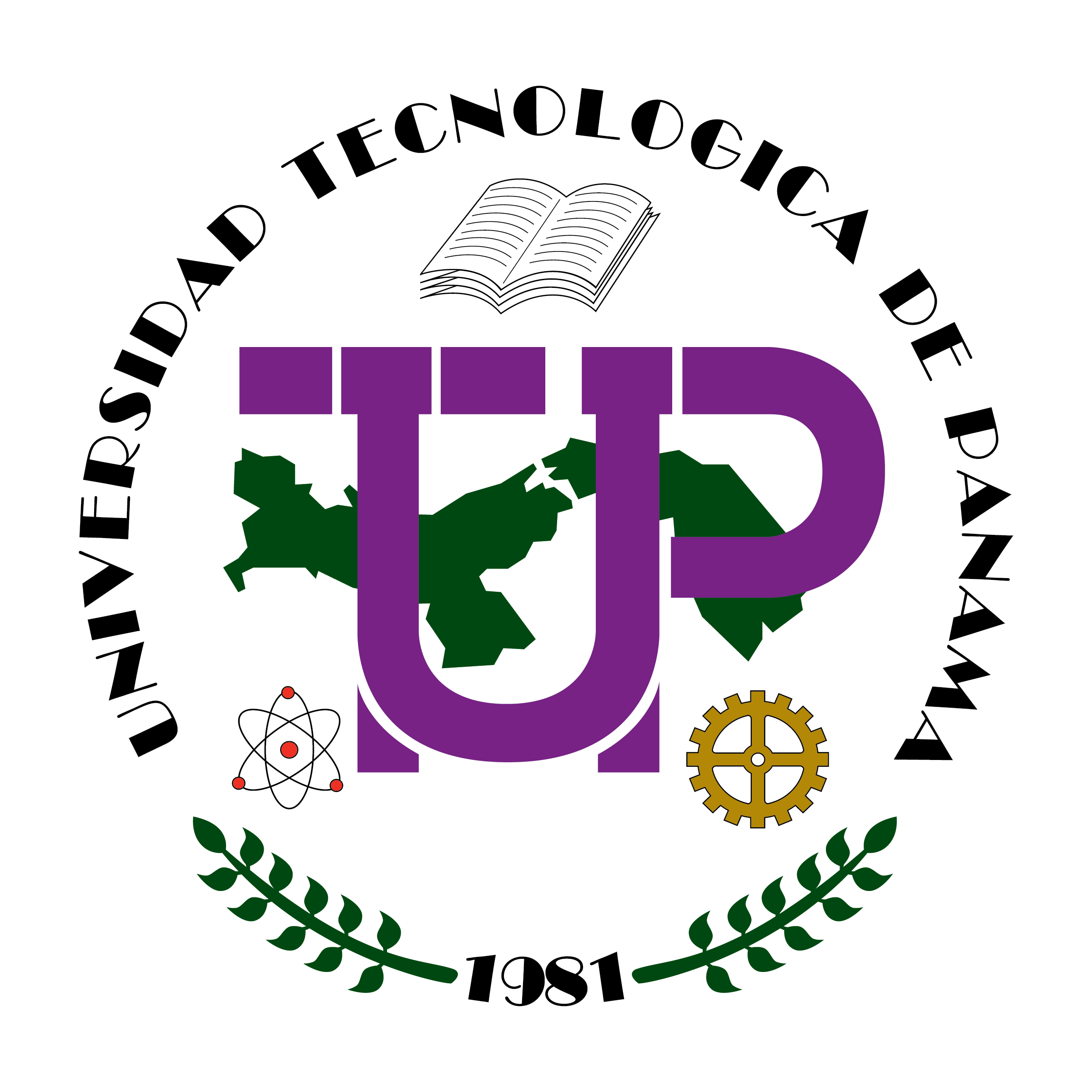
**A circular logo with text and letters

AI-generated content may be incorrect.**

**Reporte del Proyecto Final: Desarrollo de un Modelo de Probabilidad de Incumplimiento (PD).**

**Universidad Tecnológica de Panamá  
Facultad de Ingeniería sistemas  
Maestría en Analítica de Datos**

**Asignatura:**

Modelos Predictivos

**Profesor:**

Juan Marcos Castillo, PhD

**Integrantes:**

Elías Vergara

**Cédula**:

8-910-2053

**Año**

**2025**

Contents

[Introducción 3](#_Toc204621017)

[Justificación 4](#_Toc204621018)

[Antecedentes 5](#_Toc204621019)

[Definición del problema 6](#_Toc204621020)

[Planteamiento del Problema 6](#_Toc204621021)

[Análisis predictivo 6](#_Toc204621022)

[Determinación de la base de datos 7](#_Toc204621023)

[Preprocesamiento y limpieza 7](#_Toc204621024)

[Preprocesamiento de Variables categóricas 9](#_Toc204621025)

[Peso de la Evidencia (WoE) 9](#_Toc204621026)

[Valor de la Información (VI) 11](#_Toc204621027)

[Análisis descriptivo 12](#_Toc204621028)

[Selección de variables 19](#_Toc204621029)

[Selección de Modelos 19](#_Toc204621030)

[Regresión Logística: 19](#_Toc204621031)

[Random Forest: 20](#_Toc204621032)

[Curva ROC Comparativa: 21](#_Toc204621033)

[Coeficiente de Gini 22](#_Toc204621034)

[Kolmogorov Smirnov 23](#_Toc204621035)

[Conclusiones 24](#_Toc204621036)

[Recomendaciones y Futuros Estudios 24](#_Toc204621037)

# Introducción

La Probabilidad de Incumplimiento (PD) es una pieza clave en cómo las instituciones financieras entienden y manejan el riesgo de que una persona o empresa no pague sus deudas. En términos sencillos, la PD es una estimación de qué tan probable es que un prestatario (la persona que recibe el préstamo) no cumpla con sus pagos en un período de tiempo específico, por ejemplo, en los próximos 12 meses. Calcular esta probabilidad con precisión es extremadamente importante para bancos, cooperativas de crédito y otras entidades financieras. Les permite saber exactamente cuánto riesgo están asumiendo con cada préstamo, les ayuda a reservar el dinero adecuado para cubrir posibles pérdidas (lo que se conoce como provisiones), a usar su capital de la forma más inteligente ya tomar decisiones más informadas y seguras sobre a quién le conceden un crédito. Este documento se dedicará a explicar cómo desarrollamos y evaluamos un modelo para estimar esta probabilidad de incumplimiento. Para ello, utilizamos dos técnicas modernas de inteligencia artificial muy populares: la Regresión Logística y el Random Forest, aplicando estas herramientas a un conjunto de datos reales de préstamos.

# Justificación

La razón por la que es tan importante crear e implementar modelos de Probabilidad de Incumplimiento es múltiple y abarca desde el cumplimiento de reglas estrictas en el sector financiero hasta la mejora de la forma en que las empresas operan y obtienen ganancias:

**Cumplimiento de Normas y Regulaciones:** Las leyes y acuerdos que rigen el sistema bancario a nivel mundial, como los **Acuerdos de Basilea**, son la razón principal. Estos acuerdos, especialmente Basilea II y Basilea III, exigen a los bancos que midan y gestionen el riesgo de crédito de manera muy rigurosa. En particular, solicitan que se estime la PD para calcular cuánto riesgo tienen los activos de un banco y, por lo tanto, cuánto capital debe tener guardado para protegerse. Esto significa que tener un modelo de PD confiable no es solo una buena práctica, sino una obligación. A nivel local, la **Superintendencia de Bancos de Panamá (SBP)** también emite acuerdos y directrices que establecen cómo deben clasificarse los préstamos y cómo se deben crear las reservas para posibles pérdidas. Estos acuerdos de la SBP refuerzan la necesidad de que los bancos panameños cuenten con modelos internos precisos para manejar su riesgo de crédito.

**Gestión Proactiva del Riesgo:** Contar con un modelo de PD permite a las instituciones financieras adelantarse a los problemas. Si el modelo identifica a un prestatario que, con alta probabilidad, podría incumplir su pago, el banco puede tomar medidas a tiempo. Por ejemplo, podría ofrecer una reestructuración de la deuda, revisar los límites de crédito o contactar al cliente para entender su situación, antes de que el incumplimiento se concrete. Esto ayuda a reducir las pérdidas ya mantener la salud financiera de la entidad.

**Optimización de Decisiones:** Un modelo de PD contribuye a que el proceso de evaluación de créditos sea más automático, más rápido y mucho más uniforme. Esto significa que las decisiones sobre si aprobar o denegar un préstamo, y qué tasa de interés aplicar, se basan en criterios objetivos y consistentes. Al tener un sistema estandarizado, se evitan errores humanos y se mejora la eficiencia general del proceso de concesión de créditos.

**Uso Eficiente del Capital:** Al poder estimar con precisión el riesgo de cada operación de crédito, el modelo de PD influye directamente en cómo el banco asigna su capital. Si un préstamo tiene bajo riesgo, se puede asignar menos capital de reserva, lo que libera recursos para otras inversiones más productivas. Esta asignación inteligente del capital contribuye significativamente a mejorar la rentabilidad general de la institución financiera.

# Antecedentes

Los modelos de Probabilidad de Incumplimiento han tenido una evolución fascinante a lo largo de los años. Al principio, se usaron métodos estadísticos más sencillos, pero con el avance de la tecnología, se han incorporado técnicas mucho más avanzadas de inteligencia artificial (Machine Learning).

Los **Acuerdos de Basilea** , un conjunto de recomendaciones para la regulación bancaria a nivel global, han sido el principal motor de esta transformación en el sector. Basilea II, por ejemplo, fue un hito porque permitió a los bancos, bajo ciertas condiciones, usar sus propios modelos internos para calcular cuánto capital necesitaban guardar. Para ello, era fundamental estimar con precisión la PD, así como la LGD (porcentaje de pérdida si el cliente incumple) y la EAD (cantidad expuesta en el momento del incumplimiento). Basilea III, la versión más reciente, fue aún más allá, exigiendo una mayor calidad y cantidad de capital a los bancos, lo que hizo que la necesidad de modelos de PD robustos y bien validados fuera aún más apremiante. Las regulaciones de Basilea ponen mucho énfasis en que los modelos sean muy precisos, estables en el tiempo y, sobre todo, que sean "interpretables", es decir, que se pueda entender cómo llegan a sus resultados.

A nivel local, la **Superintendencia de Bancos de Panamá (SBP)** , siguiendo los estándares internacionales, ha emitido varios acuerdos que rigen cómo se debe clasificar el riesgo de los créditos y cómo deben crearse las provisiones. Por ejemplo, acuerdos como el No. 4-2013 o el No. 6-2021 detallan los criterios para clasificar los préstamos en diferentes categorías de riesgo (como Normal, Mención Especial, Subnormal, Dudoso o Incobrable). Estos criterios se basan en factores como el tiempo que lleva el cliente sin pagar, su capacidad para generar ingresos y si tiene garantías que respalden el préstamo. El Acuerdo No. 012-2019, es aún más explícito, al señalar que "La probabilidad de incumplimiento debe ser un buen indicador para predecir la frecuencia de incumplimiento para cada grupo de riesgo" y que "Una diferencia constante entre las frecuencias observadas y las probabilidades estimadas debe llevar a revisar el sistema". Esto demuestra claramente que la SBP no solo aprueba, sino que espera que los bancos utilicen modelos de PD para una gestión de riesgo más efectiva y basada en datos.

Grandes empresas de consultoría como **Deloitte** han observado y documentado la creciente tendencia de usar técnicas avanzadas de inteligencia artificial en los modelos de riesgo de crédito. Sus informes destacan cómo el Machine Learning está siendo cada vez más utilizado para desarrollar y validar modelos de PD. Aunque estos modelos de inteligencia artificial son muy potentes y precisos para predecir, también presentan un desafío importante: a menudo son como una "caja negra", difícil de entender cómo llegan a sus resultados. Esta falta de interpretabilidad puede ser un problema para cumplir con las normas regulatorias, que piden transparencia. Por ello, Deloitte subraya la importancia de tener una documentación muy clara y de realizar validaciones muy rigurosas para estos modelos más complejos.

# Definición del problema

El objetivo principal de este proyecto es construir un modelo predictivo que sea capaz de estimar la probabilidad de incumplimiento de las personas que piden préstamos. Para lograr esto, utilizaremos datos históricos de préstamos ya otorgados. Queremos descubrir cuáles son las características de los solicitantes (como su historial crediticio, ingresos, etc.) Y las condiciones del préstamo (como el plazo, la tasa de interés) que son más importantes o influyentes para determinar si un préstamo tiene un alto riesgo de no ser pagado. En este trabajo, nos hemos propuesto crear y comparar dos tipos de modelos de clasificación, la Regresión Logística y el Random Forest, para ver cuál de los dos es más efectivo a la hora de predecir si un préstamo terminará en estado de impago o "default".

## Planteamiento del Problema

Los reguladores financieros exigen que los modelos sean faciles de interpretar. por esto en línea general se suele utilizar variables Dummies para interpretar las variables categóricas ya que hacen que el entendimiento del modelo de PD sea más facil de entender.

1. Nuestra variable dependiente será binaria.

R: Asignaremos 1 a Bueno y 0 a Malo.

2. Pero si estamos modelando la probabilidad de Default ¿No deberíamos asignar 1 al Malo (Default)?

R: Recordemos: queremos un modelo fácil de interpretar

Por eso, queremos hacerlo lo más intuitivo posible

Al asignar 1 a los "Buenos," los coeficientes de las variables independientes que incrementen la probabilidad de "los buenos" serán positivos y los que la disminuya serán negativos... esto es muy intuitivo.

# Análisis predictivo

## Determinación de la base de datos

La información que hemos utilizado para este análisis proviene de una fuente pública y muy relevante: **lendingClub**. LendingClub es una plataforma en línea que conecta directamente a personas que buscan préstamos con inversionistas. El archivo original de datos que empleamos se llama “LendinClub.csv”. Este archivo es muy completo, ya que contiene una gran cantidad de características sociodemográficas de los clientes y información sobre los préstamos otorgados y lo que nos da una base sólida para el modelado.

## Preprocesamiento y limpieza

La fase de preprocesamiento y limpieza de los datos es un paso absolutamente crucial y fundamental para preparar el conjunto de datos de manera que los modelos de machine learning puedan entenderlo y trabajar con él de manera efectiva. Los pasos más importantes que realizamos en esta etapa incluyen:

**Manejo de tipos mezclados:** en algunos casos, ciertas columnas de los datos presentaban información con diferentes tipos de formatos o "tipos mixtos", lo que puede causar errores en el análisis. Identificamos y corregimos estas inconsistencias, como se observa en la columna “term” de nuestro conjunto de datos, asegurando que cada columna tenga un tipo de dato consistente.

**Extracción y conversión de variables temporales y textuales:** variables como term (que indica el plazo del préstamo, por ejemplo, "36 meses" o "60 meses") y emp\_length (que describen los años de empleo, como "10+ años" o "menos de 1 año") originalmente estaban en formato de texto o combinaciones de números y texto.

Las convertimos a formatos puramente numéricos para que los modelos puedan interpretarlas. De manera similar, la variable issue\_d (que representa la fecha en que se otorgó el préstamo) también fue procesada y transformada para ser útil en el análisis.

**Creación de variables ficticias (dummy variables):** muchas de las características en nuestros datos eran de tipo categórico, es decir, representaban categorías en lugar de cantidades (por ejemplo, el tipo de propiedad de la vivienda o el estado civil). Para que los modelos de machine learning, que generalmente trabajan con números, pudieran usar esta información, transformamos estas variables categóricas en "variables dummy". Esto implica crear nuevas columnas para cada categoría posible de una variable; si la categoría está presente, la nueva columna tendrá un 1, y si no, un 0. Este proceso se aplicó a características como home\_ownership (si es propietario, inquilino, etc.), sub\_grade (una subcategoría de la calificación crediticia), addr\_state (el estado de la dirección del prestatario), verification\_status (si los ingresos fueron verificados o no), las distintas categorías de emp\_length (años de empleo) y las categorías de annual\_inc (ingreso anual).

**Tratamiento de datos extremos:** durante la exploración de los datos, identificamos "valores extremos" o "outliers", que son datos que se desvían significativamente del patrón general (por ejemplo, un monto de préstamo extremadamente alto o una tasa de interés inusualmente baja). Estos valores pueden distorsionar los resultados del modelo, por lo que se identifican y manejan adecuadamente en variables como funded\_amnt (el monto total financiado del préstamo) y int\_rate (la tasa de interés del préstamo).

**Eliminación de columnas redundantes o irrelevantes:** se analizaron todas las columnas del conjunto de datos y se eliminaron aquellas que no aportaron información útil para la predicción o que duplicaron la información ya presente en otras columnas. Esto ayuda a simplificar el modelo ya mejorar su rendimiento. Al final obtuvimos un dataset que contenía 89 variables independiente para continuar con nuestro modelado.

Listado de las variables:

 ['Casa\_MORTGAGE','Casa\_RENT\_ANY\_OTHER\_NONE','Casa\_OWN','Direccion\_ND\_NE\_IA\_NV','Direccion\_FL','Direccion\_AL\_HI\_MO\_NM','Direccion\_CA','Direccion\_NC\_ID\_NJ','Direccion\_NY','Direccion\_KY\_LA\_MD','Direccion\_MI\_AR\_AZ\_VA\_OK\_DE\_OH','Direccion\_MN\_PA\_UT\_MA\_RI\_WA\_TN\_IN','Direccion\_CT\_IL','Direccion\_TX','Direccion\_NH\_AK\_MT\_MS\_WY\_WV\_DC\_ME',#'Verificacion\_Not Verified','Verificacion\_Source Verified','Verificacion\_Verified','Proposito\_ed\_pyme\_enerren\_mudanza','Proposito\_casa\_otros\_boda\_salud\_vacacion','Proposito\_comprasgrandes\_mejoras\_carro','Proposito\_debt\_consolidation','Proposito\_credit\_card','Grados\_B','Grados\_C','Grados\_D','Grados\_E','Grados\_F\_G','Estado\_Lista\_w','plazo\_36','plazo\_60','antiguedad\_menosde1','antiguedad\_1a4','antiguedad\_5a6','antiguedad\_7','antiguedad\_8a9','antiguedad\_10+','meses\_credito<87','meses\_credito\_87\_89','meses\_credito\_89\_90','meses\_credito\_90\_98','meses\_credito\_98\_101','meses\_credito\_101\_110','meses\_credito\_110\_126','meses\_credito\_126\_155','meses\_credito\_155+','ingresos\_anuales<20K','ingresos\_anuales20K-30K','ingresos\_anuales30K-40K','ingresos\_anuales40K-50K','ingresos\_anuales50K-60K','ingresos\_anuales60K-70K','ingresos\_anuales70K-80K','ingresos\_anuales80K-90K','ingresos\_anuales90K-100K','ingresos\_anuales100K-126K','ingresos\_anuales126K-152K','ingresos\_anuales152K-227K','ingresos\_anuales>227K','meses\_desde\_fallo\_null','meses\_desde\_fallo\_0-4','meses\_desde\_fallo\_4-30','meses\_desde\_fallo\_30-60','meses\_desde\_fallo\_60-83','meses\_desde\_fallo\_83+','morosidad\_2anios\_0','morosidad\_2anios\_1-4','morosidad\_2anios\_>=5','total\_acc\_<=6','total\_acc\_6-22','total\_acc\_22-50','total\_acc\_>50','dti<=3.2','dti\_3.2-8.8','dti\_8.8-10.4','dti\_10.4-13.6','dti\_13.6-16.0','dti\_16.0-16.7','dti\_16.7-19.9','dti\_19.9-20.8','dti\_20.8-23.2','dti\_23.2-35.2','dti>35.2','mths\_since\_last\_record\_faltantes','mths\_since\_last\_record\_0-3','mths\_since\_last\_record\_3-21','mths\_since\_last\_record\_21-31','mths\_since\_last\_record\_31-85','mths\_since\_last\_record>85']]

## Preprocesamiento de Variables categóricas

### Peso de la Evidencia (WoE)

El WoE es una medida de la fuerza de una variable independiente (o de una categoría/intervalo de esa variable) para discriminar entre dos grupos en la variable dependiente (objetivo). En el contexto de riesgo, estos grupos son compuestos "Buenos" (clientes que no default / incumplieron) y "Malos" (clientes que sí default / incumplieron).

Ejemplo de preprocesamiento para variables categóricas

Para esto, comenzaremos estimando el peso de la evidencia (PDE) PdE = ln(%Buenos/%Malos)

El peso de la evidencia indica el poder predictivo de una variable independiente (home\_ownership) en relación con la variable dependiente (buenos\_malos).

PdE se describe como una medida de la separación de buenos y malos clientes. "Clientes malos" se refiere a los clientes que incumplieron con un préstamo y "Buenos clientes" se refiere a los clientes que pagaron el préstamo.

¿Para que hacemos esto? Para combinar categorías

Es porque las categorías con PdE similar tienen casi la misma proporción

de buenos y malos. En otras palabras, el comportamiento de ambas categorías es el mismo.

Recordemos la interpretación del WoE:

    WoE positivo: Mayor proporción de "Buenos" (menor probabilidad de default).

    WoE negativo: Mayor proporción de "Malos" (mayor probabilidad de default).

    Valor absoluto más grande: Mayor poder discriminatorio.

Ejemplo de la variable “home\_ownership” archivo ipynb



Se puede analizar que los clientes con mayor probabilidad de default seria de la siguiente forma:

Clientes con MAYOR Probabilidad de Default (Alto Riesgo):

OTHER (PdE: -0.927): La categoría de mayor riesgo.

NONE (PdE: -0.601): La segunda de mayor riesgo.

Ambas con muy pocas observaciones, por lo que su confiabilidad es baja.

RENT (PdE: -0.0247): Riesgo por encima del promedio, pero moderado. (Más del 40% de las observaciones, fiable).

Clientes con MENOR Probabilidad de Default (Bajo Riesgo):

MORTGAGE (PdE: 0.1467): La categoría de mayor volumen (más del 50% de las observaciones) y con menor riesgo.

OWN (PdE: 0.0318): Riesgo ligeramente por debajo del promedio, con poca capacidad discriminatoria.

Categorías con Datos Incompletos/Problemáticos:

ANY (PdE: inf): Riesgo cero, pero basada en solo 1 observación, lo que la hace no fiable.

Los clientes con hipoteca (MORTGAGE) o propiedad total (OWN) muestran un riesgo de default inferior al promedio. Por el contrario, aquellos en categorías "OTHER", "NONE" o "RENT" presentan un riesgo mayor, siendo las dos primeras las de más alto riesgo, aunque sus pequeñas poblaciones limitan la fiabilidad de esa conclusión.

### Valor de la Información (VI)

El análisis del valor de la información es una técnica de exploración de datos que ayuda a determinar qué variables de un conjunto de datos tienen poder \*predictivo o influencia en el valor de una variable dependiente específica\*.

Evaluando el peso de las categorías dentro de la variable sino de la variable en sí VI = Sumatoria (PdE\*(proporción de los buenos en la categoría - proporción de los malos en la categoría))

Es una métrica clave para la selección de variables en el modelado de riesgo.

interpretación del IV:

    IV < 0.02: La variable tiene poco o ningún poder predictivo (inútil).

    0.02 <= IV < 0.1: La variable es un predictor débil.

    0.1 <= IV < 0.3: La variable es un predictor medio (útil).

    0.3 <= IV < 0.5: La variable es un predictor fuerte.

    IV >= 0.5: La variable podría ser demasiado fuerte y podría indicar multicolinealidad, sobreajuste, o "fuga de datos" (data leakage) del objetivo al predictor. Se debe investigar a fondo.

Tabla

El contenido generado por IA puede ser incorrecto.

Esta columna "VI" te dice cuánto "poder predictivo" aporta cada categoría específica de home\_ownership a la capacidad general de la variable para discriminar entre clientes "buenos" y "malos".

Un valor VI más alto (en magnitud) para una categoría significa que esa categoría contribuye más a la capacidad de la variable home\_ownership para predecir el default.

**División del conjunto de datos:** finalmente, después de todas las transformaciones, el conjunto de datos preprocesado se dividió cuidadosamente en dos partes principales: un "conjunto de entrenamiento” con una representación del 80% del dataset original y un "conjunto de evaluación" del 20%. El conjunto de entrenamiento se usa para que los modelos aprendan de los patrones en los datos, mientras que el conjunto de evaluación se mantiene completamente separado y se usa para probar qué tan bien funciona el modelo de regresión logística y arboles de decisión en datos que nunca antes ha visto. Esta división es crucial para asegurar que la evaluación del modelo sea imparcial y refleje su rendimiento real en el mundo exterior.

## Análisis descriptivo

Antes de construir cualquier modelo, realizamos un análisis descriptivo exhaustivo de los datos. Este paso es como hacer una "radiografía" del conjunto de datos. Implicó entender la distribución de cada variable

(por ejemplo, ¿la mayoría de los préstamos son grandes o pequeños?),

identificar cuántos valores faltantes había en cada columna y cómo manejarlos, y detectar los ya mencionados valores atípicos. Además, este análisis nos permitió obtener una comprensión general de las características de los prestatarios y los préstamos.

Exploramos visualmente y estadísticamente las relaciones entre las diferentes variables y, lo más importante, cómo estas se conectaban con nuestra variable objetivo: si el préstamo había caído en impago o no.

Esto nos dio una base sólida para entender el problema antes de aplicar cualquier algoritmo predictivo.

Las siguientes graficas muestran el peso de la evidencia por cada variable y cuál es la contribución de cada característica dentro de cada variable. Adicional se describirá el razonamiento realizado.

Interfaz de usuario gráfica, Gráfico, Gráfico de líneas

El contenido generado por IA puede ser incorrecto.

La de mayor poder predicción para nuestra variable independiente (Clientes buenos) la cual sería MORTAGE (Hipoteca).

Gráfico, Gráfico de líneas

El contenido generado por IA puede ser incorrecto.

La de mayor poder predicción para nuestra variable independiente (Clientes buenos) la cual sería DC (hace referencia a Washintong)

Gráfico, Gráfico de líneas

El contenido generado por IA puede ser incorrecto.

Es decir, entre más alto los ingresos, mayor es el peso de la evidencia En los casos en los que hay crecimientos o decrecimientos monotónicos, es decir en línea recta como el que observamos, podemos de forma segura dividir la variable en intervalos de tamaños iguales, teniendo en cuenta por supuesto, el peso de las

observaciones. Por ejemplo, podríamos dividir los ingresos desde 0 a 152,000 en intervalos de $10,000. Pero los primeros dos intervalos tendrían muy pocas observaciones. Entonces, podríamos establecer el primer intervalo de 0 a 20K o en otras palabras ingresos menores a $20K. Siguiendo el mismo razonamiento, los últimos cuatro intervalos tendrían muy pocas observaciones. Por lo tanto, podríamos establecer dos intervalos de $26K es decir de $100K a $126K sería uno y mayor a $125K a $152K sería el otro. Y los intervalos de en medio, que son los que tienen el mayor número de observaciones, a ese si los agrupáramos en intervalos de $10K

Gráfico, Gráfico de líneas

El contenido generado por IA puede ser incorrecto.

Basado en el peso de la evidencia:

0 va sólo

1-4 tienen un PdE similar y los combinaré

5-6 van en el mismo grupo

7 va sola

8-9 van en el mismo grupo

10 tiene un PdE muy superior a las demás.

Observaciones

1/3 del total, por lo tanto, tiene que ir sola

Para juntar las categorías, vamos a utilizar el método

value.isin(list) que evalúa si un valor está dentro de una lista en particular

Devuelve un verdadero o falso

Gráfico, Gráfico de líneas

El contenido generado por IA puede ser incorrecto.

La que tiene menor peso de la evidencia es G y la que tiene mayor peso de la evidencia es A. La mayoría de las calificaciones tienen un peso relevante en el número de observaciones Con excepción de las categorías G y F. Ambas tienen PdE similar y son candidatas para unir en una sola categoría

Gráfico

El contenido generado por IA puede ser incorrecto.

Generemos grupos pequeños con los intervalos que tienen un peso de observaciones significativo y luego vamos generando grupos más grandes a medida que los pesos de observaciones de vuelven irrelevantes Primer grupo: los primero cuatro intervalos del gráfico (hasta 97.63)

Gráfico, Gráfico de líneas

El contenido generado por IA puede ser incorrecto.

Del número de observaciones identificamos dos categorías muy grandes:

Consolidación de Deuda y Tarjeta de crédito, que deberían ir solos, los otros dos grupos relevantes (5%) que podrían ir solos o ser el centro de un clúster dependiendo

Otros y Mejoras al Hogar Con base al PdE y el peso de observaciones mis grupos serían los siguientes:

Educación-Pequeña Empresa-Energía Renovable-Mudanzas

Casa-Otros-Boda-Salud-Vacacion

Consolidación de Deuda

Compras Grandes - Mejoras - Carro

Tarjeta de Crédito

Gráfico, Gráfico de líneas

El contenido generado por IA puede ser incorrecto.

Observamos que la categoría de 36 tiene un PdE mucho mayor que el de 60. Pero notamos que nuestro gráfico ya no está ordenado de forma ascedente porque utilizamos la función PdE\_continua.

Gráfico, Gráfico de líneas

El contenido generado por IA puede ser incorrecto.

'verification\_status': Se recomienda no agrupar las categorías: Sólo son tres y la proporción de observaciones es significativa para las tres categorías (>30%)

## Selección de variables

La fase de selección de variables es muy importante para asegurar que los modelos se construyan con la información más relevante y útil, evitando el "ruido" que podría haber en los datos. Para la **Regresión Logística** , que es un modelo más transparente, utilizamos técnicas que nos permitieron evaluar la "significancia estadística" de cada variable. Esto significa que analizamos si una variable realmente tenía un impacto considerable en la probabilidad de incumplimiento. Por ejemplo, si usamos la clase RegresionLogistica\_con\_p\_values, esto implicaría observar los "p-valores" de cada característica: si un p-valor es bajo, la variable se considera significativa y se mantiene en el modelo.

Se utilizo una regresión logística que estima los p\_values multivariable de forma directa. La investigación se realizo en el siguiente github:

https://gist.github.com/rspeare/77061e6e317896be29c6de9a85db301d

Información de la regresión logistica

https://scikit-learn.org/stable/modules/linear\_model.html#logistic-regression

Para el **Random Forest** , el algoritmo tiene una ventaja intrínseca: al construir sus múltiples árboles de decisión, ya tiende a dar más peso a las características más importantes. Sin embargo, incluso con Random Forest, podemos analizar la "importancia de las características" que el modelo nos da al final para entender cuáles fueron las más influyentes en sus predicciones. En este proyecto, se seleccionará cuidadosamente un conjunto de claves variables para ser incluidas en la construcción de ambos modelos predictivos.

## Selección de Modelos

Para abordar el problema de predecir la probabilidad de incumplimiento, elegimos y entrenamos dos tipos de modelos de clasificación ampliamente reconocidos y utilizados en la industria financiera: la Regresión Logística y el Random Forest.

### Regresión Logística:

Este es un modelo estadístico que, a pesar de su nombre, se usa para problemas de clasificación, es decir, para predecir si algo pertenece a una categoría u otra (en nuestro caso, si hay incumplimiento o no). Es muy popular en la predicción de PD debido a su facilidad para interpretar sus resultados: podemos ver directamente la dirección y la fuerza de la relación entre cada característica y la probabilidad de incumplimiento. Entrenamos un modelo de Regresión Logística utilizando las características más relevantes de nuestro conjunto de entrenamiento y luego lo usamos para predecir las probabilidades de incumplimiento en el conjunto de datos de evaluación (los datos que el modelo nunca había visto). Al evaluar su rendimiento, el AUC (Área bajo la Curva ROC) para la Regresión Logística fue de 0.6717. Este valor, aunque aceptable, indica que el modelo tiene una capacidad predictiva moderada.

### Random Forest:

El Random Forest es un algoritmo de inteligencia artificial más avanzado, que pertenece a la categoría de "modelos de conjunto" o "ensemble". Funciona creando un gran número de "árboles de decisión" individuales, y luego combina las predicciones de todos esos árboles para llegar a una predicción final más robusta y precisa. Es muy apreciado por su alta capacidad de predicción y su habilidad para manejar relaciones complejas entre las variables que un modelo lineal como la Regresión Logística no podría captar. Además, es robusto ante datos ruidosos o atípicos. Para asegurarnos de que el Random Forest funcione lo mejor posible, utilizamos una técnica de optimización llamada RandomizedSearchCV. Este método nos permite probar eficientemente diferentes combinaciones de configuraciones internas del modelo (conocidas como "hiperparámetros") de forma aleatoria. Esto es mucho más rápido y computacionalmente eficiente que probar todas las combinaciones posibles.

Las configuraciones o "hiperparámetros" que exploramos para el Random Forest incluyen:

* n\_estimators: La cantidad de árboles de decisión que el algoritmo construye.
* max\_features: El número de características que el modelo considera aleatoriamente en cada división de un árbol.
* max\_depth: La profundidad máxima que cada árbol puede alcanzar, para evitar que se vuelvan demasiado complejos y memorizar los datos de entrenamiento.
* min\_samples\_split: El número mínimo de muestras que debe haber en un nodo para que se pueda dividir.
* min\_samples\_leaf: El número mínimo de muestras que debe haber en las "hojas" o nodos finales de los árboles.

Mejores hiperparámetros para Random Forest (obtenidos con RandomizedSearchCV):

'n\_estimators': 100, 'min\_samples\_split': 2, 'min\_samples\_leaf': 2, 'max\_features': 'sqrt', 'max\_depth': 10, 'bootstrap': True

La métrica que usamos para guiar esta búsqueda de la mejor configuración fue el AUC. El mejor modelo de Random Forest que encontramos a través de RandomizedSearchCV fue el que usamos para hacer predicciones en el conjunto de evaluación. El **AUC** para este Random Forest optimizado fue de **0.6804**. Este valor es notablemente superior al de la Regresión Logística, lo que indica que el Random Forest tiene un poder predictivo significativamente mejor para este problema.

Comparación de Modelos y Métricas de Evaluación:

Para tener una visión completa del rendimiento de ambos modelos, calculamos y comparamos varias métricas clave:

Tabla

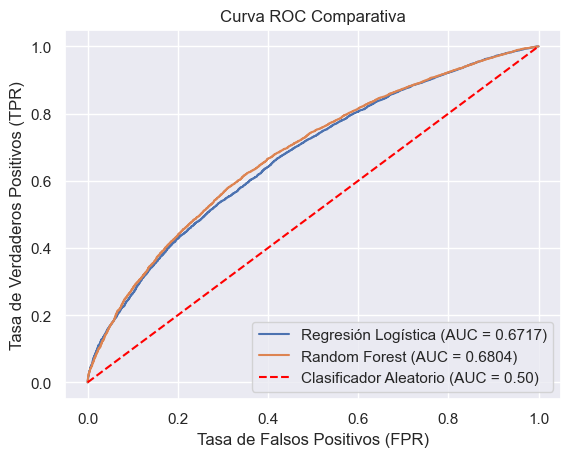
El contenido generado por IA puede ser incorrecto.

El Random Forest (Optimizado) emerge como el modelo superior para este problema de clasificación de riesgo de incumplimiento. Aunque la Regresión Logística y el Random Forest (Optimizado) muestran métricas de AUC-ROC muy similares (0.6717 vs. 0.6804, respectivamente), el Random Forest exhibe una capacidad ligeramente mejor para discriminar entre las clases. Más críticamente, el Random Forest logra un Recall (Sensibilidad) notablemente más alto (0.6936) en comparación con la Regresión Logística (0.6134). Este aumento del Recall es crucial en la detección de incumplimiento, ya que significa que el Random Forest es capaz de identificar una mayor proporción de los clientes que realmente van a incumplir, lo cual es vital para mitigar riesgos en un contexto financiero. Además, el Random Forest mantiene una precisión muy alta (0.9389), asegurando que las predicciones de incumplimiento sean fiables, y presenta un F1-Score superior (0.7978), indicando un mejor equilibrio general entre precisión y recall. Finalmente, su Log Loss significativamente menor (0.6009 vs. 13.8968 para la Regresión Logística) sugiere que las probabilidades predichas por el Random Forest son más precisas y confiables.

### Curva ROC Comparativa:

La Curva ROC (que significa "Receiver Operating Characteristic") es un gráfico muy útil que nos permite visualizar el desempeño de un modelo de clasificación a medida que cambiamos el umbral que usamos para decidir si una probabilidad predicha se convierte en un "sí" (incumplimiento) o un "no" (no incumplimiento). Un área bajo esta curva (AUC) más grande indica que el modelo es mejor discriminando entre las dos clases.

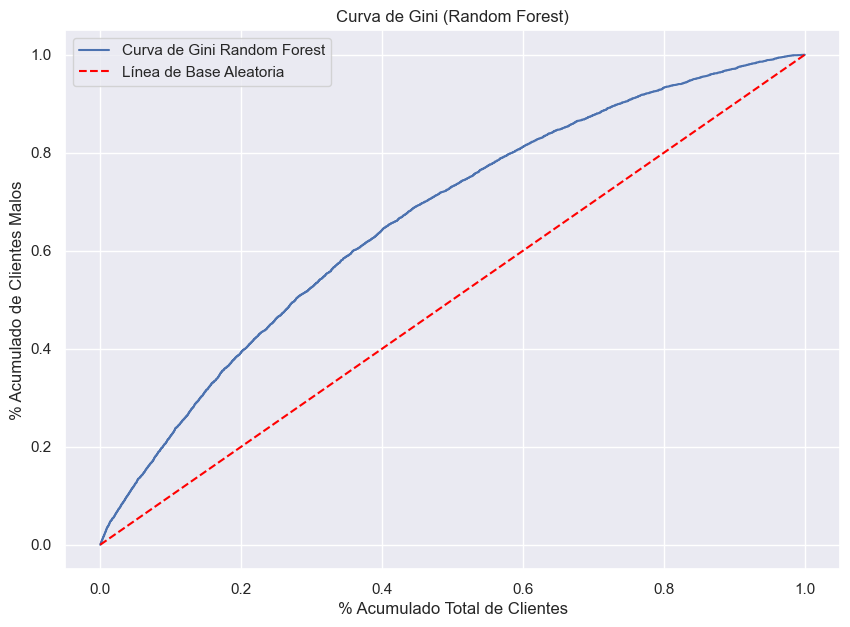
El gráfico muestra de manera clara y visual que la curva ROC del Random Forest se eleva más rápidamente y se mantiene más cerca de la esquina superior izquierda del gráfico. Esto confirma su mayor valor AUC y, por lo tanto, su superior capacidad para distinguir eficazmente entre los casos de no-incumplimiento y los de incumplimiento.



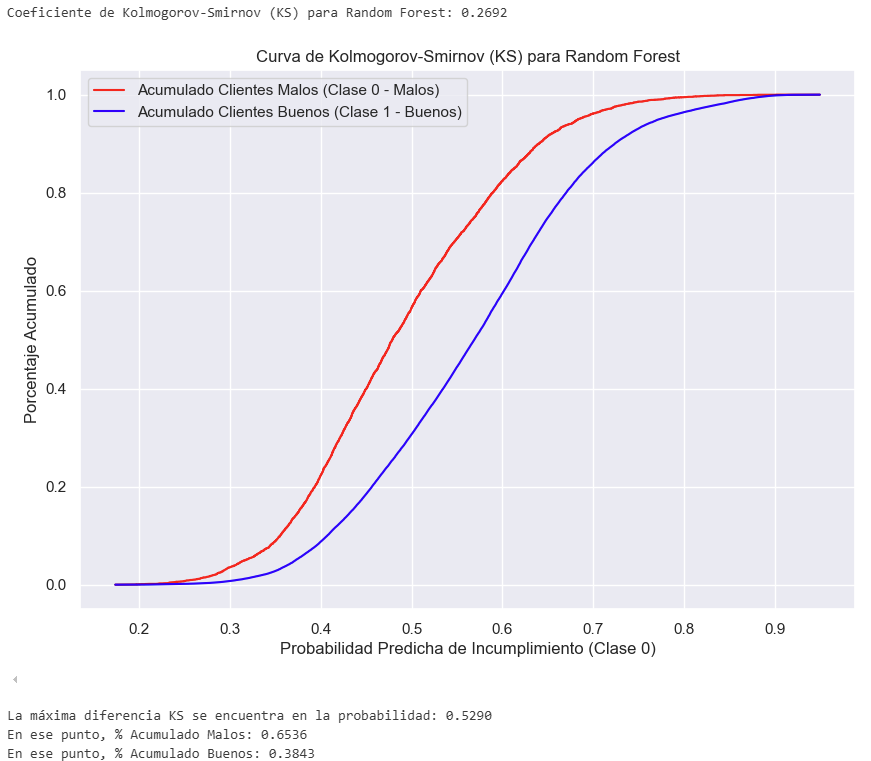
### Coeficiente de Gini

El Coeficiente de Gini es otra medida importante que se deriva directamente del AUC y nos da una idea de la capacidad de un modelo para discriminar entre diferentes grupos. Un valor de Gini de 1 indicaría un modelo perfecto (capaz de separar completamente a los que incumplen de los que no), mientras que un valor de 0 significaría que el modelo es tan bueno como adivinar al azar. Para nuestro modelo de Random Forest, el Coeficiente de Gini fue de 0.3608 (este valor se calcula como AUC \* 2 - 1). Este resultado es bastante bueno, ya que está significativamente por encima de cero, lo que sugiere que el modelo tiene una capacidad de separación considerable.

El gráfico de la curva de Gini visualiza cómo el modelo es capaz de concentrar a los "malos clientes" (aquellos que eventualmente caen en default) a medida que ordenamos a los clientes por su probabilidad de incumplimiento. Una curva de Gini que se aleja notablemente de la línea de base (la línea diagonal roja) indica que el modelo tiene un buen poder para separar y clasificar, aunque podría mejorar aplicando técnicas de modelado automatizado.



### Kolmogorov Smirnov



El Coeficiente de Kolmogorov-Smirnov (KS) para el modelo Random Forest es **0.2692**. Este valor, que mide la máxima separación vertical entre la distribución acumulada de clientes "malos" (incumplidores, Clase 0) y clientes "buenos" (no incumplidores, Clase 1), indica que el modelo posee una **capacidad de discriminación aceptable a buena**. Un KS superior a 0.20 generalmente se considera útil, y un valor cercano a 0.30 es una buena señal de poder predictivo. Para poder distinguir entre clientes buenos y clientes malos.

# Conclusiones

A lo largo de este proyecto, hemos demostrado con éxito cómo se puede construir y evaluar modelos de Probabilidad de Incumplimiento utilizando dos técnicas potentes: la Regresión Logística y el Random Forest. Los resultados han sido claros: el modelo de **Random Forest** , al que aplicamos un proceso de optimización mediante RandomizedSearchCV, mostró un rendimiento superior en clave métrica como el AUC-ROC, el Error Cuadrático Medio (MSE) y el Log Loss, en comparación con la Regresión Logística.

Esto nos confirma que el Random Forest es una herramienta muy eficaz y útil para identificar a aquellos prestatarios que presentan un mayor riesgo de incumplimiento. Si bien la Regresión Logística ofrece la ventaja de ser más fácil de entender en cuanto a cómo cada factor influye en el resultado, el poder predictivo superior del Random Forest lo convierte en la opción más adecuada para escenarios donde la precisión de la predicción es lo más importante.

La creación de estos modelos y la evaluación detallada de sus métricas no son solo ejercicios académicos; son pasos fundamentales y esenciales para cumplir con las expectativas y regulaciones de organismos tan importantes como la SBP y los Acuerdos de Basilea. Estas regulaciones exigen una gestión del riesgo de crédito que no solo sea robusta, sino también cuantificable y basada en datos sólidos.

# Recomendaciones y Futuros Estudios

Para seguir mejorando el modelo de Probabilidad de Incumplimiento y su aplicación práctica, se sugieren los siguientes pasos y áreas para futuras investigaciones:

**Validación externa y pruebas de estrés:** Es crucial que el modelo sea validado por un equipo o datos independientes para asegurar que mantiene su estabilidad y confiabilidad en diferentes situaciones. Además, se deben realizar "pruebas de estrés" para evaluar cómo se comportaría el modelo en escenarios económicos desfavorables o de crisis, siguiendo las directrices establecidas por Basilea III.

**Inclusión de Variables Macroeconómicas:** Una línea de investigación muy prometedora es incorporar al modelo variables relacionadas con la economía a nivel general, como las tasas de interés, los niveles de desempleo o el Producto Interno Bruto (PIB). Esto podría mejorar significativamente la capacidad del modelo para predecir incumplimientos en diferentes ciclos económicos. Informes de empresas como Deloitte también resaltan la importancia de integrar un enfoque macroeconómico en los modelos de PD para hacerlos más resilientes.

**Modelos de Inteligencia Artificial Más Avanzados:** Se podría explorar la aplicación de otros algoritmos de Machine Learning aún más complejos y potentes, como los modelos de Gradient Boosting (ej. XGBoost, LightGBM) o las redes neuronales profundas. Sin embargo, al considerar estos modelos, siempre se debe buscar un equilibrio cuidadoso entre la alta precisión que pueden ofrecer y la dificultad para entender cómo llegan a sus predicciones (su "interpretabilidad"). Como señala Deloitte, los modelos más complejos requieren de expertos para su desarrollo, una documentación muy detallada y procesos de validación rigurosos, especialmente en lo que respeta a explicar sus decisiones.

**Análisis de Sensibilidad y Explicabilidad (XAI):** Para modelos que son inherentemente más difíciles de interpretar, como el Random Forest, es fundamental invertir en técnicas de "Explicabilidad de la Inteligencia Artificial" (XAI, por sus siglas en inglés). Herramientas como los valores SHAP (SHapley Additive exPlanations) o LIME (Local Interpretable Model-Agnostic Explanations) pueden ayudar a entender cómo cada característica contribuye a una predicción individual. Esto es vital no solo para cumplir con las expectativas de los reguladores, sino también para que los equipos de negocio puedan confiar y usar el modelo en sus decisiones diarias.

**Optimización del Umbral de Corte:** El umbral de corte es el punto a partir del cual una probabilidad predicha se convierte en una clasificación de "incumplimiento". Es importante realizar un análisis más profundo para determinar cuál es el umbral óptimo. Esta decisión debe considerar las implicaciones de negocio: por ejemplo, el costo de clasificar erróneamente a un buen cliente como de riesgo (falso positivo) frente al costo de no detectar a un cliente que realmente va a incumplir (falso negativo).