PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ
FACULTAD DE CIENCIAS SOCIALES
MACHINE LEARNING PARA FINANZAS
SEMESTRE 2026-0

# Tarea Académica 1

| | | |
|---|---|---|
| Profesora: | Nadia Valderrama | (nvalderramav@pucp.edu.pe) |
| Integrantes: | Giovanni Martinez | (20186018) |
| | Deborah Martinez | (20220387) |
| | Isaac Benavides | (20220835) |

---

float

## Introducción

En el contexto del análisis de riesgo financiero, la capacidad de anticipar el incumplimiento de obligaciones crediticias constituye un elemento central para la gestión prudencial de portafolios, la asignación de provisiones y la toma de decisiones en instituciones financieras. En este marco, los modelos de clasificación basados en datos históricos permiten identificar patrones de comportamiento asociados a eventos de default y estimar probabilidades de incumplimiento de manera sistemática.

El objetivo de este proyecto es desarrollar un proceso integral de análisis exploratorio y preprocesamiento sobre un dataset de préstamos, con el fin de preparar adecuadamente la información para la construcción de modelos de clasificación supervisada. Para ello, se realizan etapas de diagnóstico descriptivo, evaluación de calidad de datos, análisis univariado y bivariado respecto al target, identificando variables con mayor capacidad discriminante y posibles desafíos metodológicos previos al modelamiento.

En conjunto, este trabajo busca sentar las bases técnicas y analíticas necesarias para un enfoque robusto de predicción de riesgo crediticio bajo estándares propios de analítica aplicada en banca.

## Caso de uso: Clasificacion

## Dataset y Contexto del caso

El dataset analizado corresponde a un conjunto de información crediticia orientado a describir solicitudes y resultados de préstamos personales. En particular, contiene variables asociadas tanto a las condiciones contractuales del crédito como a características financieras y socioeconómicas del prestatario.

La naturaleza del problema es típicamente bancaria: a partir de información observable al momento de originación, se busca comprender qué factores se relacionan con el cumplimiento o incumplimiento del préstamo.

**Dimensión y estructura general**

El conjunto de datos cuenta con aproximadamente:

- $19,900$ **observaciones** (clientes/préstamos individuales)
- 20 **variables originales**, posteriormente tratadas y depuradas durante el preprocesamiento
- Variables de tipo numérico y categórico, representando tanto montos continuos como atributos discretos del solicitante.

Esto implica un dataset de tamaño intermedio, adecuado para análisis exploratorio robusto y posterior entrenamiento de modelos de clasificación.

## Diccionario de variables

Tabla 1: Descripción de las variables del conjunto de datos de préstamos

| **Columna** | **Descripción** |
|---|---|
| index | Identificador único (no es una variable explicativa) |
| loan_amnt | Monto del préstamo solicitado |
| term | Plazo del préstamo en meses |
| int_rate | Tasa de interés del préstamo |
| installment | Cuota mensual estimada al originarse el préstamo |
| grade | Calificación (grade) asignada |
| emp_title | Cargo / ocupación declarada por el solicitante |
| emp_length | Antigüedad laboral (0 a 10; 0 = ¡1 año, 10 = 10+ años) |
| home_ownership | Tipo de tenencia de vivienda (RENT/OWN/MORTGAGE/OTHER) |
| annual_inc | Ingreso anual reportado |
| verification_status | Estado de verificación del ingreso (verificado / no verificado / fuente verificada, etc.) |
| loan_status | Estado actual de préstamo (Charged off, Fully Paid) |
| pymnt_plan | Indica si se estableció un plan de pagos para el préstamo |
| purpose | Motivo declarado del préstamo |
| addr_state | Estado (EE. UU.) declarado en la solicitud |
| dti | Ratio deuda/ingreso: pagos mensuales de deuda |
| delinq_2yrs | Nº de moras (30+ días) en los últimos 2 años |
| revol_util | % de utilización de crédito revolvente (uso / límite disponible) |
| application_type | Tipo de solicitud: individual o conjunta (co-prestatarios) |
| pub_rec_bankruptcies | Nº de bancarrotas registradas en registros públicos |

## Variable objetivo (Target)

El problema de clasificación desarrollado en este proyecto se centra en la variable objetivo: **Estado del préstamo (`estado_prestamo`)**

Dicha variable representa el resultado final observado para cada crédito y toma dos clases principales, **Pagado (Fully Paid) e Incumplido (Charged Off)**

Desde una perspectiva de riesgo crediticio, este planteamiento corresponde a un problema clásico de *default prediction*, en el cual se busca identificar patrones en la información financiera y socioeconómica del prestatario que permitan anticipar el evento de incumplimiento.

La distribución empírica de la variable objetivo muestra que aproximadamente:

- $\sim 85\,\%$ de los préstamos se encuentran en estado *Pagado*
- $\sim 15\,\%$ de los préstamos corresponden a casos *Incumplido*

Este comportamiento refleja un escenario realista en portafolios bancarios, donde los eventos de default suelen ser relativamente menos frecuentes. En consecuencia, se trata de un caso con **desbalance moderado de clases**, aspecto que debe ser considerado en etapas posteriores del modelamiento mediante métricas adecuadas (ROC-AUC, Recall, Precision), evitando depender únicamente de la exactitud (*accuracy*).

## Bloques conceptuales de variables explicativas

Con el fin de estructurar adecuadamente el análisis, las variables explicativas del dataset pueden agruparse en bloques conceptuales según su interpretación económica y financiera. Este enfoque es estándar en analítica aplicada a banca, ya que permite identificar con mayor claridad los determinantes del riesgo crediticio.

### Características del préstamo

Este conjunto de variables describe directamente las condiciones contractuales del crédito otorgado:

- **Monto del préstamo** (`monto_prestamo`)
- **Plazo en meses** (`plazo_meses`)
- **Tasa de interés** (`tasa_interes`)
- **Cuota mensual** (`cuota_mensual`)

En general, estas variables capturan tanto el nivel de exposición del préstamo como elementos de pricing: tasas más elevadas suelen asignarse a prestatarios con mayor perfil de riesgo.

### Perfil financiero del solicitante

Estas variables se vinculan directamente con la capacidad de pago y sostenibilidad financiera del prestatario:

- **Ingreso anual** (`ingreso_anual`)
- **Ratio deuda/ingreso** (`ratio_deuda_ingreso`)

Particularmente, el ratio deuda/ingreso es un indicador central en originación crediticia, pues aproxima la carga financiera relativa del individuo.

### Historial y comportamiento crediticio

Este bloque recoge información sobre antecedentes financieros y disciplina crediticia previa:

- **Moras en los últimos dos años** (`moras_ultimos_2y`)
- **Bancarrotas públicas** (`bancarrotas_publicas`)
- **Utilización de crédito revolvente** (`utilizacion_revolvente`)

Estas variables suelen ser altamente predictivas, dado que reflejan comportamientos históricos directamente asociados al riesgo de incumplimiento.

### Variables categóricas socioeconómicas

El dataset incluye variables cualitativas que aportan contexto sobre el solicitante:

- **Ocupación** (`ocupacion`)
- **Antigüedad laboral** (`antiguedad_laboral`)
- **Tenencia de vivienda** (`tenencia_vivienda`)
- **Verificación de ingresos** (`verificacion_ingresos`)
- **Propósito del préstamo** (`proposito_prestamo`)
- **Estado de residencia** (`estado_residencia`)

Estas variables presentan desafíos típicos en datos reales de banca, tales como alta cardinalidad (especialmente en ocupación), presencia de categorías poco frecuentes y valores no informados, lo cual exige tratamientos cuidadosos durante el preprocesamiento.

### Calificación crediticia interna

Finalmente, una de las variables más informativas es:

- **Grado crediticio** (`grado_crediticio`)

Esta variable representa una calificación ordinal del prestatario, donde categorías como $A$–$B$ reflejan menor riesgo, mientras que categorías como $F$–$G$ corresponden a perfiles con mayor probabilidad de incumplimiento. Los análisis exploratorios mostraron que este predictor presenta una relación monotónica clara con las tasas de default.

En la revisión inicial del dataset se identificaron elementos relevantes de calidad y estructura.

# Metodología y fases iniciales del proyecto

El desarrollo de un proyecto de *Machine Learning* aplicado a finanzas requiere una secuencia metodológica estructurada que garantice tanto la calidad de los datos como la validez de los hallazgos preliminares antes de proceder al entrenamiento de modelos predictivos.

En este trabajo se siguió un flujo estándar de analítica aplicada a riesgo crediticio, compuesto por las siguientes etapas:

1. Importación y revisión inicial del dataset
2. Análisis descriptivo preliminar
3. Diagnóstico de calidad de datos
4. Preprocesamiento inicial
5. Análisis exploratorio univariado
6. Análisis exploratorio bivariado respecto al incumplimiento

Cada una de estas fases cumple un rol fundamental en la construcción de una base robusta para el modelamiento supervisado posterior.

## Configuración del entorno y librerías utilizadas

El análisis fue implementado en Python, empleando librerías ampliamente utilizadas en ciencia de datos y analítica financiera. Estas herramientas permiten manipulación eficiente de datos, estadística descriptiva y visualización profesional de resultados.

Las principales librerías empleadas incluyen:

- `pandas` y `numpy`: manipulación y transformación de datos
- `matplotlib`: construcción de gráficos exploratorios
- `warnings`: control de mensajes no críticos durante ejecución

El entorno fue configurado de forma reproducible, permitiendo que el notebook pueda ejecutarse tanto localmente como en Google Colab, plataforma empleada para la evaluación académica.

## Importación del dataset

El dataset fue importado desde un archivo en formato Excel, asegurando que la lectura sea compatible tanto en ejecución local como en entornos en la nube. Posteriormente, se realizó una inspección inicial mediante funciones como `head()`, `shape` y `dtypes`, con el objetivo de verificar:

- Dimensión general del conjunto de datos
- Presencia de variables numéricas y categóricas
- Coherencia del formato de cada columna

Esta etapa inicial permite confirmar que la estructura del dataset es adecuada antes de realizar cualquier transformación.

# Análisis estadístico descriptivo inicial

Una vez importada la información, se realizó un análisis descriptivo preliminar con el objetivo de caracterizar las distribuciones generales de las variables numéricas y obtener una primera comprensión del perfil del portafolio crediticio.

Para ello, se emplearon estadísticos resumidos como:

- Media, mediana y desviación estándar
- Valores mínimos y máximos
- Cuantiles relevantes (1 %, 5 %, 95 %, 99 %)

El análisis permitió identificar patrones típicos en portafolios de préstamos minoristas, así como la presencia de asimetrías pronunciadas en variables monetarias como el ingreso anual y el monto del préstamo.

Asimismo, se confirmó que variables como `tasa_interes` y `ratio_deuda_ingreso` presentan distribuciones consistentes con literatura de riesgo crediticio, siendo potencialmente relevantes para la discriminación del incumplimiento.

# Diagnóstico de calidad de datos

El diagnóstico de calidad constituye una fase indispensable antes de aplicar cualquier algoritmo de aprendizaje supervisado, dado que inconsistencias o valores faltantes pueden distorsionar tanto el análisis exploratorio como el entrenamiento posterior de modelos.

En esta etapa se evaluaron los siguientes aspectos:

## Valores faltantes (Missing Values)

Se cuantificó el número y porcentaje de valores faltantes por variable. Se identificó que la mayoría de variables presentan niveles reducidos de missingness, con excepción de atributos categóricos como ocupación y antigüedad laboral.

Este comportamiento es habitual en datos reales de originación crediticia, donde ciertos campos pueden no estar disponibles para todos los solicitantes.

## Variables constantes y no informativas

Durante el diagnóstico se detectó la existencia de variables sin variabilidad (constantes en todo el dataset). Dichas variables fueron eliminadas, dado que carecen de capacidad predictiva y no aportan información discriminante en clasificación.

## Categorías inconsistentes y valores atípicos

Se identificaron categorías poco frecuentes o inconsistentes en variables como el estado de residencia o el tipo de solicitud.

Asimismo, en variables numéricas se detectaron rangos no plausibles, como valores extremadamente altos en la utilización de crédito revolvente, los cuales fueron tratados como valores inválidos.

# Preprocesamiento inicial

Con base en el diagnóstico previo, se aplicó un preprocesamiento inicial conservador orientado a preparar el dataset para el análisis exploratorio y el modelamiento futuro.

Las estrategias implementadas fueron:

- Eliminación de observaciones con target no definido, dado que no pueden contribuir a un problema supervisado.
- Imputación numérica mediante la mediana, por robustez frente a outliers.
- Tratamiento de variables categóricas mediante la creación de una categoría residual (*Otros*) para valores no informados.
- Corrección de valores inválidos (por ejemplo, utilización revolvente superior al 100 %).

Estas decisiones permiten preservar el tamaño muestral y, al mismo tiempo, asegurar coherencia estadística y operacional en el dataset.

# Análisis exploratorio univariado

Posteriormente se desarrolló un análisis exploratorio univariado, con el fin de estudiar la distribución individual de cada variable.

## Variables numéricas

Las variables monetarias y financieras mostraron patrones característicos:

- Distribuciones asimétricas hacia la derecha en ingreso anual y monto del préstamo.
- Concentración de tasas de interés en rangos intermedios, con colas asociadas a perfiles de mayor riesgo.
- Evidencia de outliers extremos, esperables en datos crediticios reales.

## Variables categóricas

En variables categóricas se observó:

- Dominancia de ciertos propósitos de préstamo, especialmente consolidación de deuda.
- Distribución ordenada del grado crediticio, con mayor frecuencia en categorías intermedias.
- Alta cardinalidad en ocupación, dificultando su uso directo sin agrupación posterior.

# Análisis exploratorio bivariado respecto al target

Finalmente, el análisis bivariado permitió identificar relaciones entre variables explicativas y el evento de incumplimiento.

### Variables numéricas

Se encontró evidencia clara de que los prestatarios incumplidos presentan:

- Tasas de interés significativamente superiores
- Mayor utilización de crédito revolvente
- Ratios deuda/ingreso relativamente más elevados

Asimismo, las tasas de incumplimiento aumentan monotónicamente a medida que se avanza hacia quintiles superiores en dichas variables, reflejando capacidad discriminante relevante.

### Variables categóricas

Entre las variables categóricas, destacan:

- El grado crediticio, mostrando un aumento pronunciado del default desde A hasta G.
- El propósito del préstamo, donde categorías como *small_business* presentan riesgo significativamente superior.
- Variables de contexto como vivienda o residencia aportan señales complementarias de menor magnitud.

## Principales hallazgos del análisis exploratorio

A partir de las etapas de análisis descriptivo, diagnóstico de calidad, preprocesamiento y exploración univariada y bivariada, se identificaron hallazgos relevantes desde una perspectiva de riesgo crediticio. Estos resultados constituyen evidencia preliminar fundamental para justificar el uso de modelos de clasificación supervisada en fases posteriores.

### Hallazgos asociados a la variable objetivo

En primer lugar, se confirmó que el portafolio presenta una tasa base de incumplimiento cercana al 15 %, reflejando un escenario realista en crédito minorista. Este desbalance moderado implica que la evaluación futura del desempeño predictivo debe priorizar métricas como ROC-AUC, Recall y Precision, en lugar de depender exclusivamente de la exactitud global.

### Variables numéricas con mayor capacidad discriminante

El análisis bivariado evidenció que ciertas variables numéricas presentan una relación clara y consistente con el incumplimiento:

- **Tasa de interés**: los préstamos con tasas más elevadas muestran una probabilidad significativamente mayor de default, lo cual es coherente con mecanismos de pricing basados en riesgo.

- **Utilización de crédito revolvente**: prestatarios con mayor dependencia del crédito disponible tienden a exhibir tasas superiores de incumplimiento.
- **Ratio deuda/ingreso**: se observa un incremento gradual del default en niveles altos de carga financiera relativa, sugiriendo estrés económico como factor explicativo.

Adicionalmente, la segmentación por quintiles mostró patrones monotónicos: la tasa de incumplimiento aumenta sistemáticamente en los rangos superiores de dichas variables, característica altamente deseable en modelos interpretable de scoring crediticio.

## Variables categóricas clave en riesgo crediticio

Entre las variables cualitativas, se identificaron predictores particularmente informativos:

- **Grado crediticio**: la tasa de default incrementa pronunciadamente desde categorías A–B (menor riesgo) hasta F–G (mayor riesgo). Esta variable representa uno de los determinantes estructurales más relevantes del portafolio.
- **Propósito del préstamo**: ciertas categorías como *small_business* presentan tasas de incumplimiento significativamente superiores, lo cual refleja la incertidumbre inherente a financiamientos asociados a actividad empresarial de pequeña escala.
- **Tenencia de vivienda y verificación de ingresos**: aportan señales complementarias de menor magnitud, aunque su interpretación requiere cautela debido a posibles efectos institucionales de selección en underwriting.

## Desafíos de calidad y estructura del dataset

El diagnóstico de calidad permitió identificar elementos típicos en datos reales de banca:

- **Valores faltantes moderados**: concentrados principalmente en ocupación y antigüedad laboral.
- **Alta cardinalidad en ocupación**: con más de 15,000 categorías únicas, lo que dificulta su uso directo sin estrategias de agrupación o codificación posterior.
- **Valores atípicos extremos**: especialmente en variables monetarias como ingreso anual, reflejando distribuciones altamente asimétricas y la presencia de outliers propios de portafolios heterogéneos.
- **Variables no informativas**: se detectaron columnas constantes que fueron correctamente eliminadas por carecer de valor predictivo.

Estos hallazgos resaltan la importancia del preprocesamiento previo al modelamiento, asegurando consistencia y robustez estadística.

# Limitaciones del análisis y recomendaciones metodológicas

Si bien la presente entrega se enfoca en fases iniciales del proyecto, es importante reconocer ciertas limitaciones:

- El tratamiento de variables categóricas de alta cardinalidad requiere metodologías avanzadas (Top-K, WoE, embeddings) que serán consideradas en etapas posteriores.
- El análisis exploratorio se basa en asociaciones descriptivas, sin establecer relaciones causales directas.
- La presencia de desbalance moderado en el target exige especial atención a la selección de métricas y técnicas de evaluación futura.

En consecuencia, se recomienda que la siguiente fase del proyecto incorpore:

- Codificación sistemática de variables categóricas.
- Entrenamiento de modelos comparativos (Regresión Logística, Árboles, Gradient Boosting).
- Evaluación mediante ROC-AUC, curvas Precision-Recall y análisis de umbrales óptimos.

# Diagnóstico de calidad de datos: hallazgos principales

Antes de realizar cualquier modelamiento, se efectuó un diagnóstico detallado de calidad de datos con el objetivo de detectar inconsistencias, valores faltantes y variables no informativas.

## Valores faltantes

El análisis de missing values mostró que la mayoría de variables presentan niveles reducidos de información faltante. Sin embargo, se identificaron valores faltantes moderados en variables categóricas como ocupación y antigüedad laboral, fenómeno común en bases reales de originación crediticia.

Este diagnóstico justificó la aplicación de imputación conservadora en etapas posteriores.

## Observaciones sin variable objetivo

Se detectó la presencia de un número extremadamente reducido de observaciones sin etiqueta en la variable objetivo (`estado_prestamo`). Dado que un problema supervisado requiere targets definidos, estas filas fueron eliminadas.

```
df_clean = df_clean.dropna(subset=["estado_prestamo"])
```

Listing 1: Eliminación de observaciones sin target.

Este ajuste permitió evitar la aparición de una categoría residual irrelevante en los gráficos posteriores.

### Variables constantes

El diagnóstico identificó variables sin variabilidad, las cuales carecen de valor predictivo. Estas columnas fueron eliminadas del dataset final para asegurar consistencia metodológica.

### Corrección de valores inválidos

En variables numéricas como `utilizacion_revolvente`, se detectaron valores fuera del rango esperado (mayores al 100 %). Dichos valores fueron tratados como inválidos y corregidos mediante imputación posterior.

Este tipo de anomalías es frecuente en datos financieros y requiere tratamiento explícito.

## Análisis bivariado respecto al incumplimiento

El EDA bivariado permitió identificar variables con capacidad discriminante significativa respecto al incumplimiento.

### Tasa de interés como predictor dominante

Se encontró una relación monotónica clara entre la tasa de interés y el default: los quintiles superiores presentan tasas de incumplimiento sustancialmente mayores.

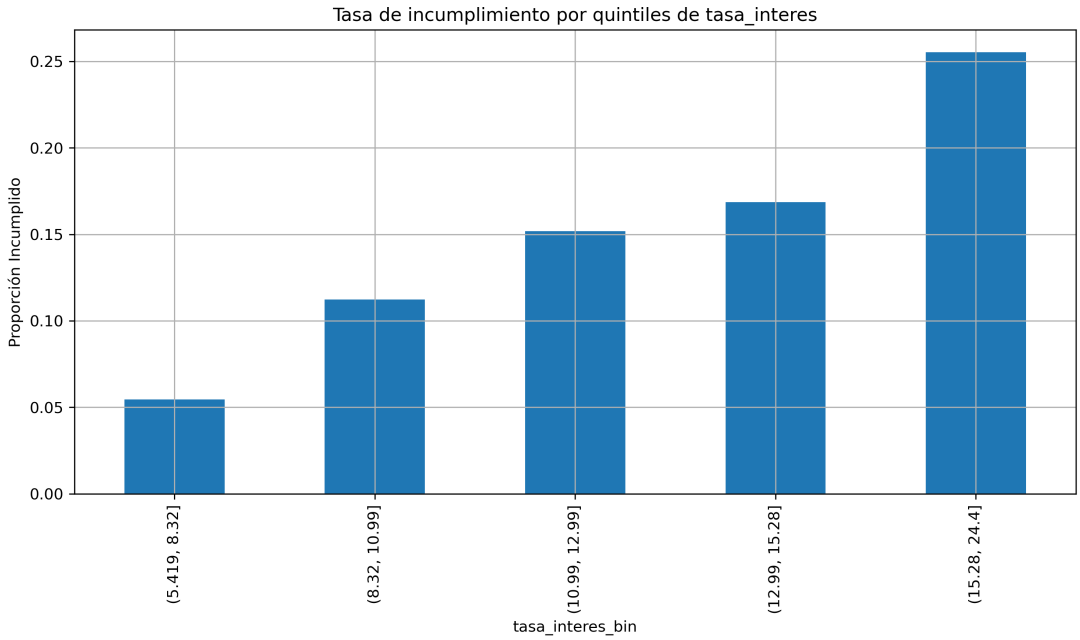


Figura 1: Tasa de incumplimiento por quintiles de `tasa_interes`.

Este resultado es consistente con mecanismos de pricing basados en riesgo: prestatarios más riesgosos reciben tasas más elevadas.

## Grado crediticio

El grado crediticio muestra el gradiente más pronunciado del análisis: las categorías de menor calificación presentan tasas de incumplimiento sustancialmente superiores.

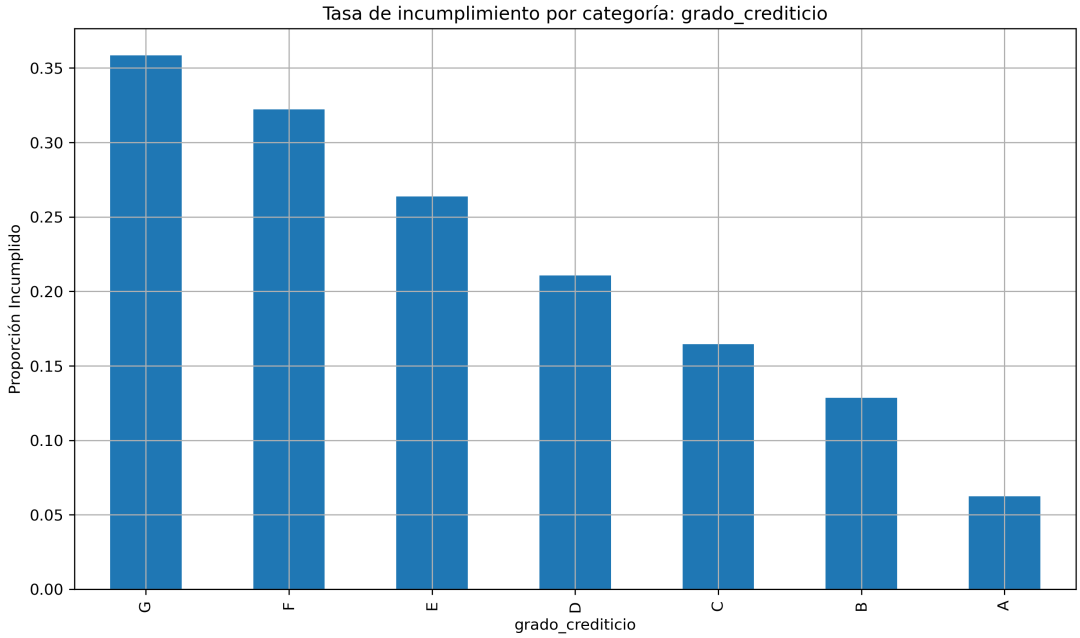


Figura 2: Tasa de incumplimiento por grado crediticio.

Este comportamiento confirma que el grade es uno de los determinantes estructurales del riesgo en el portafolio.

## Utilización revolvente y estrés financiero

Además de la tasa de interés, se identificó que la utilización de crédito revolvente constituye una señal importante de riesgo. En particular, prestatarios con valores elevados de utilización tienden a presentar mayores tasas de incumplimiento, lo cual es coherente con una dependencia creciente del crédito disponible.

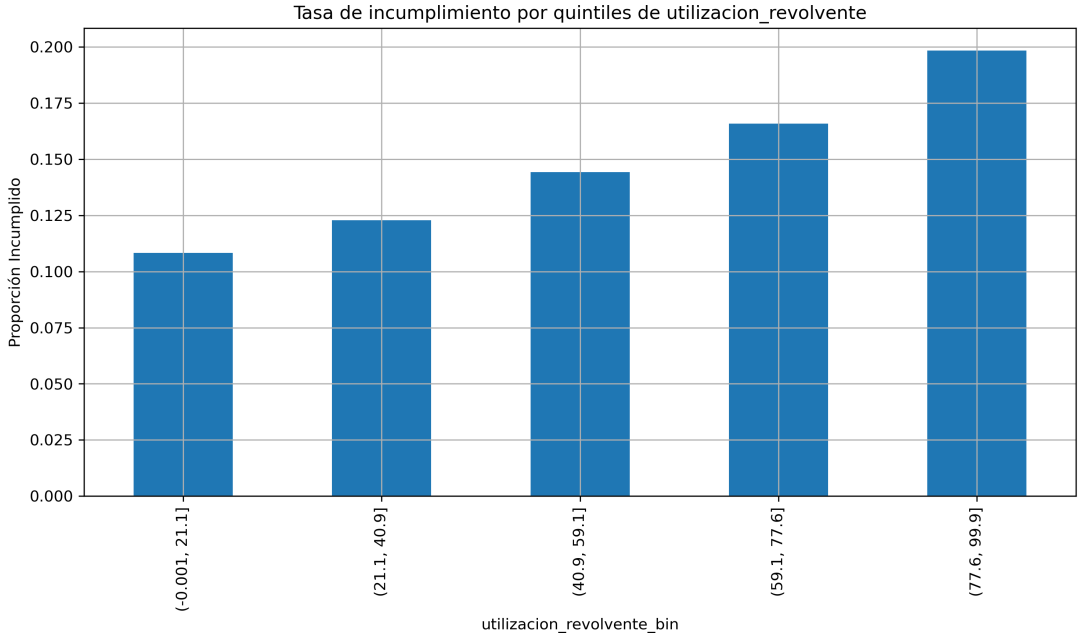


Figura 3: Tasa de incumplimiento por quintiles de `utilizacion_revolvente`.

Este resultado sugiere que la utilización revolvente actúa como proxy del apalancamiento financiero de corto plazo del solicitante.

## Ratio deuda/ingreso como indicador de carga financiera

El ratio deuda/ingreso (`ratio_deuda_ingreso`) también mostró una asociación positiva con el incumplimiento. Si bien la pendiente es menos pronunciada que en el caso de la tasa de interés, existe un incremento gradual del default conforme aumenta la carga financiera relativa.

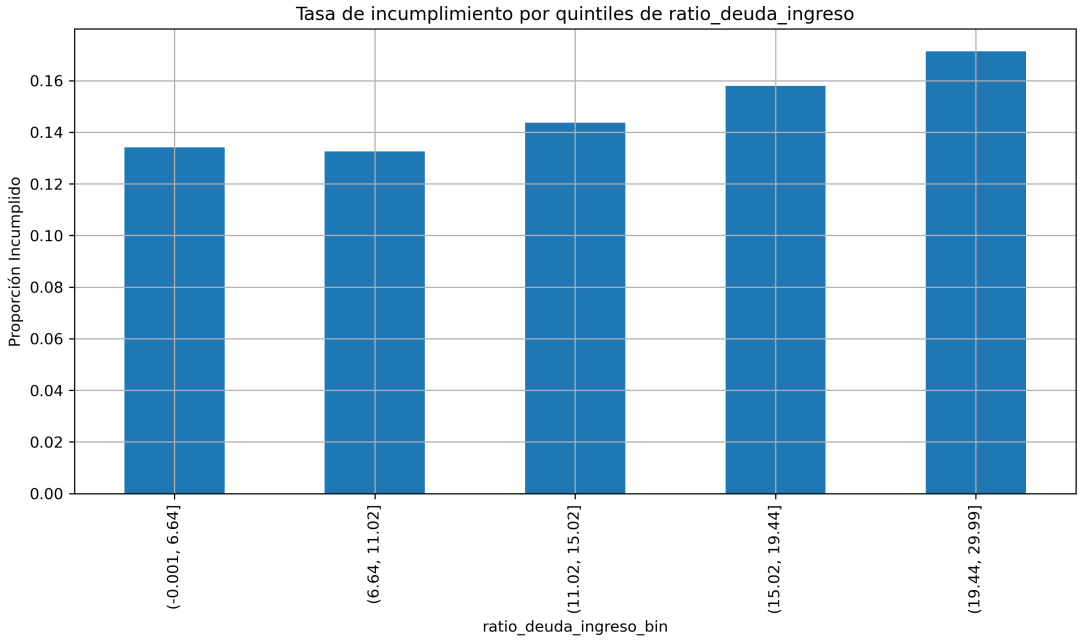


Figura 4: Tasa de incumplimiento por quintiles de `ratio_deuda_ingreso`.

Este patrón es consistente con fundamentos de riesgo crediticio: mayores niveles de endeudamiento relativo reducen la capacidad de absorción ante shocks de ingresos.

## Propósito del préstamo y heterogeneidad de riesgo

Entre los predictores categóricos, el propósito del préstamo mostró diferencias importantes en tasas de default. En particular, categorías como *small_business* presentan niveles significativamente superiores al promedio del portafolio, reflejando la mayor incertidumbre de flujos asociada a financiamiento empresarial de pequeña escala.

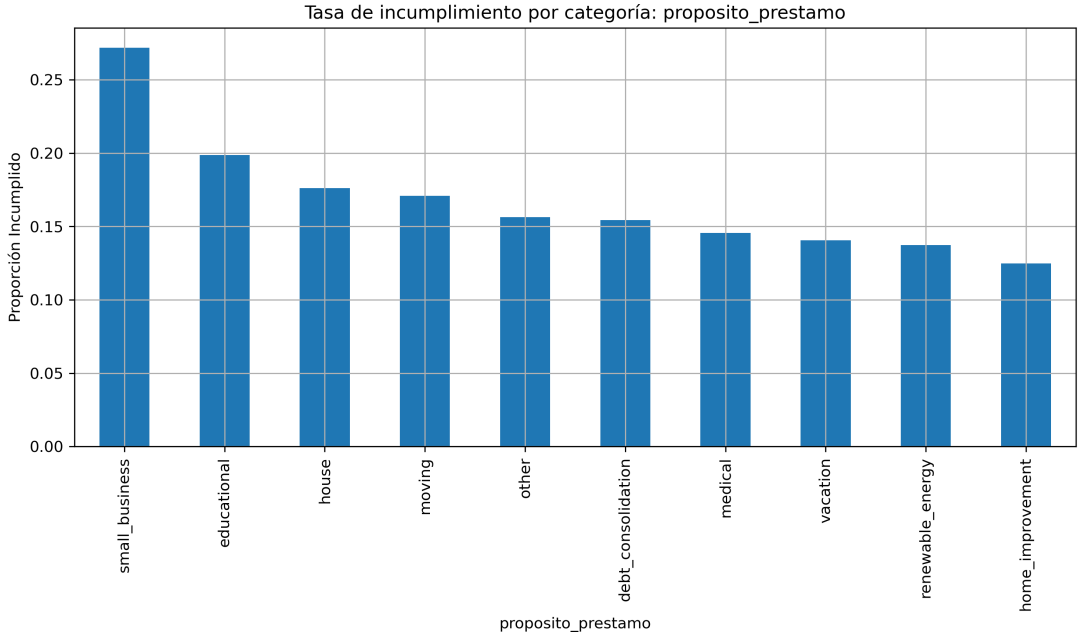


Figura 5: Tasa de incumplimiento por propósito del préstamo.

Este hallazgo confirma que la finalidad del crédito constituye una señal relevante en la segmentación de riesgo.

## Tenencia de vivienda y variables contextuales

La variable de tenencia de vivienda aporta evidencia complementaria. Se observan diferencias moderadas entre categorías como *RENT*, *MORTGAGE* y *OWN*. Aunque su poder discriminante es menor frente a variables financieras estructurales, estas características pueden capturar estabilidad patrimonial y condiciones socioeconómicas.

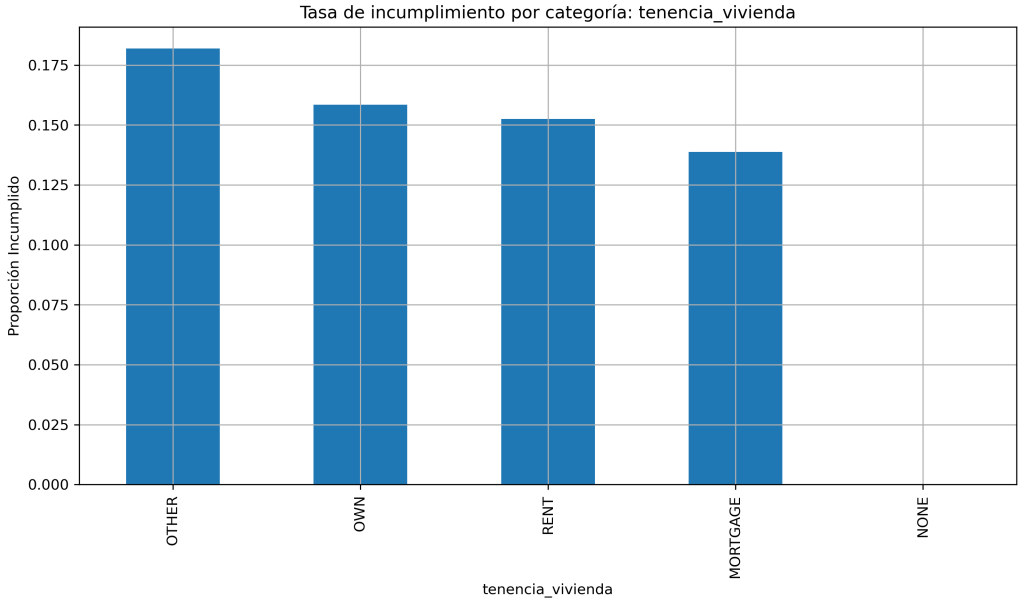


Figura 6: Tasa de incumplimiento por tenencia de vivienda.

## Efectos geográficos: estado de residencia

Finalmente, el estado de residencia muestra variaciones moderadas en incumplimiento entre los estados con mayor representación en la muestra. Estas diferencias no son extremas, pero sugieren heterogeneidad regional asociada a condiciones económicas locales.

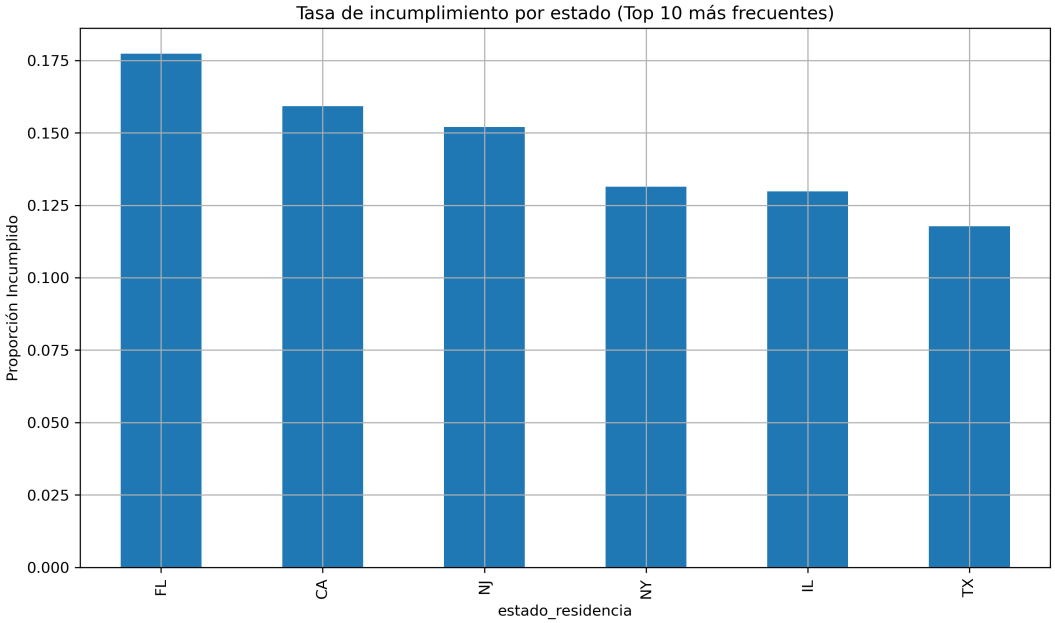


Figura 7: Tasa de incumplimiento por estado de residencia (Top 10).

En general, esta variable se interpreta como una señal secundaria, útil principalmente en combinación con indicadores financieros principales.

## Síntesis ejecutiva de hallazgos principales

A modo de consolidación, el análisis exploratorio permite destacar los siguientes hallazgos fundamentales:

- El portafolio presenta una tasa base de incumplimiento cercana al 15 %, consistente con escenarios reales de crédito minorista.
- La **tasa de interés** muestra una relación monotónica fuerte con el default, reflejando mecanismos de pricing asociados al riesgo del solicitante.
- La **utilización revolvente** y el **ratio deuda/ingreso** constituyen indicadores financieros relevantes, capturando estrés crediticio y dependencia del financiamiento.
- El **grado crediticio** es el predictor categórico más dominante, mostrando un gradiente pronunciado desde perfiles de bajo hasta alto riesgo.
- El **propósito del préstamo** introduce heterogeneidad significativa: categorías como *small_business* presentan mayores tasas de incumplimiento.
- Variables contextuales como vivienda o residencia aportan señales adicionales, pero de menor magnitud explicativa.

# Caso de uso: Regresion Lineal

# Dataset y Contexto del caso

En este caso este dataset corresponde a un conjunto de información socioeconómica y demográfica orientado a describir el perfil financiero de diversos individuos. En particular, contiene variables asociadas tanto a características personales (edad, nivel educativo, estado civil) como a factores del entorno laboral y de vivienda (ocupación, experiencia laboral y tipo de residencia).

La naturaleza del problema se enmarca en la **predicción de valores continuos**: a partir de información observable, se busca comprender qué factores influyen significativamente en la determinación del **ingreso anual (Income)**. Mediante un modelo de **regresión lineal**, el objetivo es cuantificar la relación entre estas variables predictoras y el nivel de ingresos, permitiendo identificar la sensibilidad de la remuneración ante cambios en el perfil profesional o demográfico del individuo.

### Dimensión y estructura general

El conjunto de datos utilizado para el modelado de regresión cuenta con las siguientes características:

- El dataset comprende 10,000 registros individuales.
- Se dispone de 14 variables originales (13 predictoras y 1 variable objetivo), las cuales incluyen atributos demográficos y financieros.
- El conjunto integra variables de tipo numérico (como Age y Work Experience) y categórico (como Education Level y Occupation), permitiendo un análisis híbrido entre factores cuantitativos y cualitativos.

## Diccionario de variables

Tabla 2: Definición de variables del conjunto de datos de ingresos (Regresión)

| **Variable** | **Descripción** |
|---|---|
| Age | Edad del individuo en años. |
| Education_Level | Nivel educativo más alto alcanzado (Secundaria, Bachillerato, Maestría, etc.). |
| Occupation | Ocupación o sector profesional principal (Tecnología, Finanzas, Educación, etc.). |
| Number_of_Dependents | Número de personas que dependen económicamente del individuo. |
| Location | Zona donde reside el individuo (Urbana, Suburbana, Rural). |
| Work_Experience | Años de experiencia laboral acumulada. |
| Marital_Status | Estado civil del individuo (Soltero, Casado, etc.). |
| Employment_Status | Situación laboral del individuo (Tiempo completo, Tiempo parcial, Autónomo). |
| Household_Size | Número total de personas que viven en el hogar. |
| Homeownership_Status | Condición de tenencia de la vivienda (Propia, Alquilada). |
| Type_of_Housing | Tipo de vivienda (Apartamento, Casa unifamiliar, etc.). |
| Gender | Género del individuo. |
| Primary_Mode_of_Transportation | Medio de transporte principal utilizado diariamente. |
| **Income** | **Ingreso anual del individuo (Variable objetivo del modelo).** |

## Variable objetivo (Target)

El desarrollo del modelo de regresión lineal en este proyecto se articula en torno a la variable objetivo:

**Ingreso Anual (*Income*)**

Esta variable representa la remuneración económica total percibida por el individuo de forma anual. Al tratarse de una **variable numérica continua**, su análisis permite modelar el comportamiento financiero del sujeto en una escala cuantitativa, facilitando la estimación precisa de su capacidad adquisitiva en función de su perfil sociodemográfico y laboral.

Desde una perspectiva analítica, el planteamiento busca capturar la variabilidad del ingreso para identificar los determinantes de la riqueza personal. La distribución empírica de la variable objetivo en el conjunto de datos presenta las siguientes características:

- **Rango de Valores:** Los datos muestran un espectro que abarca desde ingresos base hasta niveles de alta remuneración, reflejando la diversidad económica del perfil analizado.
- **Distribución de los Datos:** El ingreso presenta una distribución con un sesgo hacia los valores superiores, comportamiento característico en variables económicas donde existe una concentración de individuos con ingresos elevados.

- **Consideraciones de Modelado:** Debido a la naturaleza continua de *Income*, el rendimiento del modelo se evalúa mediante métricas de error de magnitud, tales como el Error Absoluto Medio (MAE) y el Coeficiente de Determinación ($R^2$), los cuales permiten cuantificar la cercanía entre los valores estimados y los ingresos reales reportados.

# Bloques conceptuales de variables explicativas

Con el fin de estructurar adecuadamente el análisis de regresión, las variables explicativas se han agrupado en bloques conceptuales según su naturaleza técnica y financiera. Esta clasificación permite identificar los determinantes del riesgo y aplicar tratamientos diferenciados durante el preprocesamiento de datos.

## Condiciones del Crédito

Son variables denominadas "de control"porque dependen de la oferta de la institución financiera y no del perfil del cliente. Estas definen el marco contractual del préstamo:

- **Tasa de interés (`tasa_interes`):** Es la variable central de este bloque. Actúa como un *proxy* del riesgo ya calculado por el banco; estadísticamente, una tasa más elevada suele correlacionar con perfiles de mayor riesgo de incumplimiento.
- **Monto del préstamo y Plazo** (`monto_prestamo`, `plazo_meses`).

## Perfil financiero

Estas variables se vinculan directamente con la capacidad de pago y la sostenibilidad financiera del prestatario a largo plazo:

- **Ratio deuda/ingreso (`ratio_deuda_ingreso`):** Es el indicador principal de solvencia. A diferencia del ingreso bruto, este ratio aproxima la carga financiera real del individuo frente a sus compromisos totales.
- **Ingreso anual** (`ingreso_anual`).

## Comportamiento crediticio

Este bloque recoge información sobre antecedentes financieros previos. En modelos de regresión, suelen ser los predictores con mayor significancia estadística:

- **Moras últimos dos años (`moras_ultimos_2y`):** Representa la variable predictora más crítica. El comportamiento de pago reciente es el mejor indicador del riesgo de *default* futuro.
- **Bancarrotas y Utilización** (`bancarrotas_publicas`, `utilizacion_revolvente`).

## Variables socioeconómicas

Incluye factores cualitativos que aportan contexto sobre el entorno laboral y personal del solicitante:

- **Antigüedad laboral (`antiguedad_laboral`):** Es la variable clave de estabilidad. Una mayor trayectoria en el empleo actual reduce la incertidumbre sobre la continuidad del flujo de ingresos.
- **Ocupación y Vivienda** (`ocupacion`, `tenencia_vivienda`).

| Bloque | Criterio de Agrupación | Variable Clave |
|---|---|---|
| Exposición | Condiciones contractuales | `tasa_interes` |
| Solvencia | Capacidad real de pago | `ratio_deuda_ingreso` |
| Disciplina | Antecedentes de pago | `moras_ultimos_2y` |
| Estabilidad | Perfil socio-laboral | `antiguedad_laboral` |

Tabla 3: Resumen de variables críticas por bloque conceptual.

## Observación: Implementación de Ejes Secundarios (`twinx`)

A lo largo del desarrollo de nuestro entorno de programación en Google Colab, se han implementado diversas técnicas de visualización, alternando entre gráficos estándar y gráficos con doble eje mediante la función `twinx`. La decisión de integrar un eje secundario responde a la necesidad de comparar variables que, aunque están relacionadas, poseen escalas de magnitud totalmente distintas (por ejemplo, comparar el número de registros frente al ingreso promedio).

- **Mejor Observación:** El uso de `twinx` permite superponer dos métricas en un mismo espacio visual sin que una opaque a la otra. Esto facilita la identificación de correlaciones directas, como verificar si los picos en el nivel de ingresos coinciden con los segmentos de población más representativos del dataset.
- **Precisión Visual:** Al tener escalas independientes (una a la izquierda y otra a la derecha), evitamos que variables con valores pequeños se vean como líneas planas al compararlas con cifras de miles de dólares.
- **Versatilidad en el Análisis:** En el *script* de regresión se utilizaron ambos métodos: gráficos simples para un análisis rápido de tendencias y gráficos con eje doble para aquellos casos donde la densidad de los datos era crítica para validar la robustez de la muestra.

## Análisis de Distribución: Variable Edad

El análisis univariado de la variable `Age` revela una distribución notablemente equilibrada entre los 18 y 70 años.

- **Madurez Laboral:** La alta concentración de individuos entre los 35 y 55 años sugiere que los datos reflejan a personas con una carrera profesional consolidada.

- **Representatividad Justa:** Al presentar una distribución casi uniforme, el análisis garantiza que las conclusiones no están sesgadas hacia una sola generación; es un reflejo fiel de la diversidad de edades en un entorno laboral real.
- **Perfil del Ciudadano:** Estamos analizando a sujetos que, por su edad, probablemente enfrentan decisiones financieras importantes, como hipotecas, educación de dependientes o planes de jubilación.

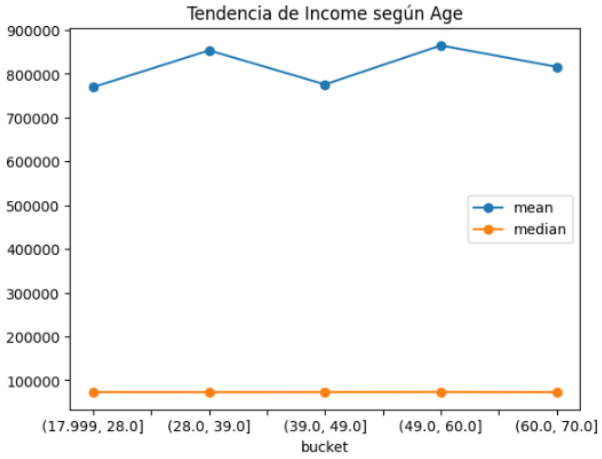


Figura 8: Tendencia de income segun intervalos de edad

# Análisis Exploratorio Bivariado (Relación con el Target)

Este analisis bivariado es fundamental para identificar cómo cada factor influye realmente en los ingresos anuales (*Income*). El objetivo es validar si las tendencias observadas en los datos coinciden con la lógica del mercado y detectar patrones que el modelo de regresión lineal deberá capturar con precisión.

## Relación entre Experiencia Laboral e Ingresos

Este análisis nos permite observar cómo evoluciona la situación económica de una persona a medida que acumula años de trayectoria en el mercado profesional.

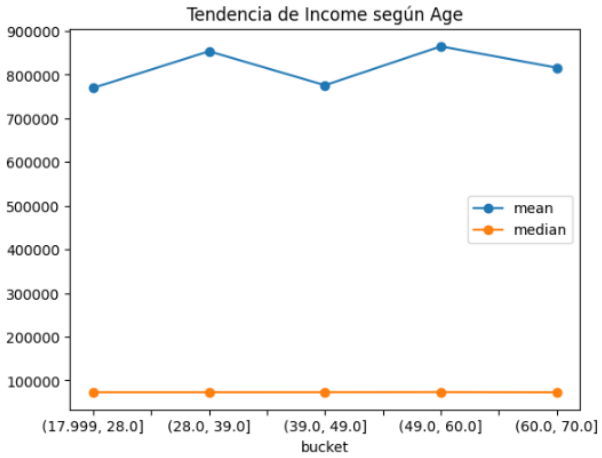


Figura 9: Relación entre rangos de experiencia laboral e ingresos.

Al analizar el comportamiento de los ingresos frente a la experiencia acumulada, se identifican matices muy interesantes sobre la vida laboral:

- **Crecimiento y Consolidación:** Se observa que los ingresos tienden a ser más dinámicos y presentar mayor variabilidad en los primeros 20 años de carrera. Es la etapa donde las personas suelen experimentar ascensos rápidos y cambios significativos de nivel salarial.
- **El Fenómeno de los Expertos:** Aunque la mediana se mantiene relativamente estable, la presencia de numerosos valores atípicos (puntos superiores) en todos los rangos demuestra que siempre existe un grupo de profesionales de alto rendimiento que logra ingresos excepcionales, independientemente de si tienen 10 o 40 años de experiencia.
- **Etapa de Retiro:** En el rango final (40 a 50 años de experiencia), se nota una leve contracción en la base de los ingresos. Esto refleja el paso a una etapa de vida más pasiva o de jubilación, donde la percepción económica principal suele estabilizarse o provenir de fondos previsionales, reduciendo la dispersión que veíamos en la juventud.

# Análisis de Composición Familiar y Recursos

Este bloque del análisis bivariado explora cómo la estructura del hogar se relaciona con el nivel de ingresos. Entender esta dinámica es vital, ya que el tamaño de la familia suele influir en la distribución del gasto y en la necesidad de generar mayores flujos de caja para mantener la sostenibilidad financiera.

## Impacto del Tamaño del Hogar en el Ingreso Anual

El análisis de la variable `Household_Size` frente al target nos permite visualizar si las familias más numerosas presentan una correlación con ingresos más altos, sea por la necesidad de mayor sustento o por la presencia de múltiples fuentes de ingreso en el hogar.

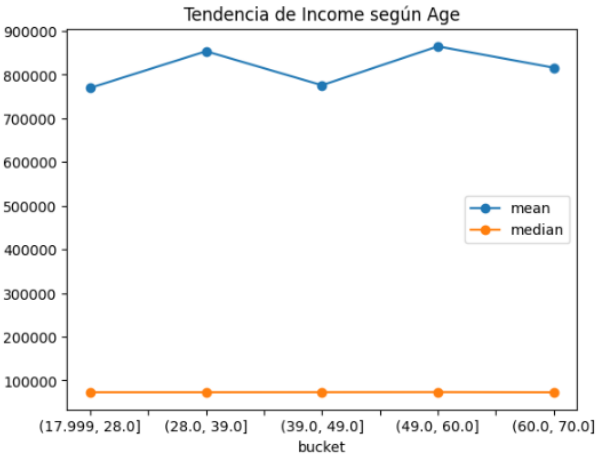


Figura 10: Relación bivariada entre el tamaño del hogar e ingresos anuales.

Al observar el comportamiento de los ingresos según el número de personas en casa, podemos extraer las siguientes conclusiones cotidianas:

- **Estabilidad en la Mediana:** A pesar de que el número de integrantes cambie, el ingreso "típico"(la mediana representada en azul) se mantiene sorprendentemente constante. Esto sugiere que, en promedio, el estilo de vida base de la población no varía drásticamente solo por tener una familia más grande.
- **El peso de los extremos:** El gráfico muestra una gran cantidad de valores atípicos (puntos negros) en todos los grupos. Esto indica que existen hogares, tanto pequeños como grandes, que logran ingresos excepcionales, lo que demuestra que el éxito financiero no está limitado por la cantidad de personas que vivan bajo el mismo techo.
- **Efecto de techo económico:** Se percibe un ligero descenso en la media (línea roja) cuando el hogar llega a los 5 o 6 integrantes, para luego repuntar en hogares de 7. Esto podría interpretarse como una etapa de presión económica donde el ingreso per cápita se diluye, hasta que en hogares muy grandes probablemente se suman nuevos aportantes (hijos adultos o familiares) que elevan nuevamente el promedio total del hogar.

# Análisis de Responsabilidades Familiares

Este apartado examina la relación entre el número de personas dependientes y el nivel de ingresos anuales. Desde una perspectiva de riesgo crediticio, esta variable es fundamental para entender el ingreso disponible real del solicitante, ya que a mayor número de dependientes, mayor es la carga de gastos fijos necesarios para el sustento familiar.

## Relación entre Número de Dependientes e Ingresos

El objetivo de este gráfico es identificar si existe una presión económica que obligue a los individuos con más responsabilidades familiares a buscar niveles de ingresos más altos o si, por el contrario, la carga familiar limita la acumulación de riqueza.

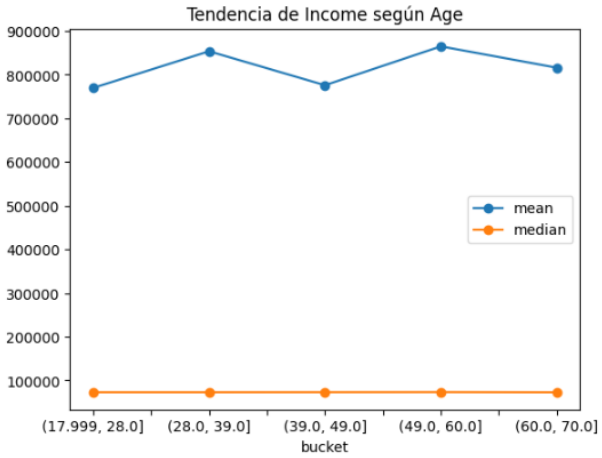


Figura 11: Impacto del número de dependientes en el ingreso anual

Al observar el comportamiento de los ingresos frente a las personas a cargo, podemos notar dinámicas muy humanas en los datos:

- **Resiliencia Financiera:** Es notable ver que la mediana de ingresos se mantiene firme sin importar si la persona tiene 0 o 5 dependientes. Esto nos dice que, en el día a día, las personas logran mantener un estándar de vida base similar, adaptando sus finanzas para que el tamaño de su familia no hunda su nivel de ingresos principal.
- **El Esfuerzo de las Familias Grandes:** Un detalle revelador es que en el grupo de 5 dependientes (familias numerosas), la media de ingresos (punto rojo) tiende a subir. Esto refleja una realidad cotidiana: los padres o jefes de familia con mucha responsabilidad suelen buscar mejores oportunidades o trabajos adicionales para cubrir las necesidades crecientes de su hogar.
- **Casos Excepcionales:** Los puntos negros (valores atípicos) presentes en todos los niveles confirman que el éxito económico no discrimina. Hay personas con familias grandes que logran ingresos altísimos, rompiendo el mito de que tener muchos dependientes es un impedimento para alcanzar niveles salariales de élite.

# Análisis del Ciclo de Vida y Generación de Ingresos

Este apartado investiga cómo evoluciona la capacidad de generar ingresos a lo largo de las distintas etapas de la vida de un individuo. En la modelación de riesgo, la edad no es solo un número demográfico, sino un indicador de estabilidad, madurez profesional y cambios en los patrones de consumo y ahorro.

## Relación entre la edad y el ingreso anual

El objetivo de este análisis es observar si existe un ”pico”de ingresos en ciertas edades o si la estabilidad financiera se mantiene constante a medida que el individuo avanza hacia la etapa de jubilación.

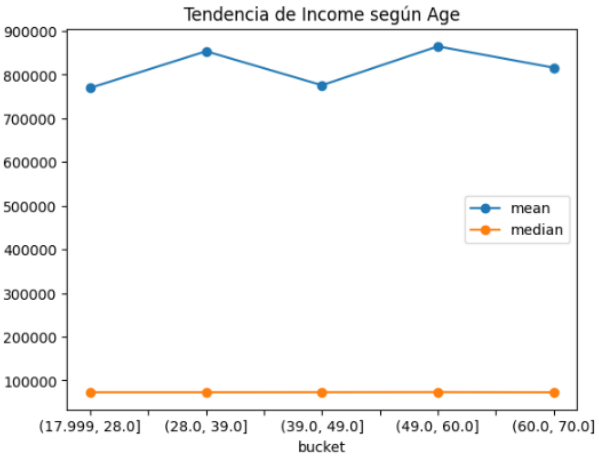


Figura 12: Tendencia de los ingresos promedio y medianos según rangos de edad.

Al analizar la tendencia de los ingresos a través de los años, se identifican patrones que reflejan la realidad de la carrera profesional:

- **Crecimiento en la juventud:** Se observa un incremento notable en los ingresos promedio entre los 18 y los 39 años. Esta es la etapa de "despegue", donde la madurez profesional y las primeras promociones laborales impactan directamente en la billetera de las personas.
- **El bache de la mediana Edad:** Curiosamente, se percibe un ligero descenso en el promedio para el rango de 39 a 49 años. Humanamente, esto suele coincidir con etapas de mayor presión familiar o transiciones de carrera donde la estabilidad se prioriza sobre el crecimiento agresivo del salario.
- **Consolidación y retiro:** Existe un repunte final hacia los 60 años, probablemente impulsado por alcanzar los puestos de mayor jerarquía o el inicio de la percepción de pensiones. Finalmente, hacia los 70 años, la tendencia se estabiliza, reflejando una transición suave hacia una vida financiera más pasiva y predecible.

# Conclusiones finales y siguientes pasos

El presente trabajo desarrolló exitosamente las fases iniciales de un proyecto de Machine Learning aplicado a riesgo crediticio. A través de un proceso estructurado de diagnóstico, limpieza, preprocesamiento y análisis exploratorio, se identificaron variables con alta relevancia predictiva y patrones consistentes con la teoría financiera del incumplimiento.

Como siguientes pasos metodológicos, se recomienda:

- Implementar codificación robusta de variables categóricas (One-Hot Encoding, WoE).
- Entrenar modelos comparativos como Regresión Logística, Árboles de Decisión y métodos *ensemble*.
- Evaluar el desempeño mediante métricas adecuadas al desbalance moderado (ROC-AUC, Precision-Recall, F1-score).
- Analizar umbrales óptimos de decisión y su impacto en políticas de riesgo.

En conjunto, los resultados obtenidos establecen una base sólida para avanzar hacia un sistema predictivo de clasificación crediticia bajo estándares consistentes con analítica bancaria moderna.