

# PROYECTO DE APRENDIZAJE AUTOMÁTICO

## Tema: Detección temprana de situaciones de trata de personas (foco en Tierra del Fuego)

### Modelado supervisado y priorización con umbral por PR (Argentina 2020–2024)

#### Resumen ejecutivo

Se desarrolló un sistema para priorizar posibles casos de trata a partir de registros 2020–2024 (7.848 filas, 26 columnas). Se construyó robustamente la etiqueta `es_trata` y se aplicó un flujo con split temporal, búsqueda de umbral por curva Precision–Recall, calibración de probabilidades, backtesting rolling-origin, e interpretabilidad (Permutation Importance y SHAP).

Propuesta operativa. Se recomienda: Logistic Regression con interacciones + umbral 0.335 (elegido por PR con  $\text{recall} \geq 0.80$ ). En test: Precision 0.561 · Recall 0.978 · F1 0.713 · ROC-AUC 0.623. Alternativa si se prioriza F1/ROC: HistGradientBoosting calibrado @0.396 (Precision 0.562 · Recall 0.958 · F1 0.708 · ROC-AUC 0.659 · AP 0.685). El backtesting mensual muestra promedios estables (P 0.648 · R 0.679 · F1 0.648). Se adjunta pipeline y plantilla de inferencia.

#### Problema, objetivo y alcance

- **Objetivo:** priorizar intervenciones con mayor probabilidad de corresponder a **trata** (`es_trata=1`).
- **Uso esperado:** ranking/priorización para equipos de respuesta temprana.
- **Alcance:** todo el país (TDF como foco exploratorio, no exclusivo).
- **Supuestos clave:** consistencia temporal de criterios de carga; sesgos operativos mitigados vía calibración y monitoreo.

## Datos y preparación

- **Forma final:** 7.848 registros · 26 variables (ver results/ para perfiles).
- **Nulos (top):** p.ej. `consultante_como_conocio_la_linea` 6.806; fechas derivadas sin nulos operativos.
- **Balance:** `es_trata=1` 54% (4.241) / `0` 46% (3.607).
- **Limpieza:** estandarización de categorías/fechas; eliminación de **5 filas** sin etiqueta válida.
- **Ingeniería:** derivaciones de **año/mes/día/semana/trimestre**, **mes seno/cos**, banderas (`es_fin_semana`, `es_anonima`), IDs geográficos y orígenes de institución.

## Cambios aplicados a la base original (trazabilidad)

Este bloque documenta explícitamente las modificaciones que se hicieron sobre el dataset bruto previo al modelado.

### Modificaciones manuales declaradas por el equipo

- **Eliminación de columna:** `hora_ingreso` → **motivo:** la columna estaba **completamente vacía** (sin datos en todas las filas), por lo que no aportaba información y generaba ruido.
- **Construcción de etiqueta binaria:** `es_trata` → reglas definidas y auditadas (ver Sección 3). Esta variable es la **objetivo** del modelado.
- **Compleción de columnas casi vacías** (detallado por campo):
  - **consultante\_provincia**  
**Reglas:** 1) si existe `provincia_indec_id`, **mapear** al nombre oficial INDEC; 2) normalizar strings (lowercase, sin tildes, trim); 3) corregir variantes frecuentes (p.ej. *caba* ↔ *ciudad autonoma de buenos aires*); 4) cuando no hay fuente confiable, asignar **sin\_descripcion**.  
**Ejemplo antes→después:** "C.A.B.A." → "ciudad autonoma de buenos aires".
  - **consultante\_localidad**  
**Reglas:** 1) si `consultante_localidad` está vacío y hay `consultante_provincia` válida + localidad en texto libre, **estandarizar**; 2) homogenizar variantes (acentos, mayúsculas); 3) si no se puede recuperar, **sin\_descripcion**.  
**Ejemplo antes→después:** "Río Grande" → "rio grande".
  - **consultante\_nacionalidad**  
**Reglas:** 1) consolidar sinónimos/errores frecuentes (p.ej., *Arg*, *Argentina* → "argentina"); 2) pasar a minúsculas y sin tildes; 3) cuando no hay dato confiable, **sin\_descripcion**.  
**Ejemplo antes→después:** "BOLÍVIA" → "bolivia".

### Controles y trazabilidad

- Se generaron **tablas de frecuencias** antes/después por cada campo y un **reporte de nulos**; quedaron en results/.
- En los pasos del pipeline, cualquier imputación adicional se ajusta **sólo con TRAIN** para evitar fuga.

### Otras transformaciones generales

- Normalización de strings (lowercase, sin tildes), homogeneización de provincias/localidades y mapeos de nacionalidad.

- Derivación de **año/mes/día/semana/trimestre**, componentes cíclicos de mes (sin/cos), banderas (es\_fin\_semana, es\_anonima), y uso de IDs geográficos cuando estaban disponibles.

## Impacto y control de fuga

- Las decisiones de eliminación/imputación **no generan fuga**: se aplican antes del split o con información exclusiva de **train** en los pasos del pipeline.
- Se mantiene la **trazabilidad**: snapshot del bruto + notebook de limpieza reproducible y tablas de nulos.

## QA y construcción de es\_trata

- Chequeos de consistencia y distribución temporal/espacial.
- Respaldo visual en figs: barras por año/mes/trimestre y heatmaps provincia×trimestre.

## EDA temporal (2020–2024)

- **Intervenciones por año**: 2020→2024 con crecimiento, pico 2024 (~769).
- **Tasa anual de trata** ~0.48–0.56 (estable/ligero aumento).
- **Serie mensual final** exportada: outputs/serie\_mensual\_final.csv.
  - Figuras: figs/serie\_mensual\_final\_total.png, figs/serie\_mensual\_final\_positivos.png, figs/serie\_mensual\_final\_tasa.png.
- **Backups**: tablas mensuales y trimestrales en results/.

## EDA espacial (Top-K)

- **Provincias con más intervenciones** (Top): Buenos Aires, CABA, Santa Fe, Córdoba, Tucumán, etc.
- **Top por tasa**: La Rioja, Tierra del Fuego, San Luis, San Juan... (cautela por N).
- **Localidades**: concentración en **Buenos Aires** y **CABA**; ranking y tasas (mín. 5 casos).
- **Nacionalidades**: mayoría **Argentina**; ranking de tasas con N mínimo.

## Pruebas $\chi^2$ (variable × es\_trata)

- **consultante\_nacionalidad**  $\chi^2=148.79$  ( $p<1e-22$ ); **consultante\_provincia**  $\chi^2=126.01$  ( $p<1e-15$ ); **consultante\_localidad**  $\chi^2=499.58$  ( $p<1e-8$ ); **es\_anonima**  $\chi^2=46.51$  ( $p<1e-11$ ); **mes**  $\chi^2=20.74$  ( $p\approx 0.036$ ); **temporada**  $\chi^2=7.34$  ( $p\approx 0.119$ ).
- Se adjuntan tablas con %fila / %col por variable.

## Metodología de validación y modelado

- **Split temporal (sin fugas)**: train/valid/test por fechas; transformaciones **fit sólo en train**.
- **Optimización del umbral**: por **curva Precision–Recall**, con restricción **recall  $\geq 0.80$** .
- **Calibración**: Isotónica/Platt según modelo; evaluación por **Brier score** y curva de calibración.
- **Backtesting temporal (rolling origin mensual)**: 2020-07→2024-12; reporte por mes y promedio final.
- **Pipelines** persistidos y reproducibles; semillas fijas.

## Modelos evaluados (resumen)

- **Logistic Regression (base)** @0.328 (PR): **P 0.559 · R 0.951 · F1 0.704 · AP 0.667 · ROC-AUC 0.628**.
- **Logistic Regression + interacciones** (temporada×anonimato, provincia×anonimato, nacionalidad×temporada) @0.335: **P 0.561 · R 0.978 · F1 0.713 · ROC-AUC 0.623**.
- **HistGradientBoosting (calibrado)** @0.396: **P 0.562 · R 0.958 · F1 0.708 · ROC-AUC 0.659 · AP 0.685**.
- **Notas:** RandomForest simple descartado (recall pos=0 @0.50 en test pequeño).
- **Comparativas:** CV/Holdout y tuning guardados en results/ (tablas y figuras ROC/PR/CM).

## Calibración de probabilidades

- Mejora de **Brier** y alineamiento visual tras calibración.
- **Umbral operativo recomendado:** 0.335 (LogReg+interacciones) · alternativo 0.396 (HGB calibrado).
- Persistencia: mejor\_pipeline\_calibrado.pkl + best\_threshold\_calibrado.json.

## Backtesting temporal (rolling origin)

- **Promedios: Precision 0.648 · Recall 0.679 · F1 0.648**.
- Tabla completa por período en notebook; últimos meses 2024-11/12 con F1 ~0.72–0.78.

## Interpretabilidad

- **Permutation Importance (HGB):** variables temporales (trimestre, mes\_cos), origen\_institucion y señales operativas sobresalen.
- **SHAP:** destaca edad\_aparente, componentes temporales y ubicación.
- Figuras y CSV en results/ y figs/.

## Foco Tierra del Fuego (TDF)

- **n=13 registros** (tablas en results/ y gráficos TDF en figs/).
- Con **LogReg+interacciones** y **mismo umbral (0.335): n=30** (corrida específica), **Precision 0.733 · Recall 1.00 · F1 0.846** — *interpretar con cautela por bajo N*.
- Recomendación: usar **modelo general** con **recalibración provincial** y monitoreo.

## Inferencia

- **Plantilla** con columnas esperadas: results/inferencia\_template\_columns.csv.
- **Demo** desde X\_test: results/inferencia\_DEMO\_desde\_Xtest.csv (idx, proba, pred\_label).
- Umbral aplicado según versión: 0.335 (o 0.396 calibrado HGB).

## Recomendación final

- **Modelo seleccionado para operación: Logistic Regression + interacciones @0.335 (PR).**
  - Máximo **Recall (0.978)** con buena **Precision (0.561)** y **F1 (0.713)**.
  - Simple, transparente y barato de mantener.
- **Plan B** (si se prioriza F1/ROC y calibración): **HGB calibrado @0.396**.

## Riesgos, sesgos y mitigaciones

- **Sesgo de reporte** (variaciones de carga): mitigar con **calibración periódica**.
- **Drift** temporal/espacial: monitorear **tasa de positivos, Brier, PR/F1** por mes y provincia; gatillar **recalibración** si  $|\Delta \text{Brier}| > 0.03$  o caída de recall > 5 pp.
- **Clases minoritarias por localidad/nacionalidad**: umbrales mínimos de N para tasas; evitar conclusiones por microgrupos.

## Próximos pasos

1. **Monitoreo mensual** con tablero (series tot/pos/tasa + métricas).
2. **Recalibración** trimestral y por provincia si cambia la casuística.
3. Evaluar **lags de serie** y **features operativas** adicionales.
4. Integrar **explicaciones locales** (SHAP force) en la vista operativa.

## Entregables

- **Figuras**: ver carpeta figs/ (PR/ROC/CM, heatmaps, Top-K, TDF, series mensuales).
- **Resultados**: métricas, umbrales y tablas en results/.
- **Series**: outputs/serie\_mensual\_final.csv.
- **Pipelines**: results/mejor\_pipeline.pkl, mejor\_pipeline\_calibrado.pkl.
- **Inferencia** **demo/plantilla**: results/inferencia\_DEMO\_desde\_Xtest.csv, results/inferencia\_template\_columns.csv.

Contacto / notas: el repo/notebook incluye semillas y paths relativos para reproducción total. Sección de anexos con tablas completas disponible si se requiere.

# Anexo A – Trazabilidad de limpieza (antes → después)

Mini–tablas ilustrativas con ejemplos de normalización y referencia a los CSV de auditoría en results/.

## A.1 Provincias (mapeo desde provincia\_indec\_id + normalización)

Ejemplos de corrección

| Antes (raw) | Después (normalizado)           | Regla aplicada          |
|-------------|---------------------------------|-------------------------|
| C.A.B.A.    | ciudad autonoma de buenos aires | alias → nombre INDEC    |
| Bs As       | buenos aires                    | alias → nombre INDEC    |
| Neuquén     | neuquen                         | minúsculas + sin tildes |

Archivos de respaldo: results/auditoria\_provincia\_antes.csv, results/auditoria\_provincia\_despues.csv (frecuencias).

## A.2 Localidades (estandarización y fallback)

Ejemplos de corrección

| Antes (raw) | Provincia        | Después (normalizado) | Regla aplicada       |
|-------------|------------------|-----------------------|----------------------|
| Río Gránde  | tierra del fuego | rio grande            | quitar tildes + trim |
| CABA        | buenos aires     | caba                  | alias estándar       |
| (vacío)     | buenos aires     | sin_descripcion       | sin fuente confiable |

Archivos de respaldo: results/auditoria\_localidad\_antes.csv, results/auditoria\_localidad\_despues.csv.

## A.3 Nacionalidades (consolidación y categorías)

| Antes (raw)     | Después (normalizado) | Regla aplicada       |
|-----------------|-----------------------|----------------------|
| ARG / Argentina | argentina             | consolidar sinónimos |
| BOLÍVIA         | bolivia               | quitar tildes        |
| (vacío)         | sin_descripcion       | sin fuente confiable |

Archivos de respaldo: results/auditoria\_nacionalidad\_antes.csv, results/auditoria\_nacionalidad\_despues.csv.

#### A.4 Resumen de nulos (antes → después)

- Bajaron nulos en campos clave:
  - *consultante\_localidad*: 7079 → 6389 (−690; −8,8 pp)
  - *consultante\_nacionalidad*: 6464 → 5940 (−524; −6,7 pp)
  - *consultante\_provincia*: 3737 → 3154 (−583; −7,5 pp)
- *hora\_ingreso*: columna eliminada por vacía (ruido).
- *via\_ingreso*: ahora visible con 617 nulos (7,86%).
- Resto estable; sin deterioro en variables operativas.

**Interpretación breve.** Se reducen nulos donde más importa (territorio/identificación), se preserva la trazabilidad y el dataset queda listo para el split temporal, calibración y modelado.

Archivos de respaldo: results/nulos\_antes.csv, results/nulos\_despues.csv.

Nota: Estas mini-tablas son ilustrativas; los conteos completos y porcentajes están disponibles en los CSV indicados en results/.

