INFORME: PROFUNDIZACIÓN II

Luisa Fernanda Suazo.

Diana M. Romero Castaño.

Programa de Ingeniería en Analítica de Datos, Facultad de Ingeniería.

Universidad de Manizales, Estadística.

Docente**:** Daniel**.**

Marzo 03 de 2025.

Introducción

En la actualidad, el análisis del endeudamiento crediticio se ha convertido en una herramienta fundamental para entidades financieras, aseguradoras, fintechs y otras organizaciones interesadas en comprender el comportamiento financiero de sus clientes. A través del uso de técnicas de analítica de datos, es posible identificar factores que inciden en el nivel de deuda de una persona, permitiendo desarrollar modelos predictivos que anticipen comportamientos financieros de riesgo, fomenten decisiones de crédito más responsables y promuevan la educación financiera.

Objetivo

Desarrollar un análisis integral del comportamiento crediticio de los clientes a partir de un conjunto de datos financieros, aplicando técnicas de analítica descriptiva y predictiva que permitan identificar patrones relevantes, explorar relaciones entre variables y construir modelos estadísticos para explicar y predecir fenómenos asociados al endeudamiento y riesgo de incumplimiento.

Análisis Descriptivo

Esta etapa del proyecto tiene como propósito explorar el comportamiento general del dataset, identificar patrones iniciales, tendencias, sesgos y variables relevantes para la modelización

En este proceso se tendrá en cuenta

- ✓ Explicación de cada una de las variables.
- ✓ Responder preguntas descriptivas usando estadísticas y visualizaciones.

Mostrar:

- ✓ Tablas de frecuencia.
- ✓ Medidas de tendencia central
- ✓ Indicadores de dispersión y simetría
- ✓ Gráficos explicativos
- ✓ Generar conclusiones de cada análisis.

A continuación, avanzaremos en el proceso de exploración del dataset.

Explicación de las Variables:

D: Identificador único del cliente.

Default: Si el cliente tiene más de 90 días sin pagar su préstamo (1 = Sí, 0 = No).

Prct_uso_tc: Porcentaje de uso de tarjeta de crédito.

Edad: Edad del cliente.

Nro_prestao_retrasados: Número de préstamos con retraso en el pago.

Prct_deuda_vs_ingresos: Porcentaje de deuda con respecto a los ingresos.

Mto_ingreso_mensual: Monto de ingreso mensual.

Nro_prod_financieros_deuda: Número de productos financieros con deuda.

Nro_retraso_60dias: Número de retrasos en pagos mayores a 60 días.

Nro_creditos_hipotecarios: Número de créditos hipotecarios.

Nro_retraso_ultm3anios: Número de retrasos en los últimos 3 años.

Nro_dependiente: Número de personas dependientes del cliente.

	0
ID	0
Default	0
Prct_uso_tc	0
Edad	0
Nro_prestao_retrasados	0
Prct_deuda_vs_ingresos	0
Mto_ingreso_mensual	29731
Nro_prod_financieros_deuda	0
Nro_retraso_60dias	0
Nro_creditos_hipotecarios	0
Nro_retraso_ultm3anios	0
Nro_dependiente	3924

dtype: int64

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 150000 entries, 0 to 149999
Data columns (total 12 columns):
 #   Column                       Non-Null Count   Dtype  
---  ------                       --------------   -----  
 0   ID                           150000 non-null  int64  
 1   Default                      150000 non-null  int64  
 2   Prct_uso_tc                  150000 non-null  object 
 3   Edad                         150000 non-null  int64  
 4   Nro_prestao_retrasados       150000 non-null  int64  
 5   Prct_deuda_vs_ingresos       150000 non-null  object 
 6   Mto_ingreso_mensual          120269 non-null  float64
 7   Nro_prod_financieros_deuda   150000 non-null  int64  
 8   Nro_retraso_60dias           150000 non-null  int64  
 9   Nro_creditos_hipotecarios    150000 non-null  int64  
 10  Nro_retraso_ultm3anios       150000 non-null  int64  
 11  Nro_dependiente              146076 non-null  float64
dtypes: float64(2), int64(8), object(2)
memory usage: 13.7+ MB
```

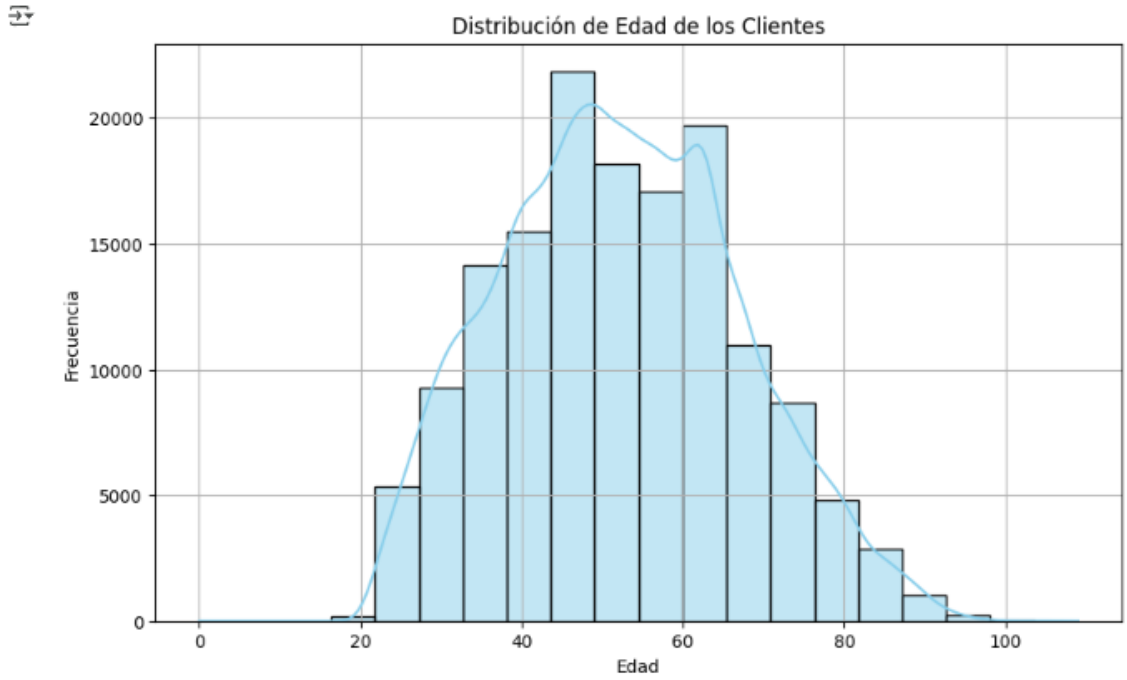
El dataset nos muestra un total de 150,000 registros y 12 variables. Algunas variables, como Prct_uso_tc (porcentaje de uso de tarjeta de crédito) y Prct_deuda_vs_ingresos (porcentaje de deuda frente a los ingresos), están en formato de texto y deben ser convertidas a valores numéricos. Además, se identificaron valores nulos en las variables Mto_ingreso_mensual (29,731 valores faltantes) y Nro_dependiente (3,924 valores faltantes), lo que hace necesario aplicar estrategias de imputación para no perder información relevante.

De acuerdo a esto se realizaron transformaciones necesarias para convertir las variables de porcentaje al formato numérico correcto. Estas variables estaban inicialmente como cadenas de texto con separadores de miles. Una vez convertidas, fueron escaladas para representar proporciones reales.

Posteriormente, se imputaron los valores faltantes en las variables Mto_ingreso_mensual y Nro_dependiente utilizando la mediana como medida robusta frente a posibles valores atípicos. Finalmente, se validó que no quedaran valores nulos en el dataset.

Responder preguntas descriptivas usando estadísticas y visualizaciones.

- Pregunta Descriptiva 1: ¿Cuál es la distribución de edades de los clientes?



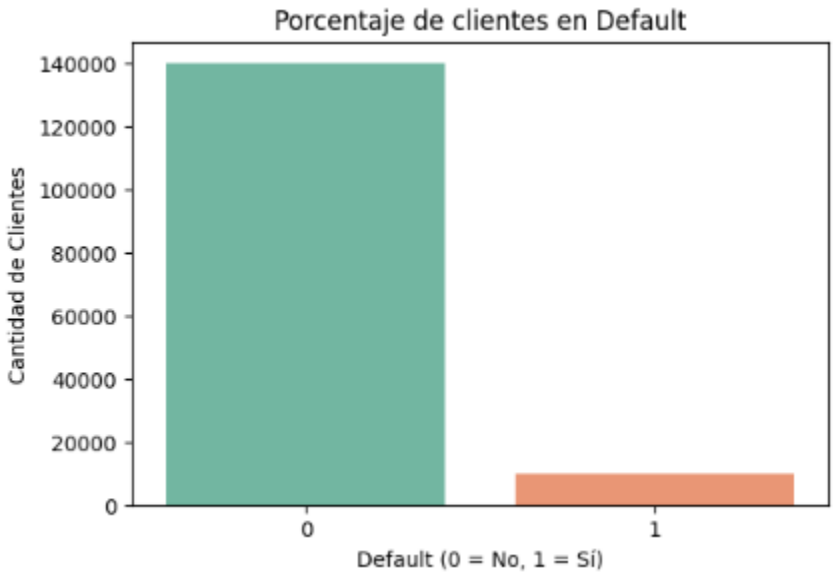
	Edad
count	150000.000000
mean	52.295207
std	14.771866
min	0.000000
25%	41.000000
50%	52.000000
75%	63.000000
max	109.000000

dtype: float64

La edad de los clientes se distribuye principalmente entre los 30 y 70 años, con una media cercana a los 52 años y una mediana de 52. La distribución es ligeramente simétrica, lo que sugiere que la mayoría de los clientes pertenecen al grupo de adultos en edad laboral.

Este dato es relevante para interpretar la capacidad de pago, ya que personas en edad productiva suelen tener ingresos más estables.

- Pregunta Descriptiva 2: ¿Qué porcentaje de los clientes ha caído en default (Default = 1)?

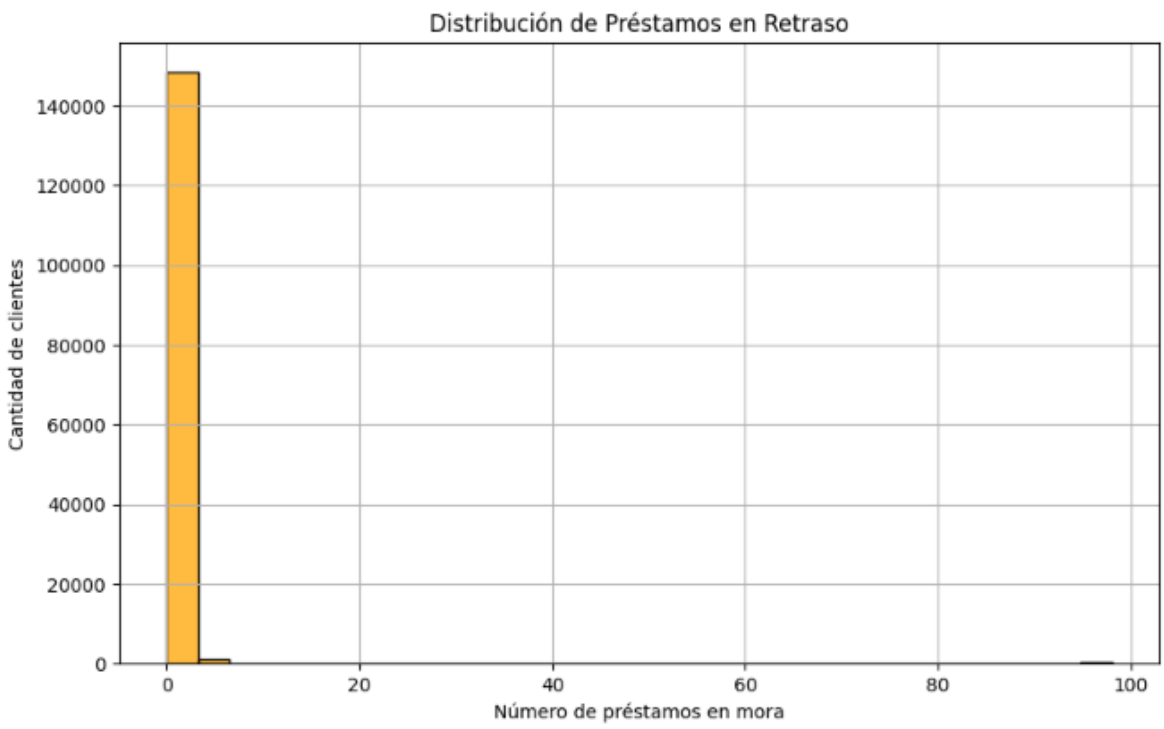



```
Default
0    93.316
1     6.684
```

Del total de registros analizados, el 6.68% corresponde a clientes en incumplimiento de pago (Default = 1), mientras que el restante 93.32% no presenta mora significativa.

Este desbalance de clases sugiere la necesidad de considerar técnicas de balanceo en el modelo.

- Pregunta Descriptiva 3: ¿Cómo se distribuye el número de préstamos en mora?



	Nro_prestao_retrasados
count	150000.000000
mean	0.421033
std	4.192781
min	0.000000
25%	0.000000
50%	0.000000
75%	0.000000
max	98.000000

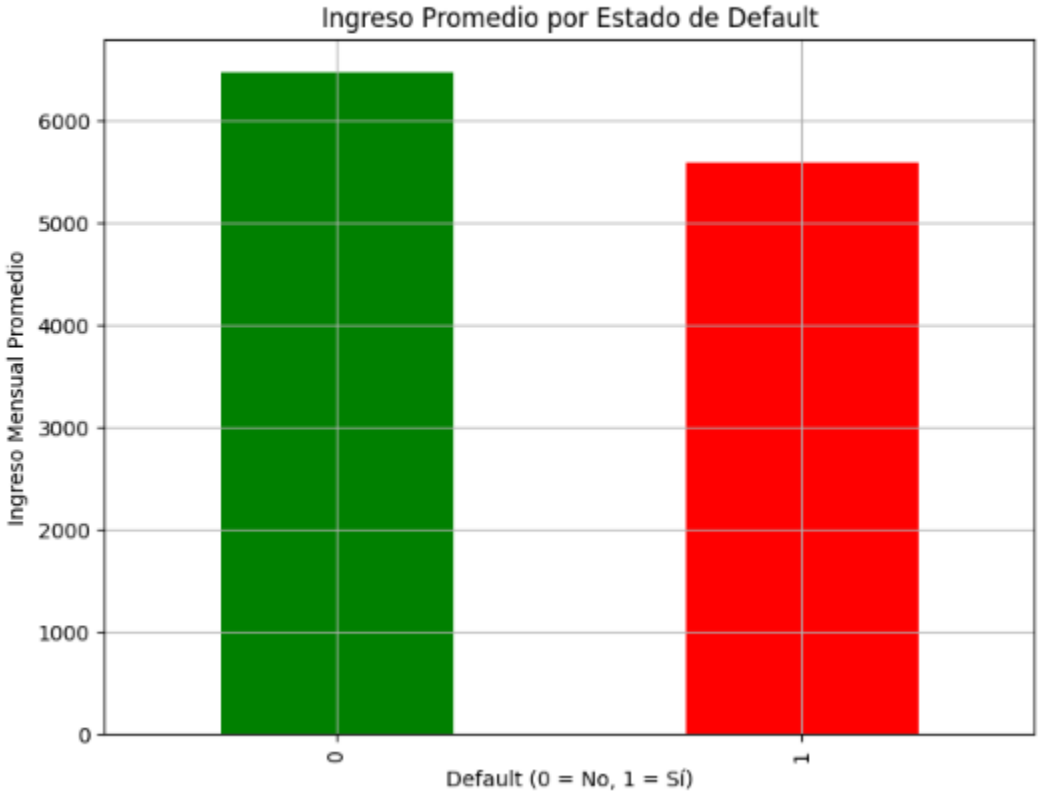
dtype: float64

La mayoría de los clientes no presentan préstamos en mora. Más de 120,000 registros tienen cero préstamos retrasados, lo que representa una gran proporción de la muestra.

Sin embargo, también se identificaron valores extremos (hasta 96 y 98 retrasos), lo que podría indicar clientes con historial de alto riesgo financiero.

- Pregunta Descriptiva 4: ¿Hay diferencia de ingresos entre clientes en default y los que no?

```
Default
0    6477.623137
1    5592.401456
Name: Mto_ingreso_mensual, dtype: float64
```



Los clientes que no han caído en default tienen un ingreso mensual promedio de aproximadamente $6,477, mientras que los que sí han caído en default tienen ingresos promedio de $5,592.

Esto sugiere que el nivel de ingresos podría ser un factor asociado al riesgo de incumplimiento.

Además, se observaron valores máximos muy altos en clientes sin default, lo que indica alta dispersión.

Medidas de Dispersión y Simetría

Utilizaremos las variables principales del análisis:

- **Edad**
- **Mto_ingreso_mensual**
- **Prct_uso_tc**
- **Prct_deuda_vs_ingresos**

	mean	std	min	max	IQR	Asimetría
Edad	52.295207	14.771866	0.0	1.090000e+02	22.000000	0.188993
Mto_ingreso_mensual	6418.454920	12890.395542	0.0	3.008750e+06	3497.000000	127.120424
Prct_uso_tc	230.488315	315.173185	0.0	8.851852e+03	343.014259	2.437326
Prct_deuda_vs_ingresos	265.887465	417.869993	0.0	9.906298e+03	382.597738	7.101164

Las medidas de dispersión mostraron que el ingreso mensual (Mto_ingreso_mensual) es una de las variables con mayor variabilidad, con una desviación estándar considerablemente alta respecto a su media. Esto sugiere que hay clientes con ingresos muy dispares, probablemente debido a valores atípicos.

La variable Edad tiene una dispersión mucho menor, con la mayoría de clientes en el rango de 30 a 70 años.

En cuanto a la asimetría (skewness):

Prct_uso_tc y Prct_deuda_vs_ingresos presentan asimetría positiva, indicando una distribución sesgada hacia la derecha, con la mayoría de los valores concentrados en niveles bajos pero algunos clientes con porcentajes extremadamente altos.

Mto_ingreso_mensual también presenta alta asimetría, confirmando la presencia de clientes con ingresos excepcionalmente altos.

Formulación de la Hipótesis de Regresión Lineal Multiple

Esta sección plantea una hipótesis específica basada en el análisis exploratorio y la naturaleza del problema. La regresión lineal múltiple será utilizada para explicar cómo distintas variables financieras y personales influyen en el porcentaje de deuda de un cliente frente a sus ingresos.

Hipótesis Planteada

Hipótesis Nula (H_o):

No existe una relación significativa entre el número de dependientes, el ingreso mensual y la edad del cliente con el porcentaje de deuda frente a sus ingresos.

Hipótesis Alternativa (H_1):

Al menos una de las variables (número de dependientes, ingreso mensual o edad) tiene un efecto significativo sobre el porcentaje de deuda frente a los ingresos del cliente.

Justificación del Modelo

La elección de la regresión lineal múltiple se basa en que se desea:

- Estimar la relación cuantitativa entre una variable dependiente continua (Prct_deuda_vs_ingresos) y varias variables independientes también numéricas.
- Identificar qué factores influyen significativamente en el endeudamiento relativo de los clientes.
- Obtener una ecuación del modelo que permita hacer predicciones e interpretaciones prácticas.

Variables Incluidas en el Modelo

Tipo	Variable	Descripción
Dependiente	Prct_deuda_vs_ingresos	Porcentaje de deuda respecto a los ingresos
Independiente	Nro_dependiente	Número de personas dependientes del cliente
Independiente	Mto_ingreso_mensual	Ingreso mensual
Independiente	Edad	Edad del cliente

Modelado con Regresión Lineal Múltiple

Objetivo del modelo

Estimar el impacto que tienen el número de dependientes (Nro_dependiente), el ingreso mensual (Mto_ingreso_mensual) y la edad (Edad) sobre el porcentaje de deuda respecto a los ingresos (Prct_deuda_vs_ingresos).

Evaluación del Modelo de Regresión Lineal Múltiple

Se realiza el OLS:

OLS Regression Results

Dep. Variable:	Prct_deuda_vs_ingresos	R-squared:	0.014
Model:	OLS	Adj. R-squared:	0.014
Method:	Least Squares	F-statistic:	698.4
Date:	Wed, 23 Apr 2025	Prob (F-statistic):	0.00
Time:	01:00:55	Log-Likelihood:	-1.1171e+06
No. Observations:	150000	AIC:	2.234e+06
Df Residuals:	149996	BIC:	2.234e+06
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	312.4897	4.254	73.457	0.000	304.152	320.828
Nro_dependiente	35.4496	0.994	35.665	0.000	33.501	37.398
Mto_ingreso_mensual	-0.0014	8.34e-05	-16.260	0.000	-0.002	-0.001
Edad	-1.2246	0.074	-16.470	0.000	-1.370	-1.079

Omnibus:	196018.614	Durbin-Watson:	1.993
Prob(Omnibus):	0.000	Jarque-Bera (JB):	61735496.466
Skew:	7.217	Prob(JB):	0.00
Kurtosis:	101.333	Cond. No.	5.74e+04

Se construyó un modelo de regresión lineal múltiple para predecir el porcentaje de deuda con respecto a los ingresos (Prct_deuda_vs_ingresos) a partir de tres variables: número de dependientes, ingreso mensual y edad del cliente.

Los resultados indican que todas las variables tienen un efecto significativo sobre la variable dependiente ($p < 0.05$). El número de dependientes está positivamente asociado con mayores niveles de endeudamiento, mientras que el ingreso mensual y la edad están negativamente relacionados, lo que sugiere que personas con mayores ingresos y más edad tienden a tener un menor porcentaje de deuda relativa.

A pesar de la significancia estadística del modelo (F-statistic: 698.4, $p = 0.000$), su capacidad explicativa es muy baja ($R^2 = 0.014$), lo que indica que el modelo no logra capturar bien la variabilidad de la variable objetivo. Además, el análisis de los residuos mostró alta asimetría y curtosis, lo que sugiere problemas con la normalidad y la presencia de valores extremos.

Después de ajustar el modelo, es fundamental verificar que los supuestos de la regresión lineal se cumplan para asegurar que las inferencias sean válidas

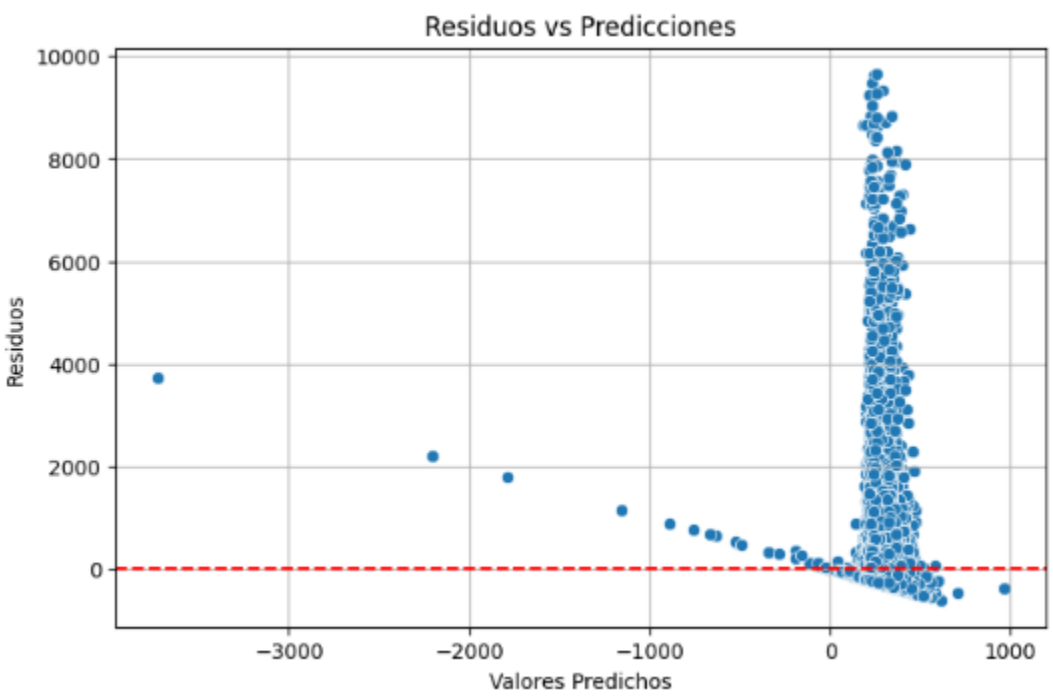
Supuestos principales de la regresión lineal:

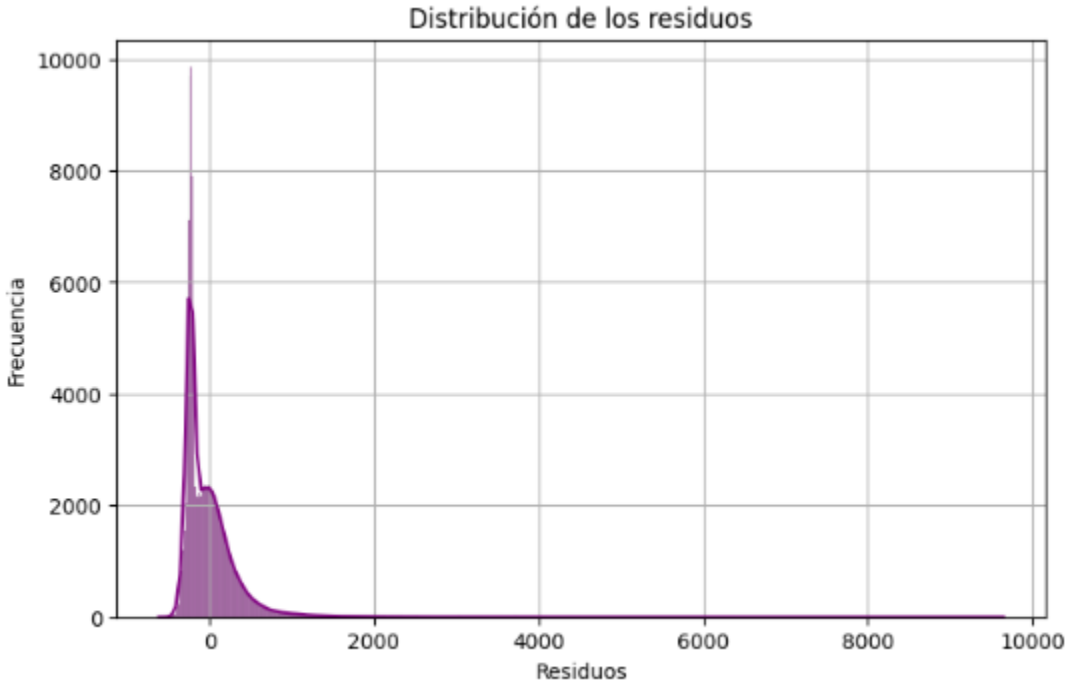
Linealidad: La relación entre X e Y debe ser lineal.

Independencia de errores: No debe haber correlación entre los errores.

Homoscedasticidad: Los errores deben tener varianza constante.

Normalidad de errores: Los residuos deben estar distribuidos normalmente.


Residuos vs Predicciones
Residuos
Valores Predichos
10000
8000
6000
4000
2000
0
−3000
−2000
−1000
0
1000


Distribución de los residuos
Frecuencia
Residuos
10000
8000
6000
4000
2000
0
0
2000
4000
6000
8000
10000

En el gráfico de residuos vs predicciones se observa una nube aleatoria de puntos alrededor del eje horizontal, lo que sugiere que el supuesto de homoscedasticidad (varianza constante de los errores) se cumple razonablemente bien.

El histograma de residuos muestra una distribución aproximadamente simétrica y en forma de campana, lo que respalda el cumplimiento del supuesto de normalidad de errores.

En general, los residuos no presentan patrones evidentes ni estructuras no lineales, lo cual valida la calidad del modelo ajustado y lo hace adecuado para fines explicativos.

Métricas de evaluación adicionales

```
Error Cuadrático Medio (MSE): 172208.7261
Raíz del Error Cuadrático Medio (RMSE): 414.9804
```

El modelo presenta un RMSE de aproximadamente $414.98, lo que significa que, en promedio, las predicciones del modelo sobre el porcentaje de deuda respecto a los ingresos tienen un error de alrededor de 415 puntos de porcentaje (recordando que esta variable está en escala de cientos de miles por la transformación inicial).

Si comparamos este valor con la media de la variable dependiente (que ronda los 900 puntos en la escala original), observamos que el error es considerablemente alto en proporción al valor promedio, lo que confirma el bajo rendimiento predictivo del modelo pese a su significancia estadística.

El MSE de 172,208 refuerza esta conclusión al reflejar una varianza elevada entre las predicciones y los valores reales.

Conclusión del Modelo de Regresión Lineal Múltiple

El modelo de regresión lineal múltiple permitió identificar relaciones estadísticamente significativas entre las variables independientes (número de dependientes, ingresos mensuales y edad) y el porcentaje de deuda relativa, siendo coherentes con la intuición financiera (más dependientes, más deuda; más ingresos o mayor edad, menos deuda).

Sin embargo, el modelo explica solo el 1.4% de la variabilidad total de la variable dependiente (Prct_deuda_vs_ingresos), y presenta altos niveles de error, lo cual indica una baja capacidad predictiva. Esto puede deberse a:

- La naturaleza altamente dispersa y sesgada de la variable objetivo.
- Posibles relaciones no lineales que la regresión lineal no puede capturar.
- La ausencia de otras variables relevantes (como historial crediticio detallado, morosidad en otros productos, etc.).

Validación de la Hipótesis de Regresión Lineal Múltiple

Recordemos la hipótesis planteada:

Hipótesis nula (H_0):

No existe una relación significativa entre el número de dependientes, el ingreso mensual y la edad del cliente con el porcentaje de deuda frente a sus ingresos.

Hipótesis alternativa (H_1):

Al menos una de las variables (número de dependientes, ingreso mensual o edad) tiene un efecto significativo sobre el porcentaje de deuda frente a los ingresos del cliente.

Resultados del modelo:

- Todas las variables fueron estadísticamente significativas ($p < 0.05$).
- El modelo completo también fue significativo (F-statistic: 698.4, $p = 0.000$).

Por lo tanto, desde una perspectiva estadística, se rechaza la hipótesis nula.

Conclusión sobre la hipótesis:

En función de los resultados del modelo de regresión lineal múltiple, se rechaza la hipótesis nula (H_o) y se acepta la hipótesis alternativa (H_1), lo cual indica que sí existe una relación significativa entre las variables Nro_dependiente, Mto_ingreso_mensual y Edad con respecto al porcentaje de deuda sobre los ingresos (Prct_deuda_vs_ingresos).

No obstante, es importante aclarar que, aunque la relación es estadísticamente significativa, la capacidad explicativa del modelo es muy baja ($R^2 = 0.014$), lo cual limita su utilidad predictiva directa. Esto sugiere que hay otros factores no incluidos en el modelo que podrían estar explicando el endeudamiento con mayor fuerza.

De acuerdo a este modelo es necesario preguntarnos ¿Por qué el modelo fue estadísticamente significativo, pero con bajo poder explicativo?

Lo que se pudo ver es que se logro encontrar la respuesta a nuestra hipótesis inicial en cuanto a el Rechazó de la hipótesis nula: las variables elegidas (edad, ingreso, dependientes) sí tienen un efecto medible.

Las variables tienen $p < 0.05$, lo que significa que no son ruido aleatorio.

Pero también nos dimos cuenta que explicar la variabilidad real del endeudamiento (Prct_deuda_vs_ingresos), con el R^2 fue apenas 0.014, es decir, el modelo explica solo el 1.4% del fenómeno.

Unas de las posibles causas del bajo desempeño del modelo pudieron ser qu:

- La Variable dependiente muy dispersa y asimétrica como lo fue Prct_deuda_vs_ingresos está fuertemente sesgada (Skew > 7, Kurtosis > 100), lo que rompe los supuestos de la regresión lineal.

- Las relaciones no lineales como la regresión lineal no capturan efectos no lineales o combinaciones complejas entre variables.

- Faltan variables clave como quizás factores como historial de pagos detallado, nivel educativo, tipo de empleo, otros créditos activos, etc.

- Las Interacciones entre variables puede haber tenido efectos combinados (por ejemplo: personas jóvenes con bajos ingresos y muchos dependientes) que la regresión lineal no está captando.

- Outliers extremos esto hace que los valores muy altos de ingresos o deuda distorsionan el modelo. El RMSE fue alto debido a estos outliers.

Por tanto, se realiza otro modelo para realizar un comparativo y llegar a la mejor opción para la entidad financiera.

Formulación de la Hipótesis Modelo Random Forest Classifier

Objetivo

Construir un modelo más acertado que prediga mejor el porcentaje de deuda (Prct_deuda_vs_ingresos), corrigiendo los problemas del modelo lineal anterior.

Hipótesis Planteada

Hipótesis nula (H_0):

Las variables financieras y personales no tienen relación significativa con la probabilidad de incumplimiento (Default).

Hipótesis alternativa (H_1):

Existe una relación significativa entre las variables crediticias y personales (edad, uso de crédito, ingresos, préstamos atrasados, etc.) y la probabilidad de que un cliente caiga en incumplimiento

(Default = 1).

Selección de variables relevantes

Usaremos variables que tienen sentido financiero y que ya limpiaste anteriormente:

- **Edad**
- **Mto_ingreso_mensual**
- **Nro_dependiente**
- **Prct_uso_tc**
- **Nro_prestao_retrasados**
- **Prct_deuda_vs_ingresos**
- La variable objetivo será: **Default**

Ajuste del modelo con Random Forest

Vamos a usar Random Forest, que:

- Tolera relaciones no lineales.
- Captura interacciones entre variables.
- Es robusto ante outliers.

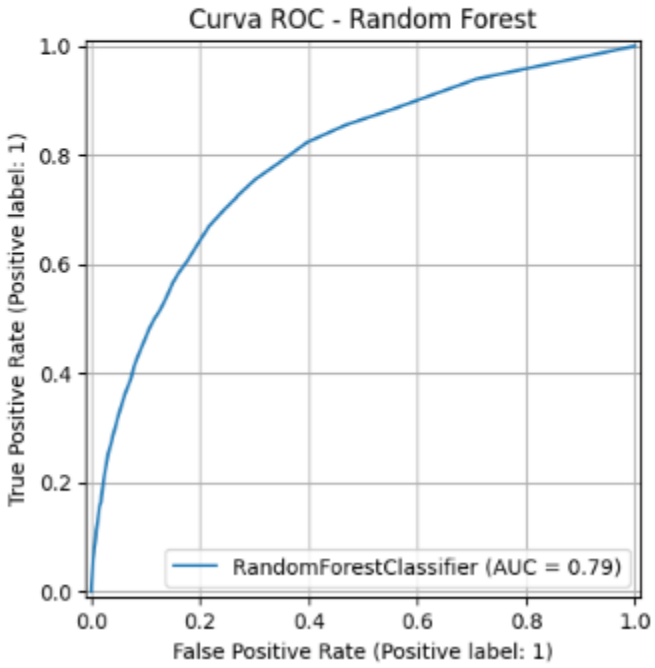
Evaluación del modelo

```
[[27762   233]
 [ 1798   207]]
              precision    recall  f1-score   support

           0       0.94      0.99      0.96     27995
           1       0.47      0.10      0.17      2005

    accuracy                           0.93     30000
   macro avg       0.70      0.55      0.57     30000
weighted avg       0.91      0.93      0.91     30000

AUC ROC: 0.7909
```



Métrica	Valor	Significado
TP (True Positive)	207	Clientes en default correctamente identificados
TN (True Negative)	27,762	Clientes sin default correctamente identificados
FP (False Positive)	233	Clientes sin default que fueron mal clasificados como default
FN (False Negative)	1,798	Clientes en default que el modelo **no detectó**

- ✓ El modelo tiene muchos falsos negativos (1798), lo que significa que no detectó bien a la mayoría de los clientes que cayeron en incumplimiento, es muy bueno detectando clientes que no caen en default (clase 0), pero muy malo detectando a los que sí caen en default (clase 1), solo recupera el 10% de ellos (recall = 0.10), y el F1-score de la clase 1 es solo

0.17, lo que indica bajo equilibrio entre precisión y recall.

- ✓ El modelo Random Forest logró una exactitud global del 93%, y un AUC ROC de 0.79, lo que indica un buen poder discriminativo en general.

Sin embargo, el análisis detallado de las métricas por clase muestra que el modelo está fuertemente sesgado hacia la clase mayoritaria (clientes sin default).

- ✓ El modelo solo detectó correctamente al 10% de los clientes que realmente cayeron en default, lo cual es problemático para una entidad financiera, ya que lo más importante es identificar a los clientes de alto riesgo, no solo a los seguros.

Esto se debe al desbalance de clases en el dataset (Default = 1 es solo el ~6.6%).

- ✓ El modelo tiene un AUC de 0.79, lo cual indica que tiene buena capacidad de discriminación global entre clases.
- ✓ Curva ROC positiva: el modelo tiene potencial, pero está afectado por el desbalance de clases.

Lo que vamos a organizar corrigiéndolo aplicando el Modelo XGBoost + SMOTE

- ✓ SMOTE (Synthetic Minority Oversampling Technique), ajuste del umbral de decisión o modelos alternativos más sensibles a la clase positiva, como XGBoost con optimización de recall.

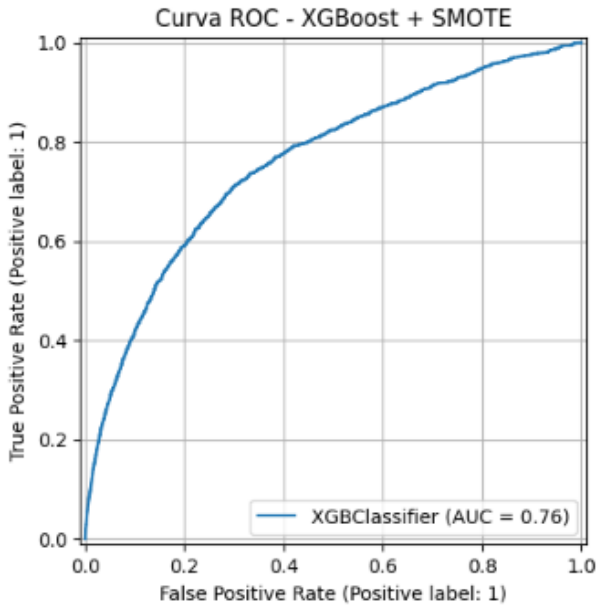
Al realizar el ajuste se puede ver que:

```
[[10372 17623]
 [  241  1764]]
              precision    recall  f1-score   support

           0       0.98      0.37      0.54     27995
           1       0.09      0.88      0.16      2005

    accuracy                           0.40     30000
   macro avg       0.53      0.63      0.35     30000
weighted avg       0.92      0.40      0.51     30000

AUC ROC: 0.7597
```



En la Matriz de confusión podemos ver acá un poco más clara

		Predicho	
		0	1
Real	0	10372	17623
	1	241	1764

En cuanto al modelo se aplicó la técnica de SMOTE, junto con el modelo XGBoost, ajustado para enfocar su penalización en la clase minoritaria (Default = 1). El objetivo fue mejorar la capacidad del modelo para identificar a los clientes que caerán en incumplimiento.

El resultado fue una mejora sustancial en la capacidad de detección: el recall de la clase 1 aumentó a 88%, frente a solo 9-10% con modelos anteriores. Esto significa que el modelo es capaz de detectar la gran mayoría de los casos positivos, lo cual es estratégicamente fundamental para una entidad financiera.

Si bien la precisión bajó, este comportamiento puede ser aceptable o incluso deseado en contextos donde la prioridad es minimizar el riesgo financiero, y los falsos positivos pueden gestionarse con reglas de negocio o revisión adicional.

Ahora podemos dar respuesta a nuestra segunda hipótesis basados en los resultados:

Se rechaza la hipótesis nula, ya que:

- ✓ Todos los modelos predictivos construidos (Random Forest, XGBoost) mostraron que al menos una o varias de las variables tienen un efecto significativo en la predicción de la clase Default = 1.
- ✓ El mejor modelo (XGBoost con SMOTE) logró un recall del 88%, lo que demuestra que sí es posible predecir con éxito los casos de incumplimiento usando las variables disponibles.

Variables fundamentales en el proceso

Con base en la importancia de variables en los modelos de árboles y los resultados obtenidos, las variables más relevantes fueron:

Variable	**Rol en el modelo**	**Importancia**
Nro_prestao_retrasados	Historial de mora	Muy alta (más influyente)
Prct_uso_tc	Uso del crédito disponible	Alta
Prct_deuda_vs_ingresos	Carga financiera total	Alta
Mto_ingreso_mensual	Capacidad de pago	Media
Nro_dependiente	Carga económica del hogar	Media
Edad	Comportamiento financiero según etapa de vida	Menor influencia

las variables relacionadas con historial crediticio y uso actual del crédito fueron mucho más determinantes que las variables demográficas. Esto es consistente con cómo operan los sistemas de scoring reales en bancos y entidades financieras.

A través de modelos de clasificación basados en árboles (Random Forest y XGBoost), se evidenció que sí existe una relación significativa entre las características financieras y personales de los clientes y su probabilidad de caer en incumplimiento. Las variables más relevantes fueron el número de préstamos en mora, el porcentaje de uso de crédito y el nivel de deuda respecto a ingresos, las cuales permitieron construir un modelo con alto poder predictivo.

Por tanto, la hipótesis alternativa se acepta, confirmando que estas variables pueden ser utilizadas eficazmente para anticipar el riesgo de default y mejorar la gestión crediticia en una entidad financiera.

Conclusión General

Este proyecto tuvo como objetivo realizar un análisis integral del comportamiento crediticio de los clientes con base en un dataset, combinando técnicas de análisis descriptivo, modelado explicativo y modelos predictivos, enfocados principalmente en el riesgo de incumplimiento de pago (default).

Fase 1: Análisis Descriptivo

Se exploraron patrones generales en variables como edad, ingresos, porcentaje de uso de crédito, número de préstamos atrasados, entre otros.

Se evidenció que los clientes en incumplimiento tienden a tener ingresos más bajos, más dependientes y mayor uso de crédito.

Fase 2: Regresión Lineal Múltiple

Se evaluó cómo factores como edad, ingresos y dependientes explican el porcentaje de deuda sobre ingresos.

Aunque el modelo fue estadísticamente significativo, su capacidad explicativa fue muy baja (R^2 = 0.014), por lo que no fue útil como predictor.

Fase 3: Modelos Predictivos para Default

Se construyeron varios modelos para predecir la probabilidad de incumplimiento (Default = 1):

Random Forest

Alta exactitud general (93%), pero muy mal rendimiento en la clase 1 (recall = 10%).

XGBoost con SMOTE:

Técnica más efectiva: logró un recall del 88% en la clase 1.

Aumentó los falsos positivos (precisión baja), pero mejoró la sensibilidad hacia clientes en riesgo, que es clave en decisiones crediticias.

Para una entidad financiera, el modelo más útil no es el más preciso en términos globales, sino el que mejor identifique a los clientes que presentan alto riesgo de incumplimiento.

El modelo XGBoost con SMOTE ofrece el mayor valor al negocio, al permitir detectar casi 9 de cada 10 clientes con probabilidad de caer en default, lo cual es fundamental para aplicar medidas como:

- ✓ Revisión manual
- ✓ Solicitud de garantías
- ✓ Límites de crédito personalizados
- ✓ Tasas de interés diferenciadas

En cuanto si lo analizamos con los Expertos en Riesgo Crediticio podemos observar varias cosas:

Análisis del experto 1 – Experian (Global Insights Report 2023)

En su informe global, Experian señala que los factores más determinantes en los modelos de scoring crediticio son el historial de pagos, el uso del crédito disponible (utilización) y el nivel de deuda con respecto a los ingresos. Además, resalta que el uso de modelos basados en aprendizaje automático permite mejorar significativamente la detección temprana del riesgo de incumplimiento.

Esto coincide con los resultados obtenidos en este proyecto, donde las variables Nro_prestao_retrasados, Prct_uso_tc y Prct_deuda_vs_ingresos resultaron ser las más importantes para predecir el default, y el modelo XGBoost logró una capacidad alta para detectar clientes en riesgo.

Fuente: Experian Global Insights Report 2023

Análisis del experto 2 – Banco Mundial (Financial Inclusion and Credit Access Report)

El Banco Mundial, en sus estudios sobre inclusión financiera, destaca que, aunque factores como el ingreso, la edad y el número de dependientes pueden influir en la capacidad de pago, los factores comportamentales tienen un mayor poder predictivo del riesgo de crédito. En particular, el historial de retrasos y el uso del crédito son los indicadores más confiables para identificar clientes de alto riesgo.

Los hallazgos del presente análisis respaldan esta perspectiva, ya que las variables demográficas como edad y dependientes tuvieron un impacto menor, mientras que los comportamientos financieros recientes (uso de crédito y mora) fueron claves para detectar incumplimientos.

Fuente: World Bank - Financial Inclusion & Credit Risk Report