	0.712004	1.925048	-1.751197	1.951172	1.574911	0.475644	-1.797278	-2.894643	-0.431788	0.196407
4	0.359098	-0.498899	0.712004	-1.312270	0.929207	-0.291554	0.622892	0.475644	0.114720	0.131869	-0.431788	0.196407

# DESARROLLO DEL MODELO

03

Construcción de los Modelos: Hacemos uso de 4 modelos, los cuales son: **Regresión Logística, Árbol de Decisión, Random Forest y Naive Bayes.**

04

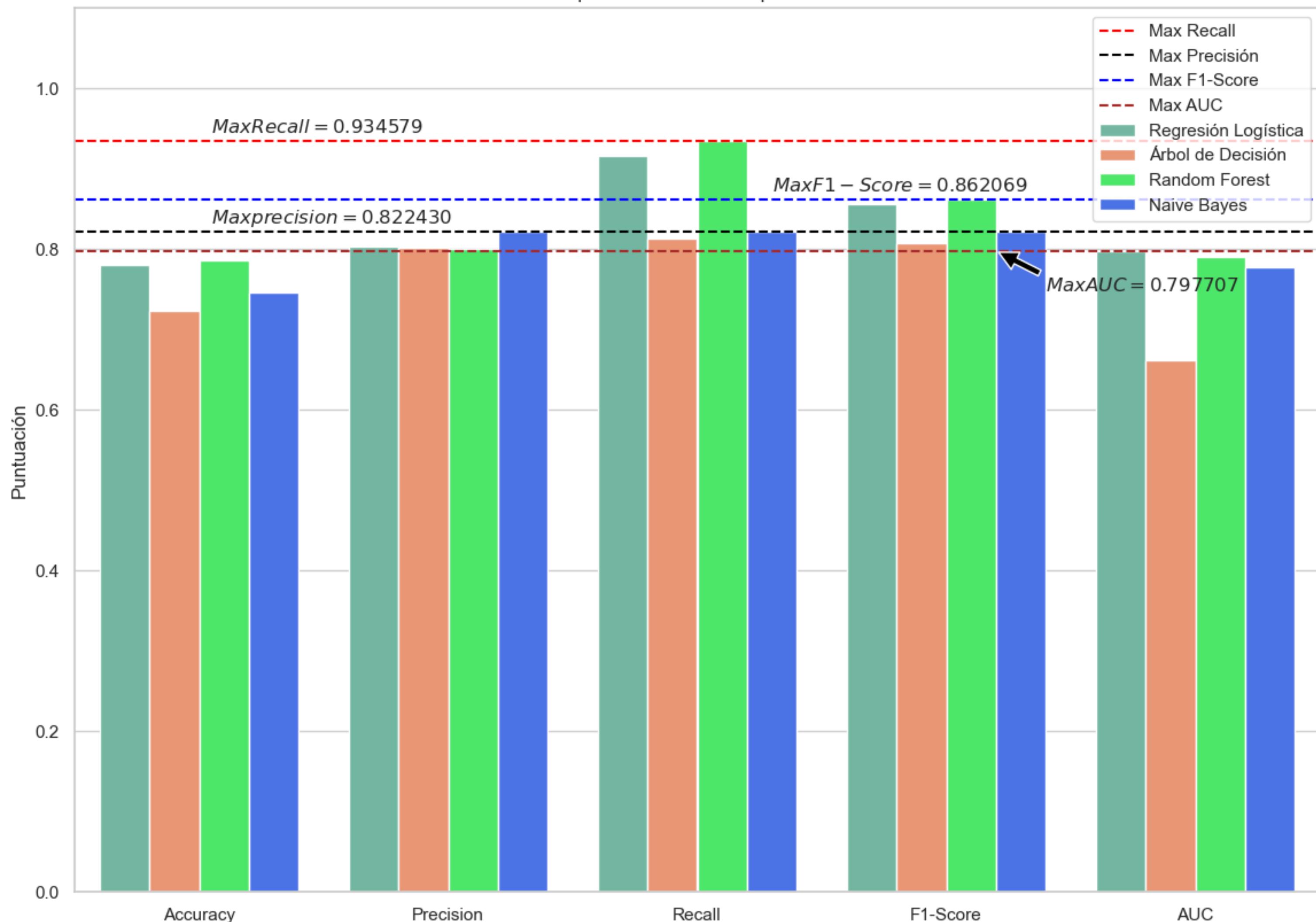
Análisis de Métricas: Utilizamos varias métricas para evaluar nuestros modelos, los cuales son: **Accuracy, Precisión, Recall, F1-Score, y AUC.**

05

Comparativo Gráfico de Métricas: También realizamos un análisis más visual de los valores de las métricas utilizadas.

Métricas	Regresión Logística	Árbol de Decisión	Random Forest	Naive Bayes
Accuracy	0.780000	0.723333	0.786667	0.746667
Precision	0.803279	0.801843	0.800000	0.822430
Recall	0.915888	0.813084	0.934579	0.822430
F1-Score	0.855895	0.807425	0.862069	0.822430
AUC	0.797707	0.661269	0.790018	0.777385

Comparación de Métricas por Modelo



# OPTIMIZACIÓN DEL MODELO

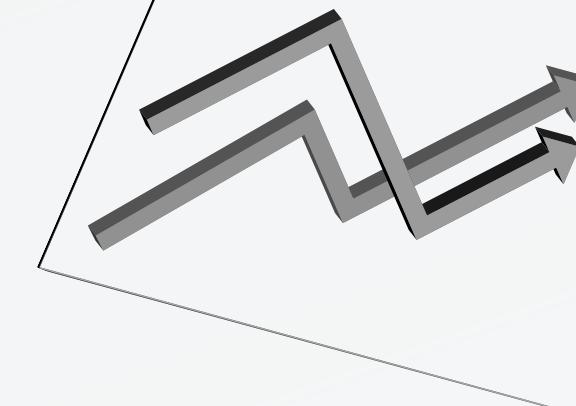
Métricas	Random Forest	Random Forest con Hiperparámetros
Accuracy	0.786667	0.790000
Precision	0.800000	0.796078
Recall	0.934579	0.948598
F1-Score	0.862069	0.865672
AUC	0.790018	0.801728

Este algoritmo ha demostrado ser robusto y preciso, lo cual es fundamental para la predicción de nuestra clase objetivo 0 (clientes buenos) en el contexto de riesgo crediticio.

Después de analizar nuestras métricas clave, las que consideramos más importantes para la predicción, priorizando la minimización de Falsos Positivos y Falsos Negativos en la clase 0, concluimos que el modelo óptimo es **RANDOM FOREST**.

Para mejorar aún más nuestro modelo seleccionado, aplicamos la técnica de búsqueda por cuadrícula, con la finalidad de encontrar los mejores hiperparámetros que nos ayuden a mejorar las métricas de nuestro modelo.

# CONCLUSIONES



- El **Modelo Random Forest**, es el que nos resulta mejor para nuestro conjunto de datos y para lo que queremos obtener, la Predicción de la clase 0 - Good Customer.
- Al hacer uso de Hiperparámetros para nuestro modelo de Random Forest, lo evaluamos para diferentes Scoring, consiguiendo mejores resultados evaluando en el Scoring F1, con lo cual obtuvimos ligeramente valores mayores de nuestras métricas. Consideramos que el análisis con los Hiperparámetros nos ayudó a mejorar nuestros valores de métricas haciendo que consigamos un mejor modelo.
- Es importante analizar todas las métricas para tomar nuestra decisión final, es por eso que consideramos como métricas más importantes el **Recall, Precisión, F1-Score y el AUC**.
- Con este modelo elegido **Random Forest**, podemos concluir que nuestras predicciones serán optimas, y que cumplirán con las condiciones que deseamos conseguir, lo cual es evaluar potenciales clientes que puedan cumplir con los créditos.



# NUESTRO EQUIPO



**David Carrillo Castillo**

Data Scientist Junior



+51957639879



[www.linkedin.com/in/davidcarrillocastillo](https://www.linkedin.com/in/davidcarrillocastillo)



Lima, Perú



**Diego Ladino Cubides**

Data Scientist Junior



+573045440896



[disgraficodiegoladino@gmail.com](mailto:disgraficodiegoladino@gmail.com)



Colombia



**Everardo**

Data Scientist Junior



+52 1 56 1647 3272



<https://github.com/EverardoFX>



México