Nombre del Proyecto

De muchas señales, algunas destacan: entendiendo el riesgo de mora con datos

Introducción

Este proyecto de minería de datos busca identificar las variables más importantes que explican el riesgo de mora de los clientes, utilizando técnicas de análisis exploratorio, limpieza de datos y selección de características descubriendo patrones clave para futuras decisiones de negocio o modelos predictivos. El dataset es obtenido de la plataforma Kaggle, el cual contiene columnas que simulan datos típicos de un buró de crédito.

El análisis de este tipo de información resulta clave para instituciones financieras, ya que permite comprender mejor el comportamiento de los clientes, predecir riesgos y tomar decisiones más informadas.

Fuente: https://www.kaggle.com/datasets/laotse/credit-risk-dataset

Objetivos

- Realizar un análisis exploratorio de datos (EDA) para comprender la estructura, calidad y distribución del dataset.
- Detectar patrones, relaciones y comportamientos relevantes entre las variables que podrían influir en el estado del préstamo.
- Aplicar técnicas de minería de datos para seleccionar las variables más importantes asociadas al riesgo de mora.
- Sentar las bases para futuros modelos o estrategias de decisión, basándose en el conocimiento obtenido sobre las variables clave.

Herramientas

- Python
- Power BI
- Google Colaboratory

Contenido

1. Planteamiento del problema

No todos los clientes que solicitan un préstamo terminan pagándolo. ¿Qué factores realmente marcan la diferencia entre cumplir y caer en mora?

Aunque se recolectan muchos datos al evaluar un crédito, no siempre está claro cuáles son las señales clave de riesgo. Este proyecto busca identificar esas variables ocultas que permiten anticipar el incumplimiento

2. Conociendo las variables y sus datos

El conjunto de datos tiene una dimensión de 32581 registros y 12 variables, además está en inglés, por lo que se traducen al español para facilitar su comprensión. Entender el significado de cada variable es fundamental para interpretar correctamente la información y orientar el análisis dentro del contexto del negocio.

	A	B	C	D	E	F	G	H	I	J	K	L
1	person_age	person_income	person_home_ownership	person_emp_length	loan_intent	loan_grade	loan_amnt	loan_int_rate	loan_status	loan_percent_income	cb_person_default_on_file	cb_person_cred_hist_length
2	22	59000	RENT	123	PERSONAL	D	35000	16.02	1	0.59	Y	3
3	21	9600	OWN	5	EDUCATION	B	1000	11.14	0	0.1	N	2
4	25	9600	MORTGAGE	1	MEDICAL	C	5500	12.87	1	0.57	N	3
5	23	65500	RENT	4	MEDICAL	C	35000	15.23	1	0.53	N	2
6	24	54400	RENT	8	MEDICAL	C	35000	14.27	1	0.55	Y	4
7	21	9900	OWN	2	VENTURE	A	2500	7.14	1	0.25	N	2
8	26	77100	RENT	8	EDUCATION	B	35000	12.42	1	0.45	N	3
9	24	78956	RENT	5	MEDICAL	B	35000	11.11	1	0.44	N	4
10	24	83000	RENT	8	PERSONAL	A	35000	8.9	1	0.42	N	2
11	21	10000	OWN	6	VENTURE	D	1600	14.74	1	0.16	N	3
12	22	85000	RENT	6	VENTURE	B	35000	10.37	1	0.41	N	4
13	21	10000	OWN	2	HOMEIMPROVEMENT	A	4500	8.63	1	0.45	N	2
14	23	95000	RENT	2	VENTURE	A	35000	7.9	1	0.37	N	2
15	26	108160	RENT	4	EDUCATION	E	35000	18.39	1	0.32	N	4
16	23	115000	RENT	2	EDUCATION	A	35000	7.9	0	0.3	N	4
17	23	500000	MORTGAGE	7	DEBTCONSOLIDATION	B	30000	10.65	0	0.06	N	3
18	23	120000	RENT	0	EDUCATION	A	35000	7.9	0	0.29	N	4
19	23	92111	RENT	7	MEDICAL	F	35000	20.25	1	0.32	N	4
20	23	113000	RENT	8	DEBTCONSOLIDATION	D	35000	18.25	1	0.31	N	4

	Variable	Traducción	Descripción
0	person_age	Edad de la persona	Edad del solicitante del préstamo. Personas más jóvenes pueden tener mayor riesgo debido a menor estabilidad financiera, mientras que personas mayores pueden tener más historial crediticio.
1	person_income	Ingreso de la persona	Cantidad de dinero que la persona gana anualmente. Un mayor ingreso generalmente indica una mejor capacidad de pago.
2	person_home_ownership	Propiedad de vivienda	Indica si la persona es dueña de una vivienda, alquila o tiene otro tipo de propiedad. Los propietarios suelen ser considerados menos riesgosos.
3	person_emp_length	Tiempo de empleo	Número de meses que la persona ha estado empleada. Una mayor duración en el empleo puede indicar estabilidad financiera y menor riesgo de impago.
4	loan_intent	Propósito del préstamo	Razón por la que se solicitó el préstamo (ejemplo: educación, automóvil, vivienda, consolidación de deuda, etc.). Algunos propósitos pueden estar más asociados con mayor riesgo de impago.
5	loan_grade	Calificación del préstamo	Clasificación del préstamo basada en su nivel de riesgo (ejemplo: A, B, C, D…). Mejor calificación = menor riesgo y menor tasa de interés.
6	loan_amnt	Monto del préstamo	Cantidad total de dinero prestado a la persona. Préstamos más grandes pueden representar mayor riesgo si la capacidad de pago del cliente es limitada.
7	loan_int_rate	Tasa de interés	Porcentaje que la persona debe pagar sobre el préstamo. Tasas de interés más altas suelen darse a clientes con mayor riesgo crediticio.
8	loan_status	Estado del préstamo	Indica si el préstamo está en **mora (1)** o **al día (0)**. Es la variable objetivo en un modelo de predicción de riesgo de impago.
9	loan_percent_income	Porcentaje de ingresos	Relación entre el monto del préstamo y los ingresos de la persona. Un porcentaje alto indica que el préstamo representa una gran parte de los ingresos, lo que puede aumentar el riesgo de impago.
10	cb_person_default_on_file	Historial de incumplimiento	Indica si la persona ha tenido un incumplimiento de pago en el pasado. Si ha tenido mora antes, es más probable que vuelva a caer en impago.
11	cb_person_cred_hist_length	Duración del historial crediticio	Número de años que la persona ha tenido crédito. Un historial crediticio más largo suele estar asociado con un menor riesgo, ya que hay más información sobre su comportamiento de pago.

3. Tratamiento de valores nulos

Indican que, por alguna razón, no se registró ningún valor, lo cual puede deberse a errores de captura, información no disponible o simplemente porque no aplicaba en ese caso. Detectar y tratar estos valores es importante, ya que pueden afectar los análisis estadísticos, visualizaciones o resultados de modelos predictivos si no se manejan adecuadamente.

	0
edad_persona	0
ingreso_persona	0
propiedad_vivienda	0
meses_empleo	895
proposito_prestamo	0
calificacion_prestamo	0
monto_prestamo	0
tasa_interes	3116
estado_prestamo	0
porcentaje_ingresos	0
historial_incumplimiento	0
duracion_historial_crediticio	0

Interpretación: se observa que las variables que presentan valores nulos son el tiempo de empleo (en meses) y la tasa de interés.

3.1 Evaluación de la mejor técnica para sustituir valores nulos

Para determinar si sustituir los valores nulos con la media, mediana u otra técnica, es importante analizar, el por qué existen esos nulos. Esto se debe a que no siempre los datos faltan por casualidad. A veces hay una razón detrás de por qué están vacíos, y si los rellenamos sin pensarlo bien, podríamos cambiar el sentido real de la información o sacar conclusiones equivocadas. Por eso, es clave revisar si esos datos se perdieron al azar o si hay algún patrón escondido. Según sea el caso, se decide la mejor forma de rellenarlos sin dañar el análisis.

Se revisan las variables que tienen datos faltantes y se comparan con las demás variables, tanto categóricas como numéricas, usando pruebas estadísticas adecuadas. Esto permite saber si los datos faltantes son aleatorios o no. A continuación, se muestran los resultados:

```
================================================================================================

RESUMEN DE VARIABLES DEPENDIENTES
================================================================================================

Las variables categóricas que afectan a meses_empleo_nulls son:
   propiedad_vivienda (p-value = 0.0000)
   calificacion_prestamo (p-value = 0.0000)
   historial_incumplimiento (p-value = 0.0002)

No se encontró evidencia de que los valores nulos en tasa_interes_nulls dependan de alguna variable categórica.


================================================================================================

RESUMEN DE VARIABLES DEPENDIENTES
================================================================================================

Las variables numéricas que afectan a meses_empleo_nulls son:
   ingreso_persona (p-value = 0.0000)
   monto_prestamo (p-value = 0.0000)
   porcentaje_ingresos (p-value = 0.0000)
   edad_persona (p-value = 0.0101)

No se encontró evidencia de que los valores nulos en tasa_interes_nulls dependan de alguna variable numérica.
```

Interpretación: los datos nulos en la variable tiempo de empleo no son casualidad, hay motivo de sus existencias:

- Personas que apenas están comenzando su vida laboral, posiblemente no tengan experiencia.

- Falta de filtro al momento de llenar el formulario, es un dato importante para la solicitud de crédito, pero no hay forma de controlar su falta de registro.

- Entre otros.

Por otro lado, está la variable tasa de interés que si demuestra casualidad.

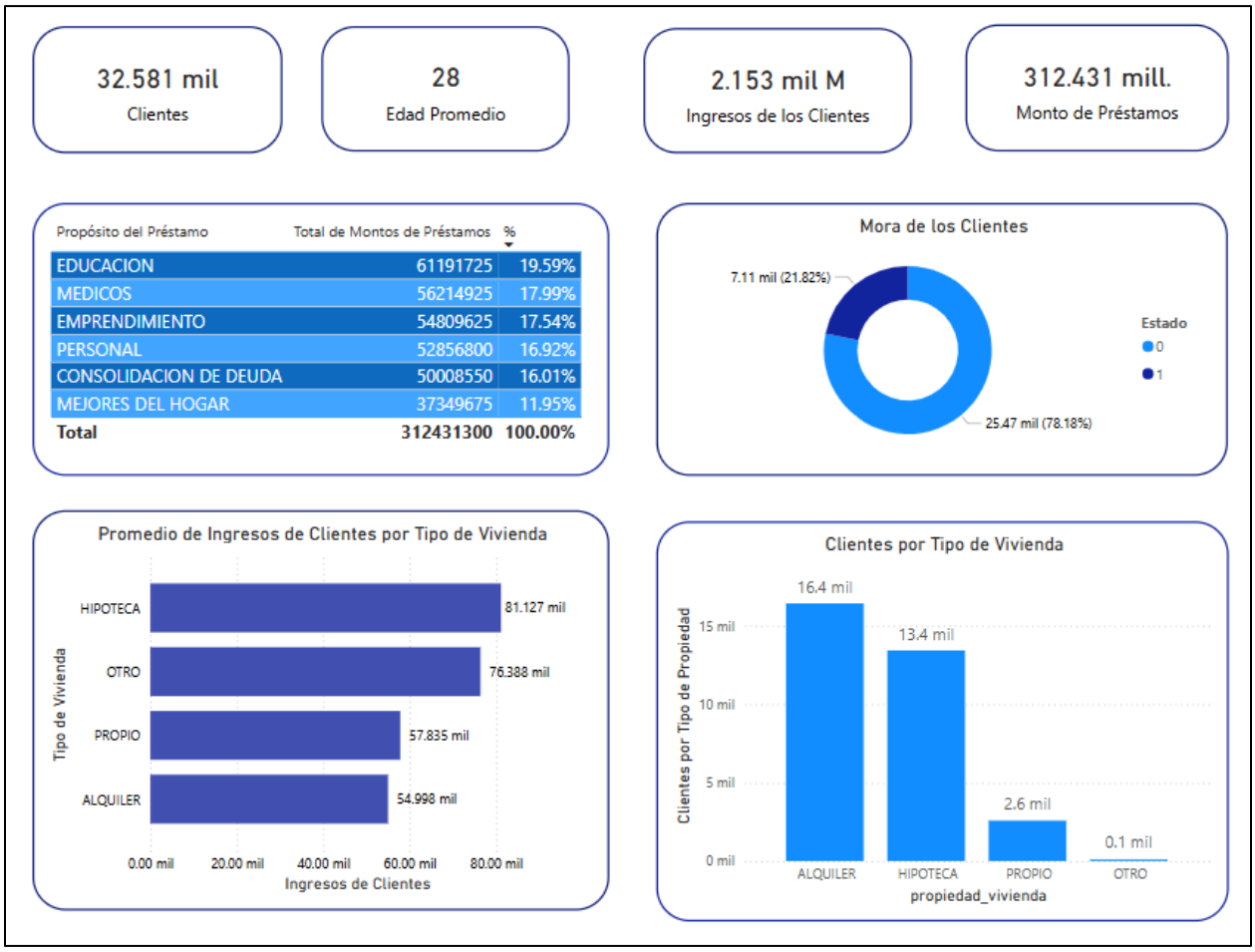
Para sustituir los valores nulos en la tasa de interés se aplicará la mediana mientras que para la variable tiempo de empleo, se aplica una técnica de aprendizaje automático llamada “Vecinos Cercanos” (KNN).

4. Análisis Exploratorio de los Datos (EDA)

Es básicamente entender cómo se comportan y descubrir cosas interesantes antes de hacer modelos o tomar decisiones. Es como revisar bien los ingredientes antes de cocinar.

4.1 Gráficos

Para esta etapa comienzo con algunas visualizaciones del comportamiento de algunas variables de interés.



Interpretación: el tablero muestra una base de datos compuesta por 32,581 clientes, con una edad promedio de 28 años, ingresos totales de aproximadamente 2.15 mil millones y un monto total de préstamos otorgados de 312.431 millones, esto refleja una cartera joven.

La inversión en educación representa el propósito más frecuente entre los clientes, abarcando el 19.59% del total del monto de préstamos. En contraste, las mejoras del hogar constituyen el propósito menos común, con un 11.95% del total.

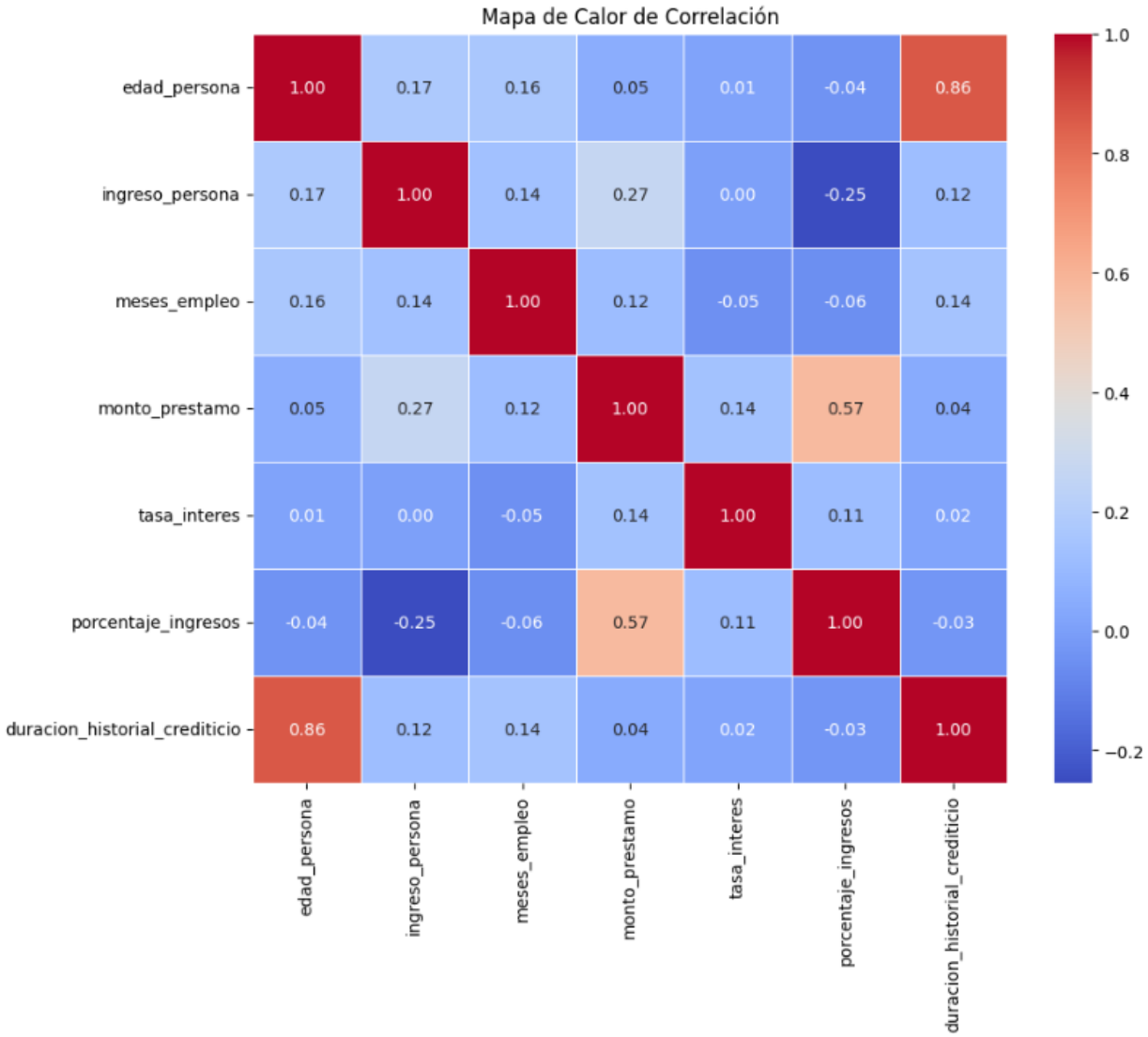
Por otro lado, los propósitos como gastos médicos (17.99%), emprendimiento (17.54%), gastos personales (16.92%) y consolidación de deudas (16.01%) presentan una distribución bastante equilibrada, sin diferencias marcadas entre ellos. Esto sugiere que los clientes destinan sus préstamos al desarrollo y cuidado personal.

Algo interesante que se observa es que, aunque solo 107 clientes tienen un tipo de vivienda clasificado como "otro", estos presentan un promedio de ingresos notablemente alto. En contraste, los clientes que viven en alquiler (16,446) o que tienen vivienda propia sin deuda (2,584) registran ingresos promedio más bajos, especialmente si se comparan con quienes tienen una hipoteca (13,444), grupo que concentra los ingresos promedio más altos entre todos los tipos de vivienda.

4.2 Estadística descriptiva

En esta parte divido las variables en numéricas y categóricas y aplico técnicas estadísticas apropiadas para ambas.

Esta es la matriz de correlación entre las variables numéricas.



Interpretación: como algunas observaciones, puedo mencionar:

- Las personas **más adultas** (edad_persona) suelen tener **más tiempo de historial crediticio** (duracion_historial_crediticio), lo cual es lógico porque han tenido más años para usar productos financieros.
- Cuando una persona pide más dinero prestado (monto_prestamo), normalmente compromete una mayor parte de sus ingresos mensuales (porcentaje_ingresos) al pago del préstamo.
- También se nota que **mientras más gana una persona** (ingreso_persona), **menor es el porcentaje de su ingreso que se va al préstamo** (porcentaje_ingresos), lo cual también tiene sentido

Para trabajar con las variables categóricas, se codifican y así poder aplicar la prueba estadística del Chi – Cuadrado que nos dice si hay dependencia entre categorías o simplemente no tienen nada que ver (son independientes).

Variable	Clave	Código
propiedad_vivienda		
	RENT	0
	OWN	1
	MORTGAGE	2
	OTHER	3
proposito_prestamo		
	PERSONAL	0
	EDUCATION	1
	MEDICAL	2
	VENTURE	3
	HOMEIMPROVEMENT	4
	DEBTCONSOLIDATION	5
calificacion_prestamo		
	A	0
	B	1
	C	2
	D	3
	E	4
	F	5
	G	6

A continuación, algunos resultados:

```
================================================================================
Prueba de Chi-cuadrado: propiedad_vivienda vs proposito_prestamo
================================================================================

Tabla de contingencia:
proposito_prestamo     0     1     2     3     4     5
propiedad_vivienda
0                   2717  3281  3430  2673  1534  2811
1                    446   528   434   786   318    72
2                   2340  2627  2190  2234  1741  2312
3                     18    17    17    26    12    17

Hipótesis:
  H0: propiedad_vivienda y proposito_prestamo son independientes.
  H1: propiedad_vivienda y proposito_prestamo son dependientes.

Resultados:
  Chi-cuadrado: 760.8721
  p-valor: 0.0000
  Grados de libertad: 15
  Conclusión: Se rechaza H0. propiedad_vivienda y proposito_prestamo son dependientes.

================================================================================
Prueba de Chi-cuadrado: proposito_prestamo vs estado_prestamo
================================================================================

Tabla de contingencia:
estado_prestamo        0     1
proposito_prestamo
0                   4423  1098
1                   5342  1111
2                   4450  1621
3                   4872   847
4                   2664   941
5                   3722  1490

Hipótesis:
  H0: proposito_prestamo y estado_prestamo son independientes.
  H1: proposito_prestamo y estado_prestamo son dependientes.

Resultados:
  Chi-cuadrado: 520.5116
  p-valor: 0.0000
  Grados de libertad: 5
  Conclusión: Se rechaza H0. proposito_prestamo y estado_prestamo son dependientes.
```

```
================================================================================
Prueba de Chi-cuadrado: calificacion_prestamo vs estado_prestamo
================================================================================

Tabla de contingencia:
estado_prestamo           0     1
calificacion_prestamo
0                      9704  1073
1                      8750  1701
2                      5119  1339
3                      1485  2141
4                       343   621
5                        71   170
6                         1    63

Hipótesis:
  H0: calificacion_prestamo y estado_prestamo son independientes.
  H1: calificacion_prestamo y estado_prestamo son dependientes.

Resultados:
  Chi-cuadrado: 5609.1842
  p-valor: 0.0000
  Grados de libertad: 6
  Conclusión: Se rechaza H0. calificacion_prestamo y estado_prestamo son dependientes.
```

Interpretación:

- El tipo de vivienda que tiene una persona (propiedad_vivienda), ya sea propia, alquilada o familiar, sí tiene relación con el motivo por el que solicita un préstamo (proposito_prestamo). Por ejemplo, las personas que alquilan podrían estar pidiendo préstamos con fines distintos a quienes ya tienen casa propia.
- El motivo por el cual una persona solicita un préstamo (proposito_prestamo) está relacionado con cómo termina ese préstamo (estado_prestamo), es decir, si fue pagado o si terminó en mora. Por ejemplo, algunos motivos como préstamos para negocios pueden estar más expuestos al impago que otros como préstamos para salud o estudios.
- La calificación que se le da a un préstamo al momento de otorgarlo (calificacion_prestamo) está muy relacionada con cómo termina ese préstamo (estado_prestamo). Quiere decir que una mala calificación inicial suele estar asociada a un mayor riesgo de que el préstamo no se pague bien, mientras que una buena calificación está más relacionada con préstamos que se pagan sin problemas.

5. Modelado

El modelado es el proceso de crear una representación estructurada de la realidad usando datos, reglas, fórmulas o relaciones entre variables. Esta representación (modelo) nos permite entender cómo funciona algo, analizar situaciones y predecir posibles resultados sin tener que experimentar directamente en el mundo real.

Este proceso nos ayuda a:

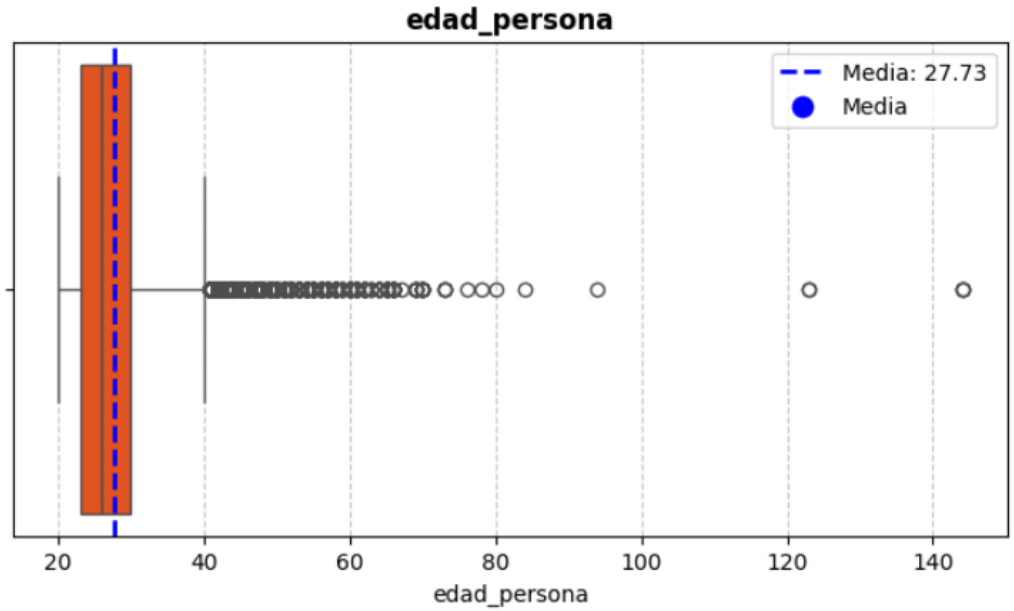
- ✓ Analizar problemas complejos de forma más sencilla.
- ✓ Explorar escenarios antes de tomar decisiones.
- ✓ Predecir comportamientos futuros (como el riesgo de que un cliente no pague).
- ✓ Tomar decisiones basadas en datos y evidencia, en lugar de suposiciones.

5.1 Tratamiento de datos atípicos

Los datos atípicos son aquellos que se alejan mucho del resto de los datos. Es decir, son valores que no siguen el patrón general del conjunto de datos y afectan el desempeño del modelo. Esto aplica solo para variables numéricas. porque pueden distorsionar los resultados del análisis y hacer que los modelos aprendan patrones incorrectos, afectando su precisión y capacidad de generalización.

Además, los atípicos pueden influir demasiado en estadísticas como la media y la desviación estándar, y hacer que las visualizaciones sean engañosas. Detectarlos y decidir si deben corregirse, transformarse o eliminarse permite obtener conclusiones más confiables y modelos más robustos.

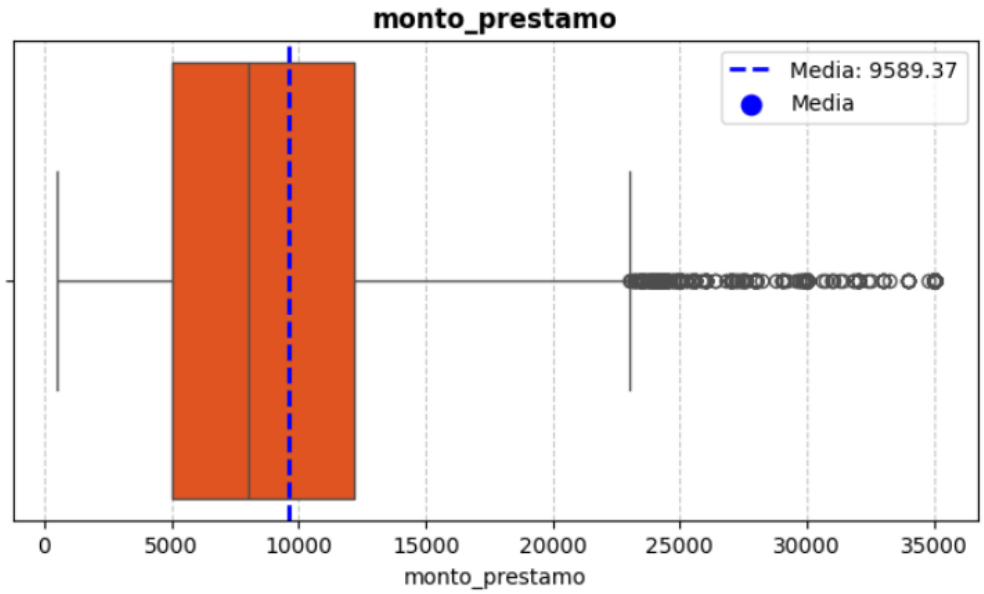
A continuación, algunos ejemplos:



Estadísticas de edad_persona:
- Media: 27.73
- Desviación Estándar: 6.35
- Min (Bigote Real): 20.00
- Q1 (25%): 23.00
- Mediana (50%): 26.00
- Q3 (75%): 30.00
- Max (Bigote Real): 40.00
- Cantidad de Outliers: 1494

Interpretación:

Interpretación: el gráfico muestra la distribución de la variable edad_persona, donde la mitad de los clientes se encuentran entre los 23 y 30 años. La edad promedio es de 27.73 años y la mediana es de 26 años, lo que indica que más de la mitad de las personas tienen menos de 26. También se observa una gran cantidad de valores atípicos (outliers) por encima de los 40 años, incluyendo algunos casos extremos que superan los 100 años, lo cual podría deberse a errores en los datos o casos poco comunes. Esto sugiere que sería conveniente revisar o tratar estos valores antes de usarlos en un análisis o modelo.

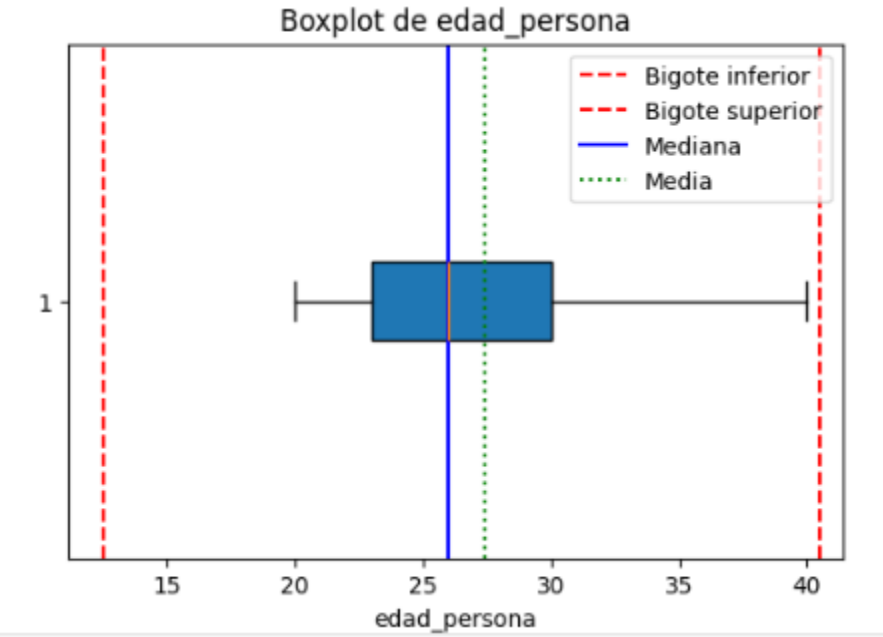


Estadísticas de monto_prestamo:
- Media: 9589.37
- Desviación Estándar: 6322.09
- Min (Bigote Real): 500.00
- Q1 (25%): 5000.00
- Mediana (50%): 8000.00
- Q3 (75%): 12200.00
- Max (Bigote Real): 23000.00
- Cantidad de Outliers: 1689

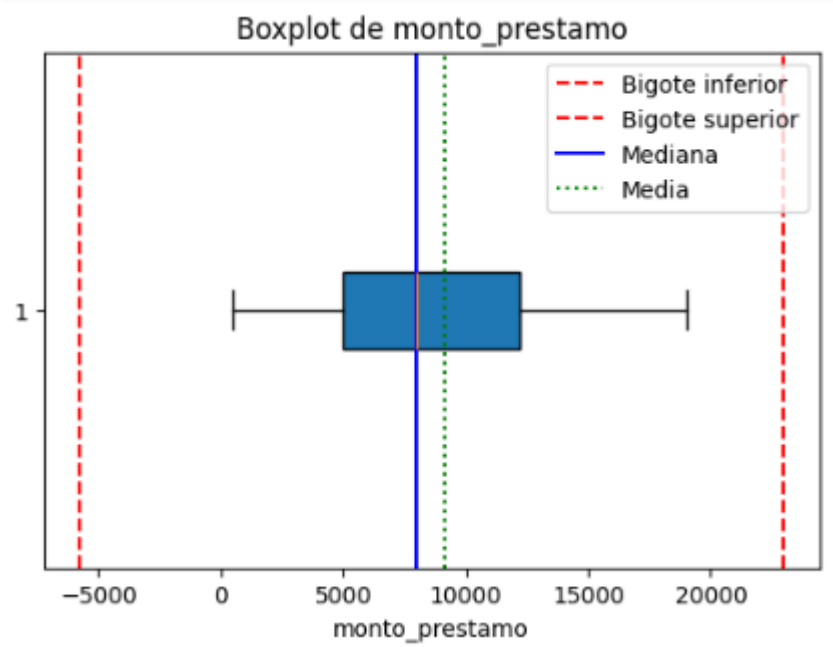
Interpretación: el gráfico de la variable monto_prestamo muestra que la mitad de los préstamos otorgados están entre 5,000 y 12,200 unidades monetarias. Se identifican muchos valores atípicos (1,689 outliers) por encima de los 23,000, algunos incluso llegando a los 35,000, lo cual sugiere que existen préstamos significativamente más altos que el resto y que podrían analizarse por separado o tratarse antes de aplicar un modelo.

Como técnica de tratamiento de datos atípicos, se usa la winsorización. Es una técnica para suavizar los datos atípicos, sin eliminarlos por completo. En lugar de borrar esos valores raros (muy altos o bajos), lo que se hace es reemplazarlos por un valor límite más razonable, como los percentiles. Para este caso se usó el percentil 95.

Resultados:




```
Análisis de la columna: edad_persona
  - Media: 27.42
  - Desviación Estándar: 5.22
  - Q1 (Percentil 25): 23.00
  - Mediana (Q2 - Percentil 50): 26.00
  - Q3 (Percentil 75): 30.00
  - IQR (Q3 - Q1): 7.00
  - Mínimo: 20.00
  - Máximo: 40.00
  - Bigote Inferior (Q1 - 1.5*IQR): 12.50
  - Bigote Superior (Q3 + 1.5*IQR): 40.50
  - Atípicos por debajo del bigote inferior: 0
  - Atípicos por encima del bigote superior: 0
  - Total de atípicos: 0
```



Análisis de la columna: monto_prestamo
- Media: 9132.03
- Desviación Estándar: 5240.77
- Q1 (Percentil 25): 5000.00
- Mediana (Q2 - Percentil 50): 8000.00
- Q3 (Percentil 75): 12200.00
- IQR (Q3 - Q1): 7200.00
- Mínimo: 500.00
- Máximo: 19000.00
- Bigote Inferior (Q1 - 1.5*IQR): -5800.00
- Bigote Superior (Q3 + 1.5*IQR): 23000.00
- Atípicos por debajo del bigote inferior: 0
- Atípicos por encima del bigote superior: 0
- Total de atípicos: 0

5.2 Estadarización

La estandarización es un proceso que convierte todas las variables numéricas a una misma escala, sin cambiar su significado. Esto se hace para que ninguna variable tenga más peso solo por tener números más grandes o pequeños. Es importante porque ayuda a que los modelos funcionen mejor, aprendan más rápido y tomen decisiones más justas al comparar diferentes tipos de datos.

Muestra antes de estandarización:

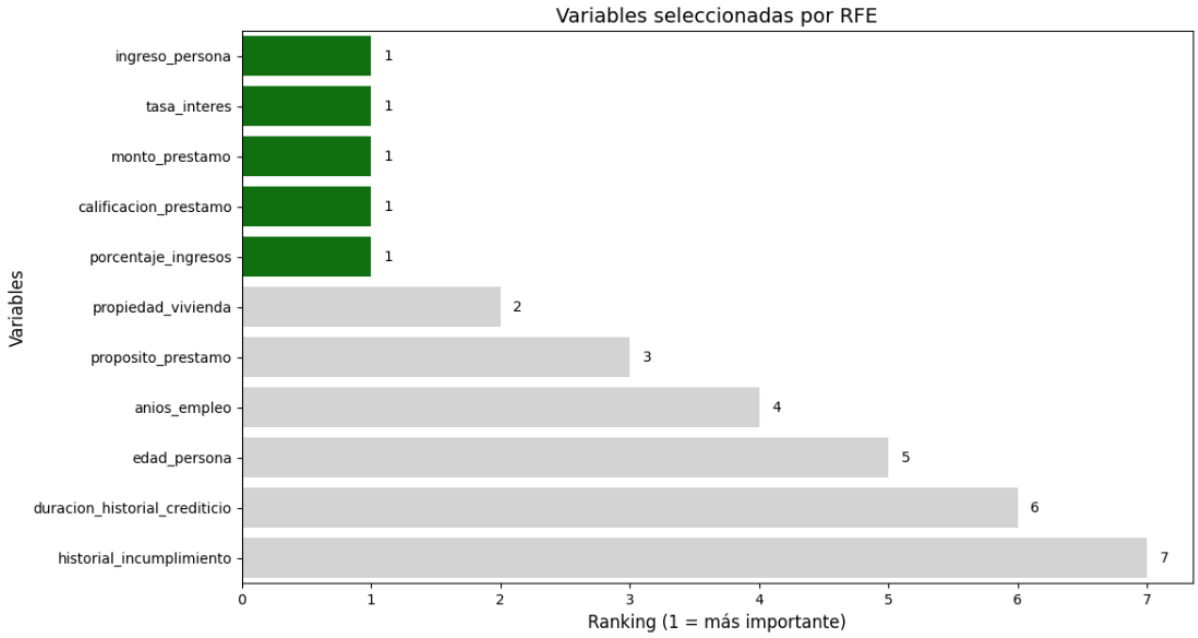
	edad_persona	ingreso_persona	propiedad_vivienda
0	22	59000.00	0
1	21	9600.00	1
2	25	9600.00	2
3	23	65500.00	0
4	24	54400.00	0
5	21	9900.00	1
6	26	77100.00	0
7	24	78956.00	0
8	24	83000.00	0
9	21	10000.00	1

Muestra después de estandarización:

	edad_persona	ingreso_persona	propiedad_vivienda
0	-0.90	-0.10	0
1	-1.06	-1.67	1
2	-0.43	-1.67	2
3	-0.75	0.10	0
4	-0.59	-0.25	0
5	-1.06	-1.66	1
6	-0.27	0.47	0
7	-0.59	0.53	0
8	-0.59	0.66	0
9	-1.06	-1.66	1

5.3 RFE (Recursive Feature Elimination)

La Eliminación Recursiva de Características es una técnica que ayuda a seleccionar las variables más importantes en un análisis. En minería de datos, su importancia radica en que permite reducir la cantidad de datos innecesarios, mejorar la precisión de los modelos y acelerar el procesamiento. Al enfocarse solo en las variables que más aportan valor, RFE facilita encontrar patrones más claros y útiles en grandes volúmenes de información, para aplicar el RFE se utilizó un modelo de Bosque Aleatorio y como variable objetivo estado_prestamo.

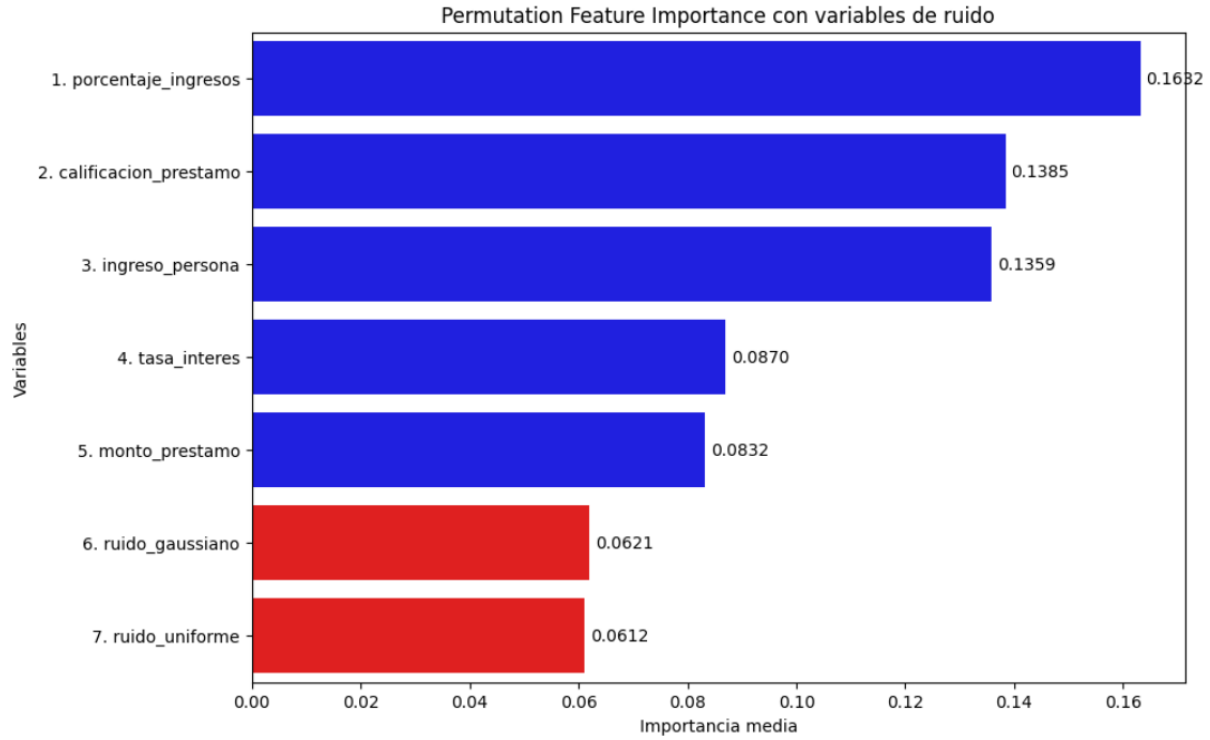


Interpretación: se considera que las 5 primeras variables son las más importantes. Se observa que esas variables están relacionadas con el ingreso y las condiciones del préstamo (como monto, tasa e ingreso) concluyendo son más útiles para predecir la mora que otras variables como edad, historial o empleo.

5.4 PFI (Permutation Feature Importance)

La Permutación de Importancia de Características es una técnica que evalúa qué tan importante es una variable midiendo cuánto empeora el rendimiento del modelo cuando se desordena aleatoriamente esa variable. En minería de datos, es importante porque permite identificar de forma clara y directa cuáles variables realmente influyen en el modelo, sin depender del tipo de modelo usado.

Se usa el modelo de Bosque Aleatorio para aplicar el PFI a las variables obtenidas en el proceso de RFE y como variable objetivo estado_prestamo.



Interpretación: las barras azules representan las variables reales del análisis, y las rojas, las variables de ruido agregadas para comparar. Vemos que porcentaje_ingresos, calificación_prestamo e ingreso_persona son las más importantes, ya que cuando se alteran, el modelo pierde precisión. En cambio, las variables de ruido (barras rojas) tienen poca influencia (valores bajos), lo cual es bueno porque significa que el modelo no está aprendiendo patrones falsos o ruido, y está enfocándose en la información útil.

Aplicar primero RFE y luego PFI trae varios beneficios, ya que se combinan dos enfoques para seleccionar variables: RFE reduce el número de variables quedándose con las más relevantes según el modelo, y luego PFI evalúa con mayor precisión cuál de esas variables seleccionadas tiene más impacto en el rendimiento. Esto permite construir modelos más

simples, rápidos y fáciles de interpretar, sin perder calidad en las predicciones, enfocándose solo en las variables que realmente aportan valor.

6. Conclusiones

- ✓ Al revisar los datos, se detectaron algunos vacíos y errores, especialmente en el tiempo de empleo y la tasa de interés. Se aplicaron técnicas adecuadas para tratarlos sin afectar la calidad del análisis.
- ✓ La mayoría de los clientes piden préstamos principalmente para educación, salud, emprendimiento y gastos personales.
- ✓ Al aplicar herramientas de minería de datos, se identificaron las variables más influyentes en la mora, siendo las principales: el **porcentaje del ingreso comprometido en el préstamo**, **la calificación del préstamo** y **el nivel de ingreso del cliente**.
- ✓ Este análisis no solo ayuda a entender mejor el comportamiento financiero de los clientes, sino que también sienta una base sólida para tomar mejores decisiones en futuros modelos de riesgo o estrategias de evaluación crediticia.