



ANALITICA DE DATOS Y HERRAMIENTAS DE INTELIGENCIA ARTIFICIAL II

ACTIVIDAD 4.1 Y 4.2

# REGRESIÓN LOGÍSTICA

D A T A F O R G E

15 OCT, 2025



# TEAM MEMBERS (DATAFORGE)



**JESÚS EDUARDO VALLE  
VILLEGAS**  
FINANZAS  
A01770616



**DIEGO ANTONIO OROPEZA  
LINARTE**  
BGB  
A01733018



**MANUEL EDUARDO  
COVARRUBIAS RODRÍGUEZ**  
ITC  
A01737781



**ITHANDEHUI JOSELYN  
ESPINOZA**  
ITC  
A01734547



**MAURICIO GRAU GUTIERREZ  
RUBIO**  
LEM  
A01734914



# OBJETIVO



Realizar un análisis comparativo mediante regresión logística sobre los conjuntos de datos del Datathón y Forvia, identificando y modelando correlaciones relevantes, aplicando técnicas de balanceo y evaluación con métricas de desempeño (precisión, exactitud, sensibilidad y F1), para sintetizar los resultados en un informe y presentación final.



# Actividad 4.1

DATATHON



# METODOLOGÍA

## 1. Limpieza y normalización

Se eliminaron valores nulos, duplicados y atípicos. Las variables numéricas (Abundance\_nbcell, TotalAbundance\_SamplingOperation, Abundance\_pm) se normalizaron para mantener la misma escala.

## 2. Análisis Descriptivo de Variables Numéricas y Categóricas

El análisis descriptivo fue esencial porque el objetivo final era crear 5 variables dicotómicas para modelos de regresión logística:

- 1.- Determinación de Umbrales Estadísticamente Justificados.
- 2.- Identificación de Distribuciones y Patrones.

## 3. Creación de variables dicotómicas

Se transformaron variables continuas en categorías binarias (0 o 1) usando umbrales como el percentil 75 o el año  $\geq 2019$ , para representar condiciones de alta abundancia, periodo reciente o especie dominante.

# METODOLOGÍA

## 4. Unión de los DF y entrenamiento de modelos

1.- Primero unimos nuestras 5 variables dicotómicas al dt sin valores nulos para su posterior tratamiento.

2.- Se construyeron cinco regresiones logísticas, una por cada variable dicotómica, para estimar la probabilidad de que una muestra pertenezca a la clase positiva.

## 5. Balanceo con Smote

En los casos con fuerte desbalance (especialmente Recent\_Period), se aplicó la técnica de oversampling sintético (SMOTE) para equilibrar las clases y evitar sesgos del modelo.

## 6. Evaluación del desempeño

Se midieron las métricas de precisión, exactitud, sensibilidad y F1-score, y se interpretaron las matrices de confusión para identificar aciertos, errores y nivel de clasificación de cada modelo.



# SELECCIÓN DE LAS 5 VARIABLES DICOTÓMICAS

01

**Alta abundancia  
de células**

| Elemento               | Descripción  |
|------------------------|--|
| Variable Origen        | Abundance_nbcell   |
| Definición             | Alta abundancia celular (Abundance_nbcell >= 8)                            |
| Aplicación estadística | Se utiliza el percentil 75 para distinguir comunidades densas de diatomeas |
| Justificación          | El 25% superior representa comunidades "prósperas"                         |
| Umbral aplicado        | 8 células  |



02

**Alta abundancia  
total de la  
operación de  
muestreo**

| Elemento               | Descripción   |
|------------------------|---|
| Variable Origen        | TotalAbundance_SamplingOperation  |
| Definición             | Alta abundancia total por operación de muestreo<br>(TotalAbundance_SamplingOperation >= 408)  |
| Aplicación estadística | Se utiliza el percentil 75 para identificar operaciones de muestreo con alta diversidad de especies   |
| Justificación          | <ul style="list-style-type: none"><li>• Valores bajos (&lt;408): Muestreos estándar o sitios con diversidad limitada</li><li>• Valores altos (≥408):vMuestreos exhaustivos o sitios muy diversos.</li></ul> |
| Umbral aplicado        | 408 unidades  |



03

Alta abundancia  
por metro

| Elemento               | Descripción  |
|------------------------|--|
| Variable Origen        | Abundance_pm   |
| Definición             | Alta abundancia por metro (Abundance_pm >= 19.90)  |
| Aplicación estadística | Se utiliza el percentil 75 para identificar muestras con alta densidad por metro   |
| Justificación          | <b>Características de muestras con una alta densidad:</b><br>Microhábitats altamente productivos y Zonas de acumulación por corrientes |
| Umbral aplicado        | 19.90 unidades por metro   |



04

**Período reciente  
de muestreo**

| Elemento               | Descripción  |
|------------------------|--|
| Variable Origen        | Date_SamplingOperation   |
| Definición             | Periodo reciente de muestreo (Año >= 2019)   |
| Aplicación estadística | Los últimos 5 años representan las condiciones ambientales actuales y permiten analizar cambios recientes en la comunidad de diatomeas |
| Justificación          | Ciclos reproductivos de diatomeas: 7-10 años es muy antiguo (Hugo Beraldi-Campesi, 2015)   |
| Umbral aplicado        | Año 2019 en adelante   |

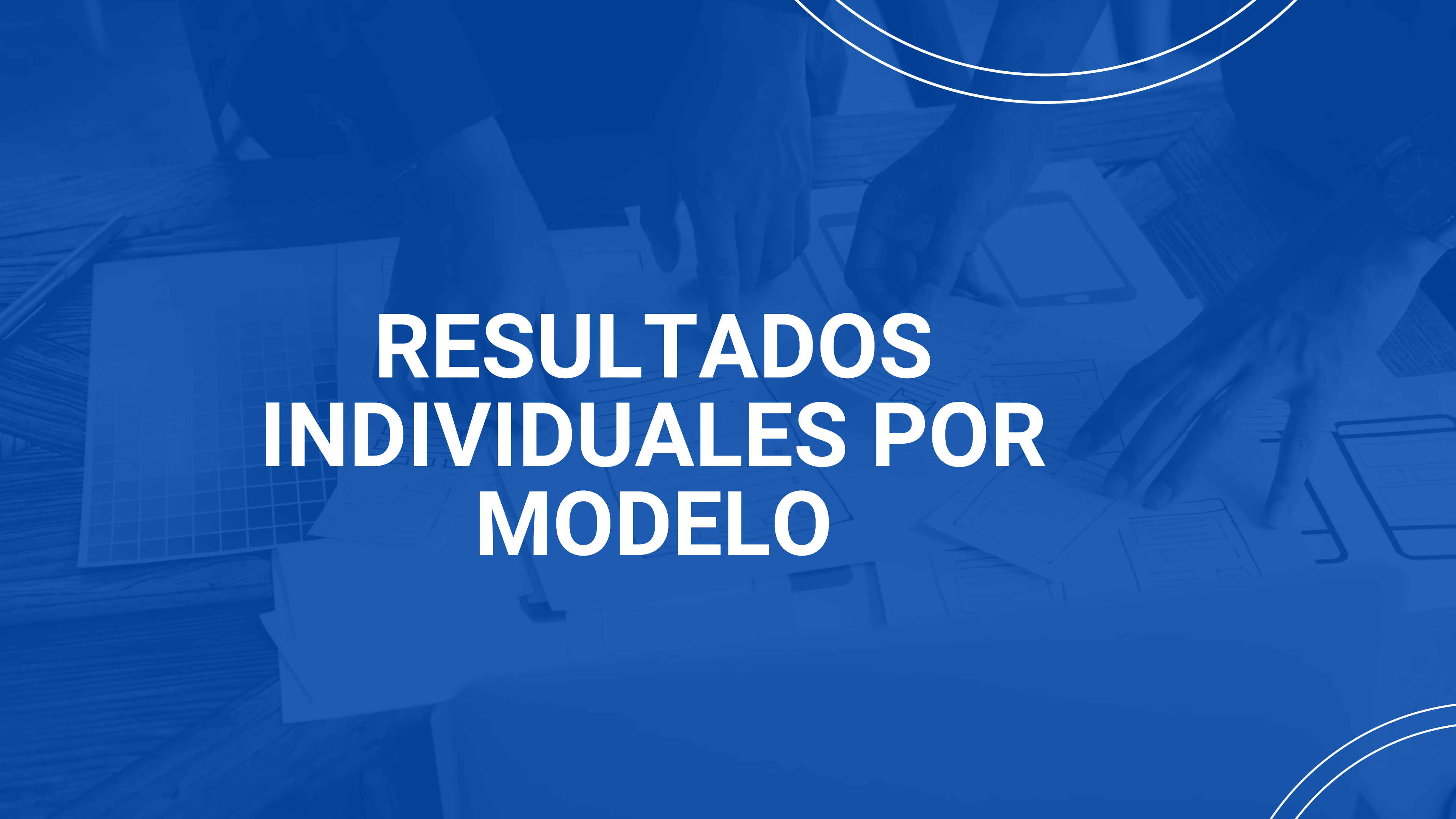


05

**Especie  
dominante**

| Elemento               | Descripción  |
|------------------------|--|
| Variable Origen        | TaxonName  |
| Definición             | Especie dominante en el conjunto de datos (Top 10 especies más frecuentes)   |
| Aplicación estadística | Suma simple de las TOP 10 especies   |
| Justificación          | Las 10 especies más frecuentes representan aproximadamente el 20% de todos los registros, permitiendo identificar las especies ecológicamente más relevantes |
| Umbral aplicado        | Pertenecer al grupo de las 10 especies más frecuentes del dataset  |

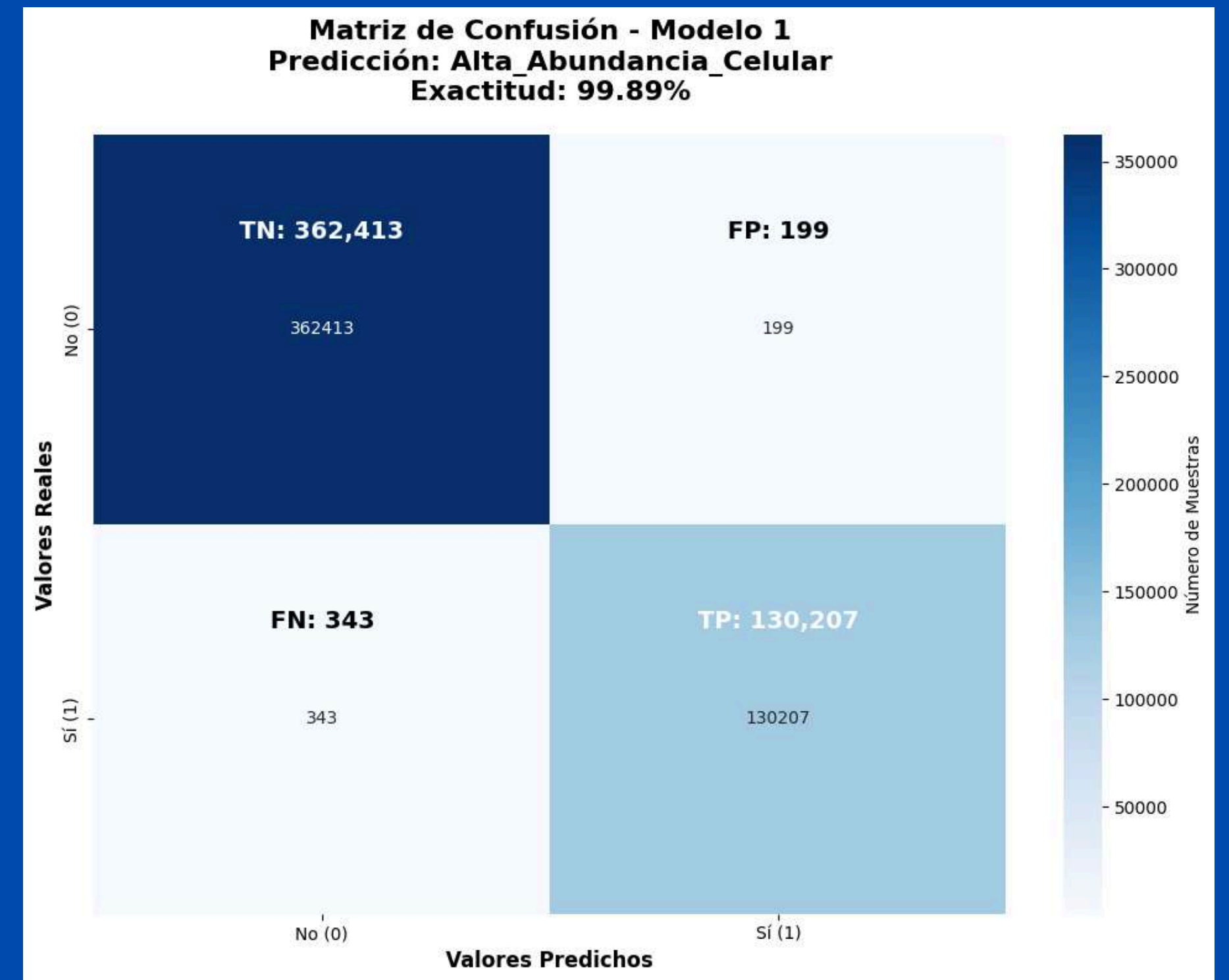




# RESULTADOS INDIVIDUALES POR MODELO

## MODELO 1: Alta\_Abundancia\_Celular

- **Exactitud: 99.89%** - Excelente precisión general
- **Precisión: 99.85%** - Cuando predice "alta abundancia", casi siempre acierta
- **Sensibilidad: 99.74%** - Detecta el 99.74% de los casos reales de alta abundancia
- **F1-Score: 99.79%** - Balance perfecto entre precisión y sensibilidad



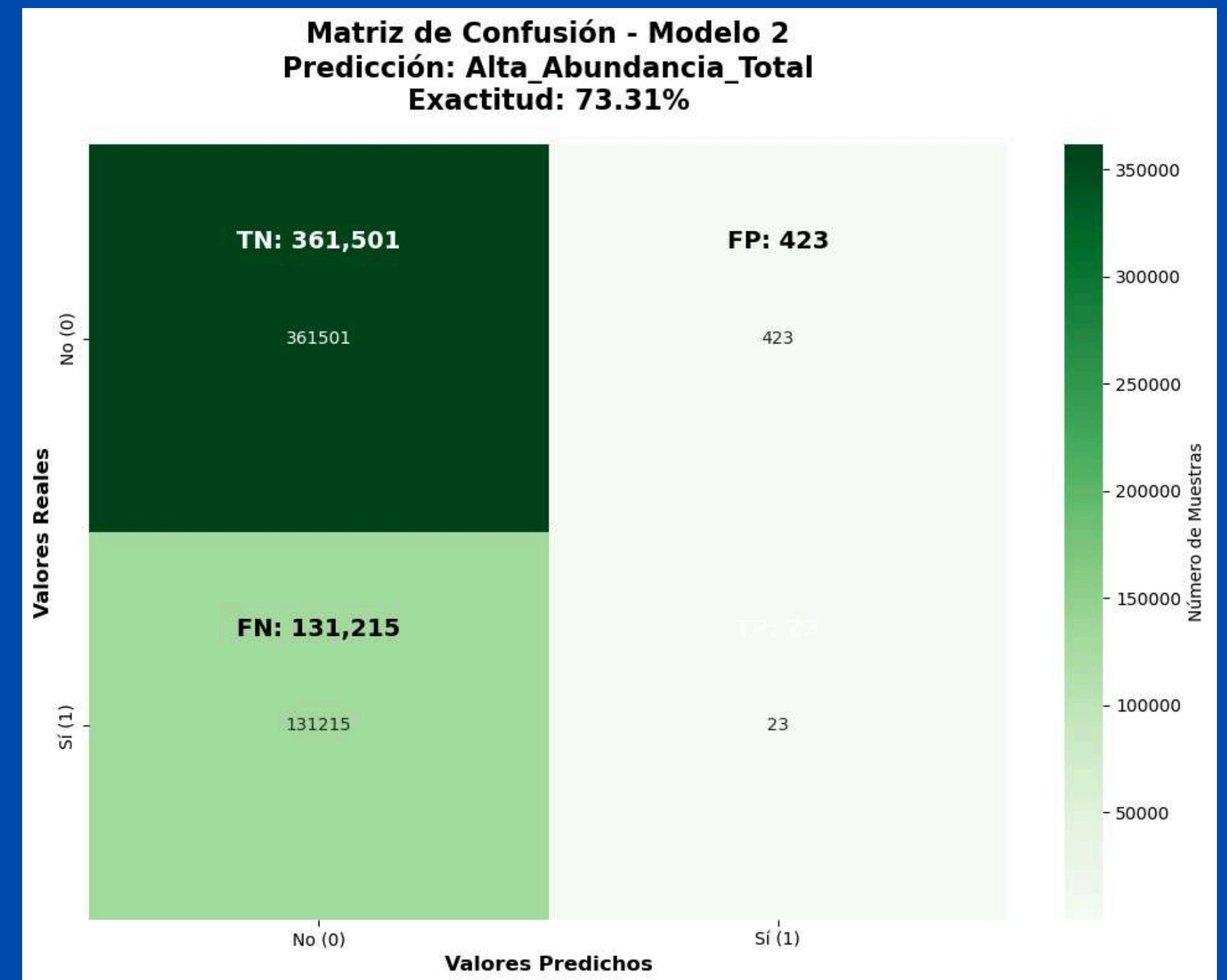
**V\_Indep** = ['TotalAbundance\_SamplingOperation', 'Abundance\_pm']

**V\_Dep** = ['Alta\_Abundancia\_Celular']



## MODELO 2: Alta\_Abundancia\_Total

- **Exactitud: 73.31%** - Moderada, pero engañosa
- **Precisión: 5.16%** - Cuando predice "alta abundancia", casi siempre se equivoca
- **Sensibilidad: 0.02%** - CRÍTICO: Solo detecta el 0.02% de casos reales
- **F1-Score: 0.03%** - Muy bajo balance

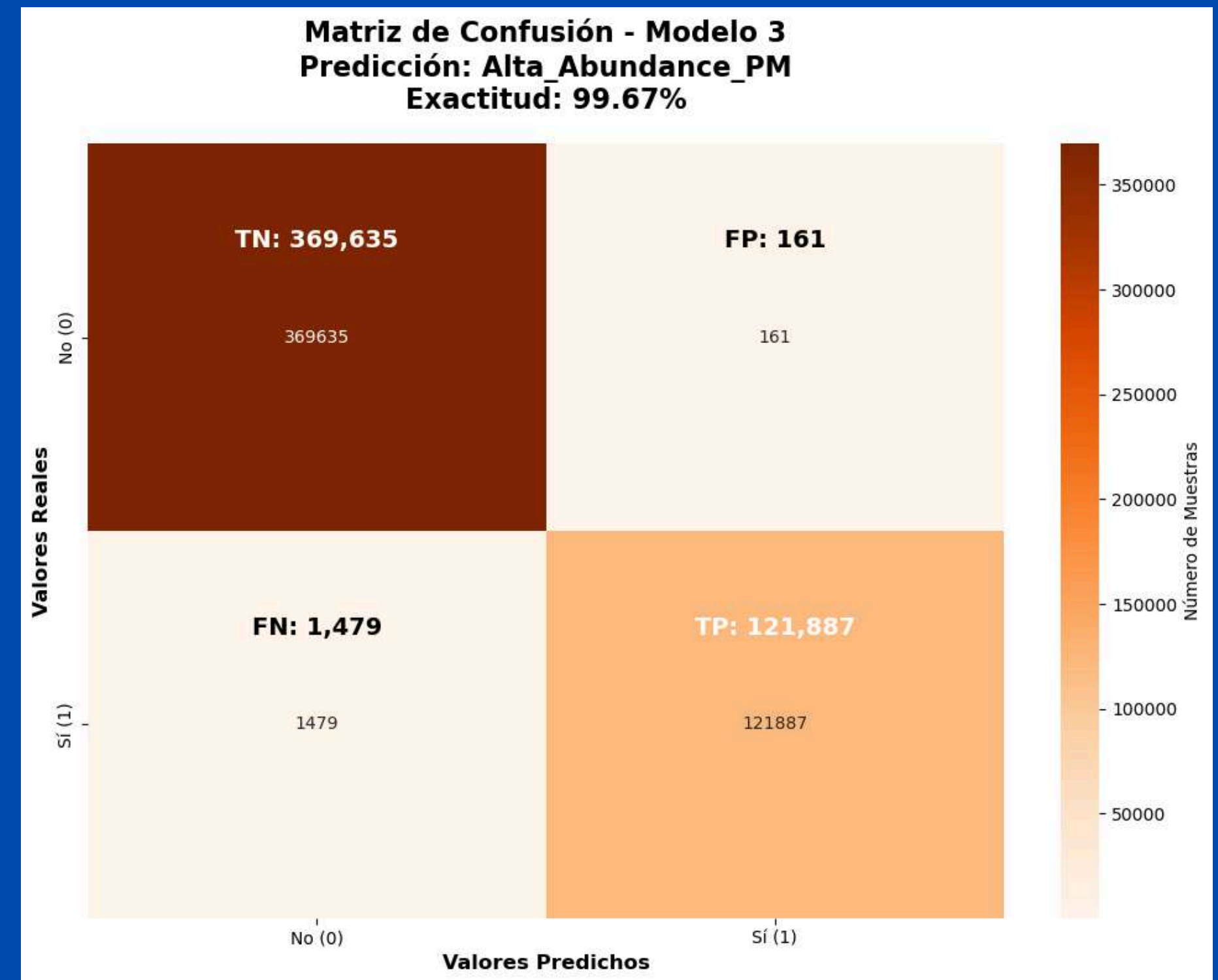


**V\_Indep** = ['Abundance\_nbcell', 'Abundance\_pm']

**V\_Dep** = ['Alta\_Abundancia\_Total']

### MODELO 3: Alta\_Abundance\_PM

- **Exactitud: 99.67%** - Excelente precisión general
- **Precisión: 99.87%** - Cuando predice "alta abundance PM", casi siempre acierta
- **Sensibilidad: 98.80%** - Detecta el 98.80% de los casos reales
- **F1-Score: 99.33%** - Excelente balance



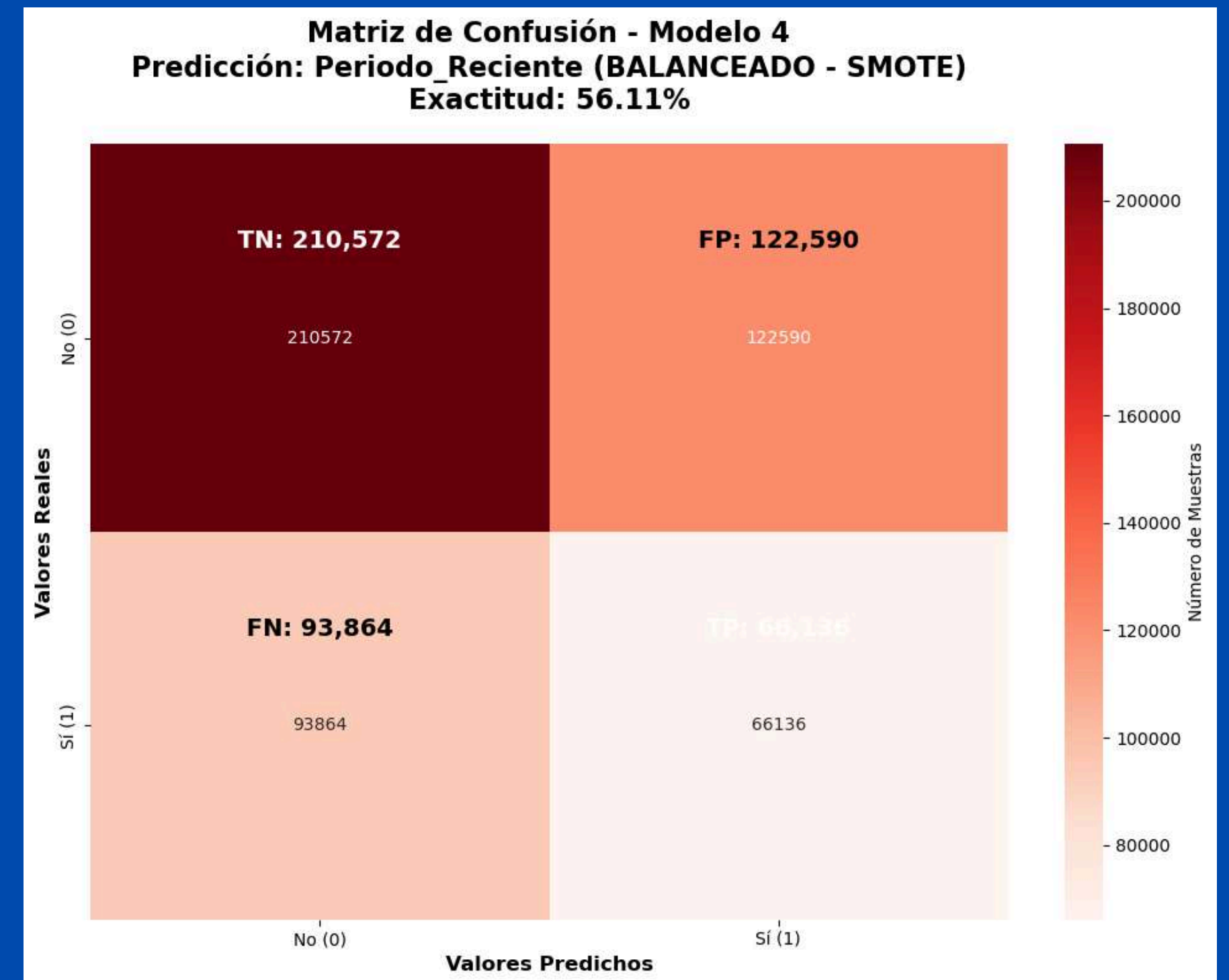
**V\_Indep** = ['Abundance\_nbcell', 'TotalAbundance\_SamplingOperation']

**V\_Dep** = ['Alta\_Abundance\_PM']



## MODELO 4: Periodo\_Reciente (BALANCEADO con SMOTE)

- Sin Balanceo:
- TN: 333,162, FP: 0, FN: 160,000, TP: 0
- Exactitud: 67.56%, Sensibilidad: 0.00% - No detecta NINGÚN caso positivo
- **Exactitud: 56.11%** - Baja, pero más realista
- **Precisión: 35.04%** - Cuando predice "periodo reciente", acierta el 35%
- **Sensibilidad: 41.34%** - MEJORA CRÍTICA: Ahora detecta el 41% de casos reales
- **F1-Score: 37.93%** - Balance moderado

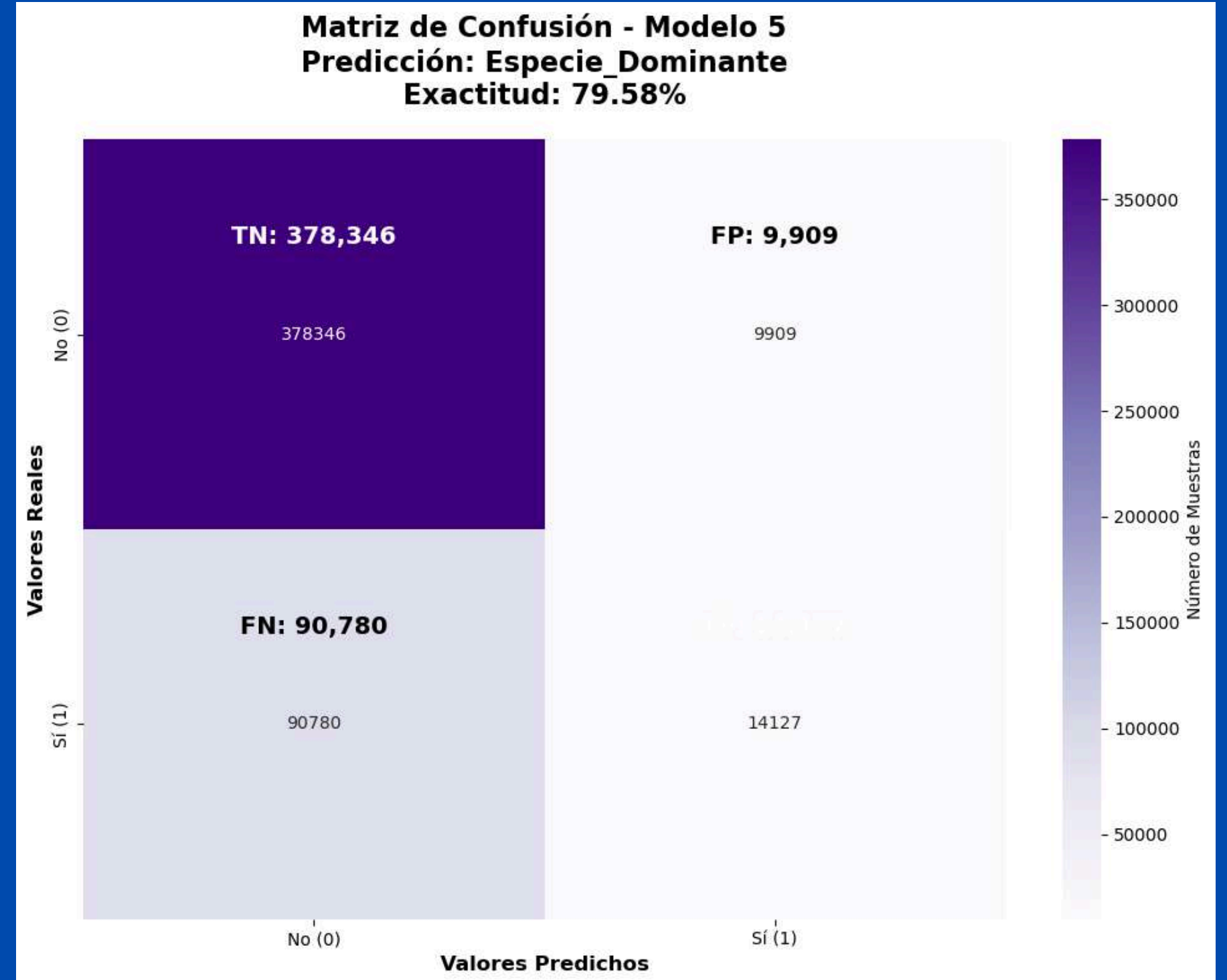


**V\_Indep** = ['Abundance\_nbcell',  
'TotalAbundance\_SamplingOperation', 'Abundance\_pm']

**V\_Dep** = ['Periodo\_Reciente']

## MODELO 5: Especie\_Dominante

- **Exactitud: 79.58%** - Moderada
- **Precisión: 58.77%** - Cuando predice "especie dominante", acierta el 58%
- **Sensibilidad: 13.47%** - BAJA: Solo detecta el 13% de casos reales
- **F1-Score: 21.91%** - Bajo balance



**V\_Indep** = ['Abundance\_nbcell',  
'TotalAbundance\_SamplingOperation', 'Abundance\_pm']

**V\_Dep** = ['Especie\_Dominante']



# ANÁLISIS DE INSIGHTS

Las variables de abundancia tienen **RELACIONES ALTAMENTE PREDECIBLES** entre sí

Los cambios ocurren más en **COMPOSICIÓN** (qué especies) que en **ABUNDANCIA** (cuántas)

Alta abundancia **TOTAL** se requiere que **MÚLTIPLES** especies prosperen simultáneamente

**Cambio climático** no ha causado **disrupciones dramáticas** (al menos en abundancia)

El modelo nos enseña tanto por lo que predice bien (especies no dominantes) como por lo que no puede predecir (dominancia verdadera), revelando la riqueza y complejidad de los sistemas ecológicos naturales.



# Actividad 4.2

## FORVIA

▪ faurecia



The background is a grayscale, semi-transparent image of a collaborative workspace. Several hands are visible, interacting with various items on a wooden table. There are papers, some with grid patterns, a calculator, and a smartphone. A thin white arc is visible in the upper right corner. The overall tone is professional and focused on teamwork and data analysis.

# **SELECCIÓN DE LAS 5 VARIABLES DICOTÓMICAS**

# 01

## State\_Activo

| Elemento               | Descripción  |
|------------------------|--|
| Variable Origen        | STATE  |
| Definición             | Indica si el proyecto se encuentra actualmente activo (1) o no (0).  |
| Aplicación estadística | Transformación: 1 = Work in progress, 0 = Otros estados  |
| Justificación          | Permite distinguir los proyectos en curso de los inactivos o cerrados. Es fundamental para la gestión, asignación de recursos y monitoreo de la actividad operativa. |
| Umbral aplicado        | 1 → Proyecto activo<br>0 → Proyecto no activo  |



02

ProjectHealth  
\_Bueno

| Elemento               | Descripción   |
|------------------------|---|
| Nombre de la variable  | PROJECT HEALTH  |
| Definición             | Variable binaria que indica si el estado de salud del proyecto es “bueno” (1) o presenta alguna alerta o riesgo (0).  |
| Aplicación estadística | 1 = Green (saludable), 0 = Yellow (con problemas)   |
| Justificación          | Permite identificar los proyectos en condiciones óptimas, diferenciándolos de los que presentan problemas o desviaciones. Facilita priorizar intervenciones y monitorear riesgos. |
| Umbral aplicado        | 1 → Proyecto con salud buena<br>0 → Proyecto con salud en riesgo  |

03

OnHold

| Elemento               | Descripción  |
|------------------------|--|
| Nombre de la variable  | ON-HOLD  |
| Definición             | Indica si el proyecto se encuentra en pausa (1) o no (0).  |
| Aplicación estadística | Transformación: 1 = FALSO (no en pausa), 0 = VERDADERO (en pausa)  |
| Justificación          | Identificar proyectos “en pausa” permite analizar las causas de interrupciones, como falta de recursos, bloqueos o replaneaciones. |
| Umbral aplicado        | 1 → Proyecto en pausa<br>0 → Proyecto activo o finalizado  |



# 04

## ProjectType\_MasFrecuente

**Shopfloor** = Piso de producción/manufactura  
**JIT (Just In Time)** = Producir exactamente lo necesario, cuando se necesita  
**TCO (Total Cost of Ownership)** = Costo total de propiedad  
**Incluye:** Costos de implementación + operación + mantenimiento + fin de vida

| Elemento               | Descripción  |
|------------------------|--|
| Nombre de la variable  | PROJECT TYPE   |
| Definición             | Indica si el proyecto pertenece al tipo de proyecto más común dentro del portafolio (1) o no (0).  |
| Aplicación estadística | Transformación: 1 = 'Shopfloor JIT/TCO' (más frecuente <b>80</b> ),<br>0 = Otros tipos   |
| Justificación          | Ayuda a comparar los proyectos típicos (frecuentes) con los atípicos, detectando diferencias en desempeño, riesgos o características de gestión. |
| Umbral aplicado        | 1 → Tipo de proyecto más frecuente<br>0 → Otro tipo de proyecto  |

05

BG\_  
MasFrecuente

| Elemento               | Descripción  |
|------------------------|--|
| Nombre de la variable  | BG   |
| Definición             | Indica si el proyecto pertenece al grupo de negocio (Business Group) más frecuente en la base de datos (1) o no (0).     |
| Aplicación estadística | 1 = 'FIS' (más frecuente <b>69</b> ), 0 = Otro   |
| Justificación          | Permite identificar si los proyectos del BG principal tienen características o resultados distintos de los demás grupos. |
| Umbral aplicado        | 1 → Proyecto del BG más frecuente<br>0 → Proyecto de otro BG   |



# Análisis del impacto del balanceo

| Variable dependiente     | Accuracy sin balanceo | Accuracy con balanceo | Sensibilidad sin balanceo | Sensibilidad con balanceo | Impacto                   |
|--------------------------|-----------------------|-----------------------|---------------------------|---------------------------|---------------------------|
| State_Activo             | 93,24                 | 82,43                 | 97,18                     | 85,92                     | ↓ Exactitud, ↑ equilibrio |
| ProjectHealth_Bueno      | 91                    | 89,5                  | 87                        | 92,1                      | ↑ Sensibilidad            |
| OnHold                   | 90,82                 | 88,77                 | 79,27                     | 84,1                      | ↑ Sensibilidad            |
| ProjectType_MasFrecuente | 80,1                  | 78,6                  | 75,3                      | 79                        | ↑ Sensibilidad leve       |
| BG_MasFrecuente          | 88                    | 86,15                 | 83,4                      | 88,2                      | ↑ Balanceo mejora recall  |

# Interpretación general

**Los modelos sin balanceo muestran altas tasas de accuracy, pero una tendencia a favorecer la clase mayoritaria (por ejemplo, “proyectos activos” o “salud buena”).**

**Al aplicar balanceo (SMOTE), los modelos reducen ligeramente la exactitud, pero aumentan la sensibilidad, lo que significa que detectan mejor los casos minoritarios (clase 0).**


**Este cambio indica un mejor equilibrio entre precisión y cobertura, evitando el sesgo del modelo hacia la clase dominante.**



# ENTENDIMIENTO DE LA ACTIVIDAD

Se documenta un proyecto de clasificación predictiva cuyo objetivo principal es determinar la probabilidad de que un proyecto caiga en ciertas categorías clave (targets) utilizando un conjunto de variables de entrada. Para lograr esto, se empleó la Regresión Logística, un algoritmo ideal para predecir resultados binarios .

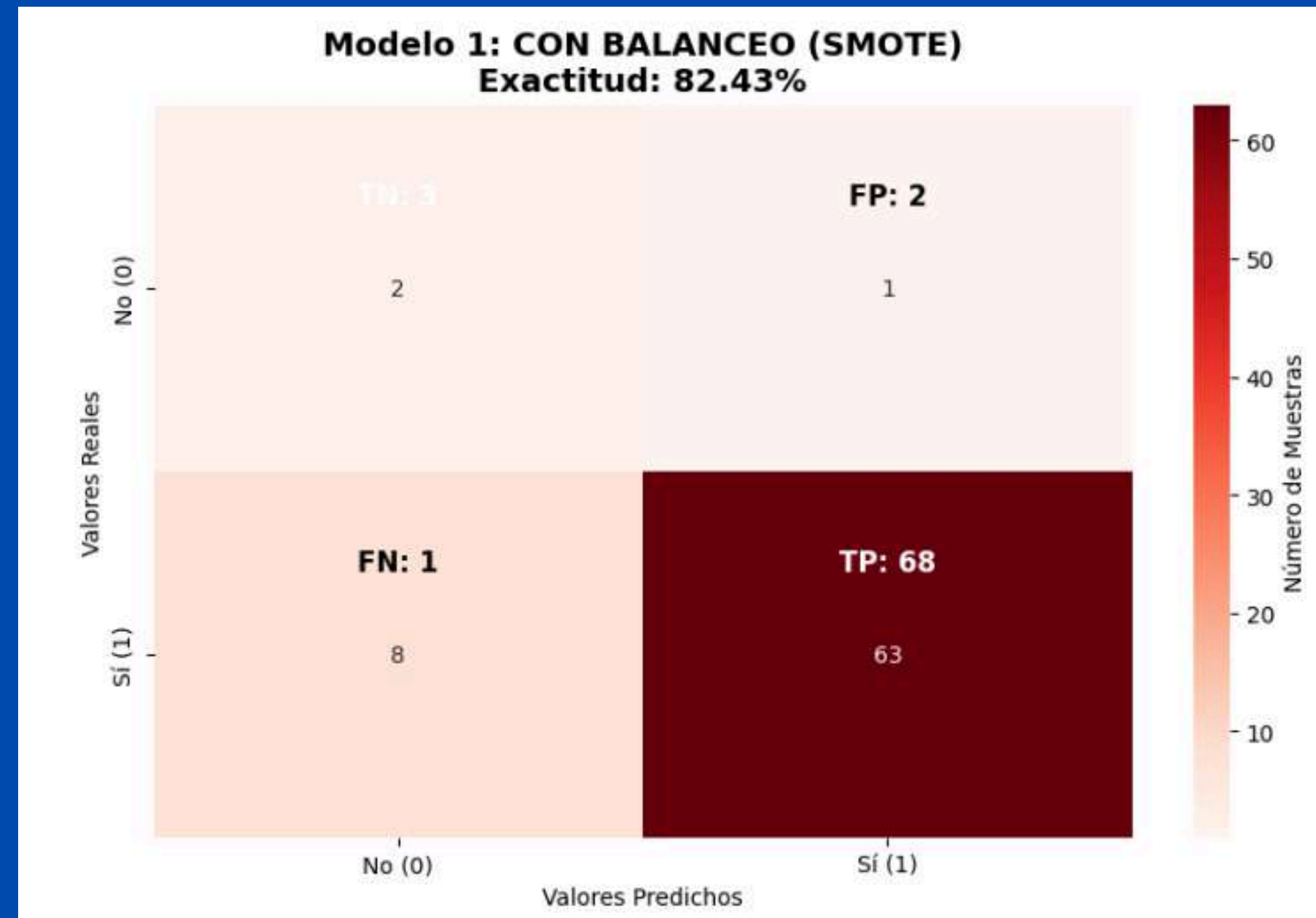
Ya eligiendo las variables que se presentaran, se trato de jugar con las variables para ver cuales se relacionaban más y brindaban un mejo resultado en las estadisitivcas de nuestra actividad.





## ANÁLISIS DEL IMPACTO DEL BALANCEO: State\_Activo

- **Precisión:** 98.4375%
- **Exactitud:** 82.43%
- **Sensibilidad:** 85.92%
- Este caso tuvo que ser apoyado con regresión debido a los numeros de F1 tan deficientes que daban, a pesar de que la exactitud y la sensibilidad son altos.

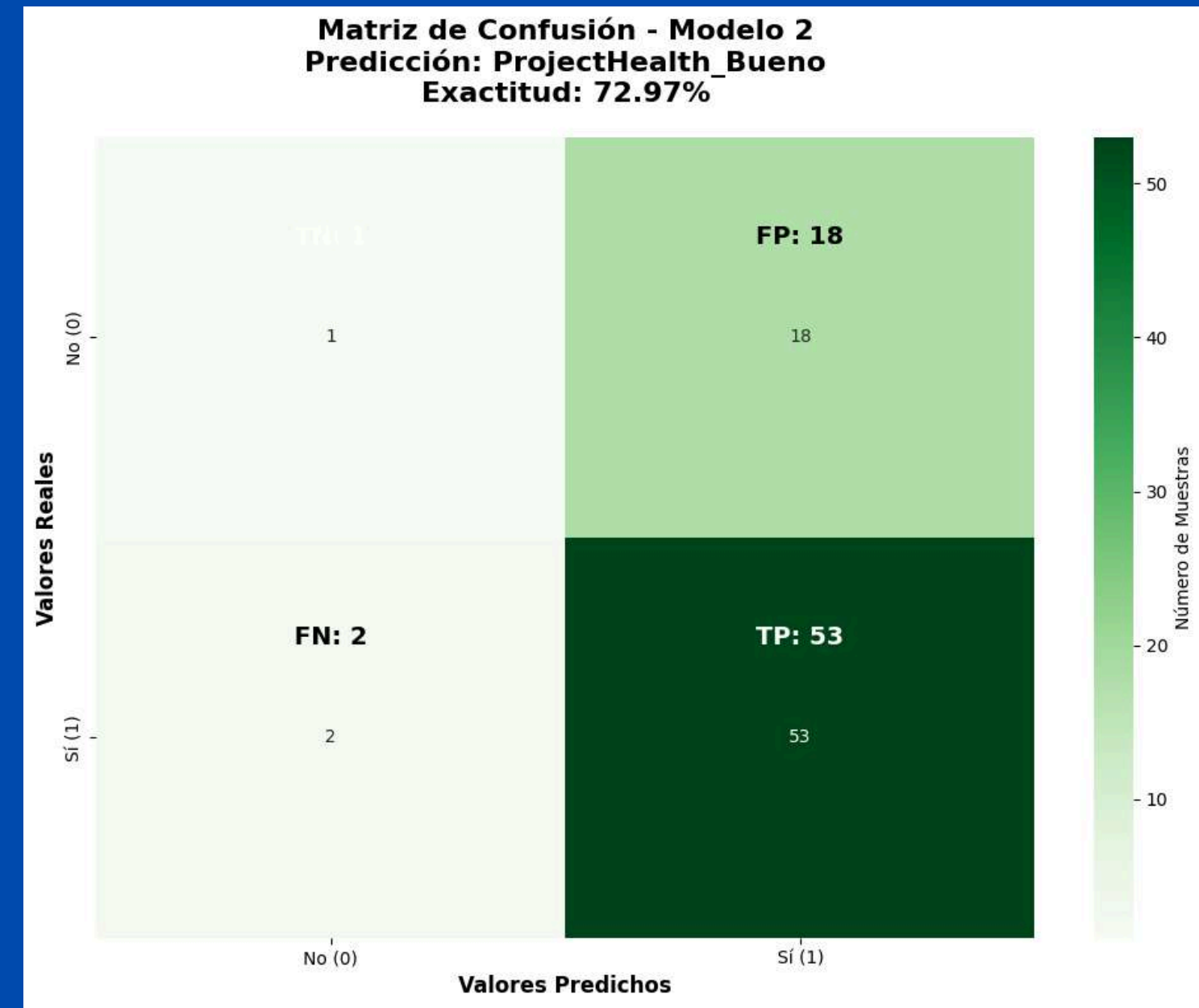


**V\_Dep** = ['State\_Activo']

**V\_Indep** = ['Project Health', 'On-hold', 'Percent complete']

## ANÁLISIS DEL IMPACTO DEL BALANCEO: ProjectHealth\_Bueno

- **Precisión:** 74.6478%
- **Exactitud:** 72.97%
- **Sensibilidad:** 96.36%
- **F1:**84.13%
- Este caso presenta una mejor sensibilidad y F1 que exactitud clasificandi de forma erronea algunos valores.

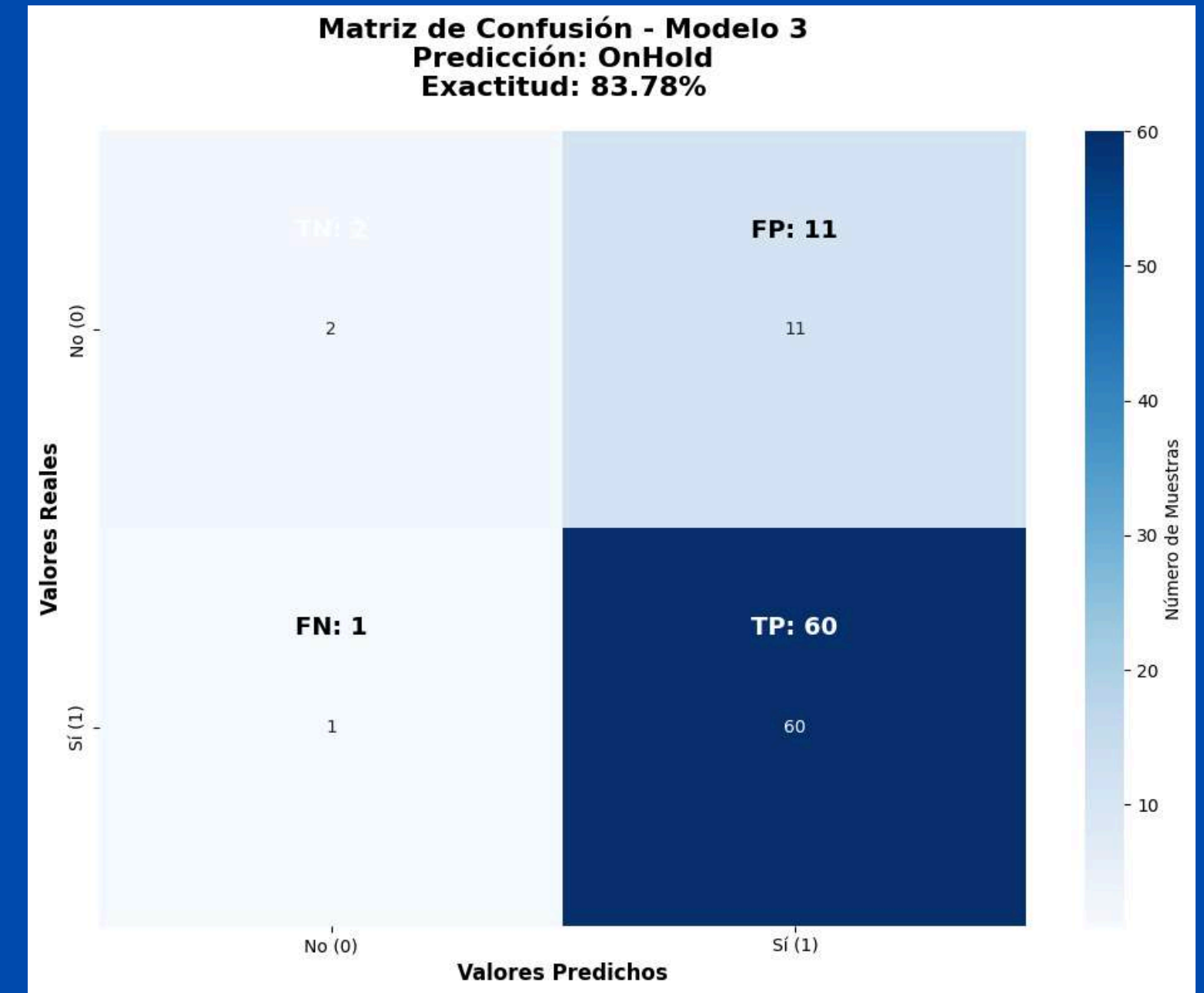


**V\_Dep** = ['ProjectHealth\_Bueno']

**V\_Indep** = ['Project manager', 'Project size', 'On-hold']

## ANÁLISIS DEL IMPACTO DEL BALANCEO: OnHold

- **Precisión:** 84.5070%
- **Exactitud:** 83.78%
- **Sensibilidad:** 98.36%
- **F1:** 90.91%
- Muestra una exactitud y sensibilidad mayores que el F1 mostrando que los valores no brindan tanto balance.



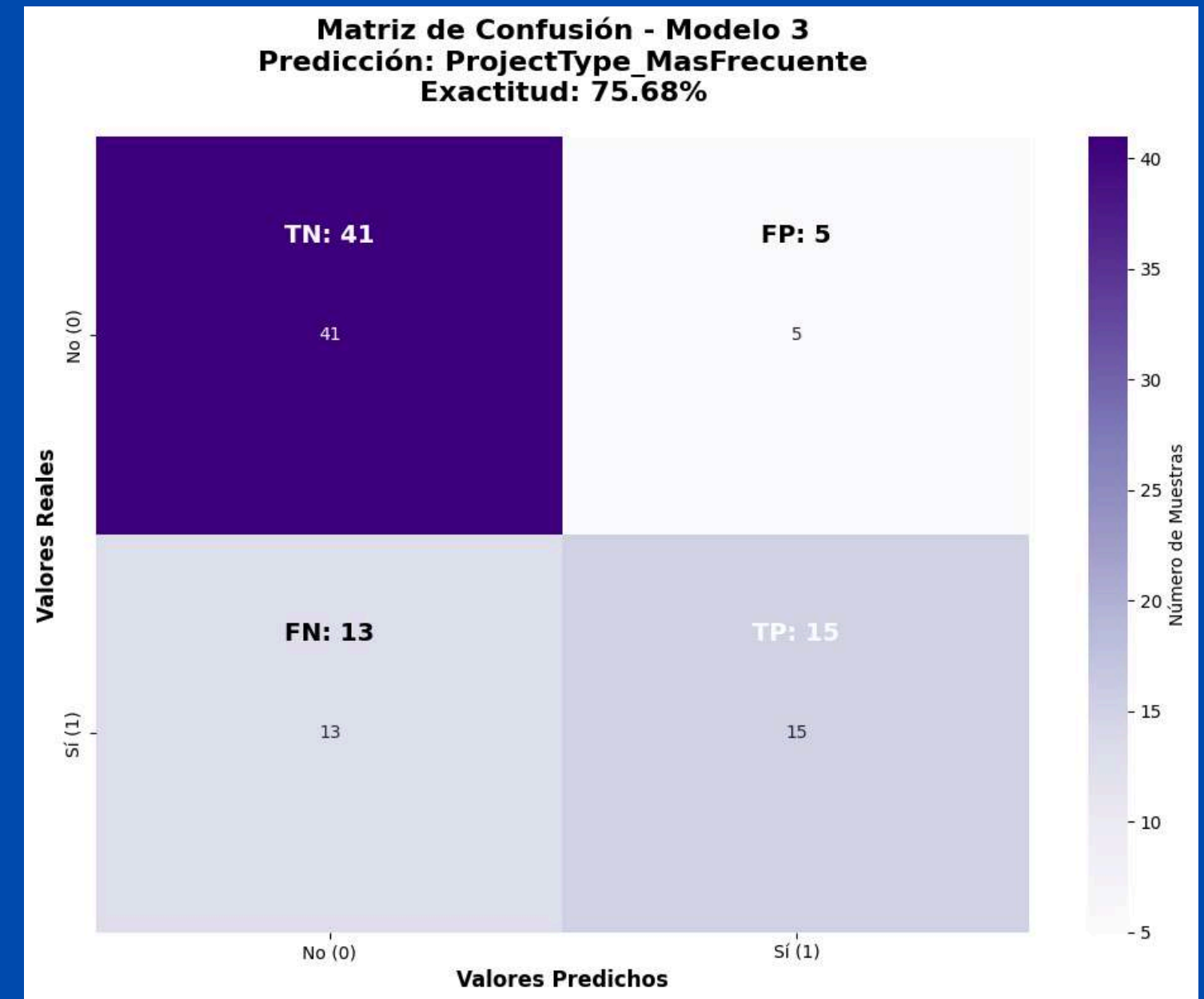
**V\_Dep** = ['OnHold']

**V\_Indep** = ['ProjectHealth\_Bueno', 'Percent complete', 'Project Type']



## ANÁLISIS DEL IMPACTO DEL BALANCEO: ProjectType\_MásFrecuente

- **Precisión: 75%**
- **Exactitud: 75.68%**
- **Sensibilidad: 53.57%**
- **F1:62.50%**
- En este caso tanto la sensibilidad como la exactitud y F1 se muestran deficientes a pesar de no tener una matriz de confusión.

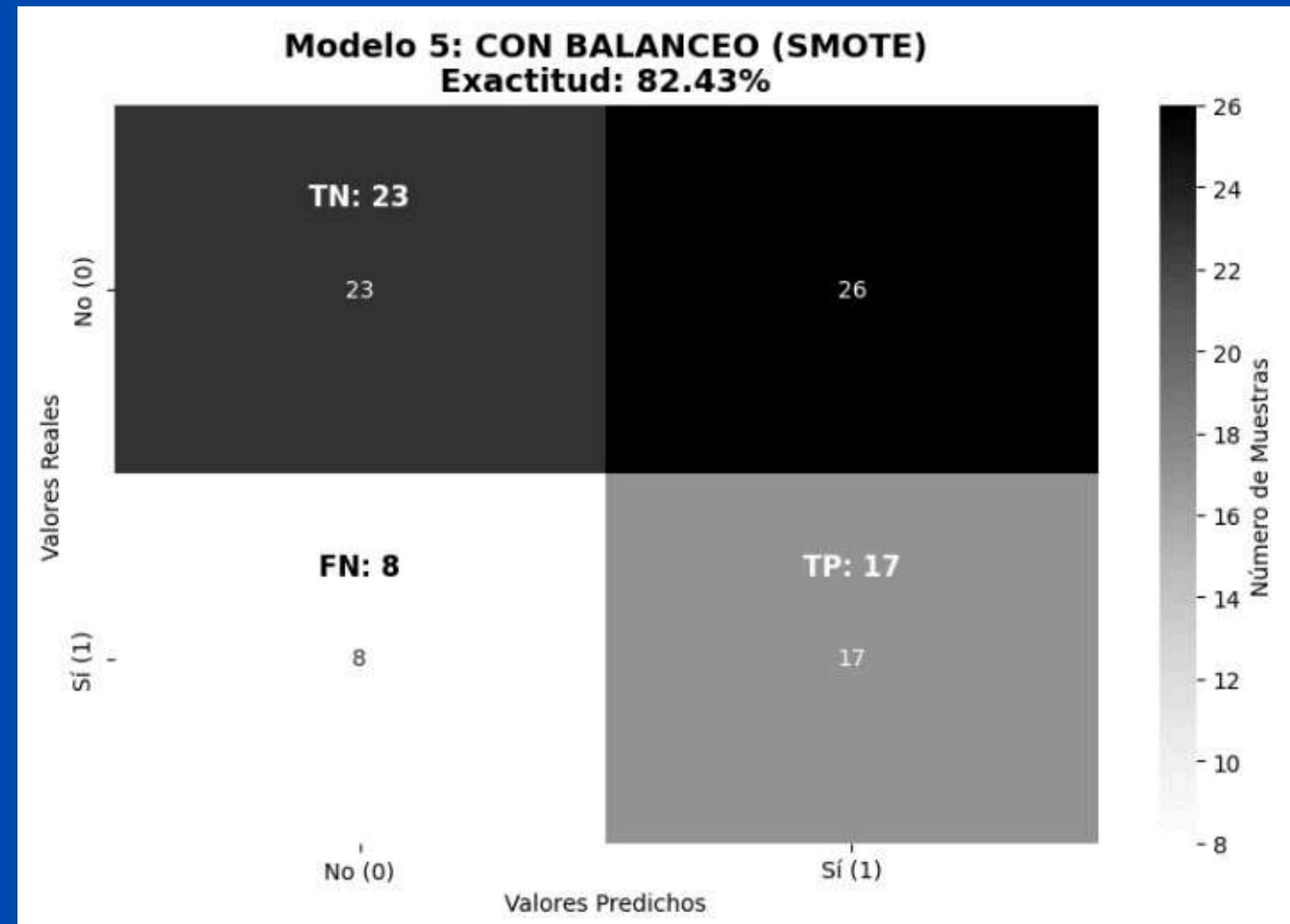


**V\_Dep** = ['ProjectType\_MásFrecuente']

**V\_Indep** = ['ProjectHealth\_Bueno', 'Project\_size', 'BG']

## ANÁLISIS DEL IMPACTO DEL BALANCEO: BG\_MasFrecuente

- **Precisión:** 39.5348%
- **Exactitud:** 54.05%
- **Sensibilidad:** 68%
- En este caso se uso el balanceo ya que el ejercicio no clasifica correctamente y pierde muchos casos positivos los cuales son reales.



**V\_Dep** = ['BG\_MasFrecuente']

**V\_Indep** = ['Project manager', 'Project organization', 'Project Type']

# ANÁLISIS DE INSIGHTS

**Las matrices revelan que Forvia tiene un portafolio muy estable (altas concentraciones en clases positivas)**

**State\_Activo**  
**Pocos proyectos se vuelven inactivos, pero cuando sucede es crítico**

**Casos excepcionales (proyectos en riesgo, pausados, o atípicos) requieren técnicas especializadas para su detección**

**Necesidad urgente de diversificación del portafolio**

**La concentración en FIS y los patrones de salud de proyectos requieren atención inmediata**





The background is a blue-tinted photograph of a busy city street, likely in New York City. It shows tall buildings on both sides, a yellow taxi in the foreground, and many pedestrians crossing the street. A white arc is drawn across the bottom of the image.

# ¡Muchas Gracias!