

PROYECTO FINAL — ENTREGA P1

Daniel Guillermo Sandoval Toro
Estudiante — Grupo 9

Curso:
Introducción a Ciencia de Datos e Inteligencia Artificial

25 de Septiembre del 2026

**CIENCIA DE DATOS E
INTELIGENCIA ARTIFICIAL**

✉ daniel.sandoval.t@utec.edu.pe

Tabla de contenidos

1. CONTEXTO DEL DATASET	4
1.1. Descripción del problema	4
1.2. Fuente	4
1.3. Objetivo del análisis	4
2. CARGA Y EXPLORACIÓN INICIAL	5
2.1. Dimensiones y estructura	5
2.2. El diagnóstico de nulos: lo que el archivo declara y lo que oculta	6
2.3. Duplicados y balance de la variable objetivo	7
3. PREPROCESAMIENTO DE LOS DATOS	8
3.1. Renombrado de columnas	8
3.2. Eliminación de duplicados	8
3.3. Tratamiento de valores faltantes	8
3.4. Recodificación del centinela 999 en pdays	9
3.5. Variable degenerada: default	9
3.6. Conversión de tipos	9
3.7. Enriquecimiento: creación de nuevas variables	10
4. ANÁLISIS EXPLORATORIO DE DATOS (EDA)	11
4.1. Análisis univariado	11
4.2. Análisis multivariado	16
4.3. Conclusiones generales del EDA	26
5. ANEXO: CÓDIGO DEL PIPELINE	28

Listado de Tablas

1. Estructura del dataset original.	5
2. Valores faltantes encubiertos.	6
3. Bitácora de preprocesamiento.	8
4. Variables derivadas en la etapa de enriquecimiento.	10
5. Medidas de tendencia central.	11
6. Medidas de dispersión.	11
7. Cuartiles y medidas de forma.	11
8. Detección de valores atípicos.	13

9.	Frecuencias agrupadas de la edad.	15
10.	Pares con riesgo de colinealidad.	16
11.	Matriz de covarianzas.	17
12.	Estacionalidad de la campaña.	22
13.	Verificación del dataset preprocesado.	28

Listado de Figuras

1.	Porcentaje de registros con valor unknown por variable. Ninguno es detectado como nulo por las funciones estándar.	6
2.	Distribución de la variable objetivo. El desbalance condiciona toda la evaluación posterior.	7
3.	Distribución de frecuencias de las diez variables numéricas. La línea naranja marca la media; la discontinua, la mediana.	13
4.	Dispersión y valores atípicos de las variables numéricas según el criterio de Tukey.	13
5.	Frecuencia de las categorías en las variables cualitativas.	14
6.	Correlación de Pearson entre variables numéricas (datos ya preprocesados).	16
7.	Correlación de Spearman. Al basarse en rangos, es robusta frente a los valores atípicos detectados.	18
8.	Correlación de cada variable numérica con la suscripción codificada como 0/1.	19
9.	Boxplots comparativos: distribución de cada variable numérica según el resultado de la llamada.	19
10.	Distribución de la duración de la llamada según el resultado, y tasa de suscripción por tramo de duración.	20
11.	Tasa de suscripción por categoría. La línea discontinua marca la media global de 11.27%.	21
12.	Volumen de llamadas y tasa de suscripción por mes. Los dos paneles comparten el eje horizontal.	22
13.	Seis combinaciones de pares de variables numéricas sobre una muestra aleatoria de 4,000 registros.	24

1. CONTEXTO DEL DATASET

1.1. Descripción del problema

Una institución bancaria portuguesa ejecutó, entre **mayo de 2008 y noviembre de 2010**, una serie de campañas de marketing directo por vía telefónica. El producto ofrecido fue un **depósito a plazo fijo**: un instrumento de ahorro en el que el cliente inmoviliza un capital durante un período pactado a cambio de una tasa de interés.

El problema de negocio es de eficiencia. Llamar es caro: consume tiempo de operador, minutos de telefonía y, sobre todo, paciencia del cliente. Con una tasa de aceptación cercana al **11.27%**, aproximadamente nueve de cada diez llamadas terminan en rechazo. Cada punto porcentual de mejora en la selección de a quién llamar se traduce directamente en costo evitado.

1.2. Fuente

Los datos provienen del repositorio *UCI Machine Learning Repository* y fueron recopilados y documentados por Moro, Cortez y Rita (2014) y Moro, Laureano y Cortez (2011) y Moro, Rita y Cortez (2014). Se trabaja con el archivo `bank-additional-full.csv`, la versión ampliada de 2014, que incorpora cinco indicadores macroeconómicos ausentes en la versión original de 2012.

Esa elección es deliberada: el período cubierto coincide con la crisis financiera global, y la disposición de un cliente a inmovilizar capital depende del contexto de tasas tanto como de su perfil personal. Sin las variables macroeconómicas, ese eje del fenómeno queda invisible.

1.3. Objetivo del análisis

El objetivo de esta primera entrega es **entender y dejar utilizable** el conjunto de datos, no todavía predecir. En concreto:

1. Diagnosticar la calidad real de los datos, distinguiendo lo que el archivo *declara* de lo que efectivamente contiene.
2. Documentar y justificar cada transformación aplicada.
3. Caracterizar el comportamiento univariado y multivariado de las variables, identificando cuáles aportan señal frente a la suscripción y cuáles introducen problemas para el modelado posterior.

La variable objetivo es `suscribio` (y en el archivo original): indica si el cliente contrató o no el depósito a plazo.

2. CARGA Y EXPLORACIÓN INICIAL

2.1. Dimensiones y estructura

```
leer_tabla("tbl_estructura.csv") |>
  rename(Variable = 1, Tipo = 2, `No nulos` = 3, `Valores únicos` = 4) |>
  tabla_bonita("Estructura del dataset original.", fs = 8)
```

Tabla 1: Estructura del dataset original.

Variable	Tipo	No nulos	Valores únicos
age	int64	41,188	78
job	str	41,188	12
marital	str	41,188	4
education	str	41,188	8
default	str	41,188	3
housing	str	41,188	3
loan	str	41,188	3
contact	str	41,188	2
month	str	41,188	10
day_of_week	str	41,188	5
duration	int64	41,188	1,544
campaign	int64	41,188	42
pdays	int64	41,188	27
previous	int64	41,188	8
poutcome	str	41,188	3
emp.var.rate	float64	41,188	10
cons.price.idx	float64	41,188	26
cons.conf.idx	float64	41,188	26
euribor3m	float64	41,188	316
nr.employed	float64	41,188	11
y	str	41,188	2

El archivo contiene **41,188 registros** y **21 variables**: 10 numéricas y 11 cualitativas, incluida la variable objetivo. Cada fila es una llamada telefónica, no un cliente: un mismo cliente puede aparecer varias veces si fue contactado en repetidas ocasiones.

El bloque de variables se organiza en cuatro grupos:

- **Perfil del cliente:** age, job, marital, education, default, housing, loan.
- **Último contacto de la campaña actual:** contact, month, day_of_week, duration.
- **Historial de contacto:** campaign, pdays, previous, poutcome.
- **Contexto macroeconómico:** emp.var.rate, cons.price.idx, cons.conf.idx, euribor3m, nr.employed.

2.2. El diagnóstico de nulos: lo que el archivo declara y lo que oculta

La documentación original afirma que el conjunto no tiene valores faltantes, y una verificación ingenua lo confirma: `is.na()` devuelve **cero** en las 21 columnas.

Esa conclusión es falsa. Los faltantes existen, pero están codificados como la cadena de texto "unknown", que pandas y R leen como una categoría válida más. La Figura 1 cuantifica el problema.

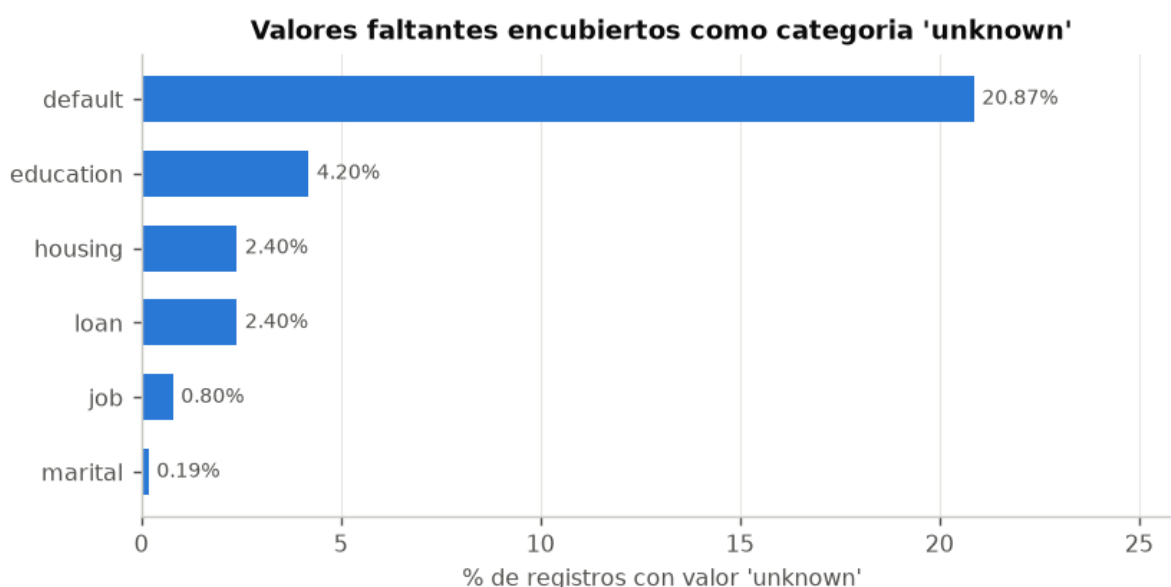


Figura 1: Porcentaje de registros con valor unknown por variable. Ninguno es detectado como nulo por las funciones estándar.

Tabla 2: Valores faltantes encubiertos.

Variable	N.º unknown	% del total
default	8,597	20.87
education	1,731	4.20
housing	990	2.40
loan	990	2.40
job	330	0.80

marital	80	0.19
---------	----	------

Seis variables están afectadas. El caso crítico es `default`, con **20.87%** de los registros sin información: uno de cada cinco clientes. Le siguen `education` (4.20%), `housing` y `loan` (2.40% cada una), `job` (0.80%) y `marital` (0.19%).

La lección metodológica es general y conviene enunciarla: **un dato faltante que no se declara como tal es más peligroso que uno que sí lo hace**, porque atraviesa silenciosamente todas las etapas posteriores. Si no se hubiese inspeccionado el contenido de cada columna categórica, el modelo de la Entrega 2 habría tratado “no sabemos si tiene deudas” como si fuera una característica del cliente equiparable a “es técnico” o “está casado”.

2.3. Duplicados y balance de la variable objetivo

Se detectan **12 filas exactamente idénticas** en las 21 columnas. Dado que cada registro documenta una llamada individual —con su duración en segundos, su día y su contexto macroeconómico—, la coincidencia simultánea en las 21 variables no es plausible como fenómeno real: son errores de carga.

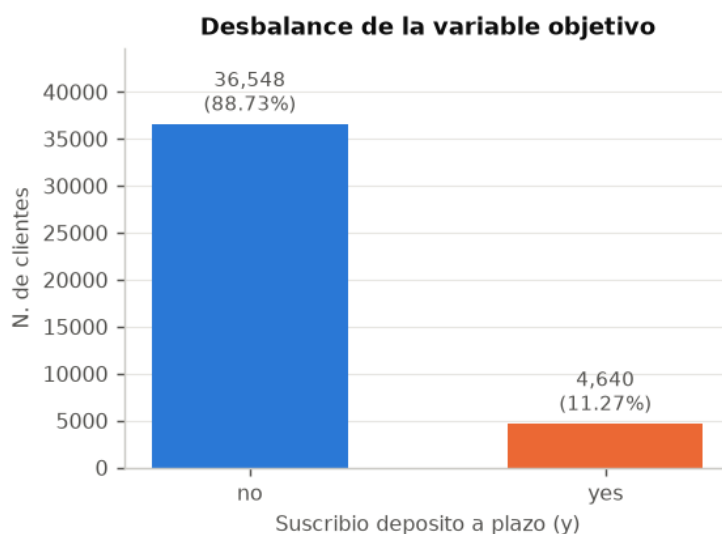


Figura 2: Distribución de la variable objetivo. El desbalance condiciona toda la evaluación posterior.

La Figura 2 muestra el hallazgo que más condiciona el resto del proyecto: solo el **11.27%** de los contactos termina en suscripción.

Esto tiene una consecuencia inmediata para la Entrega 2, que conviene dejar anotada desde ahora: **la exactitud (accuracy) será una métrica engañosa**. Un clasificador que prediga “no” para absolutamente todos los clientes acertaría el 88.73% de las veces sin haber aprendido nada. La evaluación deberá apoyarse en precisión, exhaustividad, F1 y AUC sobre la clase minoritaria.

3. PREPROCESAMIENTO DE LOS DATOS

Cada decisión de esta sección se registró en una bitácora durante la ejecución del pipeline. La Tabla 3 la reproduce íntegra.

Tabla 3: Bitácora de preprocesamiento.

Paso	Decisión y justificación
Renombrado de columnas	21 columnas a snake_case en español; se eliminan los puntos de emp.var.rate y similares, que rompen las formulas de R.
Duplicados	Se eliminan 12 filas identicas (0.029% del total); dos llamadas reales no coinciden en las 21 variables.
Nulos encubiertos	'unknown' se convierte a NaN en 6 columnas para hacer medible el faltante antes de decidir su tratamiento.
Centinela 999	pdays=999 codifica 'nunca contactado'. Se separa en 'contactado_antes' (0/1) y 'dias_ultimo_contacto' (NaN si no aplica); así la media deja de estar contaminada.
Variable degenerada	credito_impago tiene solo 3 casos 'yes' en 41,188 (0.007%) y 20.87% 'unknown'. Se reemplaza por 'impago_desconocido', que si discrimina (5.15% vs 12.88% de conversion).
Imputacion	Se descarta la imputacion por moda: el faltante es informativo (educacion 'unknown' convierte 14.50% vs 11.27% global). Se conserva como categoria explicita 'desconocido'.
Tipos de dato	educacion, mes y dia_semana pasan a factor ORDENADO; el resto a factor nominal; se crea 'suscribio_bin' (0/1) para correlaciones.
Enriquecimiento	5 variables derivadas: grupo_etario, contacto_intensivo, duracion_min, exito_previo y trimestre.

3.1. Renombrado de columnas

Se normalizaron las 21 columnas a snake_case con nombres autoexplicativos en español. La motivación no es estética: nombres como emp.var.rate o cons.price.idx contienen puntos, que en R son el operador de acceso a componentes y rompen la sintaxis de fórmulas (y ~ emp.var.rate se interpreta mal). Corregirlo ahora evita errores silenciosos en el modelado.

3.2. Eliminación de duplicados

Se eliminaron las 12 filas duplicadas (0.029% del total). El criterio es el enunciado antes: dos llamadas reales no coinciden en las 21 variables simultáneamente. El volumen es despreciable y no altera ninguna distribución.

3.3. Tratamiento de valores faltantes

El tratamiento se decidió **columna por columna**, no con una regla única. El motivo es que en este dataset el faltante no es ruido aleatorio: *es informativo*.

La evidencia es directa. Quienes no declaran su nivel educativo susciben al **14.50 %**, por encima del 11.27 % global. Quienes no declaran su situación de impago susciben al **5.15 %**, muy por debajo del 12.88 % de quienes declaran no tener deudas impagas. El hecho de no responder correlaciona con el comportamiento de compra.

Esto descarta la imputación por moda. Reemplazar unknown por la categoría más frecuente destruiría esa señal y, peor aún, **fabricaría una certeza que el dato no tiene**: afirmaría que 8,597 clientes no tienen deudas impagas cuando lo único que se sabe es que no se les preguntó o no contestaron.

La decisión adoptada fue **conservar el faltante como categoría explícita** (desconocido) en `ocupacion`, `estado_civil`, `educacion`, `prestamo_vivienda` y `prestamo_personal`.

3.4. Recodificación del centinela 999 en pdays

La variable `pdays` registra los días transcurridos desde el último contacto de una campaña anterior. El valor **999 no significa “hace 999 días”**: es un código centinela que indica “*nunca fue contactado antes*”, y afecta al **96.32 %** de los registros.

Tratarlo como un número tiene consecuencias graves y verificables:

- La media de `pdays` en el archivo crudo es **962.5 días**, un valor que no describe a ningún cliente real.
- Genera una correlación espuria de **$r = -0.588$** con `previous`. Esa relación es un artefacto del código centinela, no un fenómeno: desaparece al recodificar (baja a -0.04), como se verifica en la Figura 6.

La solución aplicada descompone la variable en dos, que es lo que en realidad estaba mezclado:

- `contactado_antes`: indicador binario (0/1).
- `dias_ultimo_contacto`: el valor real, NA cuando no aplica.

3.5. Variable degenerada: default

La distribución de `default` es **no: 32,588 · unknown: 8,597 · yes: 3**.

Tres casos positivos entre 41,188 registros (0.007 %) no permiten estimar nada: cualquier partición de entrenamiento y prueba dejaría la categoría con cero o un caso. La variable se sustituyó por el indicador `impago_desconocido`, que sí discrimina (5.15 % frente a 12.88 % de conversión) y captura lo único que la columna informa de manera fiable.

3.6. Conversión de tipos

Se asignó el tipo correcto a cada variable, distinguiendo lo nominal de lo ordinal. `educacion` (de `illiterate` a `university.degree`), `mes` y `dia_semana` son factores **ordenados**: su orden es

intrínseco y perderlo impediría detectar tendencias monótonas. El resto de cualitativas son factores nominales.

3.7. Enriquecimiento: creación de nuevas variables

Se derivaron cinco variables, recogidas en la Tabla 4, cada una con una justificación concreta:

Tabla 4: *Variables derivadas en la etapa de enriquecimiento.*

Variable	Construcción	Por qué
grupo_etario	Corte de edad en 5 tramos	Los extremos etarios se comportan de forma opuesta al centro; el efecto es no lineal y la edad continua lo diluye
contacto_intensivo	<code>n_contactos_campana > 3</code>	Marca al cliente saturado por insistencia
duracion_min	<code>duracion / 60</code>	Unidad interpretable por el área comercial
exito_previo	<code>resultado_campana_previa == "success"</code>	Aísla la señal más fuerte del conjunto
trimestre	Agrupación de mes	Sintetiza la estacionalidad detectada

El dataset resultante tiene **41,176 filas** y **28 columnas**.

4. ANÁLISIS EXPLORATORIO DE DATOS (EDA)

4.1. Análisis univariado

4.1.1. Medidas de tendencia central y de dispersión

Tabla 5: Medidas de tendencia central.

Variable	Media	Mediana	Moda
edad	40.02	38.00	31.00
duracion	258.32	180.00	85.00
n_contactos_campana	2.57	2.00	1.00
dias_ultimo_contacto	6.02	6.00	3.00
n_contactos_previos	0.17	0.00	0.00
var_empleo	0.08	1.10	1.40
idx_precios	93.58	93.75	93.99
idx_confianza	-40.50	-41.80	-36.40
euribor_3m	3.62	4.86	4.86
n_empleados	5,167.04	5,191.00	5,228.10

Tabla 6: Medidas de dispersión.

Variable	Desv. Est.	Varianza	Rango	RIC	CV%
edad	10.42	108.59	81.00	15.00	26.04
duracion	259.30	67,239.25	4,918.00	217.00	100.38
n_contactos_campana	2.77	7.68	55.00	2.00	107.88
dias_ultimo_contacto	3.83	14.63	27.00	4.00	63.59
n_contactos_previos	0.50	0.24	7.00	0.00	286.08
var_empleo	1.57	2.47	4.80	3.20	1,917.55
idx_precios	0.58	0.34	2.57	0.92	0.62
idx_confianza	4.63	21.42	23.90	6.30	-11.43
euribor_3m	1.73	3.01	4.41	3.62	47.90
n_empleados	72.25	5,220.26	264.50	129.00	1.40

Tabla 7: Cuartiles y medidas de forma.

Variable	Mín	Q1	Q3	Máx	Asimetría	Curtosis
edad	17.00	32.00	47.00	98.00	0.78	0.79

duracion	0.00	102.00	319.00	4,918.00	3.26	20.24
n_contactos_campana	1.00	1.00	3.00	56.00	4.76	36.97
dias_ultimo_contacto	0.00	3.00	7.00	27.00	1.46	2.56
n_contactos_previos	0.00	0.00	0.00	7.00	3.83	20.10
var_empleo	-3.40	-1.80	1.40	1.40	-0.72	-1.06
idx_precios	92.20	93.08	93.99	94.77	-0.23	-0.83
idx_confianza	-50.80	-42.70	-36.40	-26.90	0.30	-0.36
euribor_3m	0.63	1.34	4.96	5.04	-0.71	-1.41
n_empleados	4,963.60	5,099.10	5,228.10	5,228.10	-1.04	0.00

Las medidas solo tienen lectura conjunta: una media sin su dispersión no informa sobre su propia representatividad. Los casos que merecen comentario:

edad — Media 40.0 y mediana 38.0 años, muy próximas, con asimetría leve (+0.78). La distribución es razonablemente simétrica y la media es un buen resumen. El coeficiente de variación del 26% indica una población homogénea en edad.

duracion — Media 258 segundos frente a mediana 180. La media supera a la mediana en un 43%, señal de asimetría positiva marcada (+3.26) arrastrada por llamadas muy largas: el máximo es de 4,918 segundos (82 minutos). Aquí **la media no es representativa** y debe usarse la mediana.

n_contactos_campana — Mediana de 2 contactos, pero máximo de 56. Asimetría extrema (+4.76) y curtosis muy elevada. La cola derecha describe un comportamiento operativo concreto: un pequeño grupo de clientes fue llamado de forma insistente.

idx_confianza y var_empleo — Presentan coeficientes de variación distorsionados por tomar valores negativos, y sus distribuciones son multimodales. Esto no es un defecto del dato sino su naturaleza: son indicadores **mensuales** repetidos en todas las llamadas del mismo mes, por lo que toman pocos valores distintos.

4.1.2. Histogramas

La Figura 3 confirma tres patrones distintos:

1. **Asimetría positiva fuerte** en *duracion*, *n_contactos_campana* y *n_contactos_previos*: muchos valores bajos y una cola derecha larga. Son candidatas naturales a transformación logarítmica si el modelo posterior lo requiere.
2. **Cuasi-simetría** en *edad*, la única variable con forma aproximadamente normal.
3. **Multimodalidad** en las cinco macroeconómicas. La separación en bloques discretos refleja los regímenes de tasa que atravesó el período 2008–2010, y permite anticipar que actuarán como un *proxy* del momento temporal.

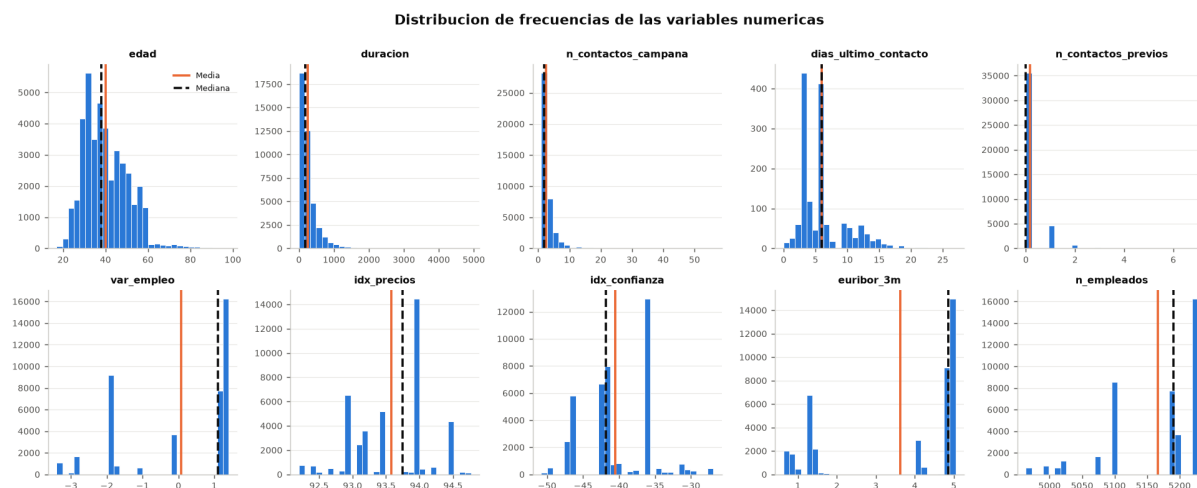


Figura 3: Distribución de frecuencias de las diez variables numéricas. La línea naranja marca la media; la discontinua, la mediana.

La distancia visible entre media y mediana en *duracion* y *n_contactos_campana* es la traducción gráfica de la asimetría ya comentada.

4.1.3. Diagramas de caja

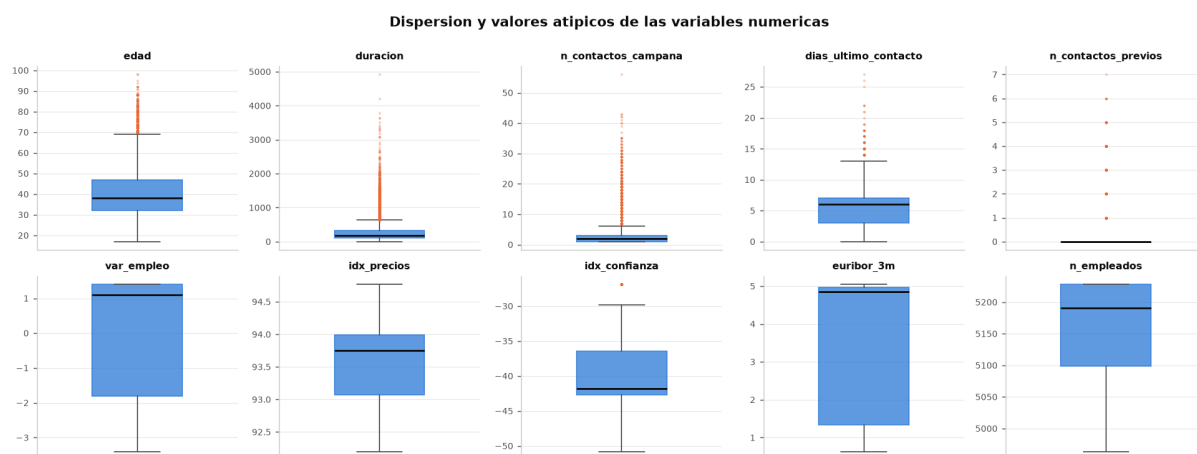


Figura 4: Dispersión y valores atípicos de las variables numéricas según el criterio de Tukey.

Tabla 8: Detección de valores atípicos.

Variable	Límite inf.	Límite sup.	N.º atípicos	% atípicos
edad	9.50	69.50	468	1.14
duracion	-223.50	644.50	2,963	7.20
n_contactos_campana	-2.00	6.00	2,406	5.84
dias_ultimo_contacto	-3.00	13.00	82	5.41
n_contactos_previos	0.00	0.00	5,625	13.66
var_empleo	-6.60	6.20	0	0.00

idx_precios	91.70	95.37	0	0.00
idx_confianza	-52.15	-26.95	446	1.08
euribor_3m	-4.08	10.39	0	0.00
n_empleados	4,905.60	5,421.60	0	0.00

El punto crítico de esta sección es **cómo interpretar los atípicos**. El criterio de Tukey (Tukey 1977) los marca como extremos respecto de la distribución, pero eso no los convierte en errores.

En duracion y n_contactos_campana los valores extremos son **genuinos y además informativos**: una llamada de 82 minutos es exactamente el tipo de conversación que termina en contratación. Eliminarlos destruiría la señal más fuerte del conjunto. En edad, los clientes por encima de 70 años tampoco son errores: la Figura 11 mostrará que el segmento de 60+ años convierte al 45.5 %, cuatro veces la media.

La decisión adoptada es **conservar todos los valores atípicos**. Son observaciones válidas de un fenómeno con cola larga, no ruido de medición.

4.1.4. Variables cualitativas

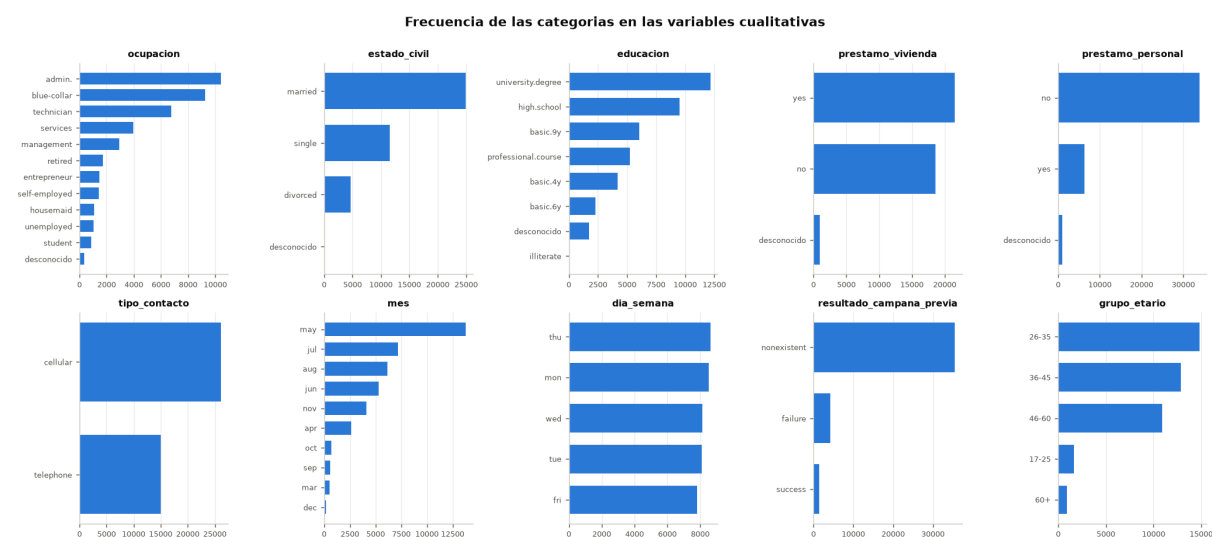


Figura 5: Frecuencia de las categorías en las variables cualitativas.

Los desequilibrios relevantes que muestra la Figura 5:

- **ocupacion** está dominada por admin. (10,422) y blue-collar (9,254), mientras student apenas reúne 875 casos. Las tasas de conversión de los grupos pequeños deberán leerse con cautela por su menor precisión.
- **mes** presenta una concentración extrema: mayo acumula **13,767 llamadas**, un tercio del total, mientras diciembre reúne 182. Además faltan enero y febrero por completo.

- **resultado_campana_previa** es nonexistent en el 86% de los casos, coherente con el 96.32% de pdays = 999.

4.1.5. Tabla de frecuencias agrupadas

Tabla 9: *Frecuencias agrupadas de la edad.*

Intervalo	Frec. absoluta	Marca de clase	Frec. acumulada	Frec. relativa%	Frec. rel. acum. %
[15, 25)	1,067	20	1,067	2.59	2.59
[25, 35)	13,684	30	14,751	33.24	35.83
[35, 45)	13,495	40	28,246	32.78	68.61
[45, 55)	8,702	50	36,948	21.14	89.75
[55, 65)	3,566	60	40,514	8.66	98.41
[65, 75)	393	70	40,907	0.95	99.36
[75, 85)	211	80	41,118	0.51	99.87
[85, 95)	55	90	41,173	0.13	100.00

La concentración se sitúa en los intervalos de 25 a 45 años, que reúnen cerca del 66% de los contactados: la base de clientes de la campaña es predominantemente adulta-joven en edad laboral activa.

4.2. Análisis multivariado

4.2.1. Mapa de calor de correlaciones

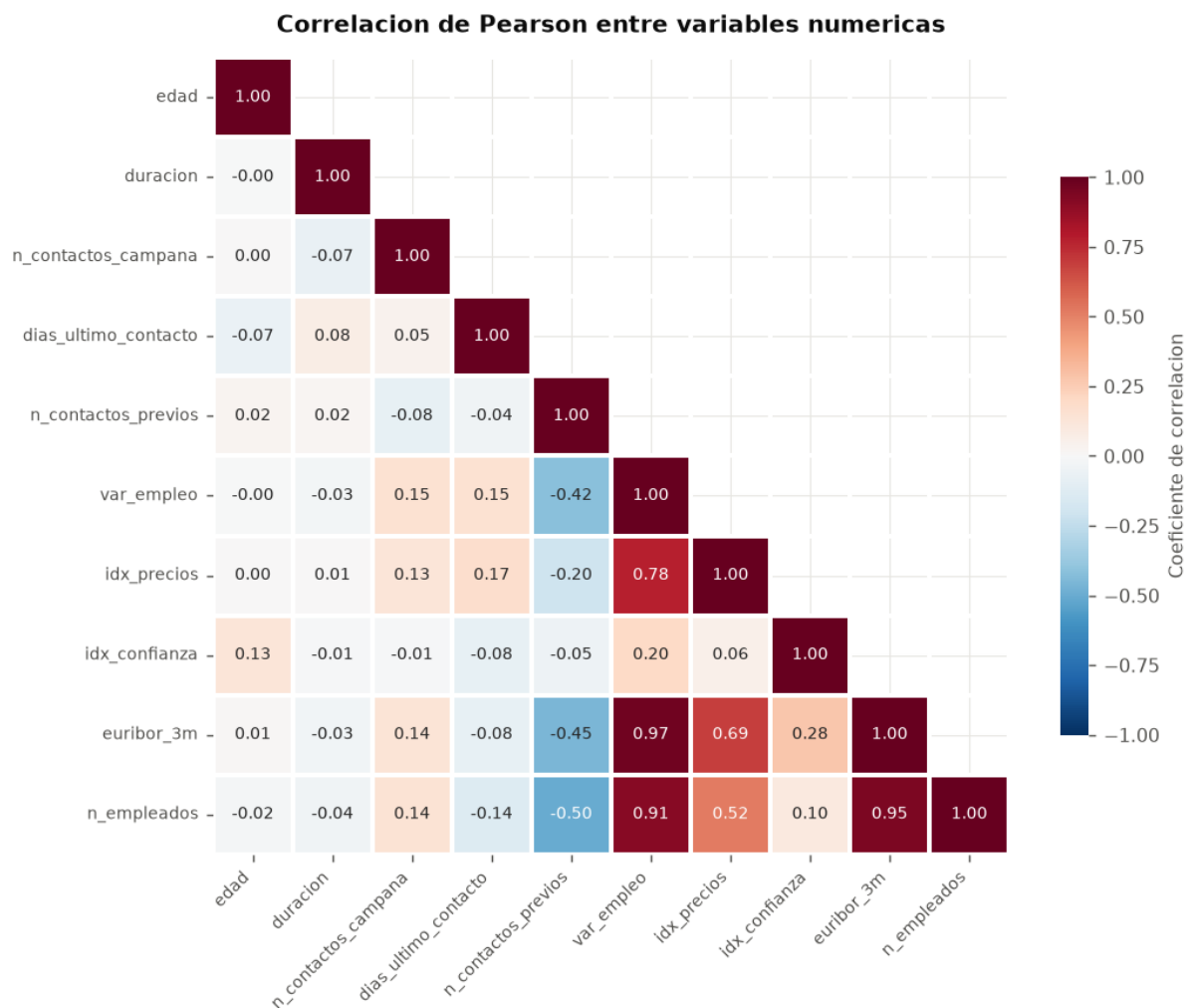


Figura 6: Correlación de Pearson entre variables numéricas (datos ya preprocesados).

Se calcularon **ambos coeficientes de forma deliberada**. Pearson mide relación lineal y es sensible a los valores extremos que el análisis univariado dejó en evidencia; Spearman opera sobre rangos, detecta cualquier relación monótona y resiste esos extremos. La coincidencia entre ambos refuerza las conclusiones; la discrepancia señalaría no linealidad.

Tabla 10: Pares con riesgo de colinealidad.

Variable A	Variable B	Pearson	Spearman	Covarianza
var_empleo	euribor_3m	0.972	0.940	2.649
euribor_3m	n_empleados	0.945	0.929	118.441
var_empleo	n_empleados	0.907	0.945	102.937
var_empleo	idx_precios	0.775	0.665	0.705

idx_precios	euribor_3m	0.688	0.491	0.691
idx_precios	n_empleados	0.522	0.465	21.829
n_contactos_previos	n_empleados	-0.501	-0.439	-17.931

El hallazgo dominante es un **bloque de colinealidad severa** entre las variables macroeconómicas:

- $\text{euribor_3m} \sim \text{var_empleo}$: $r = 0.972$
- $\text{euribor_3m} \sim \text{n_empleados}$: $r = 0.945$
- $\text{var_empleo} \sim \text{n_empleados}$: $r = +0.907$
- $\text{var_empleo} \sim \text{idx_precios}$: $r = +0.775$

Con $r = 0.972$, euribor_3m y var_empleo son **estadísticamente la misma variable**: una explica el 94.5 % de la varianza de la otra. No es casualidad económica sino estructural, ya que las cuatro son manifestaciones del mismo ciclo macroeconómico medidas mensualmente.

La implicancia para la Entrega 2 es concreta: incluir las cinco en un modelo lineal o en una regresión logística produce **inflación de la varianza de los coeficientes**, signos erráticos e interpretaciones sin sentido. La recomendación es **conservar una sola como representante del bloque** — euribor_3m , por ser la de lectura económica más directa— o aplicar reducción de dimensionalidad.

Conviene además registrar un resultado del preprocesamiento: la correlación $\text{pdays} \sim \text{previous}$ de **-0.588** presente en los datos crudos **desaparece (-0.04)** una vez recodificado el centinela 999. Era un artefacto del código, no una relación. Es la demostración más limpia de por qué el orden importa: un análisis de correlaciones ejecutado antes de limpiar habría reportado una relación inexistente.

4.2.2. Covarianza

Tabla 11: Matriz de covarianzas.

Variable	edad	duracion	n_cntcts_c	ds_ltm_cn	n_cntcts_p	var_emple	idx_precs	idx_cnfz	euribr_3m	n_empleds
edad	108.6	-2.2	0.1	-4.1	0.1	0.0	0.0	6.2	0.2	-13.3
duracion	-2.2	67,239.2	-51.6	74.9	2.6	-11.4	0.8	-9.8	-14.8	-836.9
n_cntcts_c	0.1	-51.6	7.7	0.2	-0.1	0.7	0.2	-0.2	0.6	28.8
ds_ltm_cn	-4.1	74.9	0.2	14.6	-0.1	0.5	0.5	-2.1	-0.2	-27.4
n_cntcts_p	0.1	2.6	-0.1	-0.1	0.2	-0.3	-0.1	-0.1	-0.4	-17.9
var_emple	0.0	-11.4	0.7	0.5	-0.3	2.5	0.7	1.4	2.6	102.9
idx_precs	0.0	0.8	0.2	0.5	-0.1	0.7	0.3	0.2	0.7	21.8
idx_cnfz	6.2	-9.8	-0.2	-2.1	-0.1	1.4	0.2	21.4	2.2	33.7
euribr_3m	0.2	-14.8	0.6	-0.2	-0.4	2.6	0.7	2.2	3.0	118.4
n_empleds	-13.3	-836.9	28.8	-27.4	-17.9	102.9	21.8	33.7	118.4	5,220.3

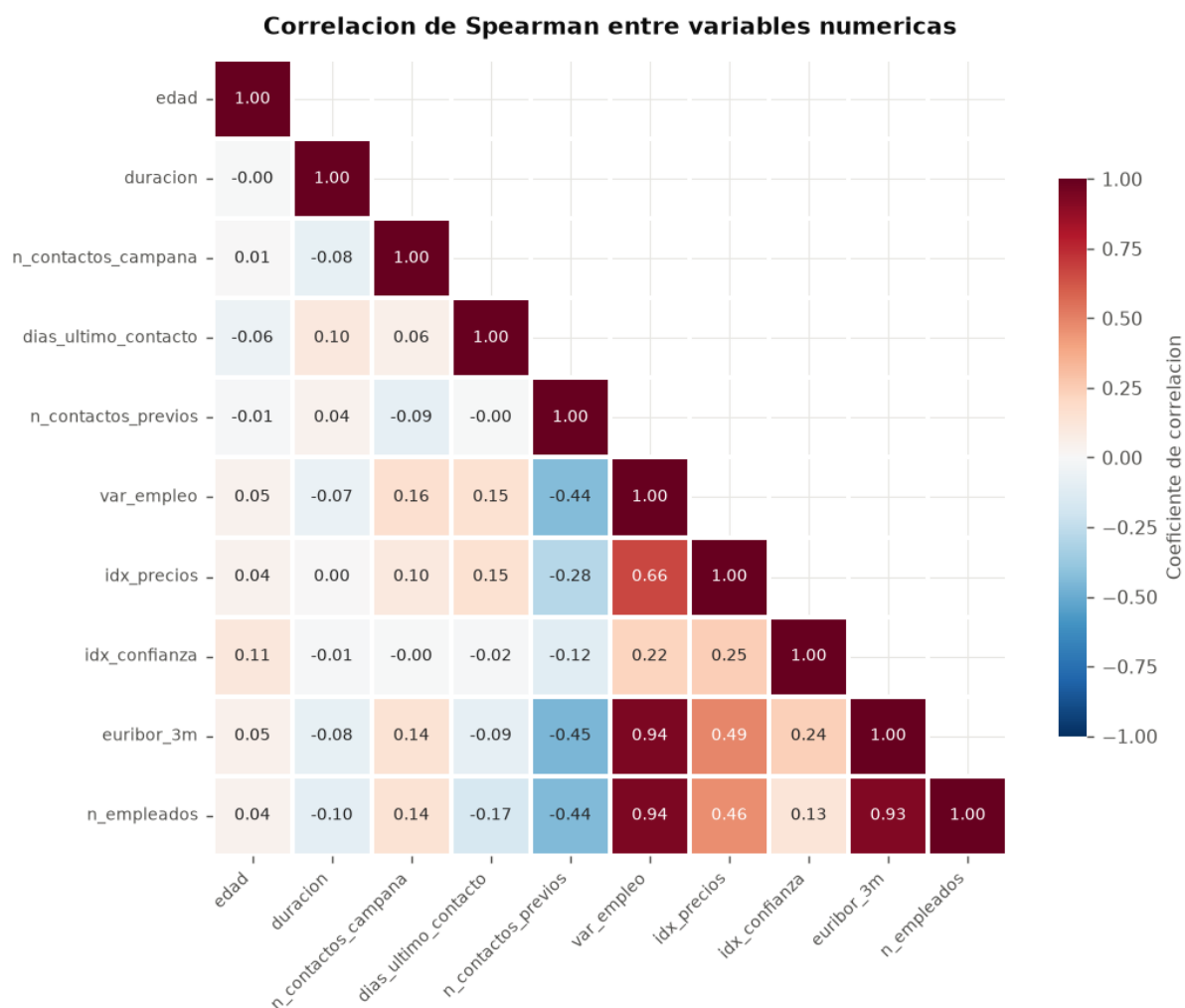


Figura 7: Correlación de Spearman. Al basarse en rangos, es robusta frente a los valores atípicos detectados.

La covarianza comparte el signo con la correlación pero **no es comparable entre pares**, porque conserva las unidades de las variables implicadas. La matriz lo ilustra con claridad: la covarianza entre duracion y n_empleados alcanza magnitudes de millares simplemente porque n_empleados se mide en miles de personas y duracion en segundos, no porque la relación sea intensa —su correlación es apenas -0.04.

Por eso el análisis de intensidad se apoya en la correlación, que es adimensional y acotada en $[-1, 1]$. La covarianza se reporta porque es el insumo del cual se deriva:

$$r_{xy} = \frac{\text{Cov}(x, y)}{s_x s_y} \quad (1)$$

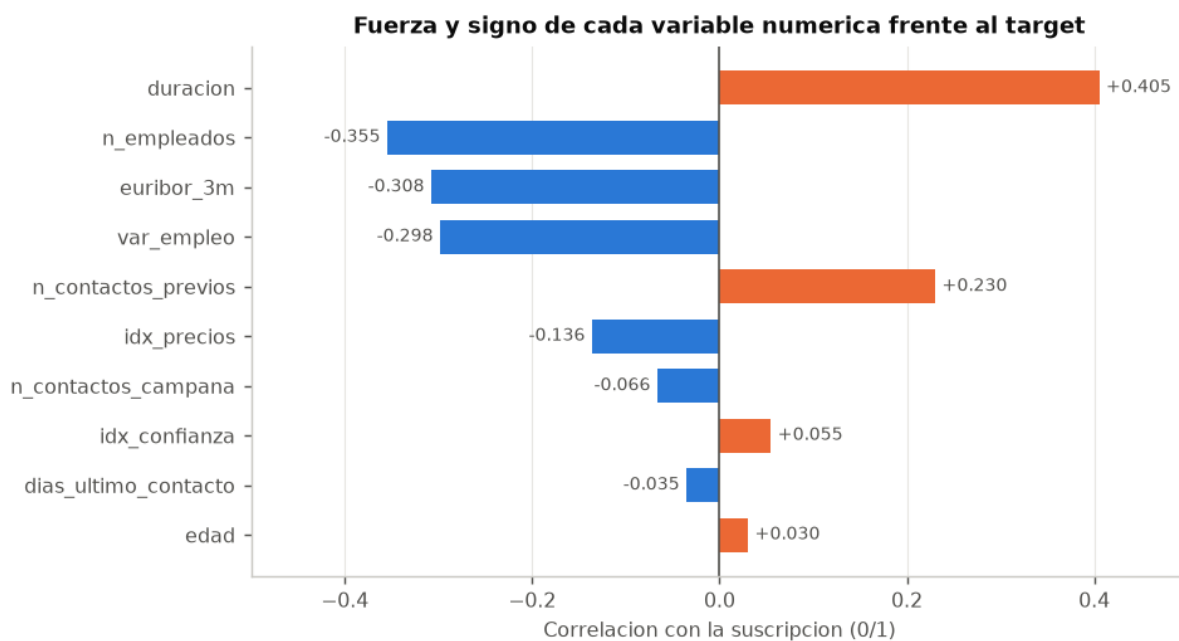


Figura 8: Correlación de cada variable numérica con la suscripción codificada como 0/1.

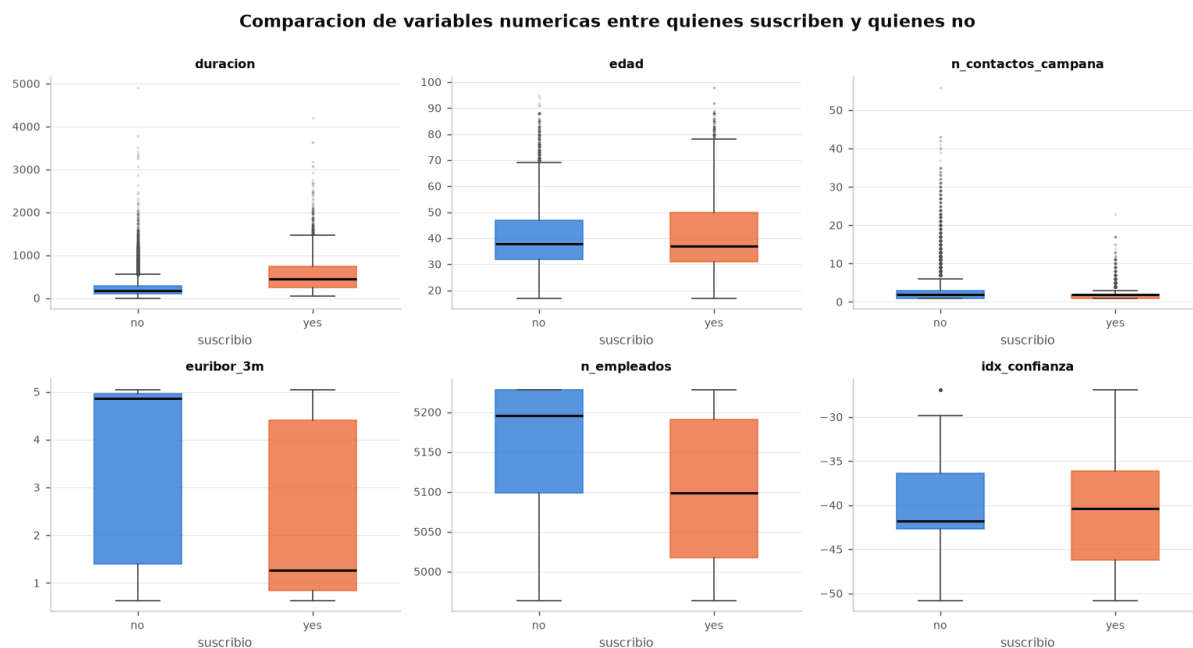


Figura 9: Boxplots comparativos: distribución de cada variable numérica según el resultado de la llamada.

4.2.3. Relación de cada variable numérica con el objetivo

La Figura 8 ordena las variables por su asociación con el objetivo, y la Figura 9 muestra la misma información como distribuciones comparadas.

duracion encabeza la lista con $r = 0.405$, y la Figura 9 lo confirma visualmente: la caja de quienes suscriben está completamente desplazada hacia arriba respecto de quienes no. La sección siguiente explica por qué este resultado, lejos de ser una buena noticia, es el problema metodológico central del dataset.

El segundo bloque lo forman las macroeconómicas, todas con signo negativo ($n_empleados -0.355$, $euribor_3m -0.308$, $var_empleo -0.298$): **la suscripción aumenta cuando la economía se deteriora**. La lectura económica es coherente. En un entorno de tasas altas y empleo en expansión, el cliente encuentra alternativas más rentables que un depósito a plazo; cuando las tasas caen y el empleo se contrae, el depósito gana atractivo relativo como refugio.

edad tiene una correlación lineal prácticamente nula ($r = +0.030$), pero esto **no significa que sea irrelevante**: su efecto es no lineal, como se verá en la Figura 11. Es un buen ejemplo de por qué la correlación de Pearson no debe usarse como único criterio de selección de variables.

4.2.4. duracion: fuga de información

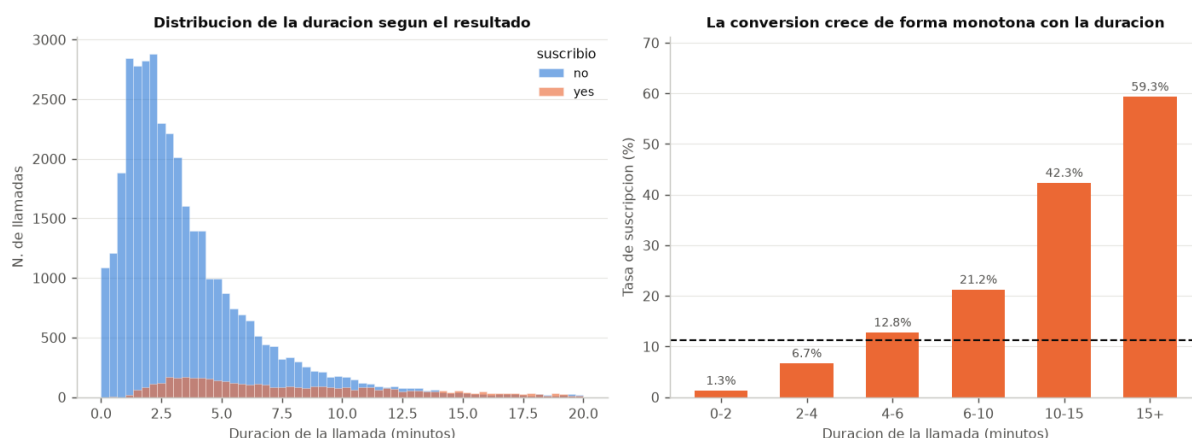


Figura 10: Distribución de la duración de la llamada según el resultado, y tasa de suscripción por tramo de duración.

La Figura 10 documenta la relación más fuerte del conjunto. La tasa de suscripción crece de forma monótona con la duración: **1.3%** en llamadas de menos de 2 minutos, **12.8%** entre 4 y 6 minutos, **42.3%** entre 10 y 15, y **59.3%** por encima de 15 minutos. La mediana de duración es de 2.7 minutos entre quienes rechazan y 7.5 minutos entre quienes aceptan.

Esta variable no puede usarse para predecir. El razonamiento es de causalidad y de disponibilidad temporal:

1. La duración de una llamada **solo se conoce cuando la llamada ha terminado**. En el momento en que el modelo debe decidir a quién llamar, el dato no existe.
2. La relación es inversa a la que aparenta. No es que hablar mucho convenza al cliente; es que el cliente interesado prolonga la conversación. La duración es **consecuencia** del interés, no su causa.
3. Una llamada de duración 0 segundos implica necesariamente $y = \text{"no"}$. La variable contiene el resultado.

Este fenómeno se conoce como **fuga de información** (*data leakage*) y está documentado de forma sistemática por Kaufman et al. (2012). Los propios autores del dataset lo advierten explícitamente (Moro, Cortez y Rita 2014). Un modelo que incluyera *duracion* alcanzaría métricas excelentes en validación y sería **completamente inútil en producción**.

Se conserva en el análisis exploratorio, porque describe correctamente la dinámica de la campaña, y se **excluirá del conjunto de predictores en la Entrega 2**.

4.2.5. Comparación entre grupos: variables cualitativas

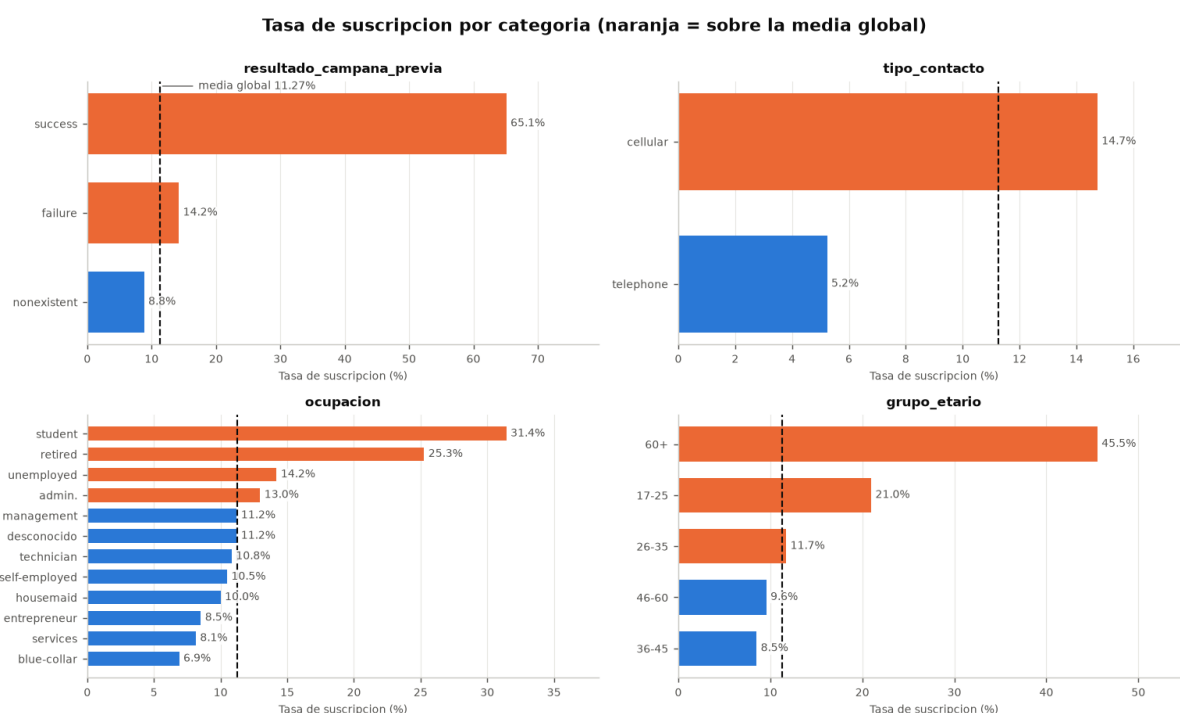


Figura 11: Tasa de suscripción por categoría. La línea discontinua marca la media global de 11.27%.

La Figura 11 concentra los hallazgos de mayor valor comercial:

resultado_campana_previa es el predictor más potente disponible. Los clientes que aceptaron una oferta anterior suscriben al 65.11%, casi seis veces la media global. Es una señal accionable de inmediato: existe un segmento de clientes receptivos, ya identificado, de apenas 1,373 personas.

El canal importa, y mucho. El teléfono móvil convierte al **14.74 %** frente al **5.23 %** de la línea fija: casi el triple. Parte del efecto es probablemente indirecto —quien tiene móvil registrado en 2008–2010 pertenece a un perfil demográfico distinto—, pero la magnitud justifica priorizar el canal móvil.

La edad tiene efecto en forma de U. Los extremos convierten muy por encima de la media: **45.5 %** en mayores de 60 años y **21.0 %** entre 17 y 25, frente a solo **8.5 %** en el tramo de 36 a 45. El patrón es interpretable: jubilados y estudiantes disponen de horizontes de ahorro y perfiles de riesgo compatibles con un depósito a plazo, mientras que la franja de mediana edad soporta hipotecas y gastos familiares. Esta forma de U es exactamente lo que la correlación lineal de +0.030 no podía capturar.

Coherentemente, en ocupacion destacan student (31.4%) y retired (25.3%), mientras blue-collar (6.9%) y services (8.1%) quedan muy por debajo.

4.2.6. Estacionalidad: esfuerzo frente a efectividad

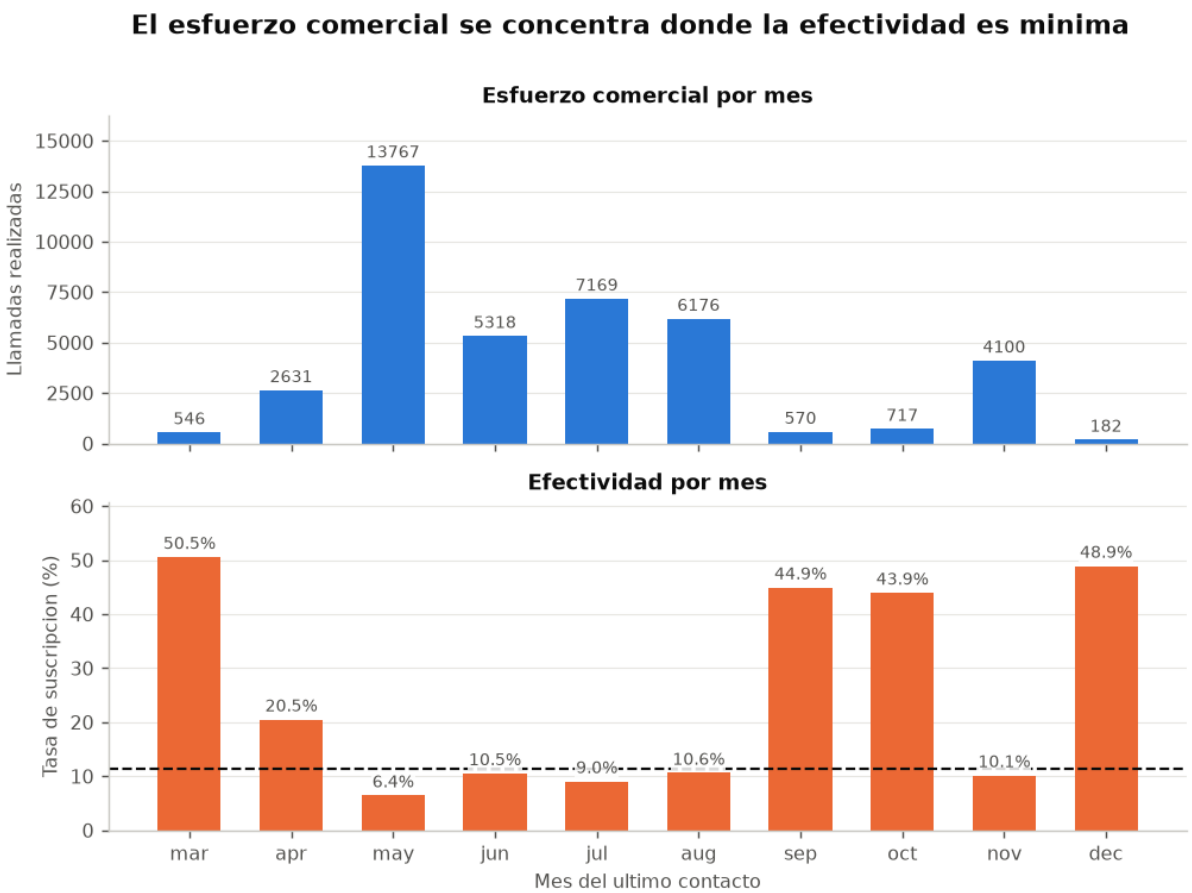


Figura 12: Volumen de llamadas y tasa de suscripción por mes. Los dos paneles comparten el eje horizontal.

Tabla 12: Estacionalidad de la campaña.

Mes	N.º llamadas	Tasa suscripción %
mar	546	50.55
apr	2,631	20.49
may	13,767	6.44
jun	5,318	10.51
jul	7,169	9.04
aug	6,176	10.61
sep	570	44.91
oct	717	43.93
nov	4,100	10.15
dec	182	48.90

La Figura 12 presenta el hallazgo de mayor impacto operativo del análisis. Se emplean **dos paneles apilados y no un gráfico de doble eje**: las unidades (número de llamadas y porcentaje) no son comparables, y superponerlas en una sola escala induciría a leer una relación entre curvas que el gráfico no puede sustentar.

El contraste es claro. **Mayo concentra 13,767 llamadas —un tercio de toda la campaña— y convierte apenas al 6.44 %**, la peor tasa del año. En el extremo opuesto, **marzo convierte al 50.55 %** con solo 546 llamadas; diciembre, septiembre y octubre se sitúan entre el 43 % y el 49 % con volúmenes igualmente reducidos.

Existe una **relación inversa entre el esfuerzo comercial desplegado y su efectividad**. Los meses de mayor inversión son los de peor rendimiento.

Este resultado admite dos lecturas que conviene distinguir con honestidad:

- **Lectura operativa**: la asignación de recursos está mal calibrada y redistribuir volumen hacia los meses de alta conversión elevaría el rendimiento agregado.
- **Lectura cautelosa**: los meses de alta conversión tienen volúmenes muy pequeños, y es posible que se contactara en ellos a segmentos previamente seleccionados. Además, mes está fuertemente confundido con las variables macroeconómicas, que también varían mensualmente.

Los datos disponibles no permiten separar ambos efectos. Lo que sí queda establecido es que **el mes de contacto es una variable con alto poder explicativo** que debe incorporarse al modelo.

4.2.7. Diagramas de dispersión

Se trabajó sobre una **muestra aleatoria de 4,000 registros** y con opacidad reducida: con 41,176 puntos el solapamiento oculta por completo la densidad. El orden de dibujo se aleatorizó para que

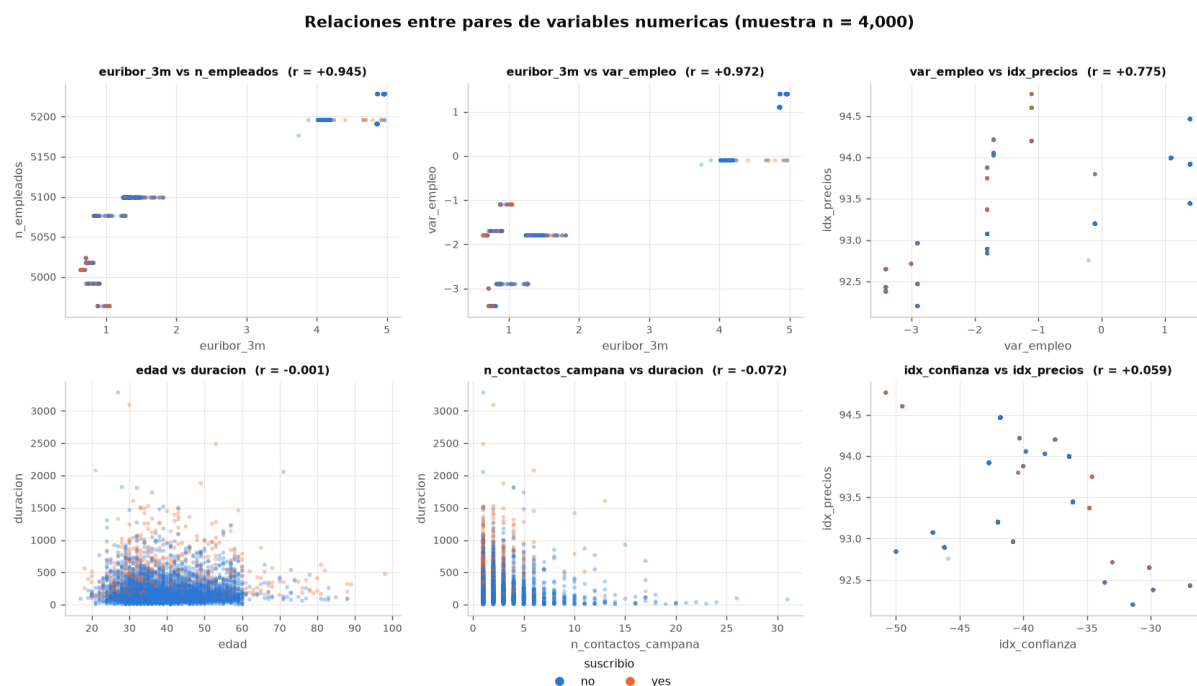


Figura 13: Seis combinaciones de pares de variables numéricas sobre una muestra aleatoria de 4,000 registros.

ninguna de las dos clases tape sistemáticamente a la otra, lo que induciría a percibir una prevalencia falsa.

Interpretación de cada panel:

1. **euribor_3m vs n_empleados ($r = +0.945$)** — Los puntos no forman una nube sino **bloques discretos**. Cada bloque corresponde a un mes del período: ambos indicadores son mensuales y se repiten idénticos en todas las llamadas de ese mes. La alineación casi perfecta confirma la colinealidad.
2. **euribor_3m vs var_empleo ($r = +0.972$)** — El caso extremo. La relación es prácticamente funcional; conocer una determina la otra.
3. **var_empleo vs idx_precios ($r = +0.775$)** — Relación positiva clara pero con más dispersión, lo que sugiere que el índice de precios aporta algo de información propia frente al bloque.
4. **edad vs duracion ($r = -0.001$)** — Ausencia total de relación lineal: una nube sin estructura. Es un resultado útil, porque descarta que la duración de la llamada dependa de la edad del cliente. Se aprecia la mayor densidad de puntos naranjas en la zona superior, reflejo de la asociación entre duración y suscripción.
5. **n_contactos_campana vs duracion ($r = -0.072$)** — Se observa un patrón en forma de L: las llamadas largas se concentran en los primeros contactos, y a partir del quinto contacto la

duración cae drásticamente. La interpretación es directa: **la insistencia agota la paciencia del cliente**. Es la justificación empírica de la variable `contacto_intensivo`.

6. **idx_confianza vs idx_precios ($r = +0.059$)** — Correlación casi nula entre dos indicadores macro, lo que muestra que el bloque no es completamente redundante: la confianza del consumidor aporta una dimensión independiente.

4.3. Conclusiones generales del EDA

4.3.1. Sobre la calidad de los datos

El dataset **no estaba limpio pese a declararse como tal**. Las funciones estándar de detección de nulos reportaban cero faltantes, cuando en realidad existían valores ausentes en seis variables, un código centinela que distorsionaba la media de pdays en tres órdenes de magnitud y una variable con tres observaciones en una de sus categorías.

La verificación manual del contenido de cada columna fue lo que permitió detectarlo. Es la conclusión metodológica principal de esta entrega.

4.3.2. Sobre la colinealidad

Existe un **bloque macroeconómico fuertemente colineal** (euribor_3m, var_empleo, n_empleados, idx_precios) con correlaciones de hasta **0.972**. Las cuatro miden el mismo ciclo. Deberá seleccionarse una representante o aplicarse reducción de dimensionalidad antes de ajustar cualquier modelo lineal.

Fuera de ese bloque, el resto de variables presenta correlaciones bajas entre sí, lo que es favorable: aportan información complementaria y no redundante.

4.3.3. Variables clave identificadas

Ordenadas por valor esperado para el modelado:

1. **resultado_campana_previa / exito_previo** — La señal más fuerte legítimamente utilizable (65.1 % de conversión).
2. **mes y trimestre** — Rango de variación del 6.4 % al 50.6 %.
3. **Bloque macroeconómico (una representante)** — Correlaciones de -0.30 a -0.36 con el objetivo.
4. **tipo_contacto** — 14.7 % frente a 5.2 %.
5. **grupo_etario** — Efecto en U que la edad continua no captura.
6. **ocupacion** — Rango del 6.9 % al 31.4 %.

4.3.4. Variables a excluir o tratar con cuidado

- **duracion** — Excluir del conjunto de predictores por fuga de información, pese a ser la variable con mayor correlación con el objetivo.
- **default original** — Excluida ya en el preprocesamiento por degenerada.
- **prestamo_vivienda y prestamo_personal** — Sus tasas de conversión (11.6 % y 10.9 %) son indistinguibles de la media global: no discriminan.
- **Cuatro de las cinco macroeconómicas** — Por colinealidad.

4.3.5. Implicancias para la Entrega 2

1. El desbalance del **11.27%** obliga a evaluar con precisión, exhaustividad, F1 y AUC, nunca con exactitud.
2. Deberán considerarse técnicas de remuestreo o ponderación de clases.
3. La partición entrenamiento/prueba debe ser **estratificada** por la variable objetivo.
4. El efecto no lineal de la edad favorece a los modelos basados en árboles frente a la regresión logística, o exige incorporar `grupo_etario` como factor.

5. ANEXO: CÓDIGO DEL PIPELINE

El análisis completo es reproducible mediante el script `src/eda_bank.py`, que genera las 13 figuras y las 14 tablas empleadas en este informe. El script `src/eda_bank.R` implementa el mismo pipeline en R con `tidyverse` y `ggplot2`.

```
# Ejecución del pipeline completo (desde la raíz del proyecto):
#
#   python src/eda_bank.py      # genera figs/ y tablas/
#   Rscript src/eda_bank.R      # versión equivalente en R
#
# El informe lee los resultados desde tablas/ y figs/, de modo que las cifras
# citadas en el texto no pueden desincronizarse del análisis.
```

A modo de verificación, se reproduce la carga y el diagnóstico inicial en R sobre los datos ya preprocesados:

```
limpio <- readr::read_csv("datos/bank_limpio.csv", show_col_types = FALSE)

tibble(
  Métrica = c("Filas", "Columnas", "Suscripciones", "Tasa de suscripción (%)" ),
  Valor = c(
    format(nrow(limpio), big.mark = ","),
    as.character(ncol(limpio)),
    format(sum(limpio$suscribio_bin), big.mark = ","),
    sprintf("%.2f", mean(limpio$suscribio_bin) * 100)
  )
) |>
  tabla_bonita("Verificación del dataset preprocesado.")
```

Tabla 13: Verificación del dataset preprocesado.

Métrica	Valor
Filas	41,176
Columnas	28
Suscripciones	4,639
Tasa de suscripción (%)	11.27

Referencias

- Kaufman, S, S Rosset, C Perlich y O Stitelman (2012). Leakage in Data Mining: Formulation, Detection, and Avoidance. *ACM Transactions on Knowledge Discovery from Data* **6**(4), 1-21.
- Moro, S, P Cortez y P Rita (2014). A Data-Driven Approach to Predict the Success of Bank Telemarketing. *Decision Support Systems* **62**, 22-31.
- Moro, S, R Laureano y P Cortez (2011). Using Data Mining for Bank Direct Marketing: An Application of the CRISP-DM Methodology. En: *Proceedings of the European Simulation and Modelling Conference (ESM'2011)*. Guimarães, Portugal: EUROSIS, pp.117-121.
- Moro, S, P Rita y P Cortez (2014). *Bank Marketing [Dataset]*. <https://archive.ics.uci.edu/dataset/222/bank+marketing>.
- Tukey, JW (1977). *Exploratory Data Analysis*. Reading, MA: Addison-Wesley.