

Octubre 2025

Análisis exploratorio de  
proyectos Datathon y FORVIA

# **REGRESIÓN LOGÍSTICA**

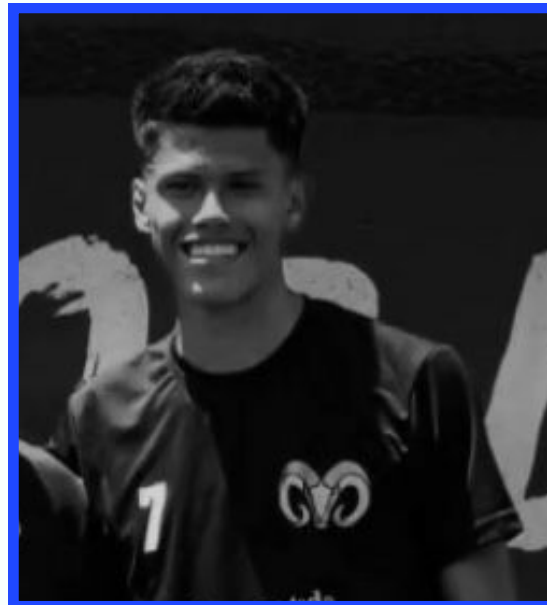
FORVIANOS.py

# FORVIANOS.PY



Maria Matanzo

A01737554



Jorge Cortes

A01736236



Marco Cornejo

A01276411



Eduardo Torres

A01734935



Laisha Puan

A01736397

# OBJETIVOS

- Aplicar Regresión Logística para analizar relaciones entre variables binarizadas.
- Evaluar el desempeño de varios modelos con distintas variables dependientes (Y).
- Medir el rendimiento con Accuracy, Precision, Recall y F1-score.

# **M E T O D O L O G Í A   D E L A S   A C T I V I D A D E S**

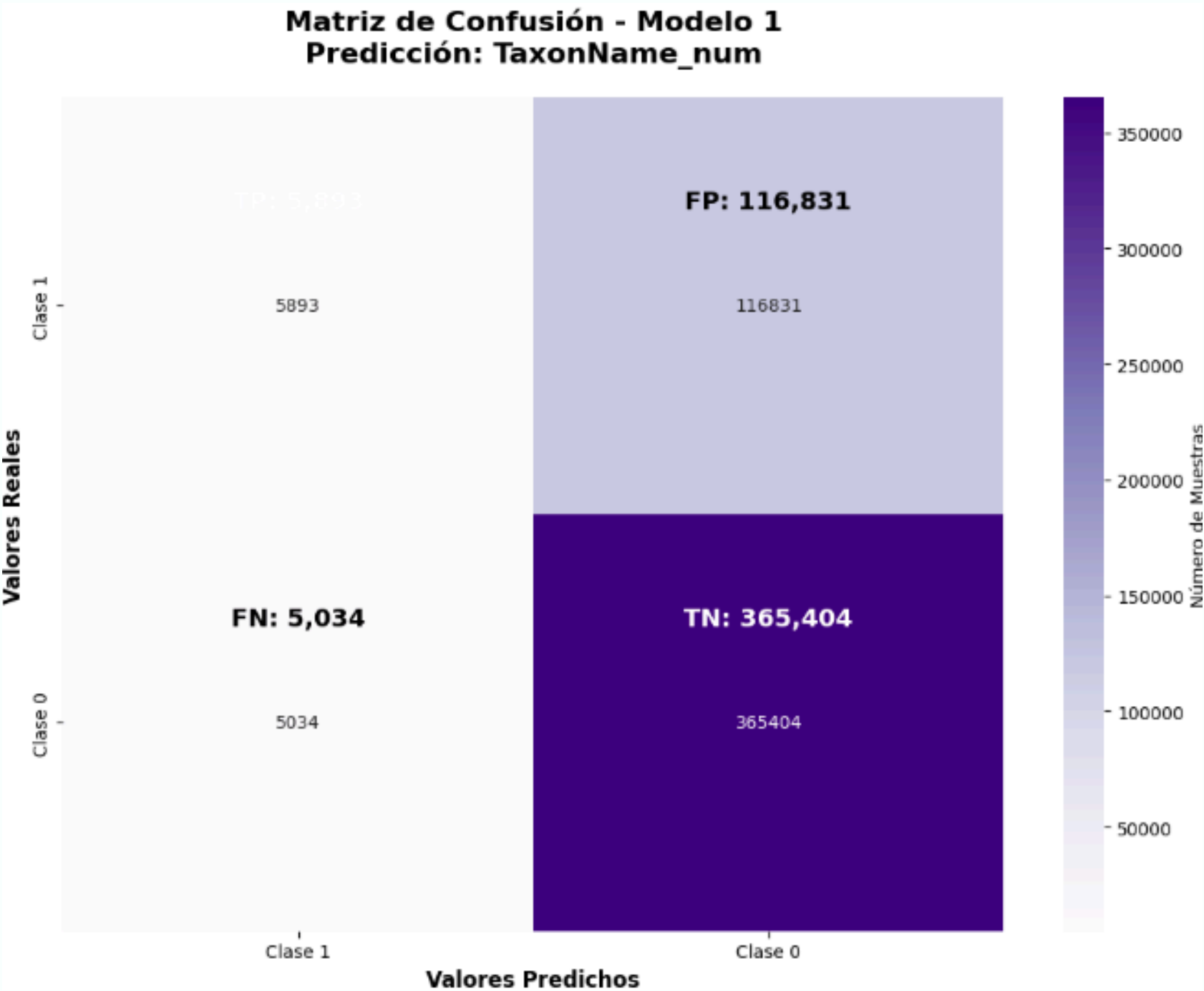
- 1.Limpieza de datos y tratamiento de valores nulos.
- 2.Conversión de variables categóricas a numéricas mediante codificación.
- 3.Binarización (0 o 1) de las variables con base en percentiles o valores centrales.
- 4.División de datos.
- 5.Entrenamiento de modelos de Regresión Logística.
- 6.Evaluación de desempeño con Accuracy, Precision, Recall y F1-score.
- 7.Interpretación de resultados y detección de sesgos o desbalance.

# MODELO PREDICCIÓN DE TAXONNAME\_NUM

- **Variable dependiente (Y):**  
TaxonName\_num
- **Variables independientes (X):**  
Abundance\_nbcell,  
TotalAbundance\_SamplingOperation,  
Abundance\_pm
- **Accuracy:** 75.29%
- **Recall Clase 0:** 4.8%

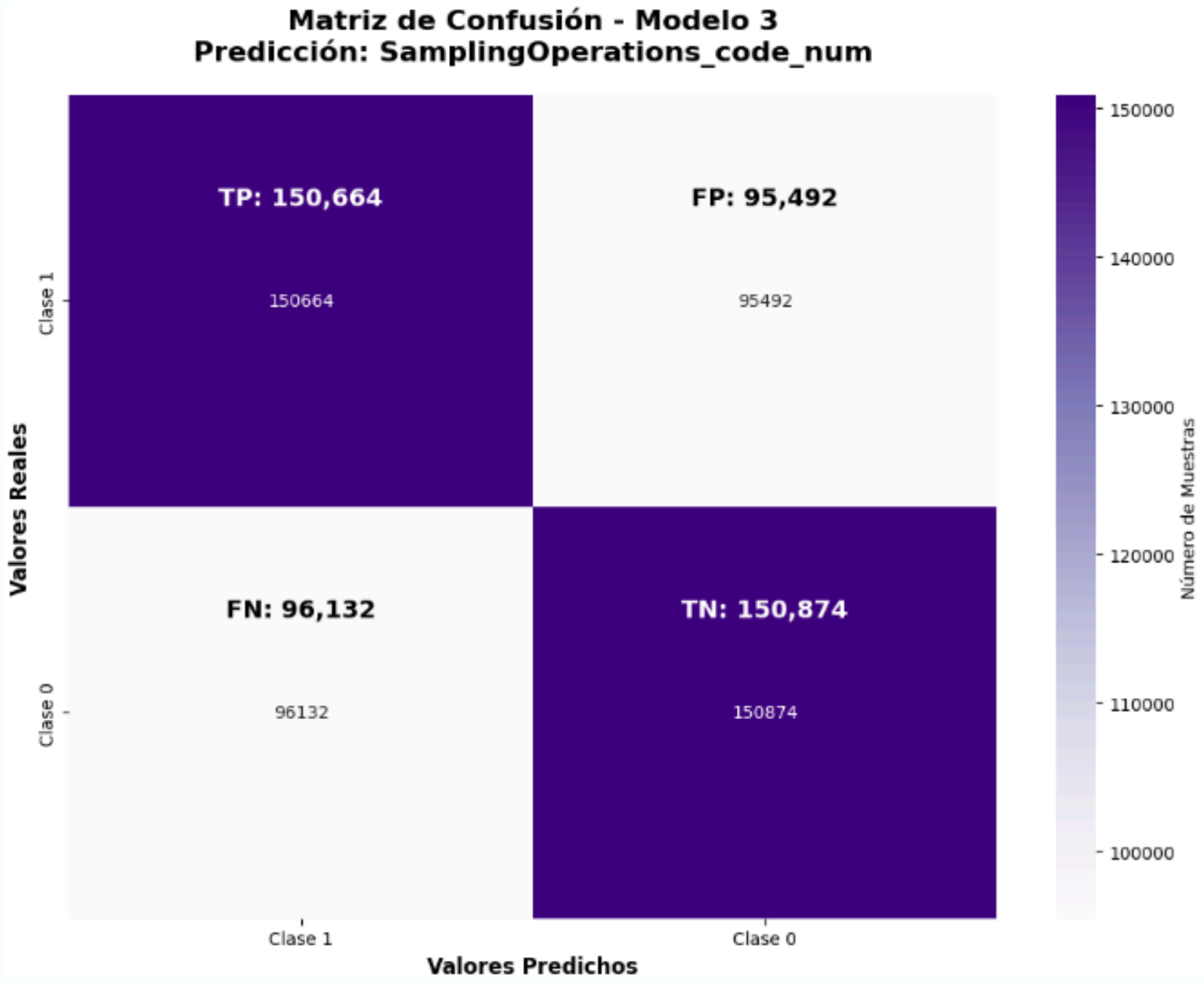
**Resultados:**

- El modelo tiene alta exactitud, pero falla al identificar la clase minoritaria.
- Gran cantidad de falsos positivos.
- Resultados similares a Taxoncode\_num.



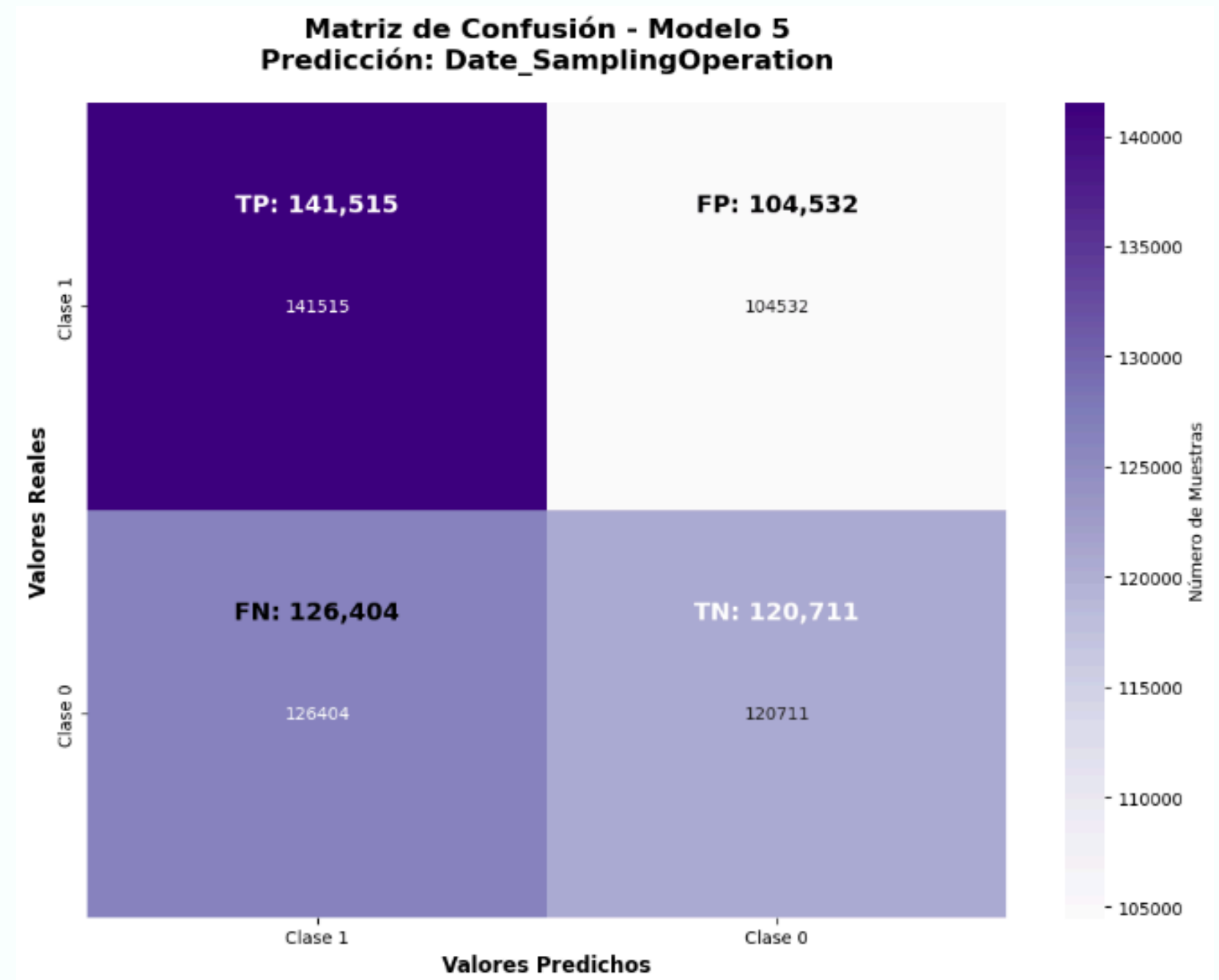
# MODELO SAMPLING OPERATIONS \_CODE\_NUM

- **Variable dependiente** Y:  
SamplingOperations\_code\_num
- **Variables independientes** X:  
CodeSite\_SamplingOperations\_num,  
Date\_SamplingOperation
- **Accuracy:** 61.14%
- **Recall:** ~61% para ambas clases.
- **Hallazgos:** Métricas equilibradas, sin sesgo notable.
- Resultados similares a  
CodeSite\_SamplingOperations\_num.



# MODELO DATE\_SAMPLINGOPERATION

- **Variable dependiente Y:**  
Date\_SamplingOperation (dicotómica por fecha central)
- **Variable independiente X:**  
TotalAbundance\_SamplingOperation,  
Abundance\_pm
- **Accuracy:** 53.17%
- **Recall Clase 0: 57.52% / Recall Clase 1: 48.85%**
- **Interpretación:** Modelo apenas mejor que una conjetura aleatoria.
- Variables de abundancia no correlacionan bien con el tiempo de muestreo.



# CONCLUSIONES DATASET DATATHON

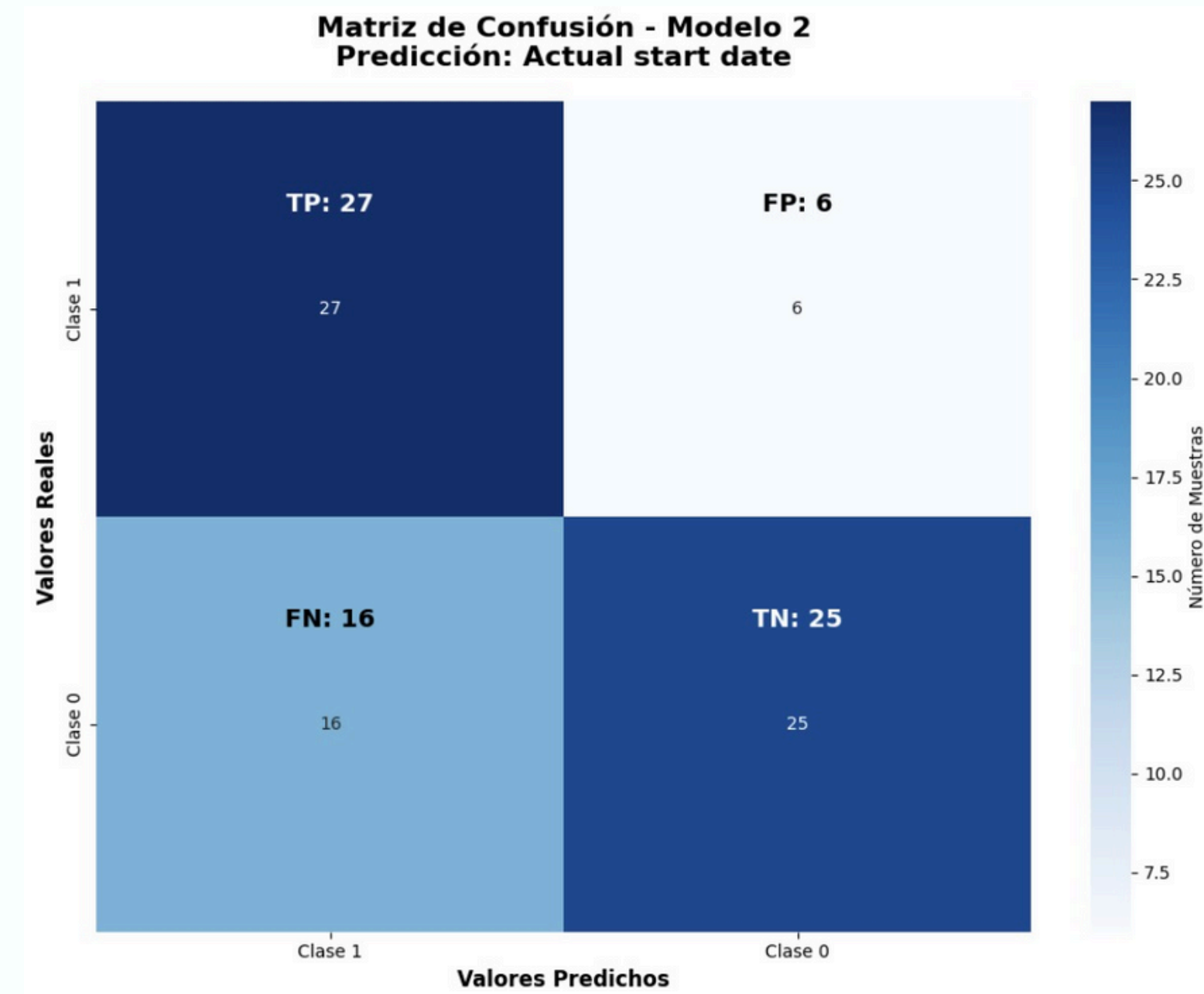
1. Modelos 1 y 2: exactitud alta pero recall mínimo → desbalance extremo.
2. Modelos 3 y 4: rendimiento balanceado (~61%).
3. Modelo 5: sin correlación temporal significativa.
4. En general: correlaciones débiles y sesgo hacia clases mayoritarias.

# ACTIVIDAD 4.2

forvianos.py

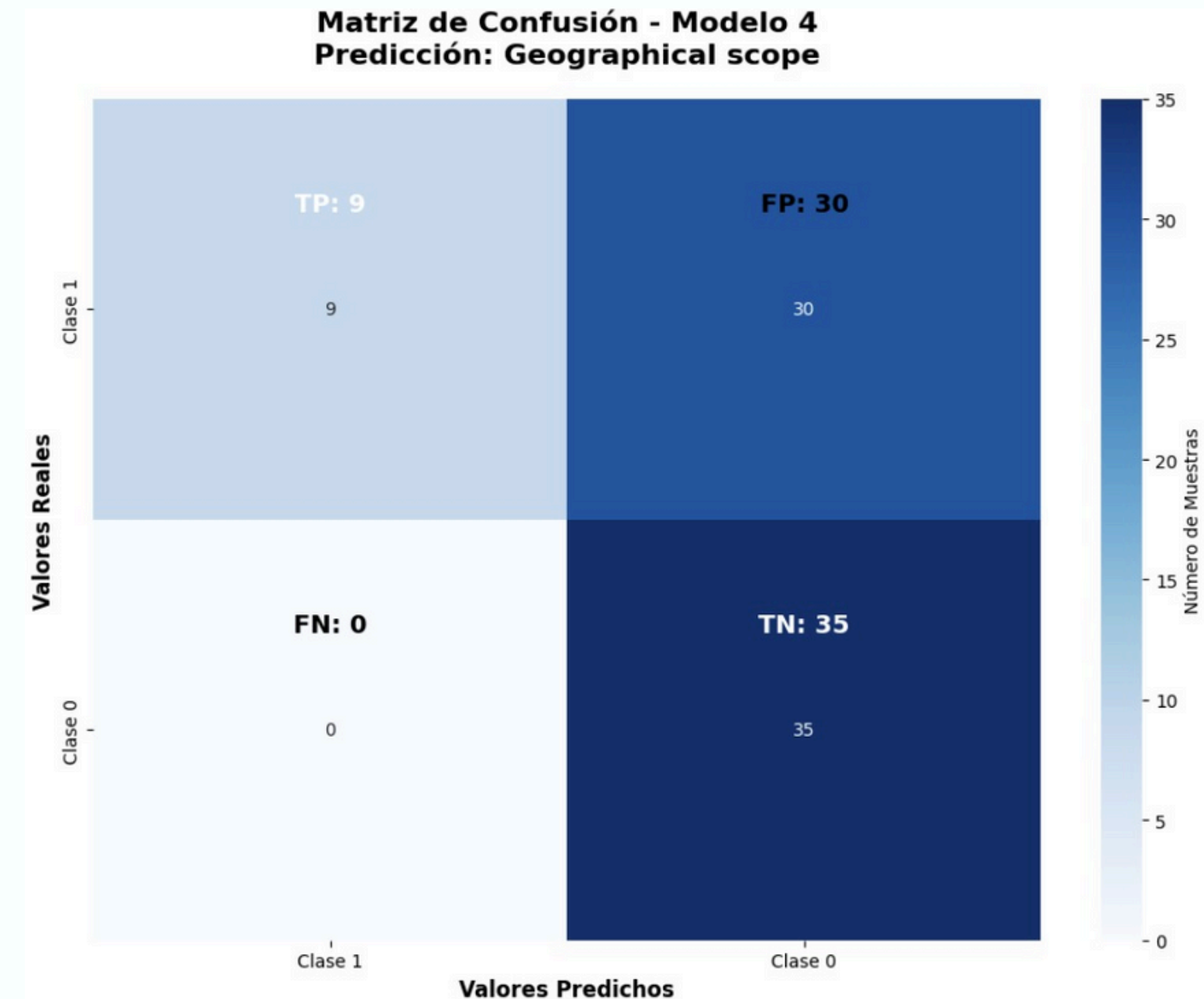
# MODELO ACTUAL START DATE

- **Variables:**
- **Y:** Actual start date (dicotómica)
- **X:** Geographical scope, Planned start date, Percent complete
- **Accuracy:** 63.51%
- **Precisión Clase 1:** 67.50%
- **Recall Clase 1:** 65.85%
- **Conclusión:** Es el modelo con mejor rendimiento general, con buen equilibrio entre precisión y sensibilidad.



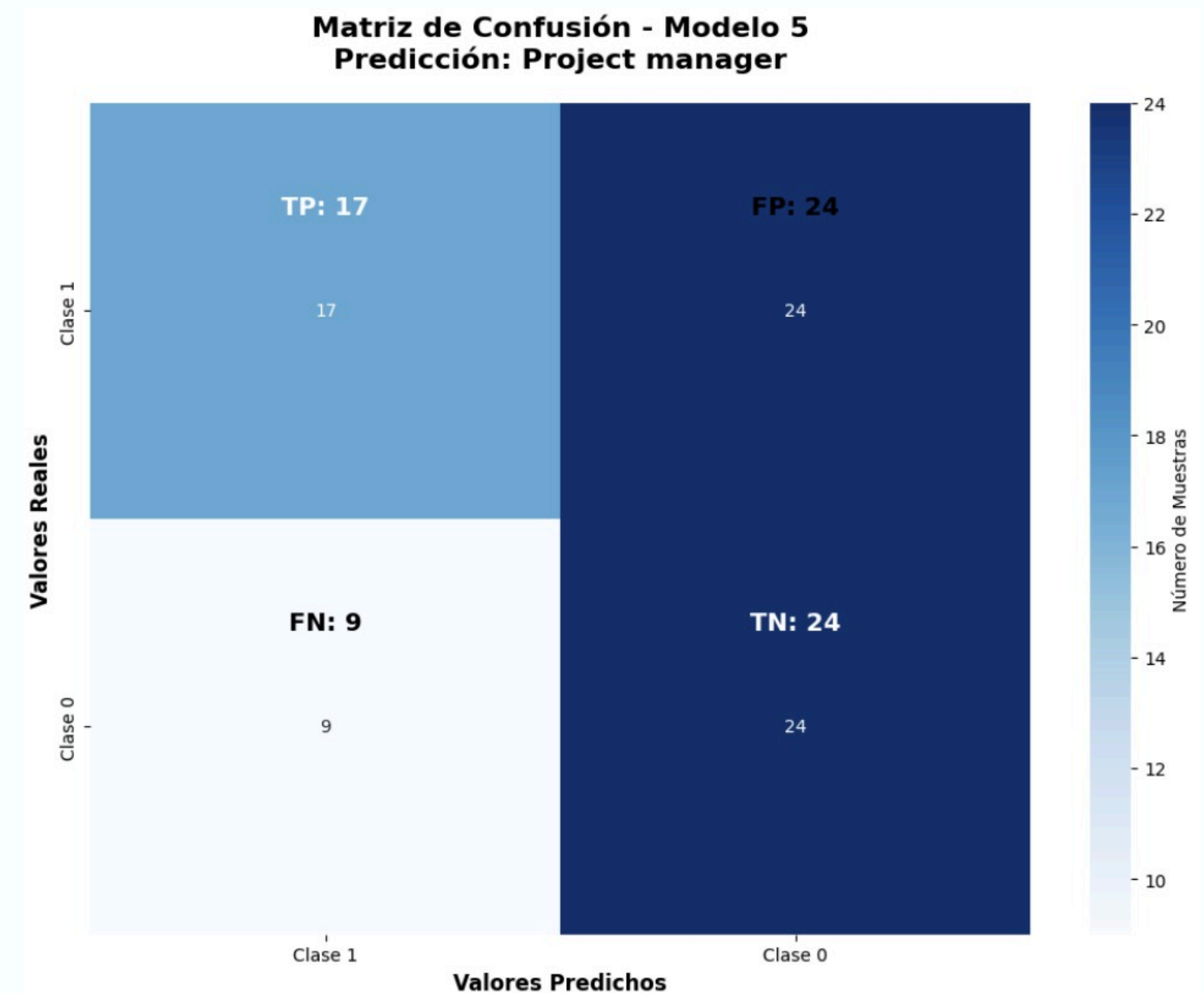
# MODELO GEOGRAPHICAL SCOPE

- **Variables:**
- **Y:** Geographical scope (dicotómica)
- **X:** Percent complete, Actual start date
- **Accuracy:** 54.05%
- **Recall Clase 1:** 80%
- **Recall Clase 0:** 30.77%
- **Conclusión:** Modelo sesgado hacia la Clase 1, con baja exactitud global.



# MODELO PROJECT MANAGER

- **Variables:**
- **Y:** Project manager (dicotómica)
- **X:** Percent complete, Actual start date
- **Accuracy:** 58.11%
- **Recall Clase 1:** 50.00%
- **Recall Clase 0:** 66.67%
- **Conclusión:** Desempeño cercano al azar; el modelo tiende a predecir más la Clase 0, mostrando bajo poder predictivo y ligero sesgo.



# CONCLUSIONES DATASET FORVIA

1. Mejor modelo: Caso 2 (Actual start date, 63.51% accuracy).
2. Casos 3 y 4 presentan sesgos fuertes entre clases.
3. Casos 1 y 5: desempeño cercano al azar.
4. Correlaciones moderadas, pero problemas de balance de clases afectan el resultado.

# CONCLUSIONES

La Regresión Logística permite explorar correlaciones binarias, pero su éxito depende del balance y calidad de datos.

- Ambos conjuntos presentan exactitud moderada (50–63%), mostrando que las correlaciones entre variables son parciales.

En general, se observó que algunos modelos logran un desempeño equilibrado, mientras que otros presentan sesgo hacia una de las clases o un comportamiento cercano al azar, lo que refleja la complejidad de los datos y la necesidad de un análisis más profundo en futuras etapas.

Octubre 2030

MUCHAS  
**GRACIAS**

Bruno Lago