Instituto Tecnológico de Monterrey

Escuela de Ingeniería y Ciencias

Reto Dentix

Entrega Final

Sara Garza Reyna A01612743

Felipe de Jesús Damián Rodríguez A01707246

María Jose Ocharte Álvarez A01707709

Juan Felipe Carmona González

Yiby Karolina Morales Pinto

Aplicación de métodos multivariados en ciencia de datos (Gpo 602)

19 de Octubre de 2025

Índice

Introducción

A lo largo del desarrollo de este proyecto se trabajará con los datos de las Clínicas Dentix, las cuales ofrecen servicios odontológicos y opciones de financiamiento para facilitar el acceso a ellos. La base de datos proporcionada contiene miles de registros de pacientes junto con la información de sus respectivos financiamientos, lo que permite analizar sus perfiles. Mediante el procesamiento y análisis de estos datos, se busca comprender las principales métricas asociadas al cumplimiento de los pagos, en especial la mora y monto de préstamo a aprobar. Este análisis permitirá identificar posibles factores que influyen en el retraso o incumplimiento de los pagos, brindando información valiosa para la toma de decisiones estratégicas de la empresa. Finalmente, se pretende utilizar el conocimiento obtenido para desarrollar modelos predictivos que permitan anticipar el riesgo de un crédito y el resultado del análisis de futuros perfiles.

Las clínicas dentales Dentix se encuentran entre las más reconocidas de Colombia, con una trayectoria de más de diez años, se posicionan como la empresa con mayor cantidad de pacientes atendidos y por lo tanto con mayores ingresos de facturación bruta. Gracias a su equipo experto y a su tecnología de vanguardia como el TAC 3D o el CAD CAM, garantizan un servicio de excelente calidad.

1 Análisis de métricas

Una de las métricas más importantes a tomar en cuenta es la mora. La cual, se refiere al retraso en el cumplimiento de una obligación de pago por parte del cliente. En el caso de los créditos ofrecidos por las Clínicas Dentix, representa el número de días que transcurren desde la fecha de vencimiento del pago hasta el momento en que éste se realiza. Este indicador es fundamental para medir el nivel de riesgo crediticio y la calidad de la cartera.

Para analizar este comportamiento, se utilizan las franjas de mora, que agrupan los créditos según su nivel de atraso. Estas franjas permiten identificar distintos grados de riesgo, en el caso de los pacientes calificados, observamos las siguientes nueve clasificaciones según los días de retraso.

1. Al día
2. 1 a 30
3. 31 a 60
4. 61 a 90
5. 91 a 120
6. 121 a 150
7. 151 a 180
8. 181 a 360
9. > 360

Con esta información se busca predecir qué clientes superarán los 30 días de atraso, es decir, aquellos que pasarán a una franja de riesgo relevante para el cobro.

El propósito del modelado predictivo es estimar la probabilidad de que un crédito caiga en mora igual o superior a 30 días. Para evaluar el desempeño del modelo se emplean distintas métricas de desempeño.

- **Coeficiente de determinación** (R^2): mide la proporción de variabilidad de los datos que el modelo es capaz de explicar para modelos de regresión.
- **Matriz de confusión:** permite visualizar los aciertos y errores en las predicciones, diferenciando entre falsos positivos y falsos negativos.
- **Precisión**: indica la proporción de las predicciones positivas fueron correctas.
- **F1-Score:** combina la precisión y la sensibilidad en una sola métrica, dando una evaluación balanceada del rendimiento del modelo.
- **ROC - AC**: métodos para evaluar el rendimiento de los modelos de clasificación binaria, se grafica la tasa de verdaderos positivos frente a la tasa de falsos positivos, posteriormente se calcula el área bajo esa curva y eso proporciona un único valor numérico que indica la capacidad del modelo para distinguir entre clases.

Estas métricas permitiran determinar qué tan bien el modelo identifica los créditos con alto riesgo de mora y ayudan a mejorar la toma de decisiones respecto a la aprobación y seguimiento de los créditos. Se utilizan conforme al modelo a evaluar.

2 Auditoría de datos

Durante la etapa de exploración y limpieza de datos, se identificaron algunas variables con presencia de valores faltantes. Entre las que cuentan con un mayor porcentaje de ausencias se encuentran "No. personas a cargo", "Otros ingresos", "Total ingresos", "Pasivos", "Operación moneda extranjera", etc. Mientras que la mayoría de variables categóricas presentan una cantidad mínima de registros faltantes, lo que sugiere una buena consistencia en el conjunto de datos.

Tomando en cuenta que la base de datos cuenta con $46,328$ registros, se muestra el siguiente cuadro con la información en concreto de las variables con más de $40\,\%$ de valores nulos.

Cuadro 1: Variables con más de $40\,\%$ de valores faltantes

Variable	**Cantidad de valores nulos**
No personas a cargo	24,286
Otros ingresos	36,663
Total ingresos	29,551
Cuota de créditos	20,787
Total egresos	29,551
Pasivos	24,213
Financiera	46,327
Operación moneda extranjera	27,292
Desistimiento	46,179

De estas variables, conservamos únicamente "Total ingresos", ya que aporta información esencial para evaluar la capacidad de pago de los clientes y resulta determinante en el análisis de riesgo crediticio. Sin embargo, el resto no se consideran relevantes y la falta de datos sugiere que el modelo tendría mejor rendimiento sin ellas. Por un lado nos parecía interesante estudiar a las personas con un desistimiento positivo, mas solo representan al $0.3\,\%$ de los datos.

Más adelante en el análisis, se identificaron y eliminaron las variables que no aportaban información relevante o presentaban problemas de calidad. En primer lugar, se borraron

las columnas “Dirección”, “Código de Confirmación”, “Pagare_id” y “Número de crédito”, ya que únicamente contenían identificadores o datos redundantes con el estrato. También se eliminaron las variables relacionadas con la financiera, el seguro y el aval, debido a que no aportaban valor analítico y presentaban registros únicos. Posteriormente, se descartaron los registros de operaciones en moneda extranjera (“Operación moneda extranjera”) y la variable “Comercial”, por considerarse irrelevantes, ya que no aportan información acerca de la salud financiera del paciente. El campo “Lugar de nacimiento” fue removido porque ya se contaba con la información de formalización, mientras que la fecha de nacimiento se transformó en edad para facilitar su análisis. Las fechas de expedición y aprobación se combinaron en una sola variable que representa el tiempo que tardó en aprobarse el crédito. También se eliminaron outliers en la variable “Tiempo de actividad”, pues existían registros con números como 1900. Finalmente, se ajustó el índice.

El siguiente paso es codificar las variables restantes. En esta etapa se identificaron las variables categóricas y se clasificaron en ordinales y nominales. Las ordinales son “Nivel de estudios”, “Franja de Mora”, “Estado civil”, “Género”, “Tipo de Vivienda” y “Tipo de Contrato”, las cuales se codificaron mediante un mapeo numérico que refleja la jerarquía.

En cuanto a las nominales, se conservaron solo las más relevantes, que son “Actividad Económica”, “Profesión”, “Ocupación”, “Barrio”, “Departamento”, “Ciudad”, “Incidencia de Formalización” y “Clínica”. Y se eliminaron las no significativas para reducir la dimensionalidad y mantener la coherencia del análisis. Así mismo, se creó un nuevo Data Frame que contiene únicamente las variables cuantitativas para el análisis de componentes principales. En el siguiente cuadro informativo se muestran las variables que fueron codificadas con su respectivo significado, para mantener el orden en el posterior análisis.

Cuadro 2: Codificación de variables ordinales

Variable	Categoría original	Valor codificado
Franja de mora	01. Al día	0
	02. 1 a 30	1
	03. 31 a 60	2
	04. 61 a 90	3
	05. 91 a 120	4
	06. 121 a 150	5
	07. 151 a 180	6
	08. 181 a 360	7
	09. >360	8
Nivel de estudios	Primaria	1
	Bachillerato	2
	Técnico	3
	Tecnólogo	4
	Universitario	5
	Licenciatura	6
	Especialización	7
	Maestría	8
	Doctorado / Postdoctorado	9
Estado civil	Soltero (a)	0
	Viudo (a)	1
	Divorciado (a) / Separado (a)	2
	Unión libre	3
	Casado (a)	4
Género	M (Masculino)	0
	F (Femenino)	1
	NB (No Binario)	2
Tipo de vivienda	Propia	1
	Familiar	2
	Arrendada	3
	Propia con crédito	4
Tipo de contrato	Indefinido	0
	No aplica	1
	Prestación de servicios	2
	Fijo	3

3 Análisis exploratorio multivariado

El análisis se realizó utilizando un conjunto de variables numéricas provenientes de la base de datos ya limpia y en variables numéricas. El objetivo principal fue examinar las relaciones lineales entre variables, estudiar la estructura de varianzas y covarianzas, y aplicar un Análisis de Componentes Principales (PCA) para reducir la dimensionalidad preservando la mayor cantidad de información posible.

Se calculó la matriz de correlación utilizando el método de Pearson. Esto permitió identificar relaciones lineales fuertes, moderadas o débiles entre las variables, así como posibles redundancias en la información.

- Las correlaciones positivas indican que ambas variables tienden a aumentar juntas.
- Las correlaciones negativas indican una relación inversa.
- Valores cercanos a 0 implican escasa relación lineal.

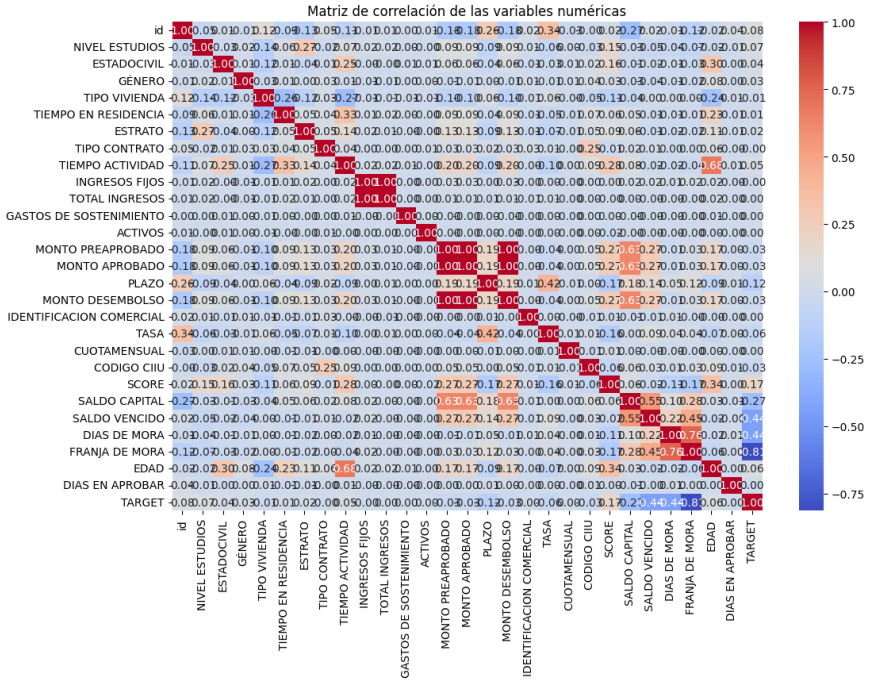

Figura 1: Heatmap obtenido

Además la matriz de covarianza se utilizó para analizar cómo varían las variables conjuntamente. A diferencia de la correlación, sus valores dependen de la escala original de las variables.

- Covarianzas altas sugieren variaciones conjuntas significativas.
- Covarianzas negativas indican que una variable tiende a aumentar mientras la otra disminuye.

Previo al PCA, se verificó la dimensionalidad del conjunto de datos y el número de variables disponibles. Todas son adecuadas para el análisis al ser continuas.

Se estandarizaron las variables para evitar que la escala afectara el cálculo de los componentes principales.

Se obtuvo la proporción de varianza explicada por cada componente principal, lo cual permite determinar cuántos componentes son necesarios para capturar la estructura subyacente del conjunto de datos.

3.1 Interpretación

El análisis multivariado permitió identificar patrones de correlación entre variables, evaluar la redundancia mediante covarianzas y comprender la estructura interna de los datos mediante componentes principales.

Variable	PC1	PC2
ACTIVOS	-0.0137273	0.1066705
CODIGO CIIU	-0.0064167	0.0654342
CUOTAMENSUAL	-0.0581409	0.1858700
DIAS EN APROBAR	0.0000001	-0.0000002
EDAD	-0.0317285	0.1345375
ESTADOCIVIL	-0.0097882	0.0082021
ESTRATO	-0.0789313	0.7743562
GASTOS DE SOSTENIMIENTO	0.0298572	-0.0542305
GÉNERO	0.0085923	-0.0011184
IDENTIFICACION COMERCIAL	0.0064544	0.0300936
INGRESOS FIJOS	-0.0657912	0.3920287
NIVEL ESTUDIOS	-0.0308308	0.2468941
PLAZO	0.2496083	-0.0972191
SCORE	-0.0660272	0.1255733
TASA	0.9456681	0.1759246
TIEMPO ACTIVIDAD	-0.0414925	0.1562957
TIEMPO EN RESIDENCIA	-0.0208477	0.0866724
TIPO CONTRATO	-0.0043459	0.0523987
TIPO VIVIENDA	0.0284197	-0.1375575
TOTAL INGRESOS	-0.1372466	0.0610580

Figura 2: Evaluación de los Componentes Principales

En términos de PC1, la variable con más peso es la tasa. A mayor tasa, mayor es PC1. Se ve como a medida que la tasa crece, también crece el plazo. En la dirección opuesta se

tiene el total de ingresos, indicando como las personas con más ingresos suelen tener la tasa ligeramente más baja. Este PC es el tamaño de la tasa.

En términos de PC2, el estrato y ingresos son lo que más pesa. Mientras mayor sean los ingresos y el estrato socioeconómico de la persona, mayor es el PC2. Los estudios, la couta mensual, la tasa, entre otros también se mueven en esta dirección con menor peso. Este PC es el del éxito económico. Mientras mayor sea esta métrica, más estabilidad y mejor salario.

Ahora que ya se conoce el significado de los PCs, se puede graficar con la mora, para validar la hipótesis que se tiene sobre la relación de los grupos con la franja de mora.

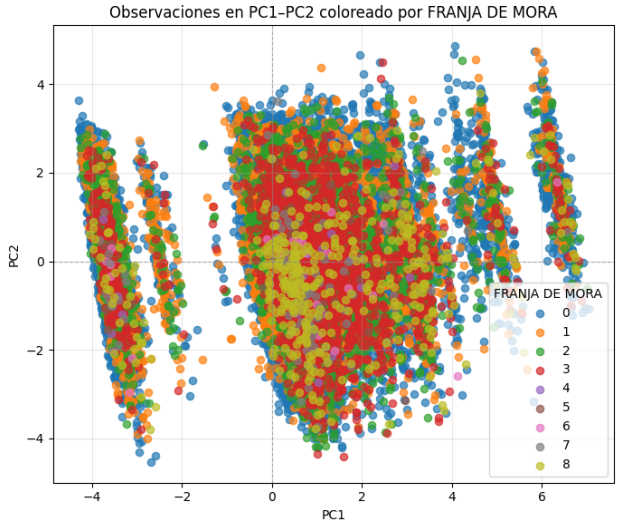

Figura 3: Componentes Principales coloreados por FRANJA DE MORA

4 Modelo de mora ≥ 30

El propósito del inciso es definir un modelo inicial que prediga la probabilidad de que un paciente tenga una mora mayor a treinta días según diferentes variables. Este modelo denominado como `df1` servirá de base para ser comparado con algunos algoritmos más complejos más adelante. Se utilizará la base de datos obtenida en el punto anterior, la cuál solo tiene variables numéricas listas para un modelo.

4.1 Preparación de datos

Con el fin de obtener un mejor rendimiento en el modelo se decidió eliminar ciertas variables que sesgan la predicción de probabilidad, entre ellas las siguientes columnas.

```
drop = ['TARGET', 'FRANJA DE MORA', 'SALDO VENCIDO', 'SALDO CAPITAL',
↪  'SCORE', 'PLAZO', 'TASA', 'MONTO PREAPROBADO', 'CUOTAMENSUAL', 'MONTO
↪  APROBADO', 'id']
```

Posteriormente definimos nuestra variable objetivo y como una columna binaria, en la que si el registro cumple con la condición de tener una mora mayor a treinta se registra como 1. De lo contrario es 0. Con estas modificaciones y considerando que nuestra base de datos ya pasó por un proceso de limpieza en el inciso anterior, se puede proceder a la definición del modelo.

4.2 Definición de modelo

El primer paso es dividir nuestros datos en dos conjuntos, el de entrenamiento y el de prueba. Esto lo hicimos utilizando el método de `stratify` para mantener un balance de clases entre las dos partes. Se obtuvieron las siguientes proporciones.

```
0=<30, 1=30
DIAS DE MORA
0    0.842
1    0.158
```

Estas proporciones de datos exponen una clara minoría entre los pacientes con más de treinta días de deuda. Lo cual, significa que la mayoría de los registros cuentan con una mora saludable. Una buena noticia para la salud financiera de los registros en las clínicas Dentix, sin embargo, puede entorpecer el rendimineto de predicción de nuestro modelo.

Ya con ambos conjuntos divididos se decidió hacer uso de un `pipeline` de la librería de `scikit-learn`. Esta herramienta permite encadenar múltiples pasos del proceso de modelado dentro de un solo objeto. En lugar de ejecutar por separado tareas como imputación de valores faltantes, escalamiento de variables o entrenamiento del modelo, el pipeline integra cada uno de estos pasos en una secuencia fija. Esto mantiene el proceso reproducible para la futura comparación y mejoramiento del algoritmo.

Se escogieron dos modelos iniciales para evaluar. La regresión logística especificada en las instrucciones y un árbol de decisión simple. Después de implementarlos, podemos revisar las métricas que definimos previamente en el reporte 3.

4.3 Métricas

Se observan los siguientes resultados.

```
=== regresión logística ===
matriz de confusión [[TN FP],[FN TP]]:
 [[6280 5295]
 [ 936 1234]]

reporte de clasificación:
              precision    recall  f1-score   support

           0      0.870     0.543     0.668     11575
           1      0.189     0.569     0.284      2170

    accuracy                          0.547     13745
   macro avg      0.530     0.556     0.476     13745
weighted avg      0.763     0.547     0.608     13745

ROC-AUC: 0.578
```

La regresión logística muestra un desempeño inicial consistente con la naturaleza desbalanceada del conjunto de datos que habíamos mencionado. El modelo logra identificar correctamente a más de la mitad de los clientes de clase 1 (con mora $\geq$ 30 días), reflejado en un recall positivo de 0.569, lo cual es adecuado para un modelo base que busca no perder demasiados casos de morosidad. Sin embargo, su precisión positiva es baja (0.189), indicando que una proporción considerable de los clientes señalados como morosos finalmente no lo son, lo que se traduce en un número elevado de falsos positivos. El AUC de 0.578 confirma que el modelo ofrece una capacidad de discriminación limitada, aunque todavía superior al desempeño aleatorio. En conjunto, estos resultados muestran que la regresión logística es útil como modelo inicial de referencia, proporcionando una línea base sobre la cual se construyen y comparan los modelos más avanzados del Punto 5.

```
=== árbol de decisión (simple) ===
matriz de confusión [[TN FP],[FN TP]]:
 [[6701 4874]
 [ 968 1202]]

reporte de clasificación:
              precision    recall  f1-score   support

           0      0.874     0.579     0.696     11575
           1      0.198     0.554     0.292      2170

    accuracy                          0.575     13745
   macro avg      0.536     0.566     0.494     13745
weighted avg      0.767     0.575     0.633     13745

ROC-AUC: 0.594
```

Para el árbol de decisión se observa un comportamiento similar al del modelo logístico, aunque con un desempeño ligeramente superior en varios indicadores clave. En términos de detección de pacientes morosos, el modelo alcanza un recall positivo de 0.554, comparable al de la regresión logística, lo que indica que también identifica a poco más de la mitad de los clientes con mora ≥ 30 días. Su precisión positiva mejora ligeramente (0.198) respecto al modelo previo, pero continúa siendo baja debido al desbalance de clases, reflejándose en un número considerable de falsos positivos. La métrica ROC-AUC asciende a 0.594, sugiriendo una capacidad de discriminación un poco mayor que la obtenida con la regresión logística. En conjunto, este árbol simple ofrece un rendimiento ligeramente superior al modelo logístico y da una buena una segunda referencia inicial para el análisis comparativo que se desarrolla en el Punto 5.

```
=== comparativo (positiva = mora 30) ===
                                precision_pos  recall_pos
                                 ↪  f1_pos  roc_auc
modelo
regresión logística                    0.189       0.569
 ↪  0.284    0.578
árbol de decisión (simple)             0.198       0.554
 ↪  0.292    0.594
```

Al comparar ambos modelos concluimos que el árbol de decisión simple presenta un desempeño ligeramente superior al de la regresión logística, especialmente en términos de su capacidad de discriminación (ROC–AUC). Aunque ambos modelos mantienen una precisión positiva baja debido al fuerte desbalance del conjunto de datos, el árbol captura mejor los patrones no lineales del df1 y logra identificar a los clientes en mora ≥ 30 días con una eficacia marginalmente mayor.
Ambos de estos resultados establecen el baseline para el problema de clasificación ya que ambos modelos ofrecen un punto de partida razonable, pero todavía limitado. Lo que permitirá explorar configuraciones más complejas e incorporar nuevas variables para determinar cuál es el más adecuado para la detección temprana de riesgo de mora.

5 Modelos de predicción categórica

Consecuente al primer modelo definido se construyó un segundo conjunto de datos `df2`, donde se incorporaron variables que habían sido descartadas previamente. SCORE y TASA, porque el análisis PCA mostró que eran relevantes para separar a los individuos. A su vez, se eliminó la variable DIAS EN APROBAR, ya que presentaba una varianza casi nula y no aportaba información útil al modelo. Finalmente, se evaluó nuevamente el Grid Search sobre este `df2` para comparar si realmente aportaba mejoras.

5.1 Reentrenamiento del modelo

Después de entrenar el primer árbol de decisión con las variables seleccionadas, aplicamos un proceso de búsqueda de hiperparámetros (Grid Search) para mejorar su rendimiento. El Grid Search confirmó que la mejor profundidad seguía siendo max_depth = 5, pero recomendó aumentar min_samples_leaf de 50 a 100. Este cambio ayudó a que el árbol fuera más estable y menos propenso al sobreajuste.

Con estos nuevos parámetros, el modelo logró un desempeño ligeramente mejor. Aunque el F1-score y el ROC-AUC ($\approx 0{,}59$) se mantuvieron similares, el modelo mostró una mejor estabilidad en sus métricas y un mejor balance entre precisión y recall. Esto indica que el reentrenamiento ayudó a obtener un modelo más consistente sin sacrificar capacidad predictiva.

5.2 Resultados con el df2 (incluyendo SCORE y TASA)

En el segundo dataset (df2) se incorporaron nuevamente las variables SCORE y TASA, ya que en el análisis previo de PCA estas dos mostraron ser relevantes para separar a los individuos con mayor probabilidad de caer en mora. Además, se eliminó la variables **DIAS EN APROBAR**, debido a que presentaba una varianza prácticamente nula y no aportaba información útil al modelo. Se mantuvieron las mismas varaibles descartadas del df1 para conservar consistencia en el experimento.

Al entrenar el árbol de decisión con este nuevo conjunto de variables, el modelo obtuvo una mejoría ligera en comparación con el df1, especialmente en las métricas de la clase positiva (mora $\geq$ 30 días), donde el recall subió a 0,506 y el F1 a 0,323. Esto indica que SCORE y TASA sí aortaron información relevante para mejorar la capacidad del modelo de identificar correctamente a los clinetes con riesgo de mora.

Posteriormente se aplicó nuevamente un Grid Search para optimizar hiperparámtero. Sin embargo, a diferencia del df1, en este caso el Grid Search no logró mejorar los resultados: el desempeño del árbol no aumentó y, en algunos casos, incluso se observó un deterioro leve en las métricas. Debido a esto, se mantuvieron los parámetros base del modelo, ya que ofrecián un equilibrio más estable entre precisión y recall.

5.3 Entrenamiento con variables relevantes del Random Forest

Al entrenar un árbol de decisión utilizando únicamente las variables más relevantes identificadas por el modelo Random Forest (por ejemplo: *SCORE*, *MONTO DESEMBOLSO*, *TASA*, *EDAD*, *INGRESOS*, *TIEMPO EN RESIDENCIA*, entre otras), se observó que el rendimiento general del modelo se mantiene prácticamente igual al obtenido empleando todas las variables originales del conjunto de datos.

La métrica clave para la clase positiva (mora $\geq$ 30 días), el *F1-score*, permaneció en un valor de 0,323, y el área bajo la curva ROC (*ROC-AUC*) se mantuvo en un nivel similar ($\approx 0{,}645$). Esto indica que la eliminación de variables con baja relevancia no afecta la capacidad predictiva del modelo, por el contrario, permite obtener un modelo más simple, más

interpretable y con menor riesgo de sobreajuste.

Por otro lado, se observó que el modelo continúa detectando adecuadamente la clase negativa (clientes sin mora), mostrando un buen desempeño en la identificación de individuos que no presentan riesgo de incumplimiento.

La precisión baja en la clase positiva se debe principalmente a que el dataset está muy desbalanceado. Hay muy pocos clientes que realmente caen en mora ≥ 30 días, y la mayoría son clientes que sí pagan. Como casi no tenemos ejemplos de la clase 1, el modelo no aprende bien cómo se comportan esos casos y por eso le cuesta identificarlos. Esto hace que se equivoque más cuando trata de predecir quién sí va a caer en mora, aunque su recall sea decente. En resumen, el modelo aprende bien a reconocer a los que no caen en mora, pero batallan más en detectar a los que sí tienen riesgo porque casi no hay datos de ese tipo.

6 Modelo de monto a aprobar

Para este punto se construyó un modelo de regresión lineal con el objetivo de estimar el monto desembolsado a partir de las variables disponibles en el dataset. El propósito principal de este ejercicio es evaluar qué tan predecible es el monto y en qué medida las características de los pacientes o de las solicitudes ayudan a explicar su variabilidad. Este modelo se utiliza como referencia, ya que el monto aprobado suele depender también de políticas internas y criterios comerciales que no necesariamente están reflejados en los datos.

Antes de entrenar el modelo se realizó una auditoría de la variable objetivo. Se identificaron valores extremadamente altos (hasta 25 millones COP) que afectaban severamente el ajuste de la regresión. Con el fin de obtener un modelo estable y evitar la distorsión causada por estos valores atípicos, se decidió acotar el análisis a montos menores o iguales a un millón de pesos colombianos. Este rango es consistente con los montos reales observados en tratamientos odontológicos de las clínicas Dentix y permite trabajar con una distribución más coherente.

Posteriormente se definió un pipeline que incluye imputación de valores faltantes y escalamiento de variables numéricas. También se eliminaron las columnas con fuga de información, al igual que en el modelo de mora ≥ 30 días. Se entrenaron dos variantes del modelo: una regresión lineal estándar y otra aplicando transformación logarítmica al objetivo (log1p), con el fin de evaluar si dicha transformación ayudaba a estabilizar la relación entre el monto y las variables explicativas.

```
=== lineal ===
RMSE: 189,916.67
R²:   0.052

=== lineal (objetivo log-transformado) ===
RMSE: 191,934.86
R²:   0.032
```

Los resultados muestran que la regresión lineal explica una proporción limitada de la variabilidad del monto desembolsado. El modelo estándar obtiene un R^2 de 0.052, lo cual indica que solamente logra capturar alrededor del 5 % de la variación total. El modelo con transformación logarítmica presenta un desempeño ligeramente menor, lo que sugiere que la relación entre las variables y el monto no sigue una estructura multiplicativa. El error cuadrático medio (RMSE) se mantiene cercano a 190 mil COP en ambos casos, un nivel de error razonable considerando la magnitud típica de los desembolsos y la heterogeneidad entre clínicas.

El análisis de residuos evidencia una dispersión amplia y una ligera heterocedasticidad, lo cual es común en datos financieros donde los montos más altos tienden a presentar mayor variabilidad. A pesar de esto, el modelo permite identificar tendencias generales y sirve como un punto de partida para entender qué tan predecible es el monto de desembolso a partir de la información disponible.

En conclusión, la regresión lineal no alcanza un poder predictivo elevado debido a la alta variabilidad del monto y a la ausencia de variables clave que probablemente influyen en la decisión de desembolso. Sin embargo, los resultados obtenidos son consistentes con la naturaleza del problema y ofrecen una referencia útil para comprender las limitaciones y posibles mejoras futuras en el modelado del monto.

7 Modelo refinado de aprobación

Con base en el análisis previo de correlaciones y la reducción dimensional mediante *PCA* inicial, se seleccionaron las siguientes variables explicativas debido a su contribución conjunta a la variabilidad del sistema y su relevancia predictiva detectada en primeras iteraciones del modelo:

Variables usadas:

- SCORE
- MONTO_DESEMBOLSO
- EDAD
- NIVEL_ESTUDIOS
- INGRESOS_FIJOS
- GASTOS_DE_SOSTENIMIENTO
- ACTIVOS
- TIEMPO_ACTIVIDAD
- TIEMPO_EN_RESIDENCIA
- CODIGO_CIIU
- ESTADOCIVIL
- ESTRATO

Estas variables fueron evaluadas mediante un *MANOVA*, con el fin de determinar si existían diferencias multivariadas significativas entre individuos con mora ≥ 30 y sin mora.

Cuadro 3: Resultados MANOVA

Estadístico	**Valor**	**F**	**p-value**
Wilks' Lambda	0.9622	638.47	$< 0{,}001$
Pillai's Trace	0.0378	638.47	$< 0{,}001$
Hotelling–Lawley Trace	0.0393	638.47	$< 0{,}001$
Roy's Greatest Root	0.0393	638.47	$< 0{,}001$

7.1 Interpretación

- Los cuatro criterios rechazan con alta significancia estadística ($p < 0{,}001$).
- Esto indica que la combinación conjunta de las variables seleccionadas genera diferencias entre clientes morosos y no morosos.
- Por lo tanto, el vector de características tiene capacidad discriminante real y justificó su entrada posterior en modelos predictivos y LDA.

En términos prácticos, el perfil financiero, laboral y sociodemográfico sí influye en el comportamiento de pago, lo que valida continuar con XGBoost, LDA y segmentación KMeans usando este subconjunto.

7.2 MANOVA por ciudad

Para evaluar si existen diferencias estadísticas significativas entre los perfiles de los clientes según la ciudad donde se ubica la clínica, se construyó la variable `CLINICA_CIUDAD`. Esta se obtuvo extrayendo el código de ciudad presente en el texto de la columna `CLINICA` (por ejemplo: BOG, BAQ, MDE, CLO, CTG). Este enfoque permite agrupar las observaciones de forma coherente sin generar una categoría por cada clínica individual, lo cual sería inestable debido al elevado número de sedes y tamaños muestrales desiguales.

El modelo MANOVA se estimó utilizando como variables dependientes los indicadores económicos, demográficos y laborales del solicitante (SCORE, Monto desembolso, Edad, Ingresos fijos, Activos, Gastos, Estrato, etc.), y como variable independiente la ciudad asociada a la clínica. Los resultados muestran efectos multivariados significativos:

- Wilks' lambda = 0.9868
- Num DF = 228, Den DF = 371024
- $F = 2{,}11$, $p = 0{,}0000$

Dado que el valor-p es prácticamente cero en todas las pruebas (Wilks, Pillai, Hotelling–Lawley y Roy), se concluye que **existen diferencias estadísticamente significativas entre las ciudades** en al menos una combinación lineal de las variables de interés.

En términos prácticos, esto indica que los solicitantes presentan perfiles distintos dependiendo de la ciudad en la que se atienden: variaciones en ingresos, estabilidad laboral, estrato socioeconómico, score crediticio o composición financiera pueden diferir sistemáticamente entre ciudades como Bogotá, Medellín, Barranquilla, Cali o Cartagena. Estas diferencias deben considerarse en el análisis operativo, pues reflejan que el riesgo crediticio y las características del cliente no son uniformes a nivel nacional.

Aunque MANOVA confirma la existencia de diferencias globales entre ciudades, este análisis no determina cuáles variables específicas producen estas diferencias; únicamente

valida que, en conjunto, los perfiles multivariados no son iguales. Para una caracterización más detallada sería necesario complementar con análisis post-hoc por variable o técnicas de segmentación adicionales.

7.3 LDA

Debido al desbalance observado entre clases ($\sim 85\,\%$ no moroso vs $\sim 15\,\%$ moroso), el modelo LDA sin tratamiento previo tendía a clasificar casi todos los casos como clase 0, obteniendo recall ≈ 0 para morosos.

Por ello fue necesario aplicar SMOTE, permitiendo que el discriminante aprendiera diferencias reales entre segmentos.

Una vez reentrenado con SMOTE, se evaluó sobre el test original:

Cuadro 4: Resultados del modelo LDA con Threshold = 0.45

Clase	**Precision**	**Recall**	**F1-score**	**Support**
0	0.90	0.52	0.66	38584
1	0.22	0.70	0.33	7232
Accuracy			0.55	45816
Macro Avg	0.56	0.61	0.49	45816
Weighted Avg	0.79	0.55	0.61	45816

Se pueden obtener las siguientes conclusiones.

- El modelo LDA recupera bien a los morosos (recall 0.70), lo cual es valioso si la prioridad es detectar riesgo.
- Sin embargo, su precisión en morosos es baja (0.22), lo que implica un número considerable de falsos positivos.
- LDA sirve más como herramienta de interpretación y perfiles de riesgo, no como modelo operativo final.

Cuadro 5: Ranking de variables según coeficiente LDA

ID	Variable	Peso_LDA
3	NIVEL_ESTUDIOS	-1.268808e-01
10	ESTADOCIVIL	-2.904267e-02
2	EDAD	-6.717172e-03
11	ESTRATO	-5.645346e-03
0	SCORE	-2.180585e-03
8	TIEMPO_EN_RESIDENCIA	1.757982e-03
7	TIEMPO_ACTIVIDAD	8.398138e-04
9	CODIGO_CIIU	1.775232e-05
1	MONTO_DESEMBOLSO	9.824594e-08
4	INGRESOS_FIJOS	1.706622e-11
5	GASTOS_DE_SOSTENIMIENTO	-5.287062e-13
6	ACTIVOS	-2.075211e-15

Los coeficientes del modelo LDA permitn identificar qué variables empuja a un cliente hacia mayor o menor riesgo de mora. Valores negativos en el discriminante como mayor nivel educativo, estado civil más estable, edad y score se asocian con una menor probabilidad de incumplimiento, mientras que variables con peso positivo como mayores montos desembolsados y mayor tiempo operativo, incrementan el riesgo de atraso. En conjunto, LDA resulta valioso para interpretación y segmentación de riesgo.

7.4 Segmentación no supervisada con K-Means

Después del modelo discriminante (LDA), el objetivo fue descubrir perfiles naturales de clientes, sin necesidad de la variable MORA. Es decir, identificar grupos con patrones similares de comportamiento financiero para posteriormente contrastarlos con la morosidad observada.

Primero definimos el set de variables numéricas relevantes para segmentación: SCORE, CUOTAMENSUAL, INGRESOS_FIJOS, TIEMPO_ACTIVIDAD, TIEMPO_EN_RESIDENCIA,EDAD Y PLAZO.

Debido a que trabajan en escalas distintas (miles de pesos, años, tasas, puntajes), se aplicó normalización con StandardScaler, para evitar que variables con valores grandes dominen el clustering.

Para mejorar la separabilidad de los clusters y reducir ruido, aplicamos un PCA a 2 componentes principales, manteniendo la mayor varianza posible del sistema. Esto permitió:

- visualizar los segmentos en 2D
- acelerar la convergencia del K-Means
- suavizar correlaciones entre atributos financieros

El PCA se trabajó exclusivamente para clustering, no para clasificación en esta parte.

Probamos $K = 2$ a 6 y calculamos el Silhouette Score, métrica que evalúa qué tan separado y cohesivo está cada grupo, en donde el mejor desempeño se obtuvo con $K = 3$, por lo que se tomó como segmentación final.

K	Silhouette
2	0.409
3	**0.462** ← Mejor
4	0.447
5	0.397
6	0.399

Cuadro 6: Resultados del coeficiente Silhouette para distintos valores de K.

A partir de K-Means con $k = 3$, se identificaron tres grupos con características financieras diferenciadas. El análisis se realizó sobre variables estandarizadas y reducidas mediante PCA para asegurar separación geométrica y evitar dominancia por escala.

Cluster 0 — Perfil joven, bajo score y alta carga de pago (Alto riesgo)

Medianas observadas:

- Score: **632**
- Cuota mensual: **$142,539**
- Ingresos fijos: **$3,000,000**
- Tiempo en actividad: **4 años**
- Tiempo en residencia: **5 años**

- Edad: **29 años**

Descripción: Cliente joven, menor historial laboral y residencial, cuota elevada respecto al ingreso. Mayor sensibilidad al estrés financiero → segmento con riesgo.

Cluster 1 — Cliente consolidado, alto score y mayor solidez económica (bajo riesgo)

Medianas observadas:

- Score: **735.5**
- Cuota mensual: **$189,949**
- Ingresos fijos: **$4,000,000**
- Tiempo en actividad: **19 años**
- Tiempo en residencia: **16 años**
- Edad: **55 años**

Descripción: Clientes maduros, con estabilidad laboral y residencial. Mayor capacidad de pago y score más alto → segmento sano y con baja probabilidad de impago.

Cluster 2 — Ingresos altos, carga de pago media y estabilidad intermedia (riesgo medio)

Medianas observadas:

- Score: **726**
- Cuota mensual: **$214,825**
- Ingresos fijos: **$4,000,000**
- Tiempo en actividad: **8 años**
- Tiempo en residencia: **8.5 años**
- Edad: **40 años**

Descripción: Segmento intermedio con buena capacidad económica, pero menor historia que Cluster 1. Potencial equilibrado: no riesgoso como el Cluster 0 ni tan sólido como el Cluster 1.

7.5 Interpretación de los clusters

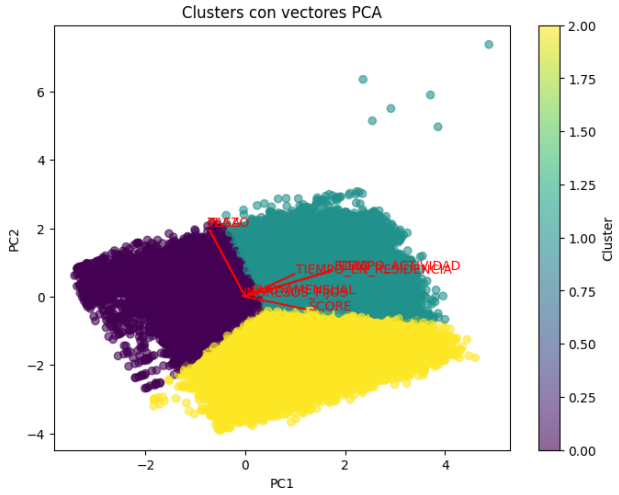

Figura 4: PCA+Kmeans

- PC1 explica principalmente variación asociada a SCORE, TIEMPO_ACTIVIDAD, EDAD y TIEMPO_EN_RESIDENCIA.Valores más altos en PC1 corresponden a clientes con mayor estabilidad laboral/residencial y mejor perfil crediticio.
- PC2 captura principalmente variación explicada por TASA y PLAZO, variables asociadas a condiciones financieras del crédito.Valores elevados en PC2 representan clientes con créditos más largos y/o tasas más altas.

Con las componentes principales (PC1 y PC2), los grupos formados por KMeans quedan bien separados y reflejan tres perfiles distintos de solicitantes:

- **Cluster morado (0) – Perfil joven y vulnerable.** Se ubica en valores bajos de PC1: menor score, menor antigüedad laboral y residencial, y mayor presión financiera. Representa el grupo de **mayor riesgo**.
- **Cluster verde (1) – Perfil consolidado y estable.** Presenta valores altos de PC1: mejor score, mayor estabilidad laboral y residencial. Es el segmento de **bajo riesgo**.

- **Cluster amarillo (2) – Perfil intermedio.** Se sitúa en una zona media de PC1 y PC2: buenos ingresos y condiciones razonables, pero menor historial que el grupo estable. Corresponde a un **riesgo medio**.

En conjunto, el análisis PCA + KMeans permite visualizar cómo Dentix puede distinguir tres perfiles claros de solicitantes: riesgo alto (morado), riesgo medio (amarillo) y riesgo bajo (verde). Esta segmentación complementa la política de umbrales y apoya decisiones más precisas en la evaluación crediticia.

8 Refinamiento

8.1 Modelo Final de Aprobación

Después de probar modelos iniciales como regresion logistica y arbol de decision, se decidió trabajar con XGBoost como modelo final. La razón principal fue que incorpora regularización, permite controlar sobreajuste y captura relaciones no lineales entre las variables, algo que los modelos base no lograban con la misma eficiencia. Además, fue posible ajustar el umbral de decisión para priorizar la correcta detección de clientes morosos.

Uno de los retos del modelo era el desbalance observado entre clientes morosos y no morosos. Para corregirlo se aplicó SMOTE, técnica que genera observaciones sintéticos similares a los morosos reales y permite equilibrar el conjunto de entrenamiento. Gracias a esto, el modelo tuvo suficientes ejemplo de incumplimiento para aprender patrones reales de mora y evitó caer en la predicción casi exclusiva de la clase 0 (no moroso)

Cuadro 7: Resultados después del ajuste de umbral (Threshold = 0.15)

Métrica	**No moroso (0)**	**Moroso (1)**
Precision	0.91	0.21
Recall	0.47	0.74
F1-Score	0.62	0.32

AUC-ROC = 0.645

Estos valores muestran que el modelo es más sensible para detectar morosos. Aunque la presición baja en esta clase, el recall aumenta de forma importante, lo que significa que el sistema identifica a siete de cada diez clientes con riesgo de incumplimiento. Para Dentix, esta característica es útil porque el costo de no detectar un moroso puede ser mayor que el de rechazar a un cliente sano.

Cuadro 8: Matriz de confusión del modelo

		Predicción	
		No moroso (0)	**Moroso (1)**
Real	**No moroso (0)**	TN = 5412	FP = 6163
	Moroso (1)	FN = 556	TP = 1614

El número de falsos negativos se reduce, lo que implica que el modelo deja escapar menos clientes con alto riesgo. A cambio, se incrementan los falsos positivos, pero esto puede ser aceptable si el objetivo principal es prevenir pérdidas por mora.

El modelo resultante muestra un equilibrio razonable entre sensibilidad y estabilidad. No es el más preciso, pero sí el más efectivo para detectar casos que representan riesgo financiero. Esto lo convierte en una herramienta viable para análisis preventivo y políticas de aprobación bajo criterios de riesgo.

8.2 Robustez del modelo

Para validar estabilidad y la robustez, el modelo XGBoost se evaluo separando el desempeño por género, estrato socioconómico y nivel educativo. Las métricas se mantuvieron consistentes en todos los grupos: precisión alta para clientes cumplidos y buen recall para morosos ($\approx 0{,}70 - 0{,}8$ en la mayoría de subpoblaciones), lo que indica que el algoritmo identifica riesgo de forma estable incluso cuando los datos se fragmentan.

En conjunto, estos resultados muestran que el modelo generaliza bien y no depende de un tipo específico de cliente para funcionar. XGBoost conserva capacidad predictiva en todos los perfiles evaluados.

9 Política de decisión y umbrales costo-sensibles

Tras el entrenamiento del modelo XGBoost con balanceo mediante SMOTE, se evaluó el desempeño a distintos valores de umbral de decisión. El objetivo fue identificar un punto operativo que permitiera maximizar la detección de morosos sin elevar en exceso los falsos positivos.

Los resultados mostraron que el valor 0.15 ofrece el mejor compromiso entre sensibilidad y estabilidad del modelo: incrementa de forma notable la identificación de clientes con riesgo (recall alto) aun cuando su precisión disminuye ligeramente respecto a valores más conservadores. Por ello, 0.15 se establece como umbral recomendado para operación, privilegiando

la alerta temprana de incumplimiento.

La política final se complementa con la segmentación obtenida mediante PCA + KMeans, permitiendo decisiones diferenciadas por perfil:

- **Cluster 0 (Morado) — Riesgo alto:** aprobación sólo si presenta documentación adicional sólida o garantías comprobables.
- **Cluster 1 (Verde) — Riesgo bajo:** aprobación directa con requisitos estándar.
- **Cluster 2 (Amarillo) — Riesgo medio:** aprobación con verificación adicional de ingresos y estabilidad.

Este enfoque combina el umbral óptimo con la segmentación obtenida, permitiendo controlar el riesgo sin frenar la colocación. Además, el umbral 0.15 puede ajustarse estratégicamente según los objetivos del negocio:

Umbral más bajo (0.10–0.15) → detecta más clientes riesgosos (mayor recall), reduce el riesgo de mora, pero disminuye la cantidad de créditos aprobados. **Umbral más alto (0.20–0.30)** → se escapan más clientes riesgosos (menor recall), aumenta la aprobación, pero también la exposición al riesgo de incumplimiento.

10 Conclusiones

El análisis realizado permitió entender a profundidad los factores financieros, sociodemográficos y operativos que influyen en el riesgo de mora dentro de las Clínicas Dentix. A partir de la auditoría de datos, el análisis multivariado, la construcción de modelos predictivos y la segmentación de clientes, se obtuvieron hallazgos sólidos que sustentan la propuesta final de aprobación crediticia.

En primera instancia, el análisis de correlaciones, covarianzas y PCA permitió identificar relaciones relevantes entre variables, destacando que la Tasa, el Plazo, los Ingresos y el Estrato socioeconómico ejercen una fuerte influencia en la estructura dimensional del sistema. La prueba MANOVA demostró que el perfil financiero y demográfico sí genera diferencias significativas entre morosos y no morosos, validando su uso posterior en modelos discriminantes.

Los modelos iniciales mostraron limitaciones importantes debido al fuerte desbalance de clases, sin embargo, sirvieron como baseline para el desarrollo de modelos más robustos. La

incorporación de variables como SCORE y TASA confirmó que estos atributos aportan el mayor poder predictivo observado en nuestro dataset. Asimismo, la evaluación con Random Forest reveló que un subconjunto reducido de variables es suficiente para mantener la capacidad predictiva, simplificando el modelo.

El uso de LDA permitió interpretar con claridad qué factores empujan a un cliente hacia mayor o menor riesgo, aunque su desempeño operativo resultó limitado por la baja precisión en la clase positiva. Por su parte, el clustering con PCA + KMeans identificó tres perfiles naturales de solicitantes—riesgo alto, medio y bajo—que complementan la toma de decisiones y permiten diferenciar segmentos de pacientes.

El modelo final seleccionado, XGBoost con balanceo mediante SMOTE, alcanzó el mejor compromiso entre sensibilidad y estabilidad. Con un umbral de decisión de 0.15, el modelo logra detectar aproximadamente 74 % de los clientes con riesgo de mora, reduciendo significativamente los falsos negativos. Aunque esto incrementa los falsos positivos, representa una estrategia adecuada para Dentix, donde el costo de no detectar un moroso supera el de aplicar verificaciones adicionales a clientes sanos.

En conjunto, el proyecto sustenta un sistema predictivo efectivo y soportado estadísticamente para la toma de decisiones crediticias. La combinación de preprocesamiento riguroso, análisis multivariado, modelado avanzado y segmentación, permite a Dentix mejorar la identificación de clientes riesgosos, optimizar su política de aprobación y reducir la exposición a pérdidas futuras. El umbral recomendado puede ajustarse estratégicamente según los objetivos del negocio para ofrecer flexibilidad y control en la gestión del riesgo crediticio.

A Apéndice

Todo el análisis, código utilizado, reporte técnico, presentación ejecutiva y hoja de instrucciones se puede encontrar dentro del siguiente repositorio que consolida el proyecto.
Link - Github