Imagen que contiene reloj, señal

Descripción generada automáticamenteA logo with text and symbols

Description automatically generated**Universidad Tecnológica De Panamá**

**Facultad De Ingeniería Industrial**

**Campus Víctor Levi Sasso**

**Licenciatura En Ingeniería Industrial**

**Materia:**

**Modelos predictivos 1AN214**

**Profesor:**

**Juan Marcos Castillo Guerra, PhD**

**Tema:**

**Aplicación de Modelo de regresión logística en Python predecir la probabilidad de impago (default)**

**Presentado Por:**

**Ramsés Ariel Mong Rojas**

**8-919-1737**

**Panamá, 2025**

Contenido

[**Introducción** 2](#_Toc195217627)

[Motivación 3](#_Toc195217628)

[**Metodología** 4](#_Toc195217629)

[1. Comprensión del Negocio 5](#_Toc195217630)

[2. Comprensión de los Datos 5](#_Toc195217631)

[3. Preparación de los Datos 9](#_Toc195217632)

[4. Modelado Predictivo 11](#_Toc195217633)

[5. Evaluación del Modelo 15](#_Toc195217634)

[6. Implementación 19](#_Toc195217635)

[Conclusiones 20](#_Toc195217636)

[Recomendaciones 20](#_Toc195217637)

[Retos y acciones a futuro 21](#_Toc195217638)

[Bibliografía 21](#_Toc195217639)

## 

## **Introducción**

El presente estudio se enfoca en aplicar un modelo predictivo para anticipar el incumplimiento de préstamos personales (default) utilizando algoritmos de clasificación supervisada, específicamente regresión logística. Este tipo de análisis permite a las instituciones financieras tomar decisiones informadas sobre el otorgamiento de crédito, reduciendo riesgos financieros y mejorando la rentabilidad. Para este proyecto se siguió la metodología CRISP-DM (Cross-Industry Standard Process for Data Mining) es un enfoque estructurado para desarrollar proyectos de minería de datos. Se compone de seis fases principales: comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación e implementación.

El conjunto de datos seleccionado proviene de la plataforma Kaggle y contiene información sobre más de 70,000 préstamos otorgados. Entre las variables disponibles se incluyen ingresos, edad, historial crediticio, propósito del préstamo y estado laboral del solicitante.

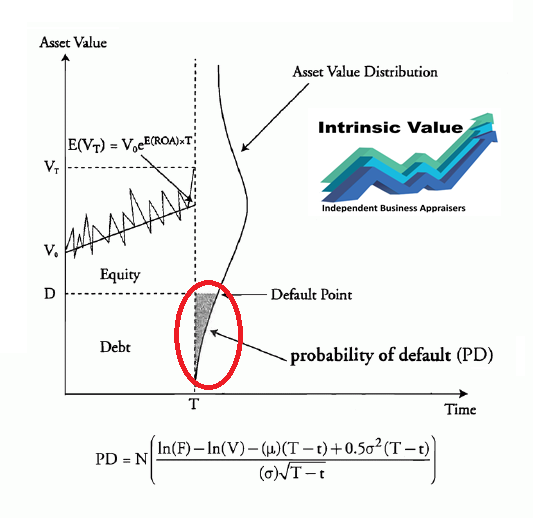


Figura 1: Este gráfico representa visualmente el modelo de **Merton** para estimar la **probabilidad de incumplimiento (PD, por sus siglas en inglés)**

## Motivación

Elegí trabajar en la predicción del incumplimiento de préstamos personales porque actualmente formo parte del equipo de **riesgos en banca de consumo**. En mi rol diario, tengo acceso a información con características similares a las presentes en este conjunto de datos, como ingresos, historial crediticio, antigüedad laboral y monto del préstamo.

Al seleccionar este tema, busqué conectar el contenido académico de la maestría con **casos reales y aplicables a mi entorno profesional**. Esta experiencia me ha permitido profundizar en técnicas analíticas modernas y evaluar su potencial de implementación en el análisis de riesgo crediticio que enfrentamos en la institución donde laboro. Además, al trabajar con un conjunto de datos comparable, puedo contrastar la efectividad de los modelos predictivos y generar ideas para optimizar nuestras estrategias de aprobación y monitoreo de riesgo.

## **Metodología**

**3.1 Proceso CRISP-DM**

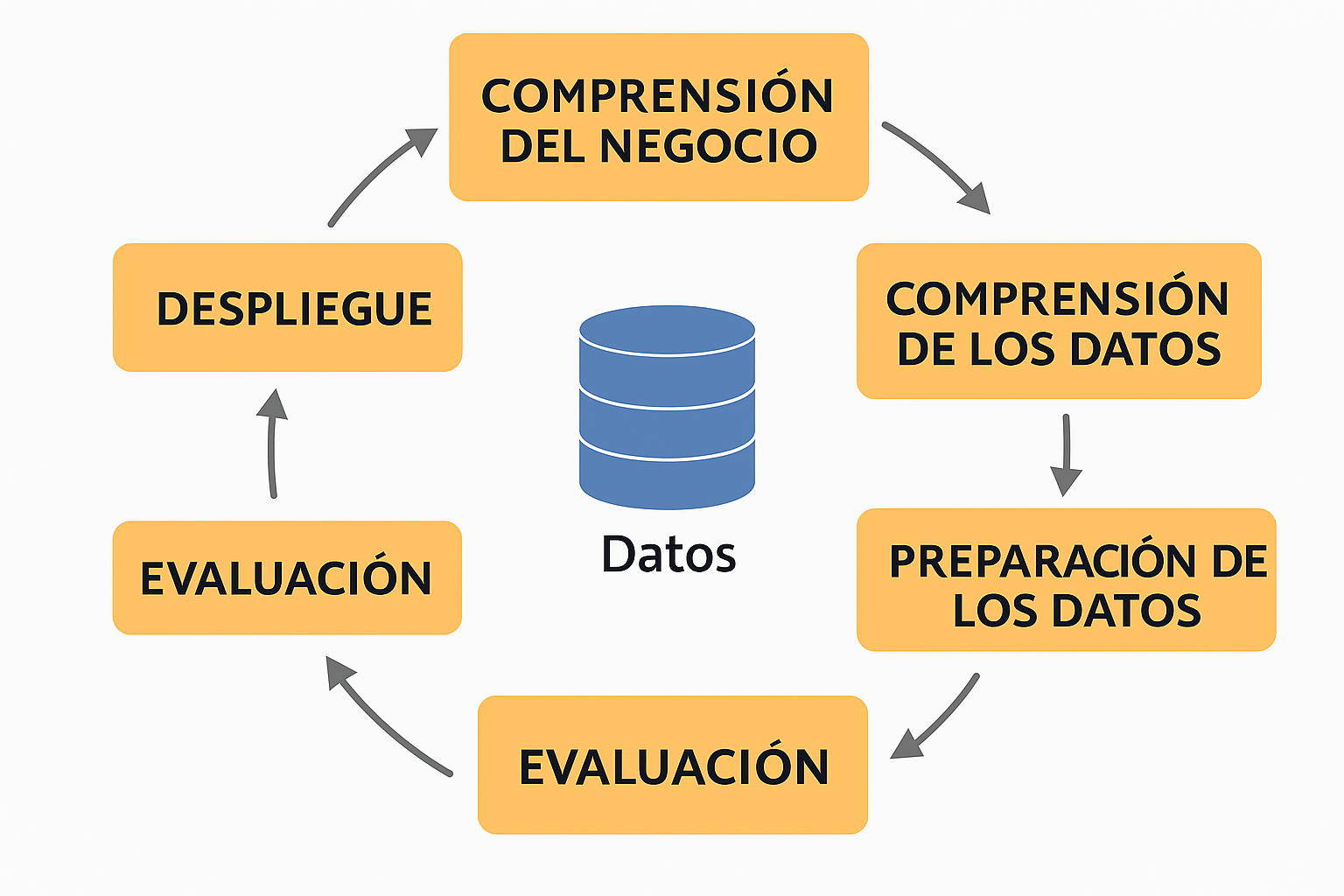


Figura 2: Los 6 pasos de la metodología CRISP-DM.

La metodología CRISP-DM (Cross-Industry Standard Process for Data Mining) es un enfoque estructurado y ampliamente adoptado para desarrollar proyectos de minería de datos. Se compone de seis fases principales: comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación e implementación. Este proceso iterativo permite a los analistas y científicos de datos abordar problemas complejos al alinear los objetivos técnicos con los objetivos del negocio, asegurando así que los resultados generen valor real. La flexibilidad del modelo facilita su adaptación a diversos dominios y tipos de proyectos, desde análisis predictivos hasta segmentaciones de mercado (Chapman et al., 2000).

Durante su ejecución, CRISP-DM promueve una retroalimentación constante entre las etapas, permitiendo ajustar y refinar tanto los datos como los modelos desarrollados. Por ejemplo, si en la fase de evaluación se detecta un bajo rendimiento del modelo, puede ser necesario regresar a la preparación de datos o reconsiderar la selección de variables. Esta naturaleza cíclica distingue a CRISP-DM como una metodología práctica y orientada a la mejora continua. Su popularidad en la industria se debe a que proporciona una guía clara y adaptable que ayuda a mitigar riesgos comunes en proyectos de ciencia de datos (Wirth & Hipp, 2000).

siguió el marco CRISP-DM, adaptado al análisis:

1. Comprensión del Negocio

En el área financiera, la probabilidad de incumplimiento (o default) se refiere a la posibilidad de que una persona que recibió un préstamo no lo pague según lo acordado. Esta medida es muy importante para los bancos y otras instituciones, ya que les ayuda a calcular el riesgo que representa cada cliente. Si una persona no paga su préstamo, la entidad financiera debe activar procesos para intentar recuperar ese dinero, lo que genera costos adicionales. Por eso, cuando se estima que alguien tiene más riesgo de no pagar, normalmente se le cobra un interés más alto para compensar ese posible riesgo (Altman & Saunders, 1998).

Existen dos formas principales de estimar la probabilidad de incumplimiento. Una se basa en modelos estructurales, que usan información del mercado y de estados financieros, y suelen aplicarse a empresas o gobiernos. La otra, que es la que usamos en este proyecto, se basa en modelos empíricos, que analizan datos históricos del comportamiento de pago de personas. Esta segunda opción es la más utilizada en créditos personales y se relaciona con mi experiencia profesional (Merton, 1974; Thomas, Crook, & Edelman, 2017).

El conjunto de datos seleccinado incluye miles de préstamos con información como edad, ingresos, monto solicitado, historial crediticio, tiempo trabajando, número de créditos activos, tasa de interés, plazo del préstamo y relación deuda-ingreso. Se utilizó un modelo de regresión logística binaria, para predecir de que la persona vaya a incumplir o no.

1. Comprensión de los Datos

Exploración inicial ayudo a entender las variables dentro del data set.

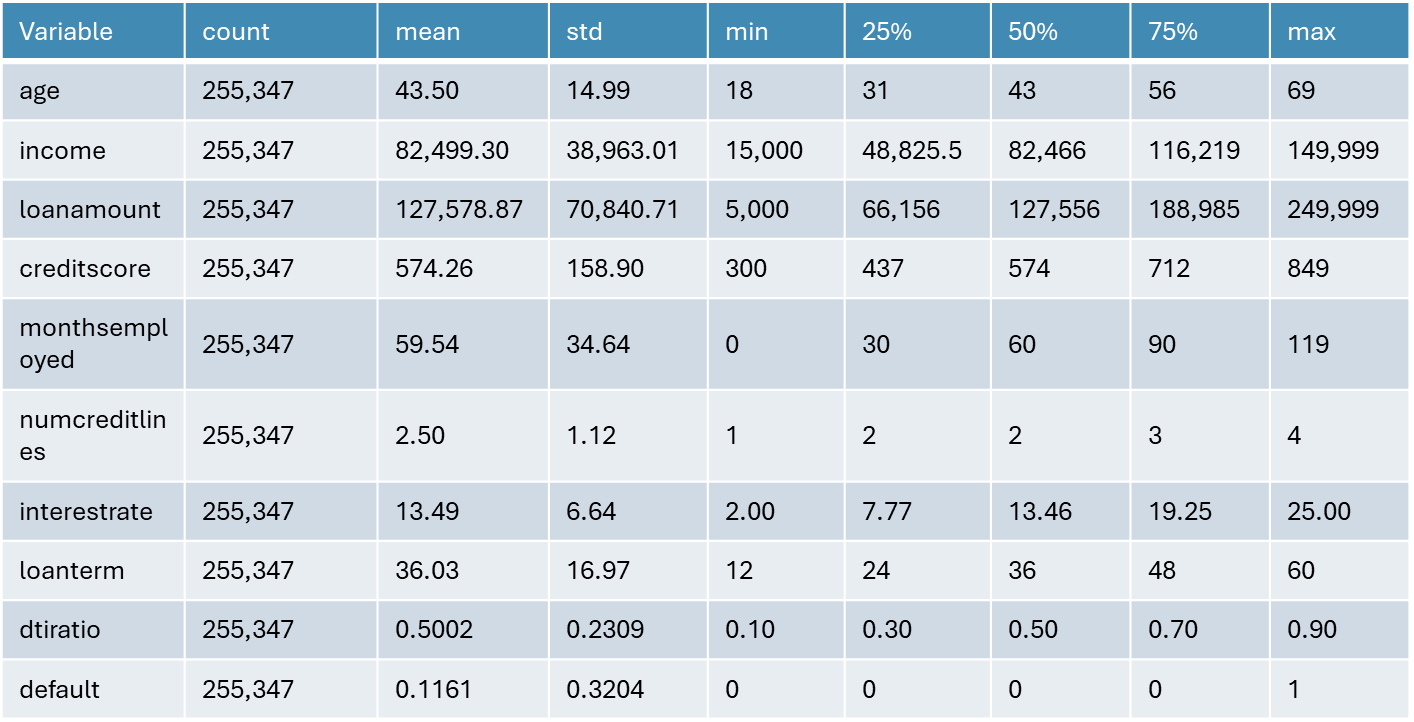


Tabla 1. Resumen estadístico de elaboración propia.

Al realizar el análisis descriptivo EDA se observó que la mayoría de las personas que solicitaron préstamos tienen alrededor de 43 años, lo que indica que son adultos en plena edad laboral. Sus ingresos varían bastante: algunos ganan muy poco (desde $15,000), mientras que otros ganan mucho más (hasta $149,000). Los montos de los préstamos también son variados, con un promedio de unos $127,000, pero pueden ir desde $5,000 hasta casi $250,000. Esto significa que hay clientes que piden préstamos pequeños y otros que solicitan montos grandes, lo que puede indicar distintos niveles de riesgo. El puntaje de crédito promedio está en un nivel medio (574), pero hay personas con mal historial (300) y otras con excelente puntaje (hasta 849), lo que muestra una gran diversidad entre los clientes.

También se observó que, en promedio, las personas han trabajado unos 5 años con su empleador actual, lo que indica que un número importante de clientes con estabilidad laboral. La tasa de interés promedio que se cobra es alta (13.5%) y en algunos casos puede llegar hasta el 25%, lo que podría aumentar el riesgo de que una persona no pueda pagar. La mayoría de los préstamos se otorgan a 3 años (36 meses), aunque hay desde 1 hasta 5 años. En cuanto a la relación entre deuda e ingresos (DTI), la mayoría está en un nivel razonable (0.5), pero hay casos más extremos donde la deuda casi iguala los ingresos, lo cual es más riesgoso. Finalmente, solo el 11.6% de los clientes no pagaron sus préstamos, lo que no es muy alto, pero sí lo suficiente como para analizarlo con más detalle y prevenir futuros casos.

**Visualización de datos**

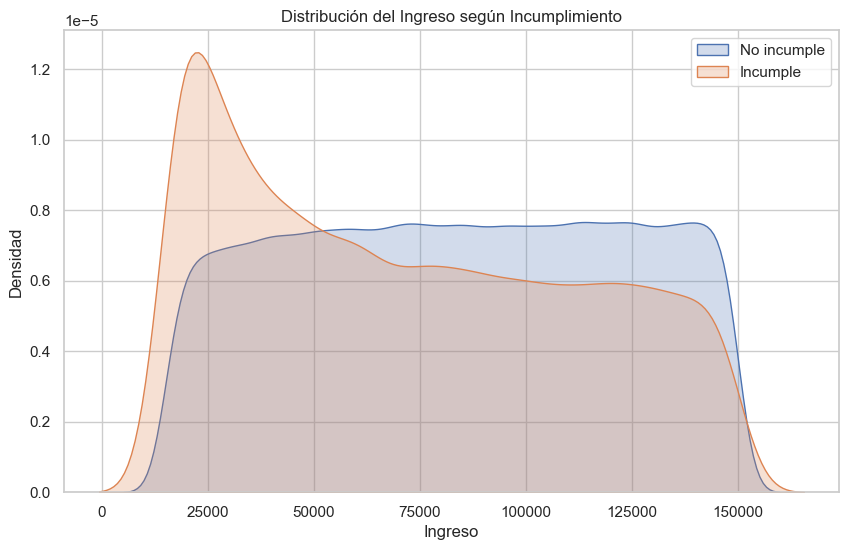


Figura 3. Gráfico vista ingresos de elaboración propia.

En esta gráfica se observa que los clientes que **incumplen** (línea naranja) tienden a concentrarse en los niveles **más bajos de ingreso**, especialmente por debajo de los $25,000.

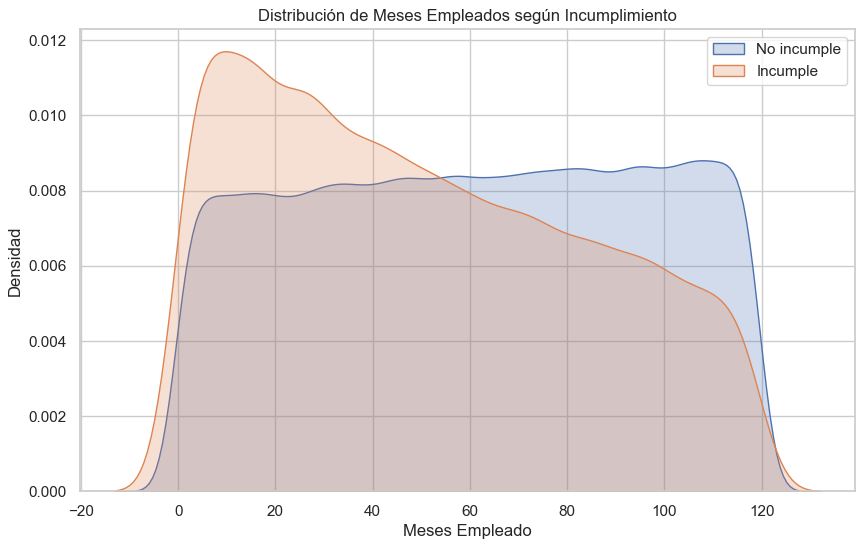


Figura 4. Gráfico vista estabilidad laboral de elaboración propia.

los clientes en **incumplimiento** suelen tener **menos tiempo empleados**. La mayor concentración de incumplidores se da en rangos por debajo de los 40 meses

**Resultados de la prueba de T**

Prueba t para age: estadístico=86.003, p-valor=0.000

Prueba t para income: estadístico=50.335, p-valor=0.000

Prueba t para loanamount: estadístico=-43.956, p-valor=0.000

Prueba t para creditscore: estadístico=17.275, p-valor=0.000

Prueba t para monthsemployed: estadístico=49.440, p-valor=0.000

Prueba t para numcreditlines: estadístico=-14.321, p-valor=0.000

Prueba t para interestrate: estadístico=-66.914, p-valor=0.000

Prueba t para loanterm: estadístico=-0.275, p-valor=0.783

Prueba t para dtiratio: estadístico=-9.722, p-valor=0.000

Los resultados de las pruebas t indican que la mayoría de las variables analizadas presentan diferencias significativas entre los clientes que incumplen y los que no, lo cual respalda su relevancia como posibles predictores en el modelo. Variables como la edad, ingreso, monto del préstamo, puntaje crediticio, antigüedad laboral, número de líneas de crédito, tasa de interés y relación deuda/ingreso (DTI) muestran p-valores muy bajos (menores a 0.05), lo que sugiere que existen diferencias reales en sus promedios entre ambos grupos. Esto es consistente con lo que se espera en modelos de riesgo crediticio, donde estos factores suelen tener fuerte influencia en la probabilidad de default. La única excepción es la variable **loanterm**, cuyo p-valor alto (0.783) indica que el plazo del préstamo no varía significativamente entre clientes cumplidores y morosos, por lo que podría no aportar valor al modelo predictivo.

**Sobremuestreo con SMOTE**

Se aplicó la técnica **SMOTE** (*Synthetic Minority Over-sampling Technique*). Esta herramienta permite equilibrar las clases en el conjunto de entrenamiento generando nuevos casos sintéticos de la clase minoritaria (clientes en default), en lugar de simplemente duplicar los existentes.

El funcionamiento de SMOTE se basa en identificar los vecinos más cercanos de cada observación de la clase minoritaria y, a partir de uno de ellos, crear ejemplos artificiales que mantienen la estructura del dato original, pero con ligeras variaciones. Esto ayuda al modelo a aprender mejor los patrones asociados al incumplimiento, sin caer en sobreajuste por replicar datos. Esta técnica fue implementada en Python como parte de la preparación de datos antes del entrenamiento del modelo predictivo.

### Preparación de los Datos

Se realizó comparativa entre los modelos de Regresión logística y Random Forest.

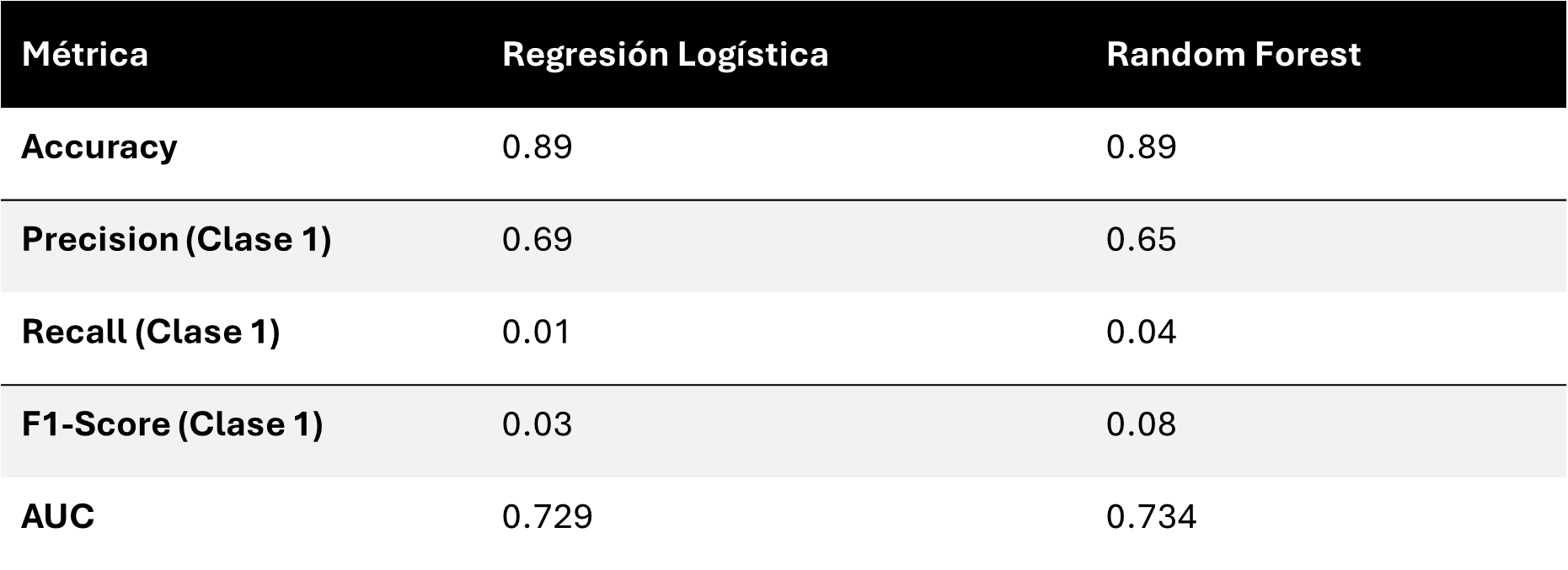


Tabla 2. Resumen estadístico de elaboración propia.

La tabla 2 presenta dos métricas clave para evaluar el desempeño en la predicción de la **clase 1 (clientes en incumplimiento)**: **precisión** y **F1-score**. La **precisión** de ambos modelos es similar: 0.69 para regresión logística y 0.65 para random forest. Esto indica que, cuando el modelo predice que alguien caerá en default, acierta aproximadamente el 65–69% de las veces.

Sin embargo, el **F1-score**, que equilibra precisión y recall, revela una diferencia importante: mientras que la regresión logística logra apenas 0.03, el modelo de random forest alcanza 0.08. Aunque ambos valores son bajos, esto refleja que ambos modelos tienen dificultades para detectar correctamente a los clientes que incumplen (bajo recall), pero random forest es ligeramente mejor en este aspecto.

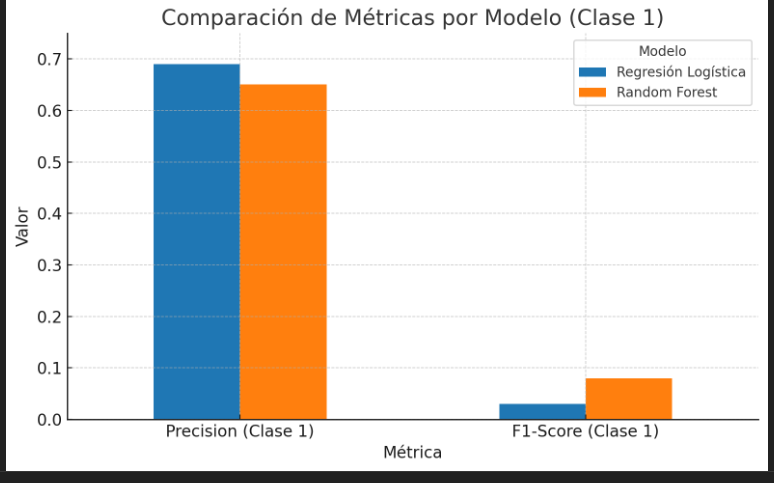


Figura 5. Comparación de métricas por modelo.

Al aplicar ambos modelos inicialmente ambos lograron una exactitud al 89%, porque la mayoría de los casos son "no-default". Ambos modelos mostraron gran dificultad para detectar correctamente a los incumplidores (bajo recall).

* + La regresión logística tiene ligeramente mejor precisión.
  + El Random Forest mejora el F1-score, indicando mejor balance entre precisión y recall.

Finalmente se seleccionó el modelo de regresión logística por su facilidad de interpretación con el entrenamiento de modelo se buscó compensar el balance para la población de bajo recall.

### Modelado Predictivo

**Implementando el modelo**

import statsmodels.api as sm

logit\_model=sm.Logit(y,X)

result=logit\_model.fit()

print(result.summary2())

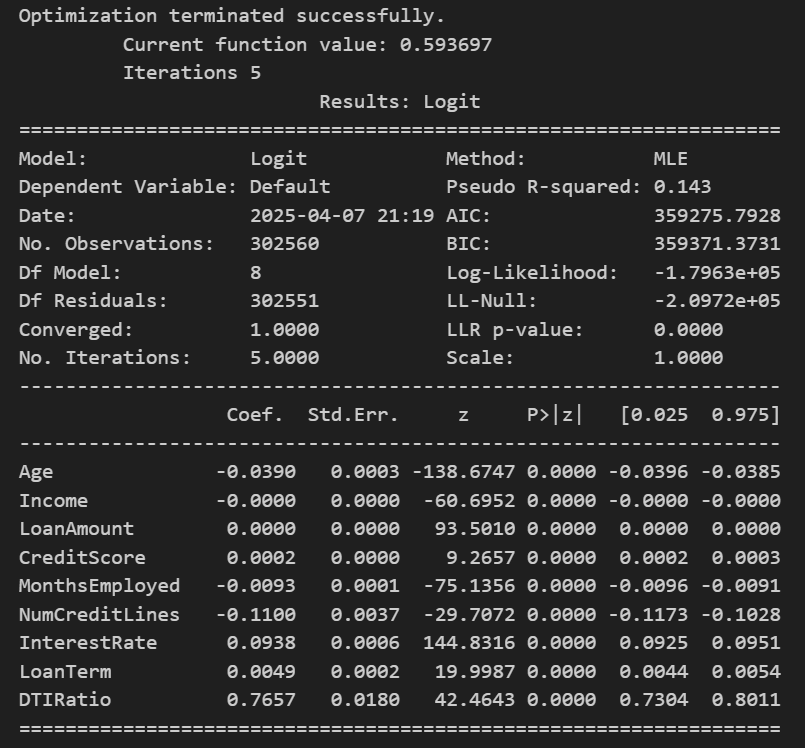


Figura 6. Los valores p de todas las variables son menores de 0.05.

# Extraer p-valores del resultado del modelo

pvalue = pd.DataFrame(result.pvalues, columns=['p\_value'])

pvalue

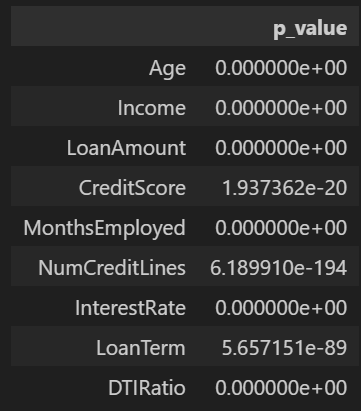


Figura 7. P-value de las variables de interés.

X = os\_data\_X[pvs]  # pvs: lista de variables seleccionadas previamente

y = os\_data\_y['Default']  # ✅ nombre correcto de la variable objetivo

import statsmodels.api as sm

logit\_model = sm.Logit(y, X)

result = logit\_model.fit()

print(result.summary())

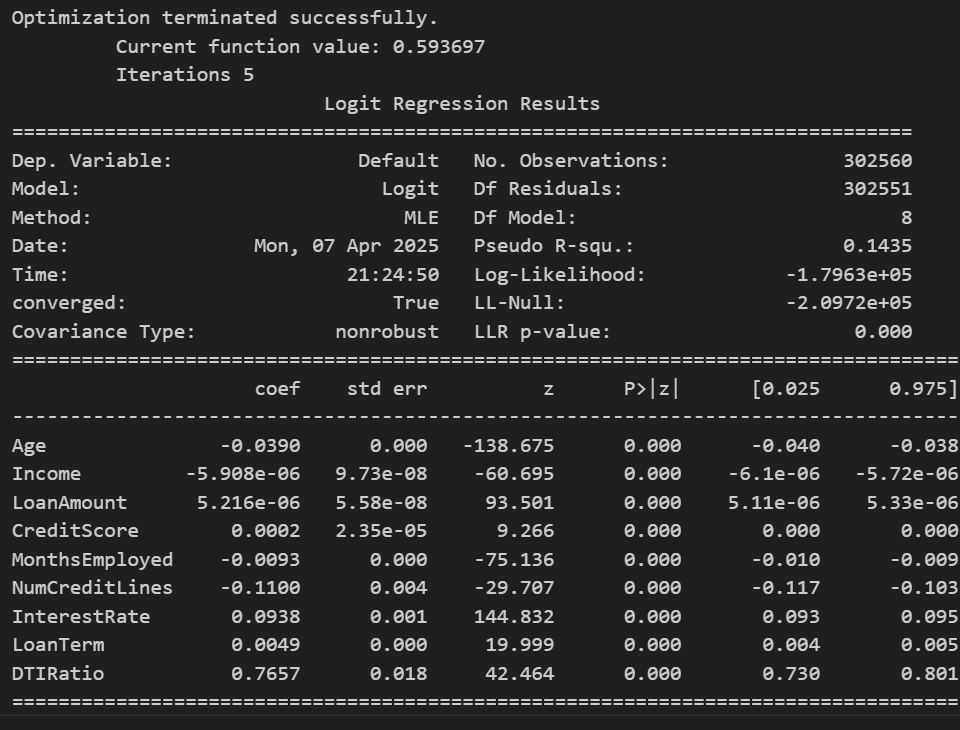


Figura 7. Optimización del modelo.

**Ajuste del modelo de regresión logística**

from sklearn.linear\_model import LogisticRegression

from sklearn import metrics

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.33, random\_state=42)

logreg = LogisticRegression()

logreg.fit(X\_train, y\_train)

import joblib

# Guardar el modelo entrenado

joblib.dump(logreg, "logistic\_model.pkl")

# Guarda el escalador

joblib.dump(scaler, "scaler.pkl") # O comenta esta línea si no estás usando uno

# Guarda los nombres de las variables

joblib.dump(X.columns.tolist(), "feature\_names.pkl")

1. Evaluación del Modelo

El modelo mostró resultados sólidos (ver sección de resultados).

**Evaluación**

**Predecir los resultados del conjunto de pruebas y calcular la precisión**

from sklearn.metrics import accuracy\_score

y\_pred = logreg.predict(X\_test)

print('Accuracy of logistic regression classifier on test set: {:.2f}'.format(accuracy\_score(y\_test, y\_pred)))

***Precisión del clasificador de regresión logística en el conjunto de prueba: 0.65***

**Matriz de confusión**

from sklearn.metrics import confusion\_matrix

confusion\_matrix = confusion\_matrix(y\_test, y\_pred)

print(confusion\_matrix)

***[[33525 16583]***

***[18080 31657]]***

El resultado de de la matriz de confusión se puede interpretar de la siguiente manera.

Verdaderos Negativos (33,525): El modelo acertó al predecir que estos clientes no incumplían y efectivamente no incumplieron.

Verdaderos Positivos (31,657): El modelo acertó al predecir que estos clientes iban a incumplir y efectivamente incumplieron.

Falsos Positivos (16,583): El modelo se equivocó al predecir incumplimiento para clientes que sí pagaron su préstamo. Esto puede generar rechazos innecesarios o condiciones más duras para buenos clientes.

Falsos Negativos (18,080): El modelo no identificó correctamente a muchos clientes que sí incumplieron. Esto es riesgoso para la institución financiera, ya que son clientes en default que no fueron anticipados por el modelo.

print("\033[1m The result is telling us that we have: ",(confusion\_matrix[0,0]+confusion\_matrix[1,1]),"correct predictions\033[1m")

print("\033[1m The result is telling us that we have: ",(confusion\_matrix[0,1]+confusion\_matrix[1,0]),"incorrect predictions\033[1m")

print("\033[1m We have a total predictions of: ",(confusion\_matrix.sum()))

***The result is telling us that we have: 65182 correct predictions***

***The result is telling us that we have: 34663 incorrect predictions***

***We have a total predictions of: 99845***

**Reporte de clasificación del modelo**

from sklearn.metrics import classification\_report

print(classification\_report(y\_test, y\_pred))

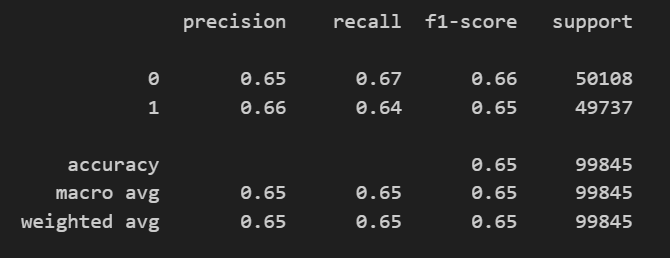


Figura 8. Resultado del modelo entrenado.

***Interpretaciones:***

* La precisión de la clase 1 en el conjunto de prueba (0.66) nos indica que, entre todos los solicitantes que el modelo identificó como "malos" (incumplidores), el 66 % realmente lo eran. Es decir, 2 de cada 3 predicciones positivas fueron acertadas. Este valor refleja qué tan confiable es el modelo cuando detecta a un cliente como potencial incumplidor.
* La recuperación (recall) de la clase 1, que fue de 0.64, indica que el modelo fue capaz de identificar correctamente al 64 % de todos los solicitantes "malos" que realmente estaban en el conjunto de prueba. En otras palabras, el modelo logró encontrar aproximadamente 6 de cada 10 casos reales de incumplimiento, lo cual representa una sensibilidad aceptable, aunque todavía hay margen para mejorar.

**Curva ROC**

[La curva ROC (curva característica operativa del receptor)](https://en.wikipedia.org/wiki/Receiver_operating_characteristic) es otra herramienta común en los clasificadores binarios. La línea punteada representa la curva ROC de un clasificador puramente aleatorio; un buen clasificador se mantiene lo más alejado posible de esta línea (hacia la esquina superior izquierda).

from sklearn.metrics import roc\_auc\_score

from sklearn.metrics import roc\_curve

sns.set(style="whitegrid", color\_codes=True)

logit\_roc\_auc = roc\_auc\_score(y\_test, logreg.predict(X\_test))

fpr, tpr, thresholds = roc\_curve(y\_test, logreg.predict\_proba(X\_test)[:,1])

plt.figure()

plt.plot(fpr, tpr, label='Logistic Regression (area = %0.2f)' % logit\_roc\_auc)

plt.plot([0, 1], [0, 1],'r--')

plt.xlim([0.0, 1.0])

plt.ylim([0.0, 1.05])

plt.xlabel('False Positive Rate')

plt.ylabel('True Positive Rate')

plt.title('Receiver operating characteristic')

plt.legend(loc="lower right")

plt.savefig('Log\_ROC')

plt.show()

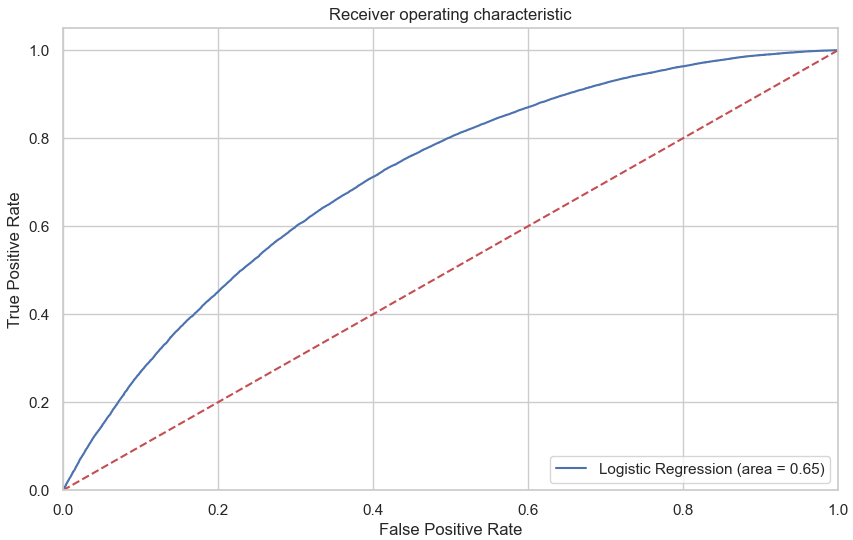


Figura 9. Curva ROC.

**Área bajo la curva (AUC)**

* El modelo tiene un **AUC = 0.65**, lo que indica que tiene una **capacidad moderada para diferenciar** entre incumplidores y no incumplidores.
* Un AUC de 0.65 significa que, si tomamos aleatoriamente un cliente que incumple y uno que no, el modelo tiene un 65 % de probabilidad de asignar una probabilidad mayor de incumplimiento al cliente correcto.

1. Implementación

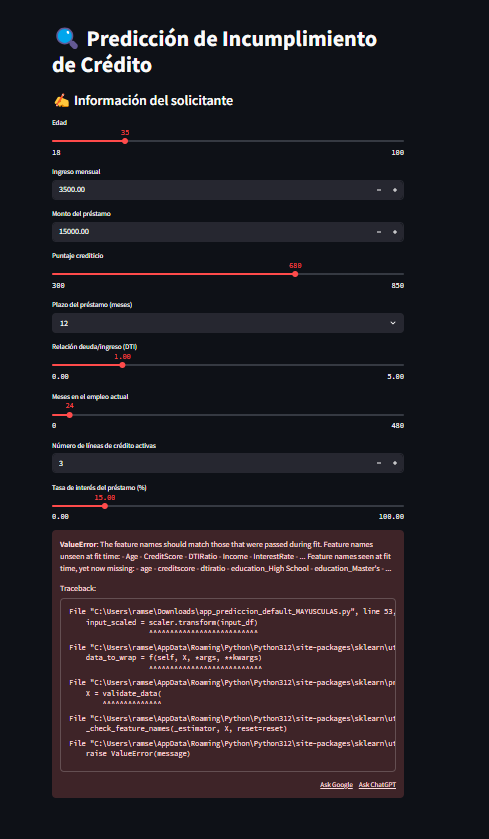


Figura 10. Interfaz, aplicativo del simulador que se alimenta del modelo predictivo.

Se construyó una app interactiva en Streamlit que permite predecir la probabilidad de default con base en los datos introducidos por un usuario. Esta herramienta puede integrarse a flujos de trabajo bancarios para facilitar decisiones crediticias en tiempo real.

# Conclusiones

* El modelo de regresión logística es una herramienta efectiva, transparente y útil para la predicción de default en préstamos.
* Las variables más relevantes identificadas fueron: puntaje crediticio, ingreso, monto del préstamo, tasa de interés, edad y estabilidad laboral.
* La regresión logística mostró mayor precisión en la clase incumplidora, mientras que Random Forest logró un mejor F1-score, es decir, un mejor equilibrio entre precisión y recall.
* El modelo inicial tuvo un bajo recall, lo que significaba que detectaba pocos incumplidores. Esto fue resuelto parcialmente al balancear las clases con técnicas como SMOTE (Synthetic Minority Over-sampling Technique), mejorando el desempeño.
* Al usar un umbral de 0.5, el modelo tendía a ser muy conservador: clasificaba como "incumplidores" solo a aquellos con probabilidades bastante altas.
* Hay casos en los que se espera que el modelo se descalibre, por ejemplo, los modelos de cobranza cuando la variable es que tan probable que no se recupere el monto adeudado. En este caso descalibración del modelo es un resultado positivo, porque me dice que las estrategias que estoy implementando generan un resultado en contravía de lo que predecía inicialmente el modelo.

# Recomendaciones

* Es mejor no fijarse solo en el porcentaje de aciertos. También hay que revisar otras medidas que nos dicen si el modelo realmente está detectando bien los casos importantes.
* Cuando hay muy pocos casos de una clase, como los clientes que no pagan, conviene equilibrar los datos para que el modelo aprenda a reconocerlos mejor.
* El valor que se usa para decidir si alguien pertenece a una clase u otra se puede cambiar. Esto ayuda a que el modelo se enfoque más en lo que queremos priorizar, como detectar más riesgos, aunque se cometan algunos errores.
* Usar solo las variables más importantes hace que el modelo sea más rápido, más fácil de entender y más confiable.
* Hacer gráficos y visualizar los resultados ayuda mucho a entender cómo está funcionando el modelo y a tomar mejores decisiones.

# Retos

En base al modelo predictivo un reto fue poder generar un aplicativo 100% funcional, por las limitantes del tiempo, solo se pudo llegar al desarrollo conceptual y de aplicativo con el código. No se puede considerar un simulador completo hasta resolver los errores.

# Bibliografía

Suárez González, J. D. (2012). *El modelo de Merton para la estimación del riesgo de default*. Universidad del Rosario. Recuperado de <https://repository.urosario.edu.co/handle/10336/4064>

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0: Step-by-step data mining guide. SPSS Inc.

Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining. In Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining (pp. 29–39).

Altman, E. I., & Saunders, A. (1998). Credit risk measurement: Developments over the last 20 years. Journal of Banking & Finance, 21(11–12), 1721–1742. https://doi.org/10.1016/S0378-4266(97)00036-8

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied logistic regression (3rd ed.). Wiley.

Merton, R. C. (1974). On the pricing of corporate debt: The risk structure of interest rates. The Journal of Finance, 29(2), 449–470. https://doi.org/10.1111/j.1540-6261.1974.tb03058.x

Thomas, L. C., Crook, J. N., & Edelman, D. B. (2017). Credit scoring and its applications (2nd ed.). SIAM.