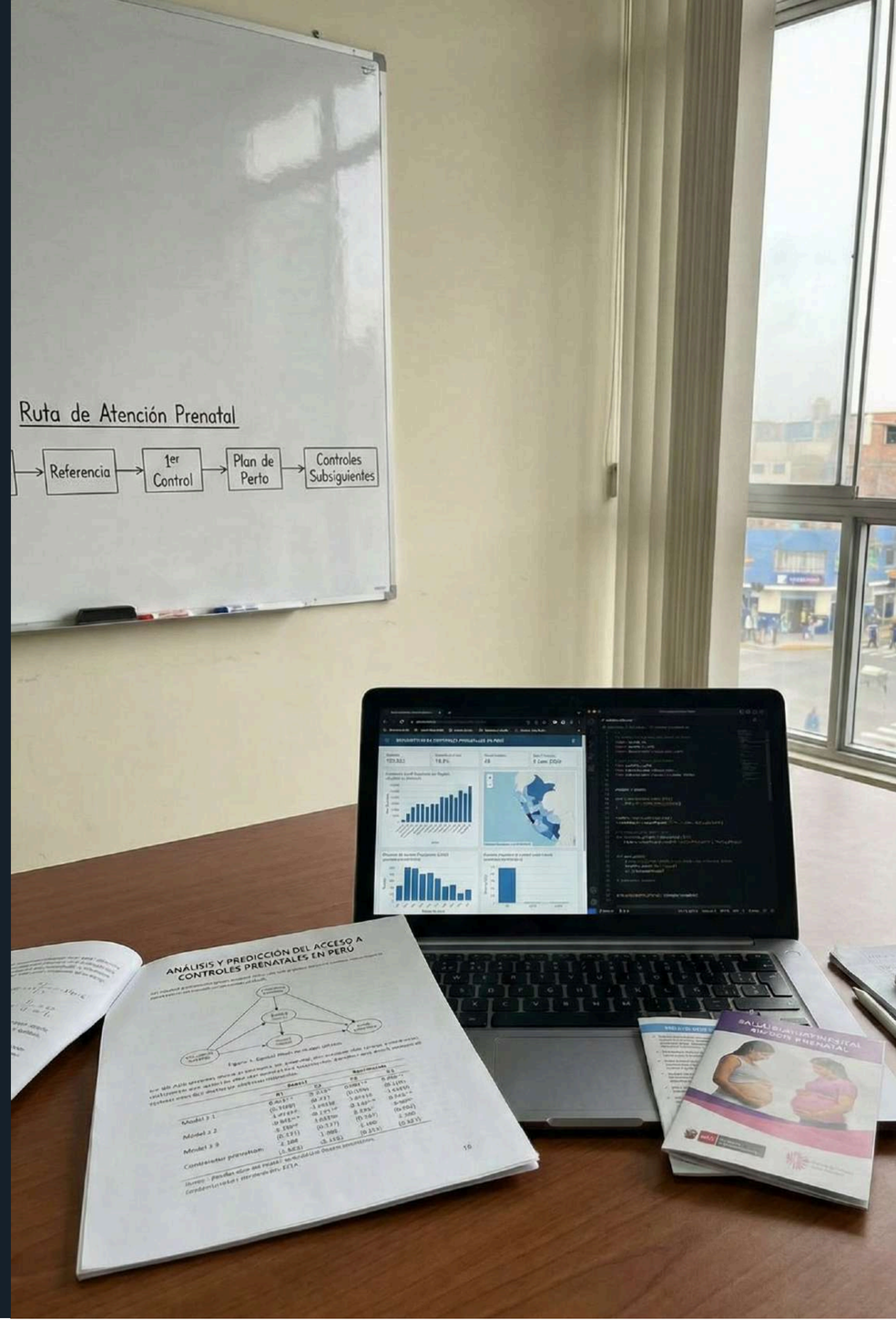
Grupo 5

INTRODUCCIÓN A CIENCIA DE DATOS Y MACHINE LEARNING CON PYTHON

Análisis y Predicción del Acceso a Controles Prenatales en Perú.

INTEGRANTES:

Fabricio Calle Cardoza \\ Bianca Nicole Jimenez Vargas \\
Aracely Lalupu Lalupu \\ Jose Alonso Naira Carmen.



MOTIVACIÓN

- La atención prenatal ayuda a prevenir riesgos durante el embarazo y garantiza un parto más seguro.
- En el Perú se recomiendan 6 o más controles para una atención adecuada.
- No todas las madres alcanzan este estándar, especialmente en zonas rurales y en hogares con menos recursos.
- La educación puede influir en la información, las decisiones y la búsqueda de atención de salud.

Grupo 5

¿CÓMO SE RELACIONA EL NIVEL EDUCATIVO DE LA MADRE CON EL ACCESO A CONTROLES PRENATALES ADECUADOS EN EL PERÚ?

Hipotesis: Mayor nivel educativo aumenta la probabilidad de recibir un control prenatal adecuado.

DATA

Grupo 5

Fuente de datos: ENDES (2018 - 2023)

Variables:

- Control Prenatal Adecuado: Variable dependiente.
- Edad de la Madre
- Nivel Educativo de la Madre
- Quintil de Riqueza
- Ubicación geográfica (urbano/rural)
- Región

	prenatal_adecuado	Edad	Nivel_riqueza	Urbano	SREGION	educacion_grupo
0	1	42.0	3.0	1	3	2
1	0	20.0	2.0	1	3	3
2	0	25.0	2.0	1	3	3
3	1	35.0	5.0	1	3	3
4	1	22.0	5.0	1	3	2
5	0	48.0	5.0	1	3	3
6	1	26.0	2.0	1	3	2
7	0	39.0	4.0	1	3	3
8	1	18.0	3.0	1	3	2
9	1	23.0	1.0	1	3	2

Grupo 5

TRABAJO 1:

Exploración de datos (EDA)

Figura 1: Mapa de Missing Values

Grupo 5

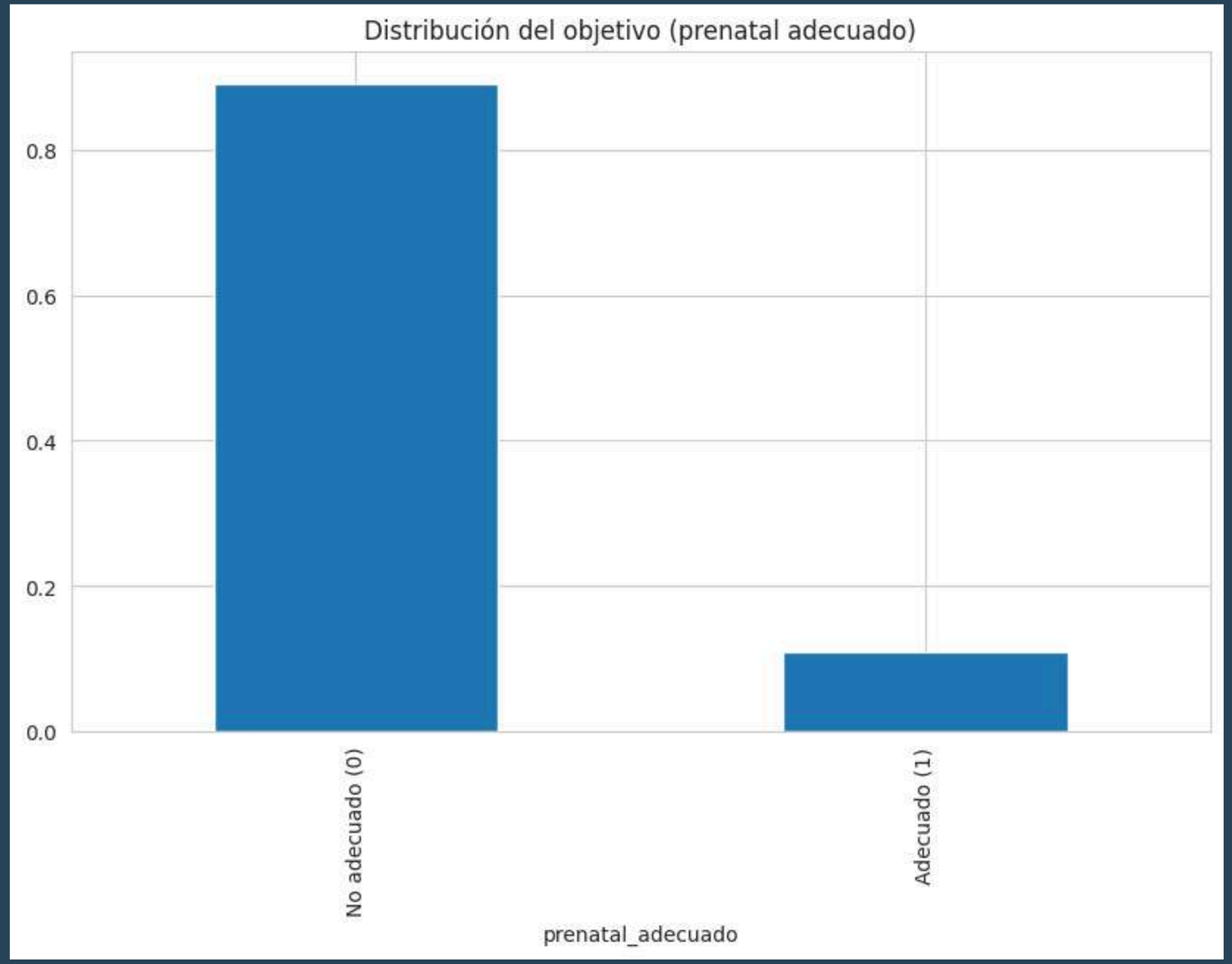


Figura 2: Distribución del objetivo

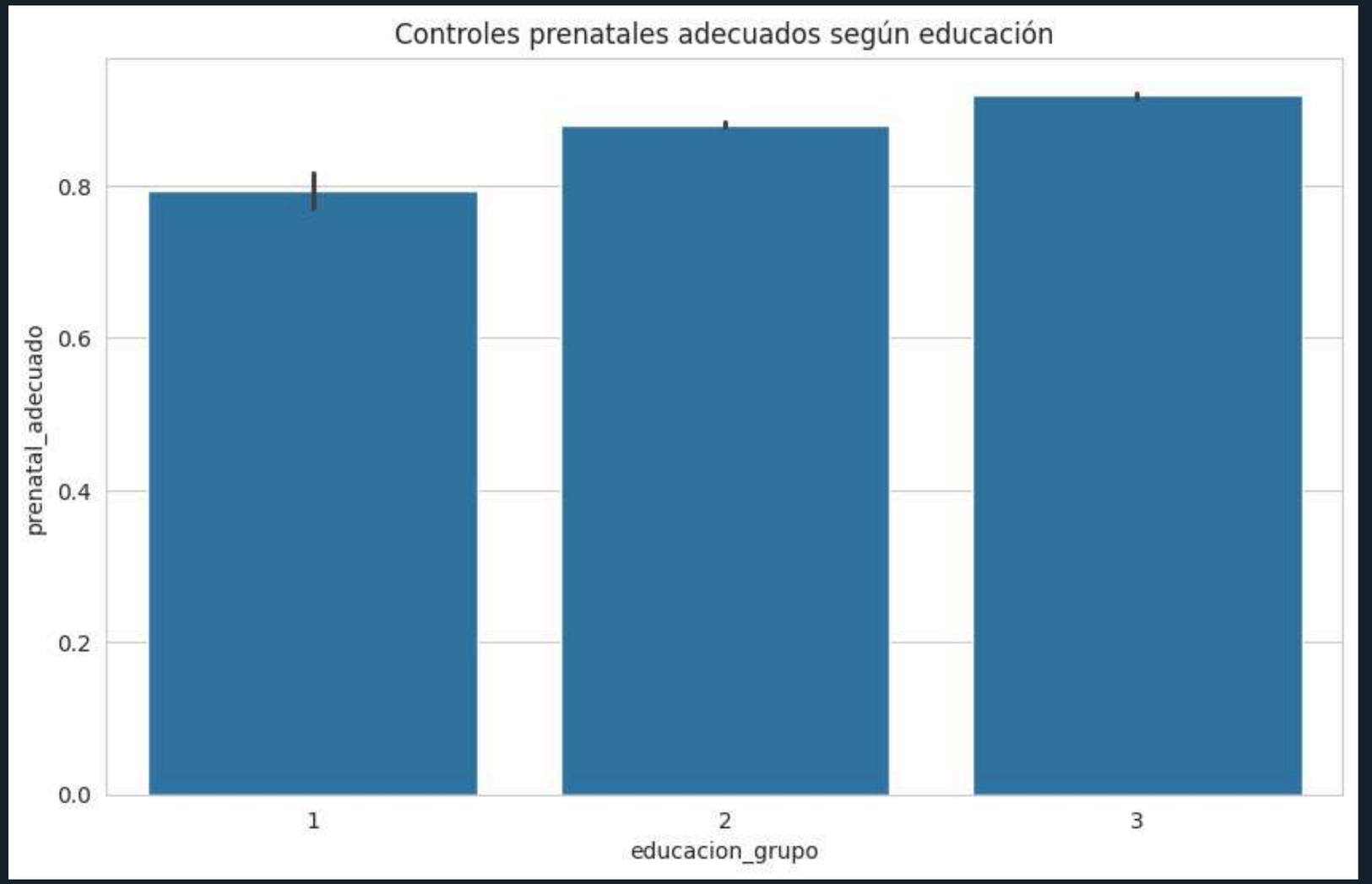


Figura 3: Controles prenatales adecuados según educación

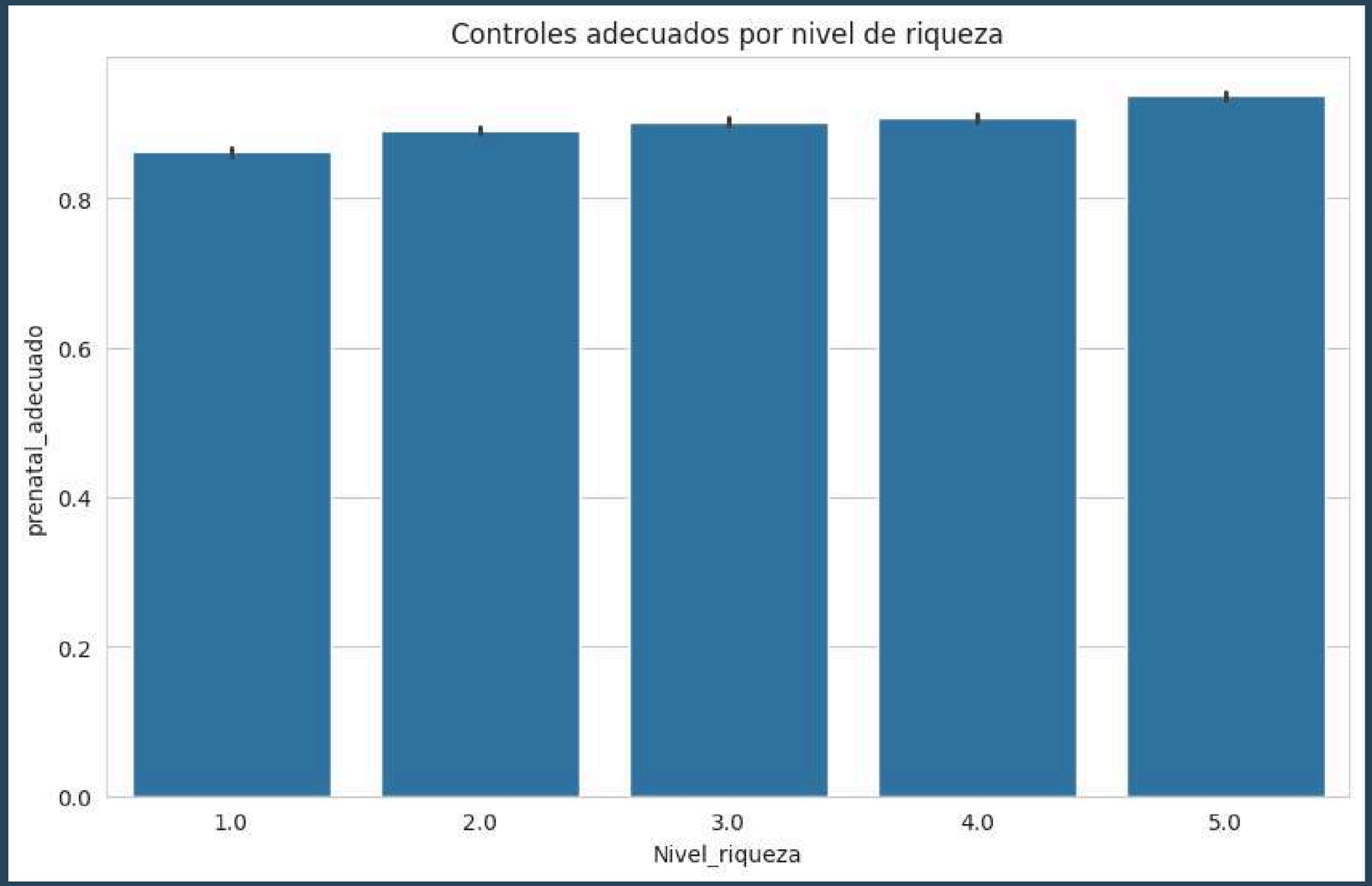


Figura 4: Controles adecuados por nivel de riqueza

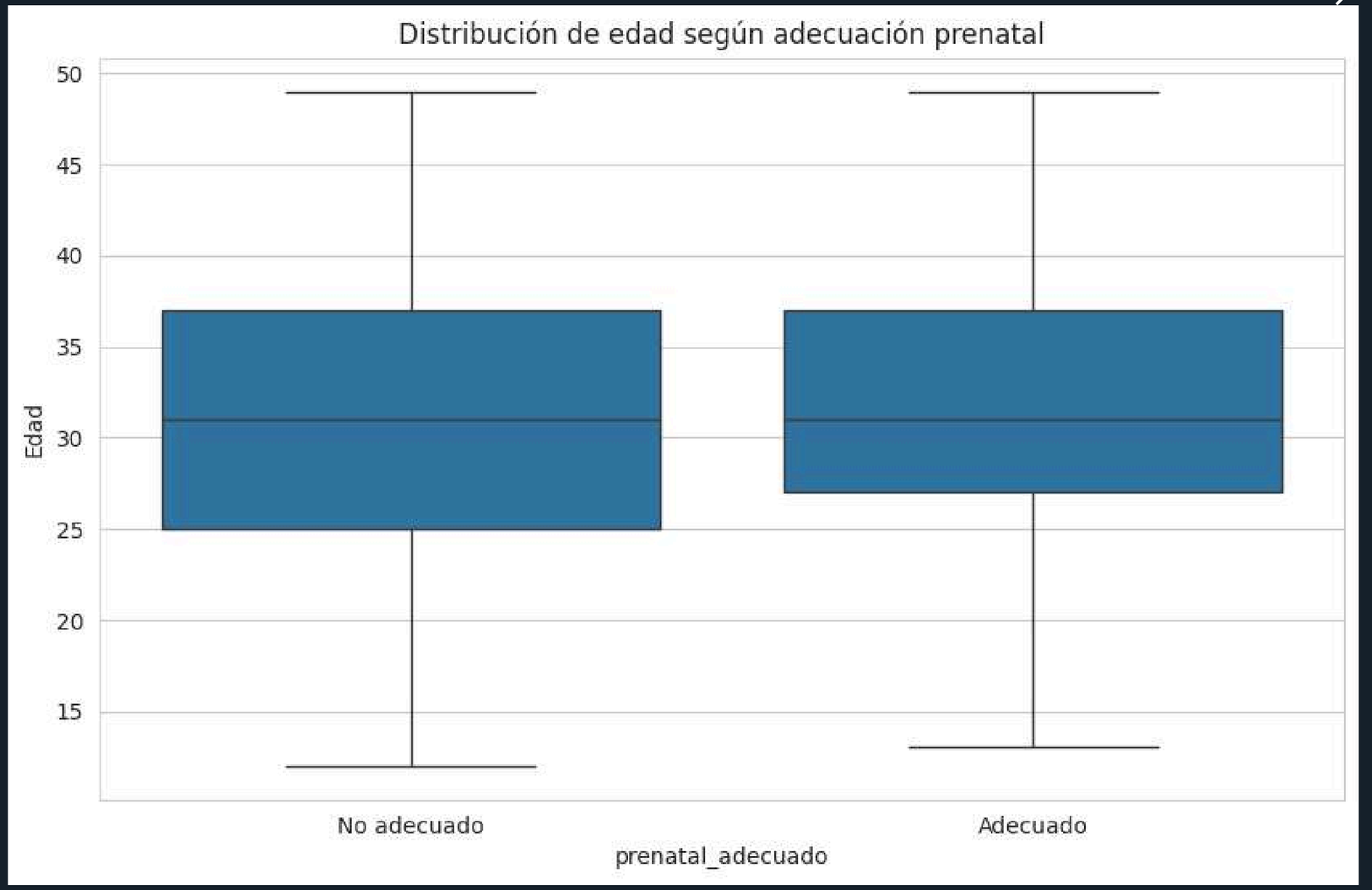


Figura 5: Dsitribución de edad según adecuación prenatal

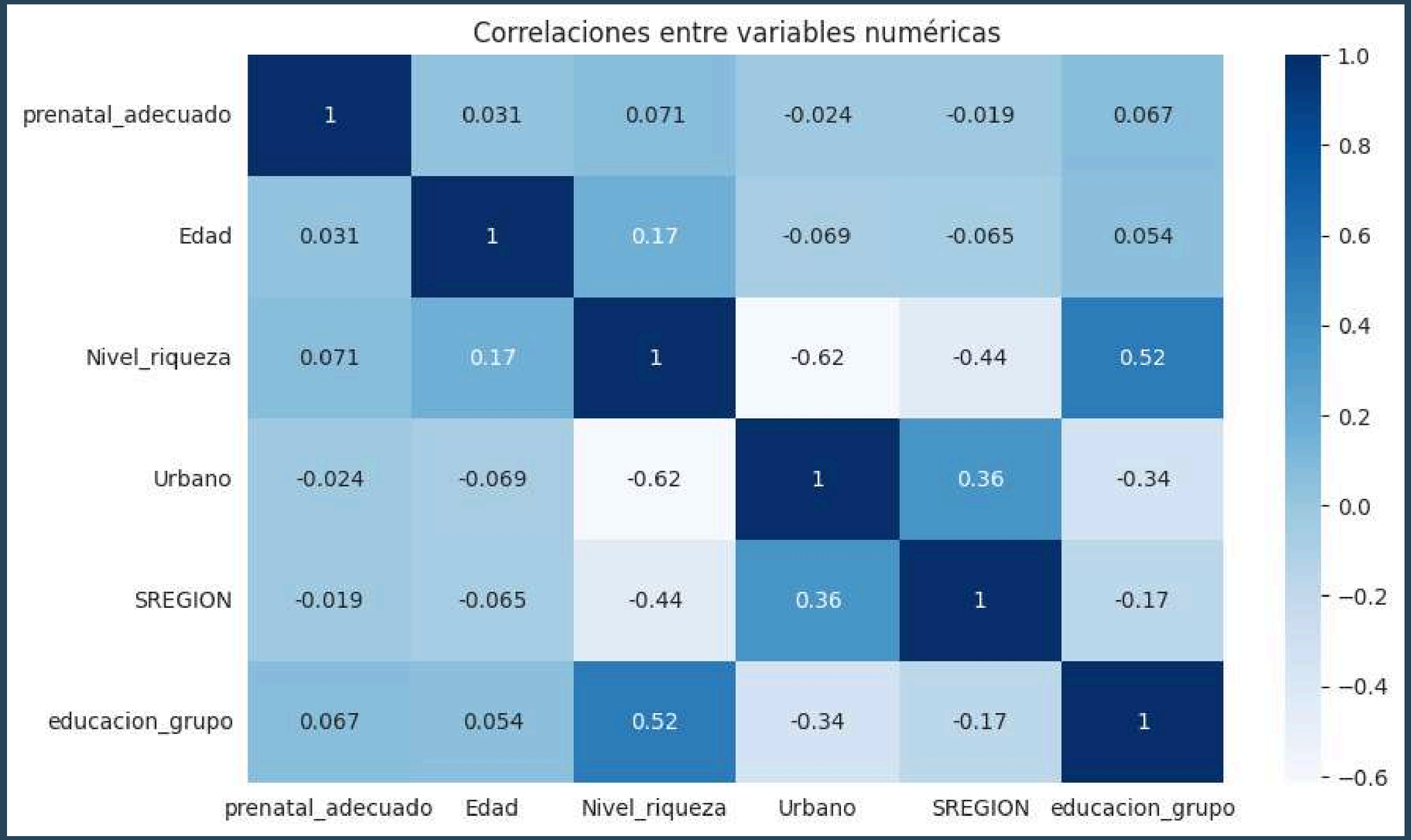
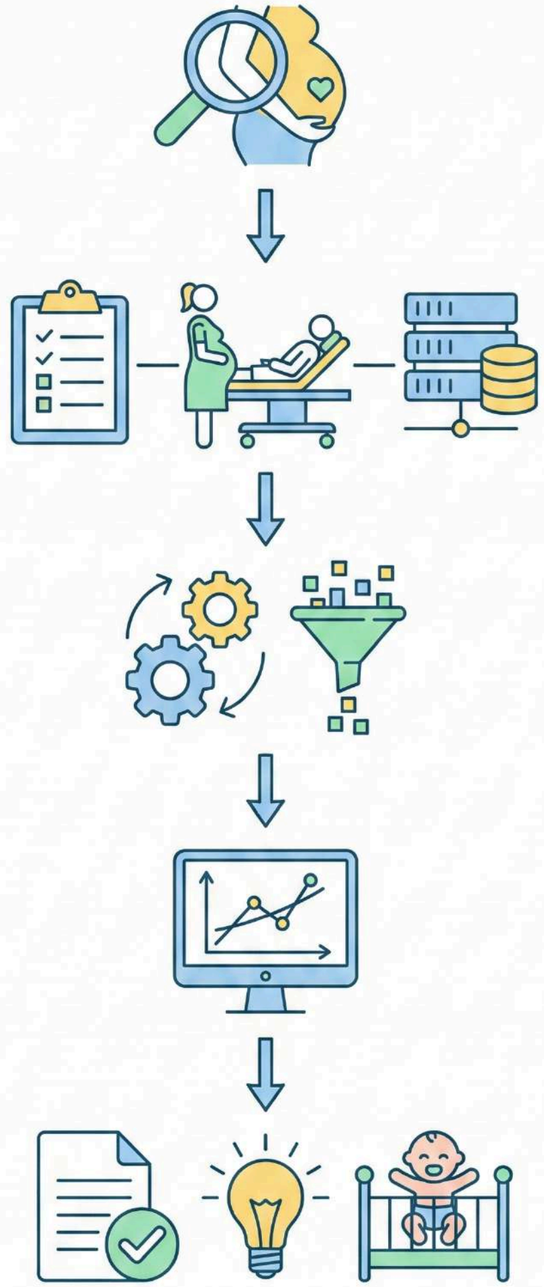


Figura 6: Tabla de correlaciones

Grupo 5

Estrategia Empírica (El Modelo)



Grupo 5

TRABAJO 2:

Modelo Base (Baseline)

MODELO BASE: REGRESIÓN LOGÍSTICA

Grupo 5

Ecuación:

$$Y_{it} = \beta_0 + \beta_1 \, Educ_i + \beta_2 \, Edad_i + \beta_3 \, Quintil_i + \beta_4 \, Reg_i + \beta_5 \, Urb_i + \epsilon_{it}$$

Variables:

- Educación
- Edad, Riqueza
- Región
- Zona Urbana

Se eligió Logit por ser el estándar para clasificación binaria (Asiste/No Asiste) y permitir interpretación de probabilidades.

RESULTADOS E INTERPRETACIÓN (ODDS RATIOS)

Hallazgos Clave:

- **Educación: Odds Ratio > 1. A mayor educación, mayor probabilidad de asistencia.**
- **Riqueza: Odds Ratio 1.20. El dinero es un factor protector fuerte.**
- **Zona Urbana: Odds Ratio 0.79. Curiosamente, vivir en zona urbana redujo la probabilidad en este modelo (posible saturación de servicios).**

Variable	Odds Ratio
Constante	2.3382
Educación_grupo	1.3215
Edad	1.0078
Nivel_riqueza	1.2092
SREGION	1.0342
Urbano	0.7936

EVALUACIÓN DE DESEMPEÑO Y LIMITACIONES

Validación Cruzada

Accuracy Promedio (CV): 0.8910 (+/- 0.0000)
F1-Score Promedio (CV): 0.9424 (+/- 0.0000)

Evaluación Final del Modelo de Clasificación:

Clase	Precisión	Recall	F1-score	Soporte
0	0.00	0.00	0.00	1879
1	0.89	1.00	0.94	15352
Accuracy			0.89	17231
Macro avg	0.45	0.50	0.47	17231
Weighted avg	0.79	0.89	0.84	17231

Hallazgos Clave:

- **Accuracy del 89%. Parece un modelo excelente a primera vista.**
- **(El problema): La tabla de métricas por clase. Clase 0 (No asiste): Recall = 0.00, F1-Score = 0.00.**

El 89% de exactitud es una ilusión. El modelo básicamente decidió 'apostar a lo seguro' diciendo que todas las madres cumplen con sus controles. El resultado es un algoritmo ciego: acierta en la mayoría, pero falla exactamente donde más lo necesitamos: en detectar a las madres que no van a asistir.

Necesitamos modelos que no solo acierten en promedio, sino que encuentren los casos difíciles.

Grupo 5

MODELOS MÁS COMPLEJOS Y OPTIMIZACIÓN

Grupo 5

INGENIERÍA DE CARACTERÍSTICAS

Nuestro objetivo en este apartado es enriquecer la data para capturar no linealidad

INGENIERÍA DE CARACTERÍSTICAS

Grupo 5

Nuestro objetivo en este apartado es enriquecer la data para capturar no linealidad:

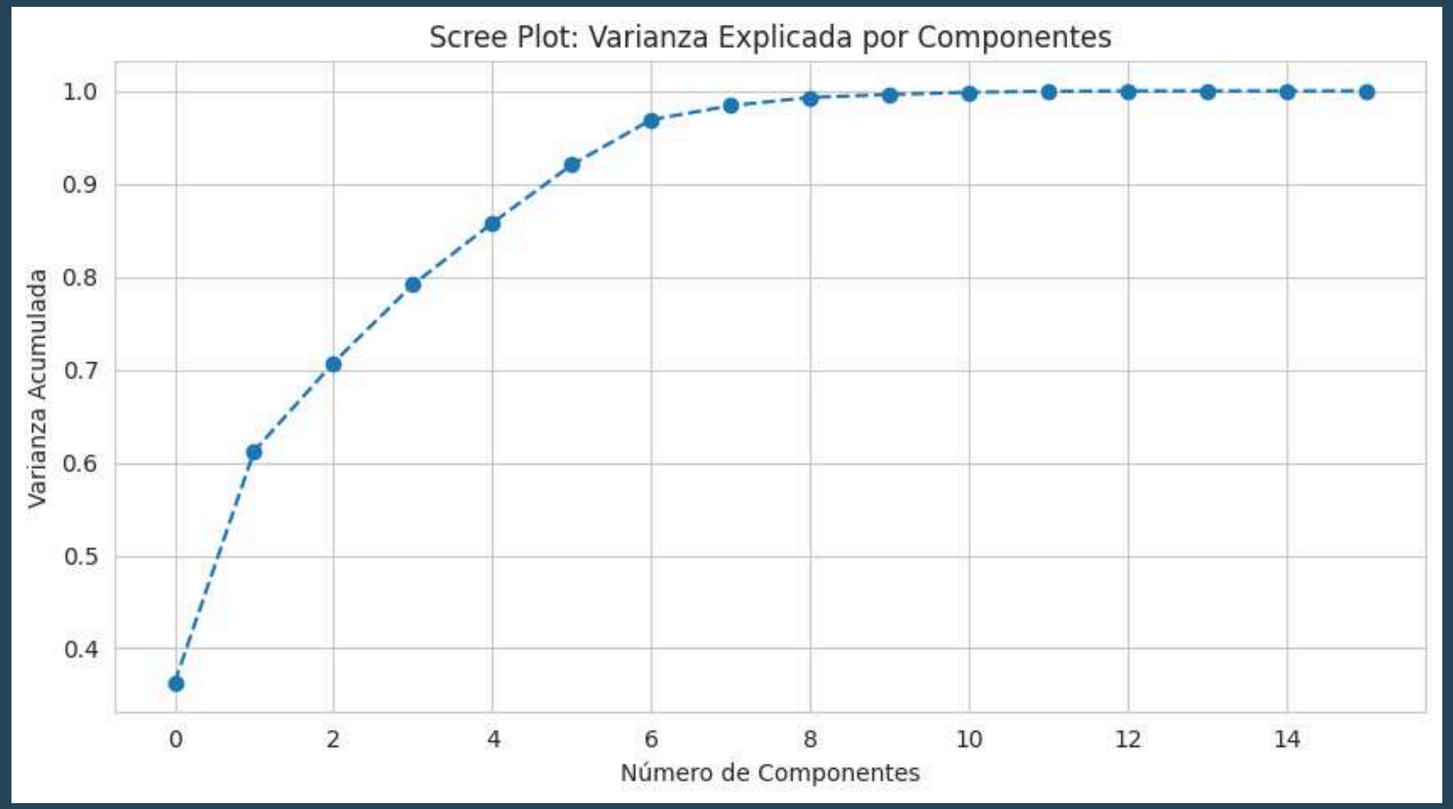
- No Linealidad (Edad^2): Elevar la edad al cuadrado permite modelar relaciones curvas (ej. mayor riesgo en extremos de edad).
- Interacciones (A × B): Variables combinadas que revelan cómo el efecto de la riqueza cambia según el nivel educativo.
- Flags de Riesgo: Indicadores binarios (0/1) explícitos para grupos vulnerables (adolescentes <18 o mayores >35).
- One-Hot Encoding: Convierte variables categóricas (como 'Región') en múltiples columnas binarias matemáticas.

Interacciones

- Edad condicionada por Riqueza
- Educación × Riqueza
- Riqueza × Urbano

INGENIERÍA DE CARACTERÍSTICAS

Grupo 5



Definición PCA: Técnica matemática que reduce la dimensionalidad de los datos buscando patrones de varianza.

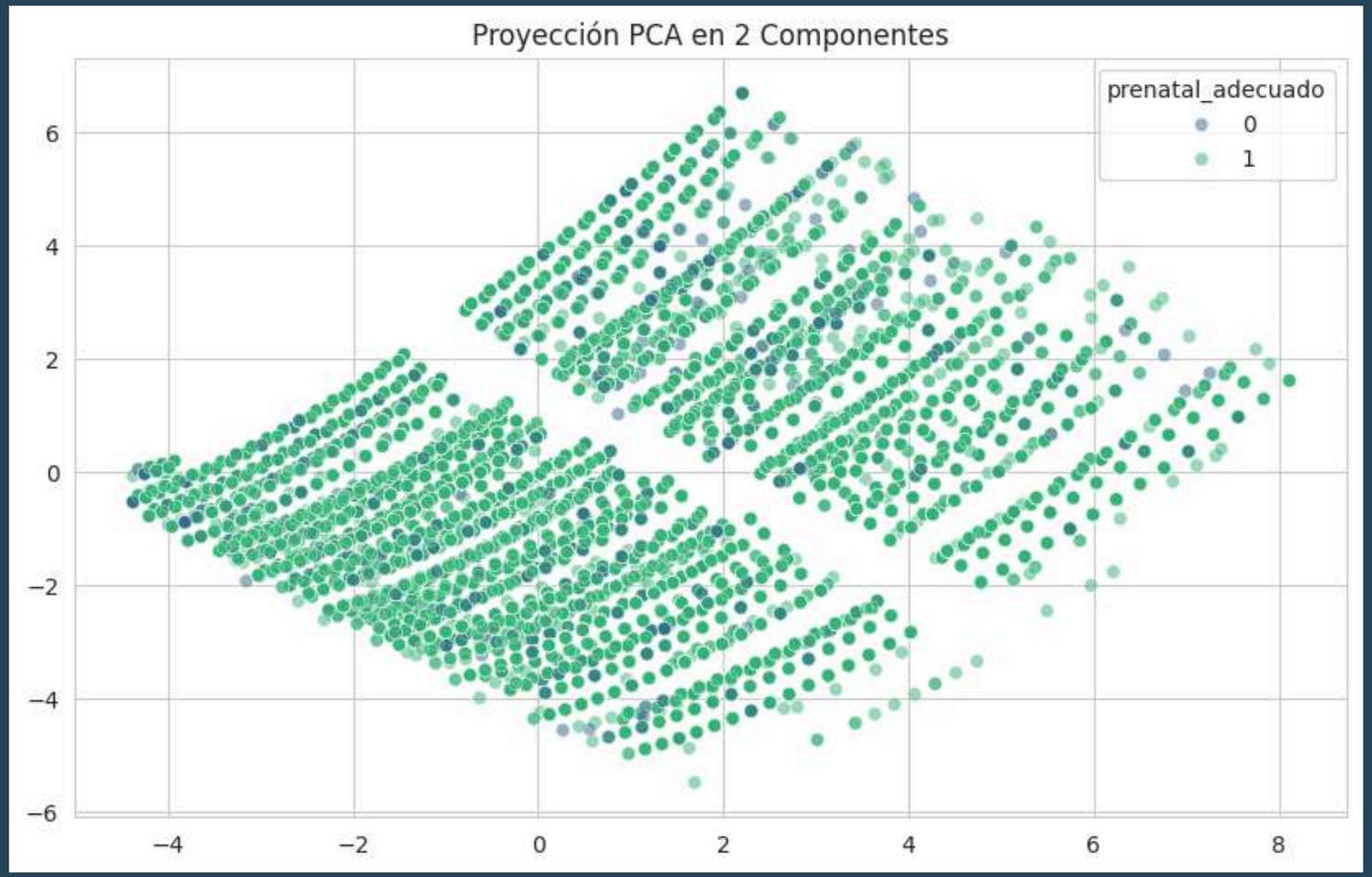
Hallazgo: La estructura es 'Combinatoria', lo que significa que no hay grupos separados limpiamente en el espacio.

INGENIERÍA DE CARACTERÍSTICAS

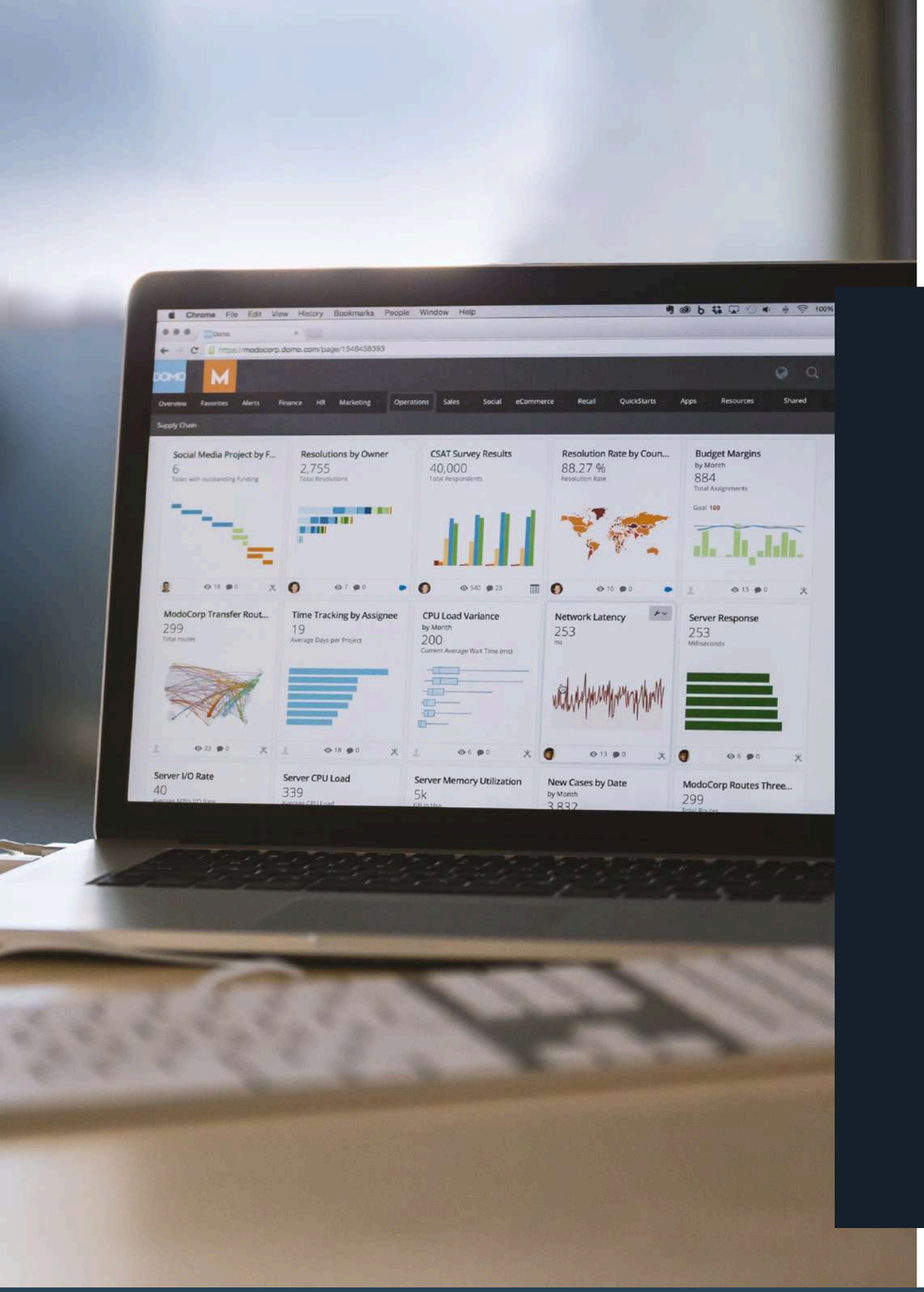
Grupo 5

Separabilidad Lineal: Capacidad de dibujar una línea recta para dividir clases. Aquí NO existe.

Conclusión: Se requieren modelos no lineales (flexibles) capaces de dibujar fronteras de decisión complejas.



Grupo 5



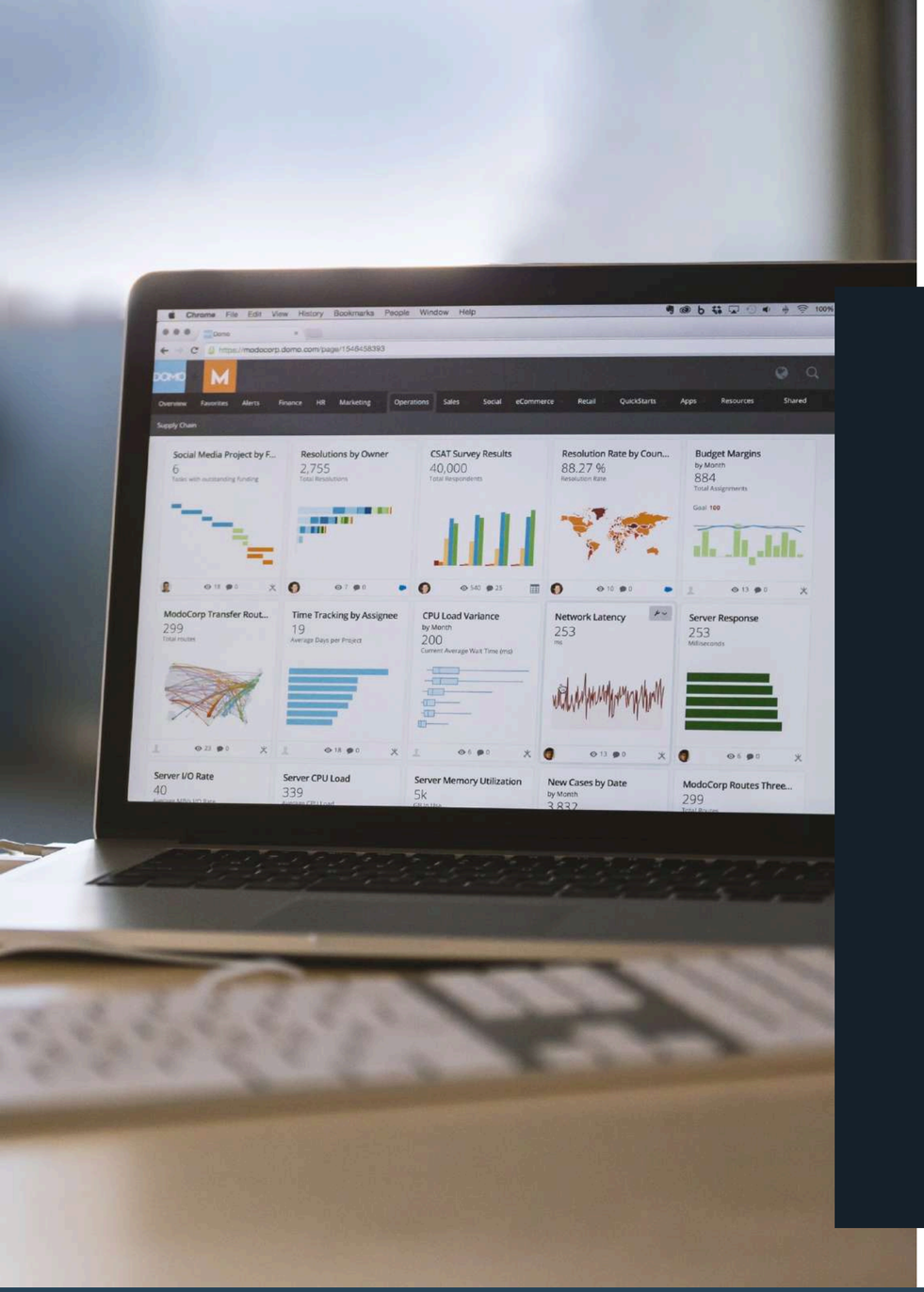
LOGIT MEJORADO

Pipeline: Secuencia automatizada de pasos (SMOTE → Escalado → Modelo) para evitar fugas de información.

Regularización L2 (Ridge): Penalización matemática a coeficientes grandes para reducir el sobreajuste (overfitting).

Métrica PR-AUC: Área bajo la curva de Precisión-Recall. Es el estándar de oro para datos desbalanceados.

Problema del Umbral 0.5: Usar el punto de corte estándar ignora que las clases no pesan lo mismo.



Grupo 5

Falso Negativo (FN): El error más costoso. Una madre que necesitaba control pero el modelo predijo que NO.

Fallo Inicial (Umbral 0.5): Con el umbral por defecto, el Recall (Sensibilidad) era solo 0.619, dejando sin detectar al ~38% de los casos de riesgo.

Falso Positivo (FP): Error menor. Una madre alertada innecesariamente (gasto de recursos, pero salva vidas).

Costo Asimétrico: Definimos que 1 FN cuesta más del doble que 1 FP (0.7 vs 0.3).

Umbral Óptimo (~0.01): Bajamos la vara de decisión drásticamente para capturar todos los casos de riesgo (Recall=1.0).

Table 1: Comparación de Métricas del Modelo Logit Mejorado por Umbral

Métrica	Umbral = 0.5	Umbral Óptimo
PR-AUC	0.9205	0.9205
ROC-AUC	0.5958	0.5958
F1	0.7386	**0.9424**
Recall	0.6190	**1.0000**
Precision	0.9155	0.8910
Accuracy	0.61	0.89

Matriz de Confusión

Clase (Real vs. Pred.)	Umbral = 0.5	Umbral Óptimo
Real 0 → Pred. 0 (Verdaderos Negativos)	1201	**0**
Real 0 → Pred. 1 (Falsos Positivos)	1053	**2254**
Real 1 → Pred. 0 (Falsos Negativos)	7020	**0**
Real 1 → Pred. 1 (Verdaderos Positivos)	11403	**18423**

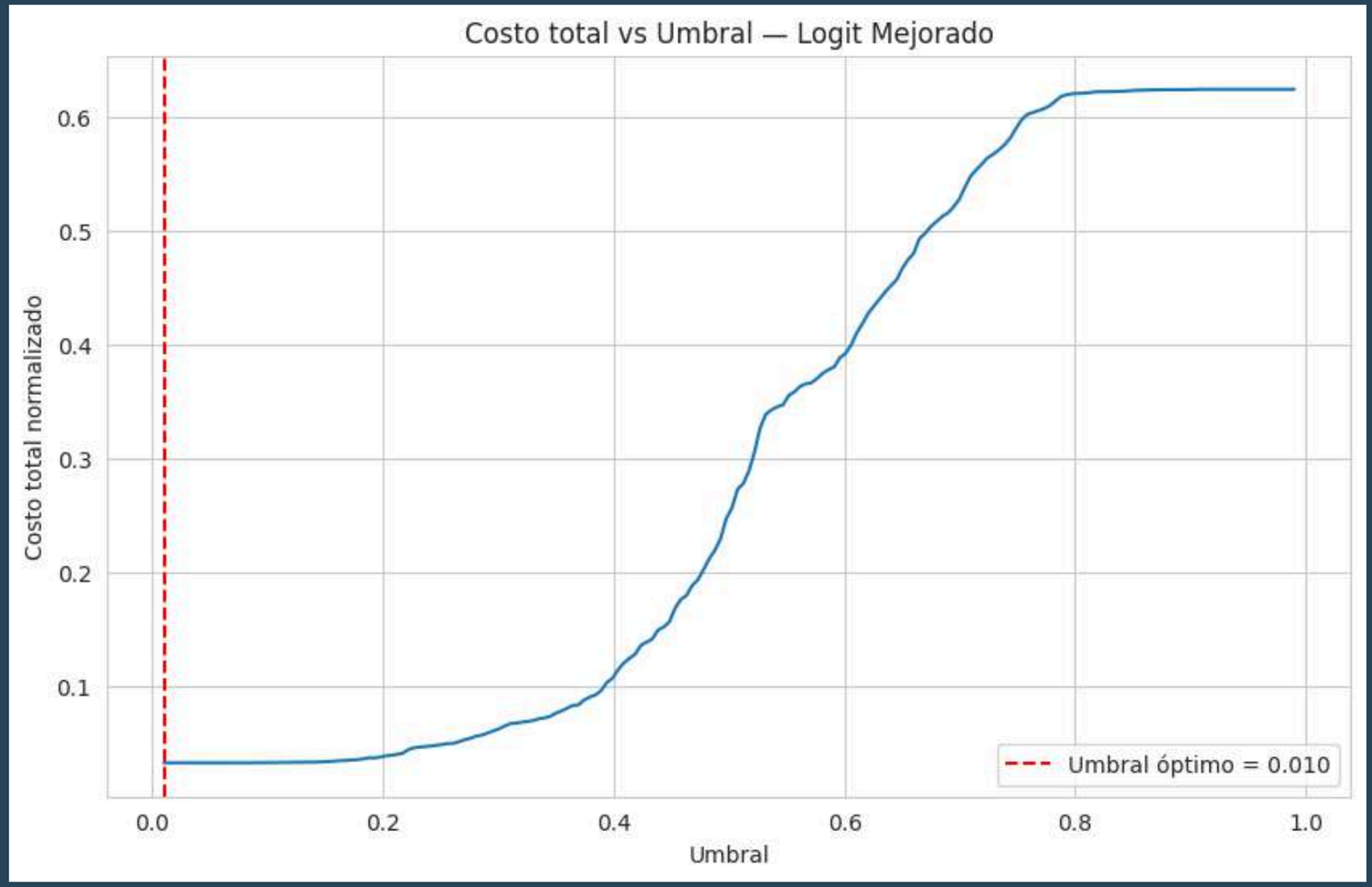


Figura 7: Costo Total vs Umbral

Grupo 5

XBOOST Y UMBRAL

El XGBoost es el modelo de alta potencia que se usa para asegurarse de que la predicción sea lo más precisa posible, y para confirmar la jerarquía de los factores de riesgo.

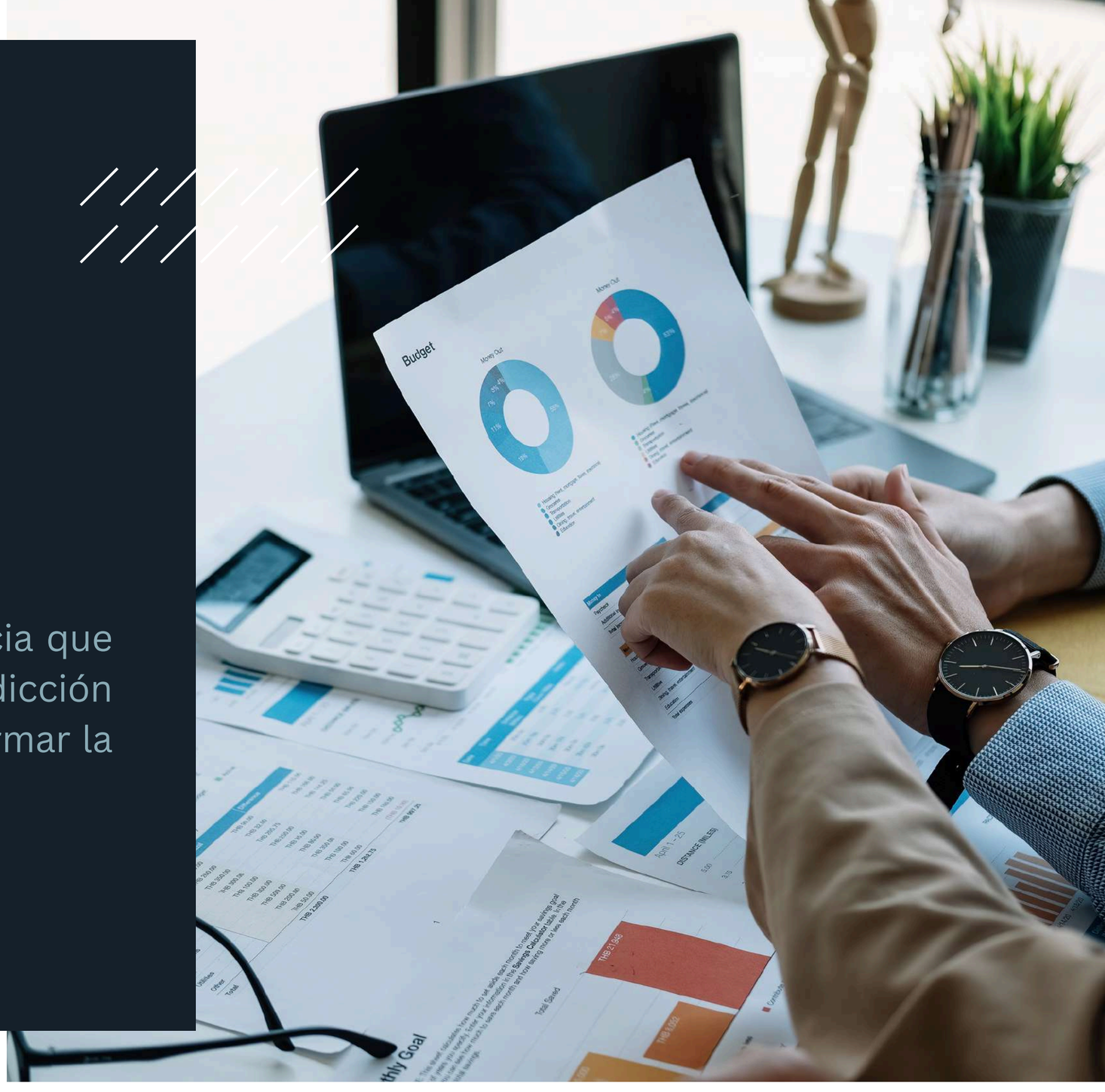


Table 2: Comparación de Métricas del Modelo XGBoost por Umbral

Métrica	Umbral Óptimo	Umbral = 0.5
PR-AUC	0.9187	0.9187
ROC-AUC	0.5966	0.5966
F1	**0.9424**	0.7686
Recall (Clase 1)	**1.0000**	0.6642
Precision (Clase 1)	0.8910	0.9118
Accuracy	0.89	0.64

Matriz de Confusión

Clase (Real vs. Pred.)	Umbral Óptimo	Umbral = 0.5
Real 0 → Pred. 0 (Verdaderos Negativos)	**0**	1070
Real 0 → Pred. 1 (Falsos Positivos)	**2254**	1184
Real 1 → Pred. 0 (Falsos Negativos)	**0**	6186
Real 1 → Pred. 1 (Verdaderos Positivos)	**18423**	12237

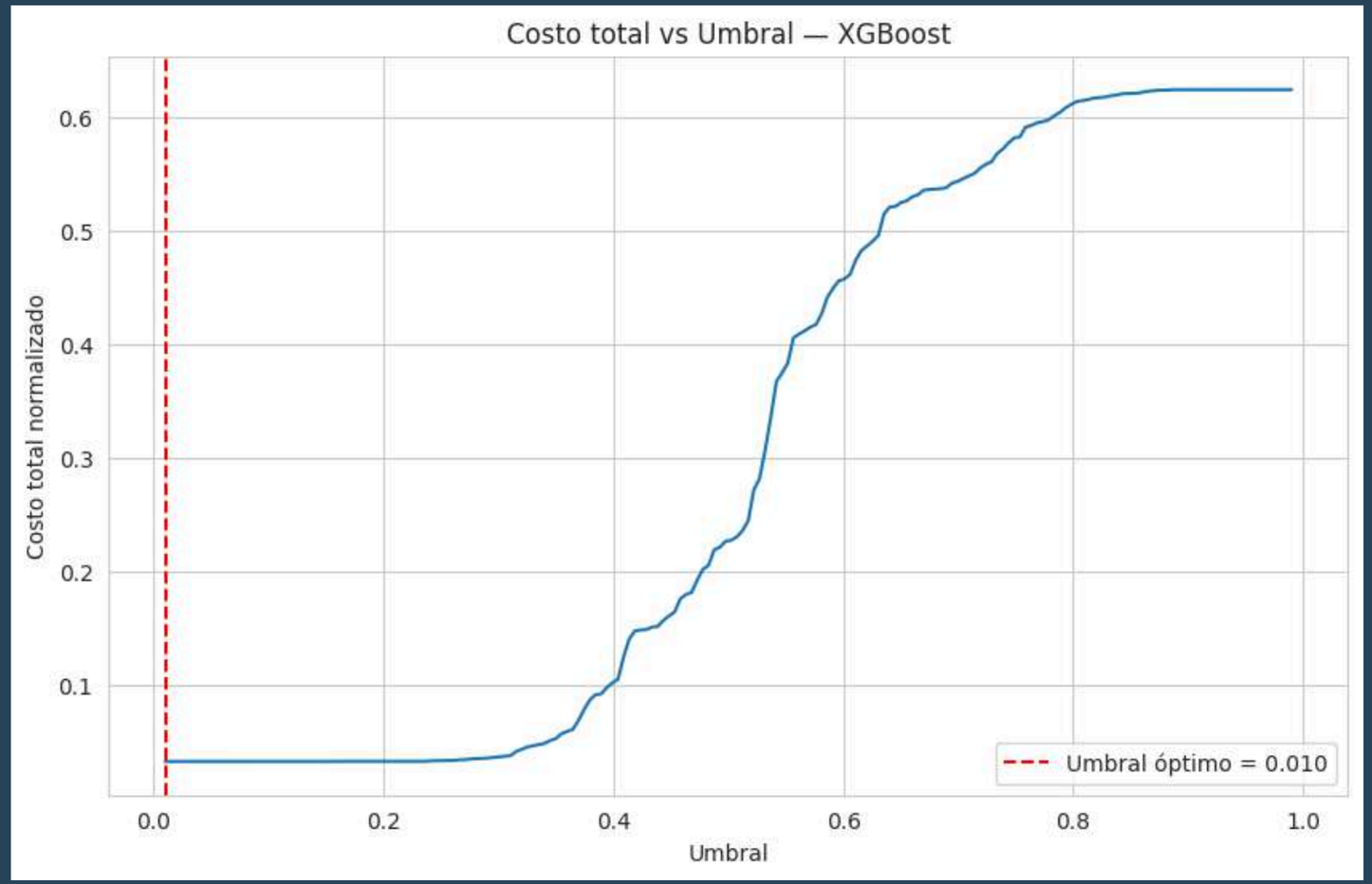
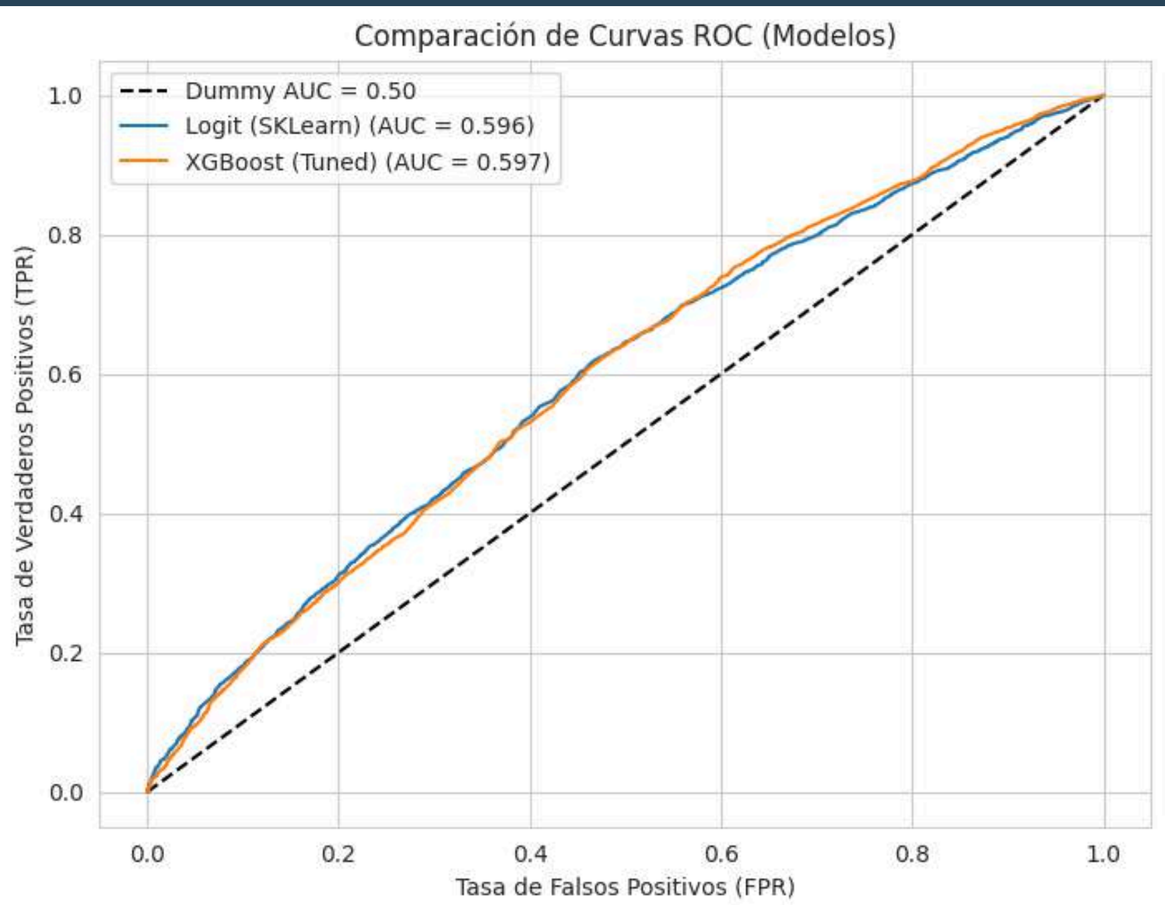


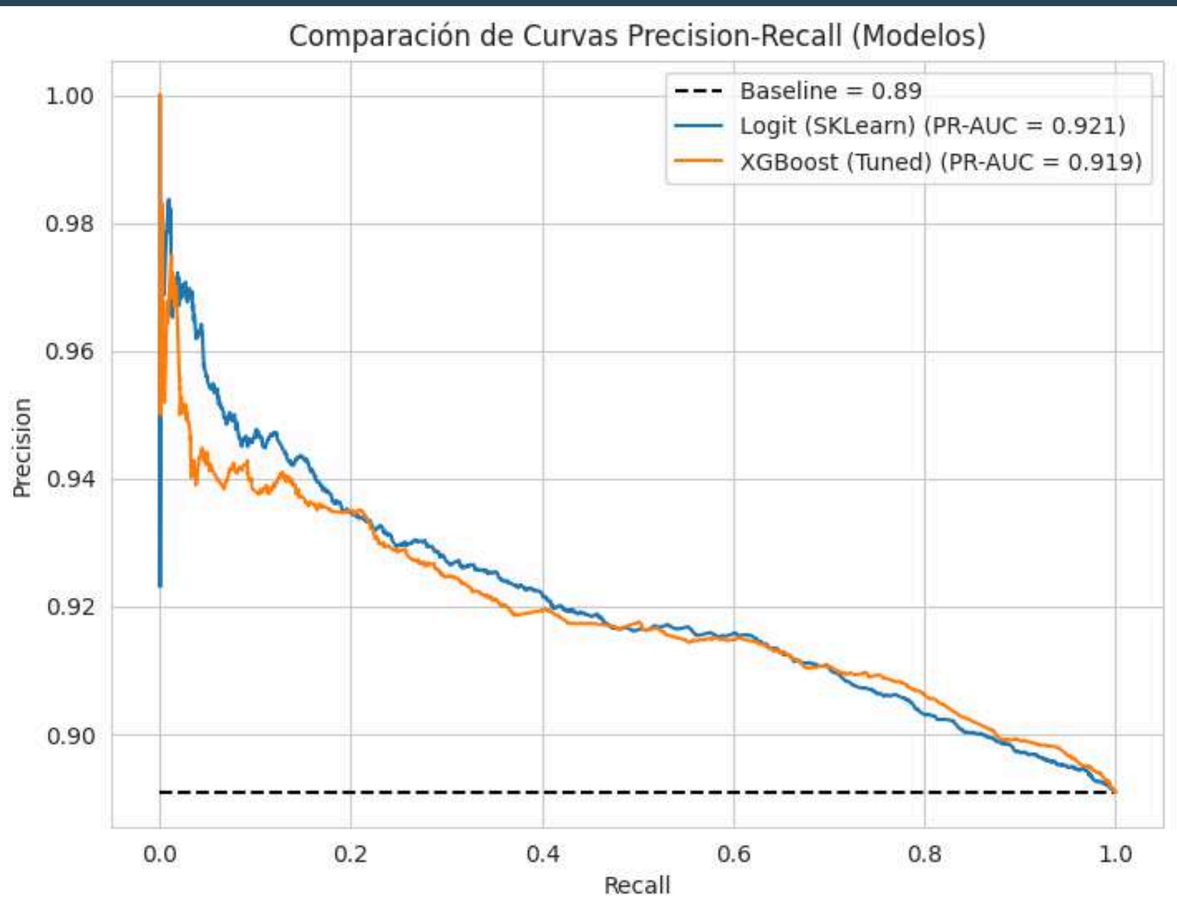
Figura 8: Costo Total vs Umbral

Table 3: Comparación de Modelos: Rendimiento con Umbral Óptimo

Métrica	Logit Mejorado	XGBoost
F1	**0.9424**	**0.9424**
Recall (Clase 1)	**1.0000**	**1.0000**
Precision (Clase 1)	0.8910	0.8910
ROC-AUC	0.5958	**0.5966**
PR-AUC	**0.9205**	0.9187
Accuracy	0.89	0.89
Macro Avg F1	0.47	0.47
Recall (Clase 0)	0.00	0.00

Grupo 5


Comparación de Curvas ROC (Modelos)
Dummy AUC = 0.50
Logit (SKLearn) (AUC = 0.596)
XGBoost (Tuned) (AUC = 0.597)
Tasa de Verdaderos Positivos (TPR)
Tasa de Falsos Positivos (FPR)
0.0
0.2
0.4
0.6
0.8
1.0


Comparación de Curvas Precision-Recall (Modelos)
Baseline = 0.89
Logit (SKLearn) (PR-AUC = 0.921)
XGBoost (Tuned) (PR-AUC = 0.919)
Precision
Recall
1.00
0.98
0.96
0.94
0.92
0.90
0.0
0.2
0.4
0.6
0.8
1.0

Grupo 5

DAG EXPLÍCITO

El DAG propone que el acceso adecuado a controles prenatales depende directamente de factores estructurales como nivel de riqueza, educación, edad y área urbana, que influyen en la capacidad real de una gestante para acceder a servicios de salud.

Grupo 5

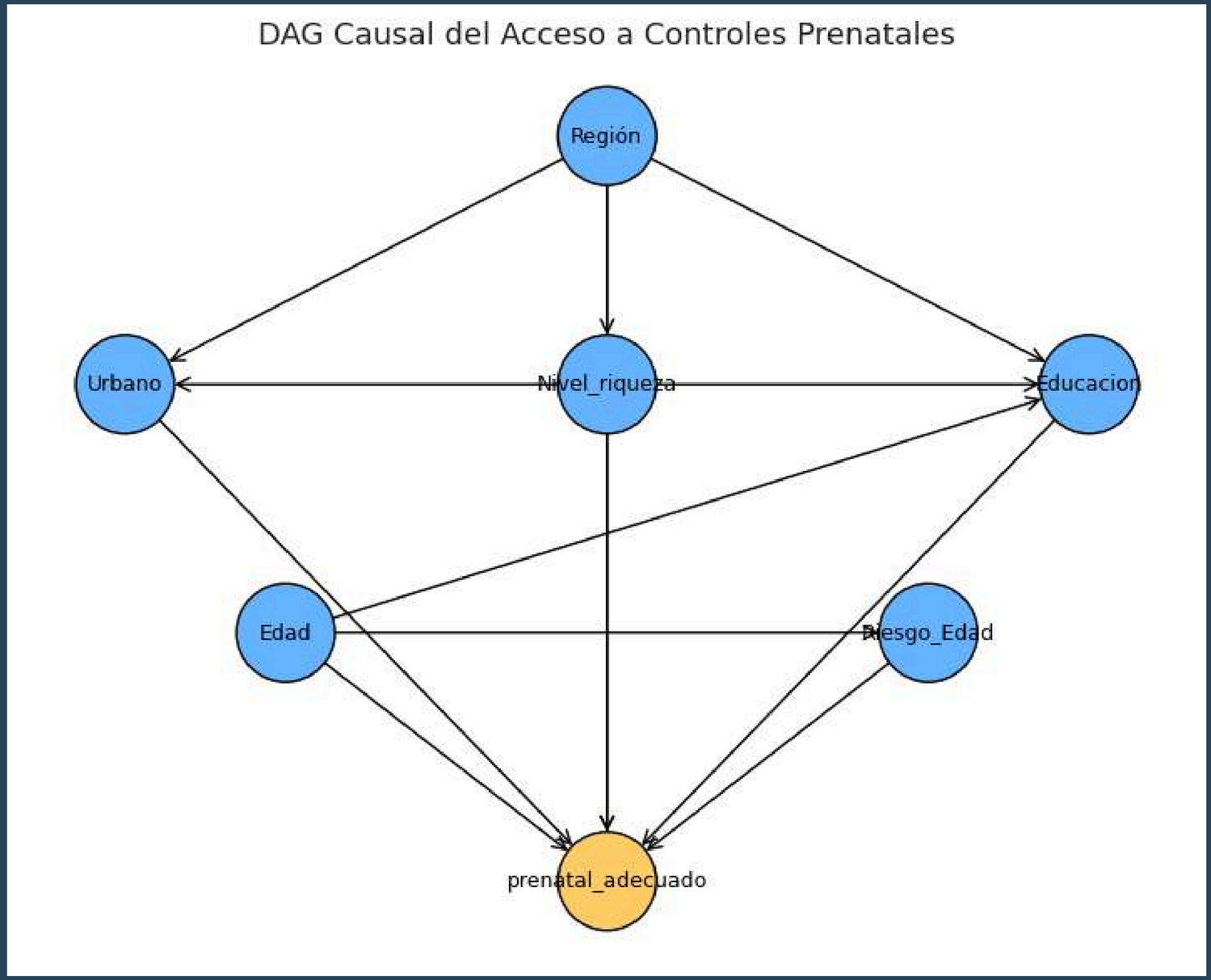


Figura 6: DAG CAUSAL

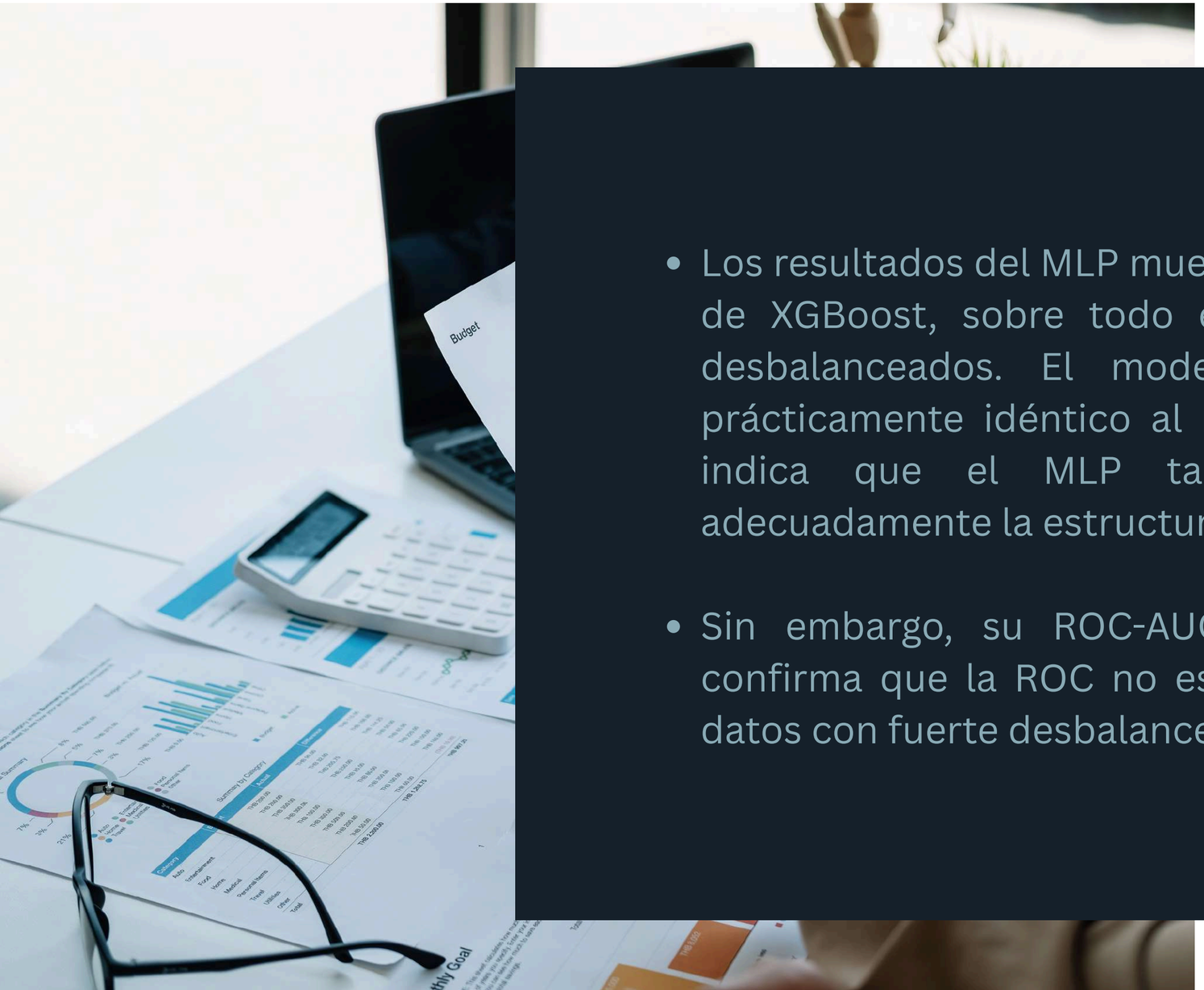
Grupo 5

REDES NEURONALES

Grupo 5

```
=== MÉTRICAS DEL MLP ===
ROC-AUC: 0.6004
PR-AUC: 0.9176
F1: 0.9424
Precision: 0.8911
Recall: 0.9999

Matriz de Confusión:
[[    2  2252]
 [    1 18422]]

Reporte:
              precision    recall  f1-score   support

           0       0.67      0.00      0.00      2254
           1       0.89      1.00      0.94     18423

    accuracy                           0.89     20677
   macro avg       0.78      0.50      0.47     20677
weighted avg       0.87      0.89      0.84     20677
```

Figura 6: Métricas y matriz de confusión



- Los resultados del MLP muestran un desempeño muy similar al de XGBoost, sobre todo en métricas robustas para datos desbalanceados. El modelo logra un PR-AUC = 0.918, prácticamente idéntico al mejor modelo del trabajo, lo que indica que el MLP también es capaz de capturar adecuadamente la estructura subyacente del problema.

- Sin embargo, su ROC-AUC (0.600) es moderado, lo que confirma que la ROC no es una métrica útil en este tipo de datos con fuerte desbalance.

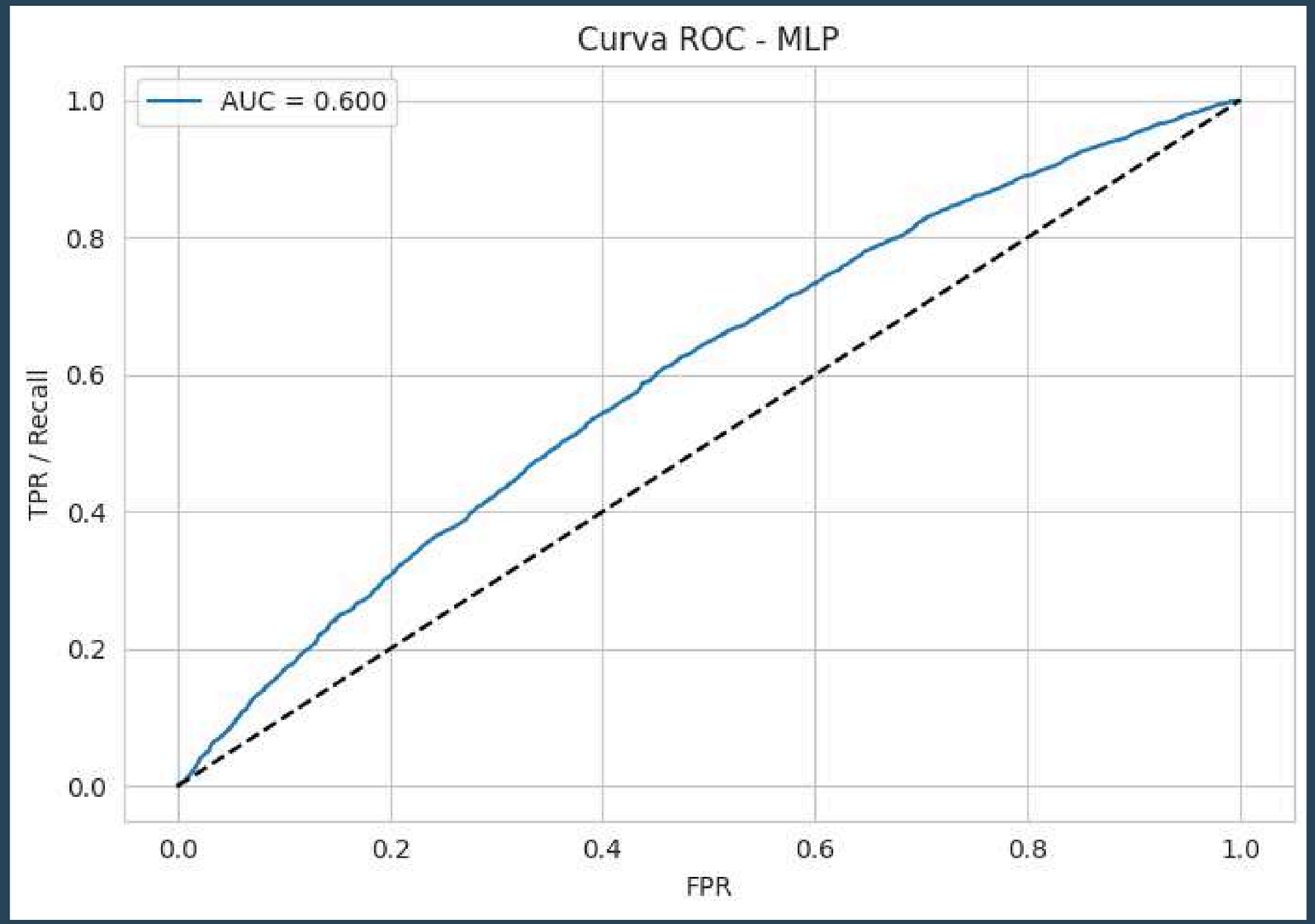


Figura 6: Curva ROC - MLP

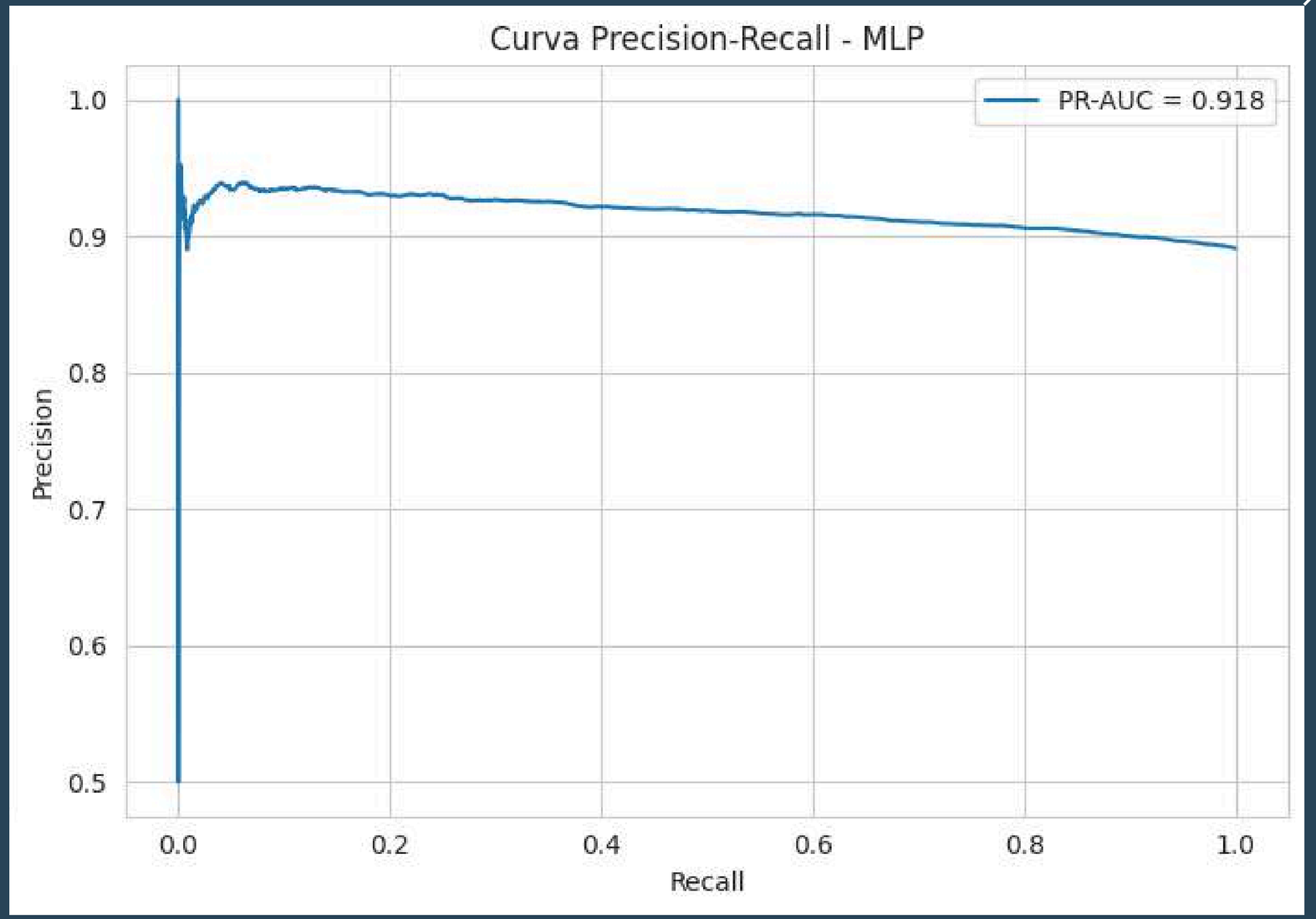
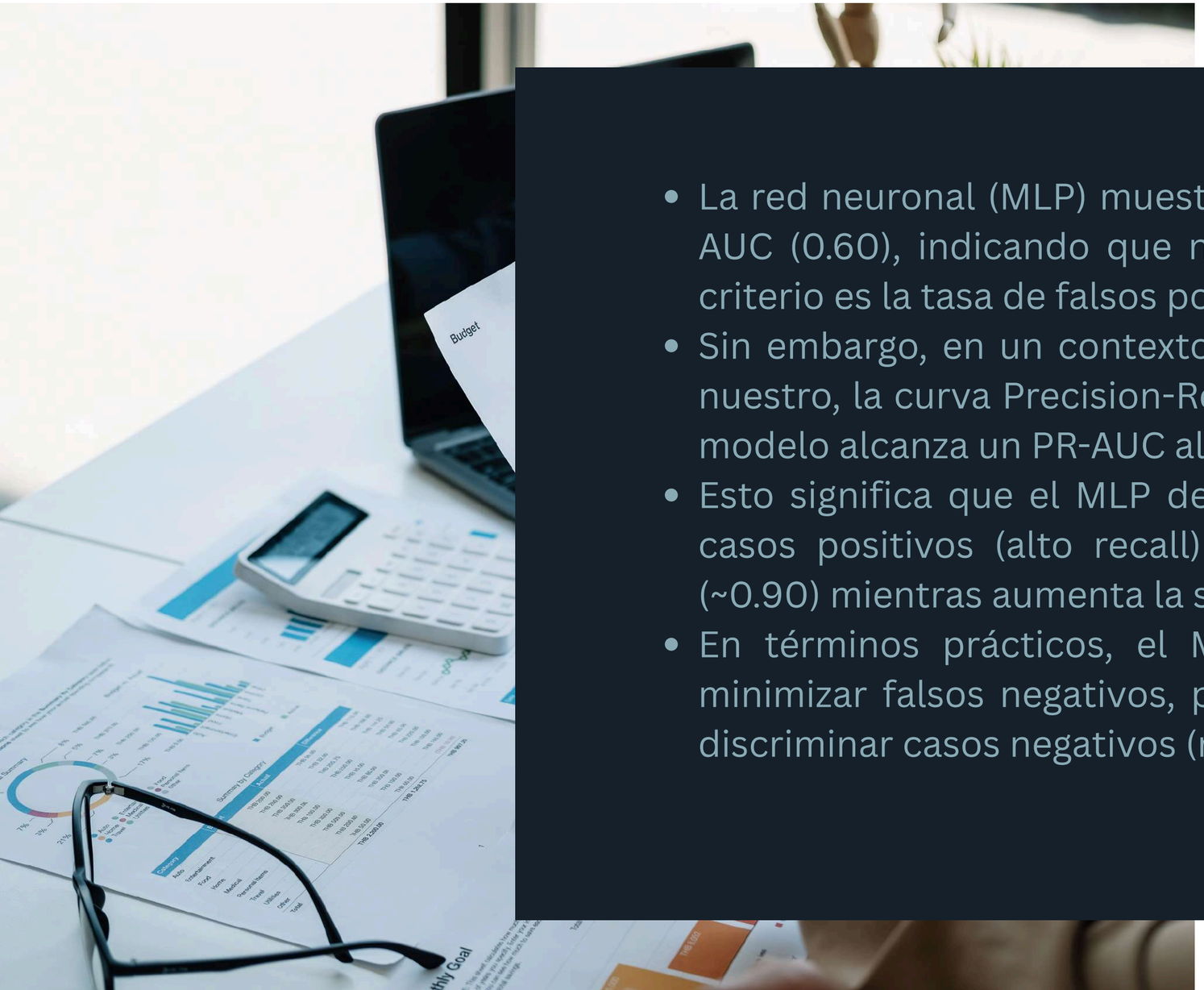


Figura 6: Curva MLP

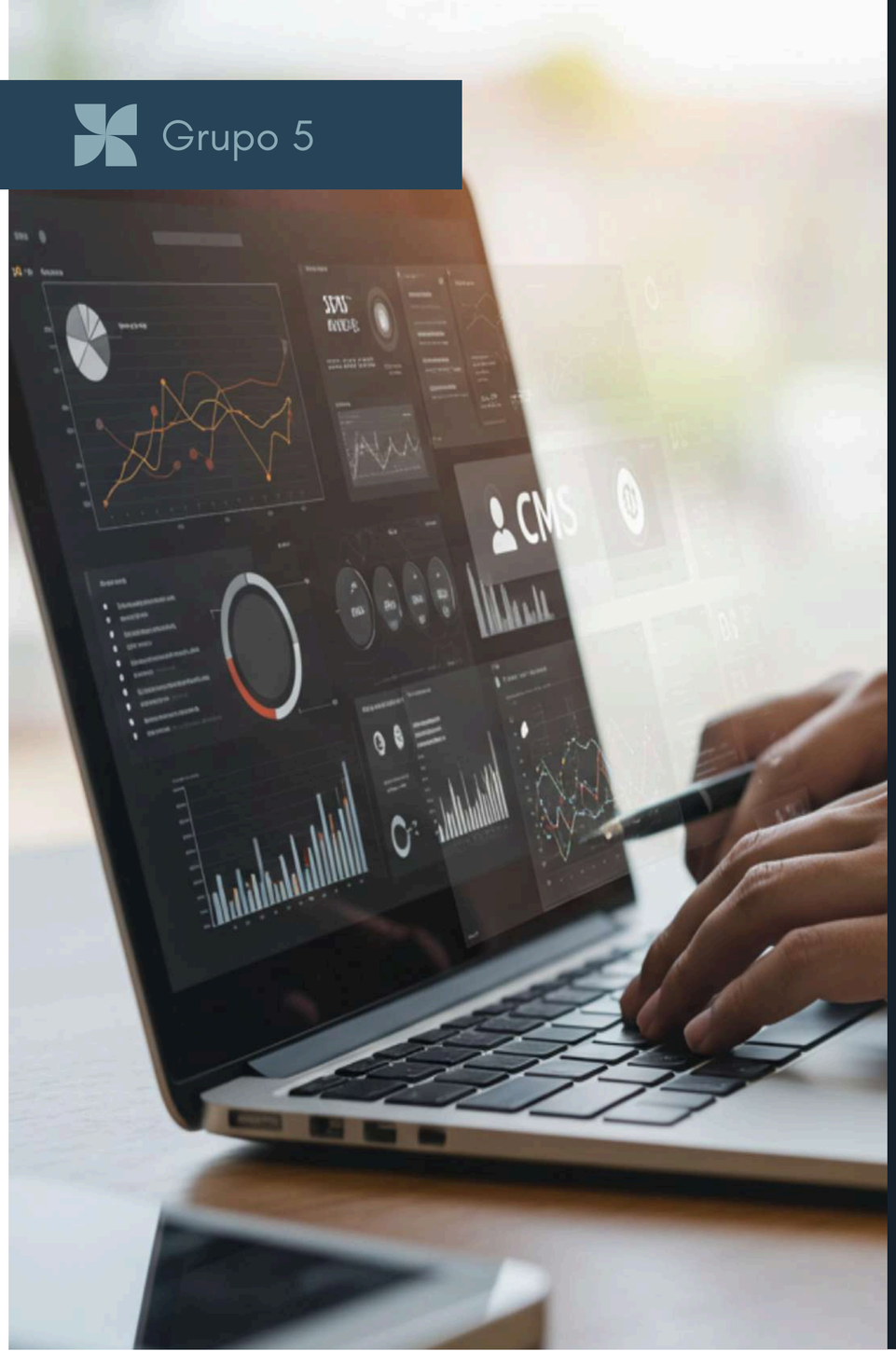
Grupo 5



- La red neuronal (MLP) muestra un desempeño modesto en ROC-AUC (0.60), indicando que no separa bien las clases cuando el criterio es la tasa de falsos positivos.
- Sin embargo, en un contexto altamente desbalanceado como el nuestro, la curva Precision-Recall es mucho más informativa, y el modelo alcanza un PR-AUC alto (0.918).
- Esto significa que el MLP detecta correctamente casi todos los casos positivos (alto recall) y mantiene una precisión estable (~0.90) mientras aumenta la sensibilidad.
- En términos prácticos, el MLP es útil cuando el objetivo es minimizar falsos negativos, pero aún presenta limitaciones para discriminar casos negativos (reflejadas en el ROC).

Grupo 5

INTRODUCCIÓN A CIENCIA DE DATOS Y MACHINE LEARNING CON PYTHON



Gracias

Fabricio Calle Cardoza \\ Bianca Nicole Jimenez Vargas \\ Aracely Lalupu Lalupu \\ Jose Alonso Naira Carmen.