Tecnológico de Monterrey

Analítica de datos y herramientas de inteligencia artificial II

Grupo 101

4.1 Regresión Logística Datathon

DataForge

Jesús Eduardo Valle Villegas | A01770616

Manuel Eduardo Covarrubias Rodríguez | A01737781

Diego Antonio Oropeza Linarte | A01733018

Ithandehui Joselyn Espinoza Mazón | A01734547

Mauricio Grau Gutierrez Rubio | A01734914

Última edición el 21 de Octubre de 2025

Objetivo

Aplicar modelos de Regresión Logística para analizar y predecir la relación entre distintas variables ecológicas del conjunto de datos *DiatomInventories_GTstudentproject_B.csv*, mediante la transformación de variables continuas en dicotómicas y la evaluación de su capacidad predictiva. Asimismo, ajustar los modelos en caso de desbalance de clases empleando técnicas de reponderación u oversampling (SMOTE), con el fin de mejorar la sensibilidad y la estabilidad de los resultados. Finalmente, comparar el desempeño de los modelos obtenidos a través de métricas de precisión, exactitud y sensibilidad, identificando los predictores más relevantes en la dinámica ecológica de las diatomeas.

Metodología

1. **Revisión y limpieza inicial de datos:**
   Se eliminaron registros duplicados y se verificó la coherencia de los tipos de datos (numéricos y categóricos).
   Se identificaron los valores faltantes y atípicos para garantizar la integridad del conjunto de datos.

2. **Normalización de variables numéricas:**
   Se aplicaron técnicas de normalización y detección de límites superiores e inferiores para asegurar que las variables cuantitativas (como la abundancia por célula y por mililitro) tuvieran una escala comparable.

3. **Creación de variables dicotómicas:**

   Se construyeron **cinco variables binarias (0 y 1)** con base en distintos criterios ecológicos y temporales, de la siguiente forma:

   - **Periodo_Reciente:** 1 si la fecha del muestreo es igual o posterior a 2019.
   - **Especie_Dominante:** 1 si la especie pertenece al grupo de las 10 más frecuentes.
   - **Alta_Abundancia_celular:** 1 si la abundancia celular (Abundance_nbcell) supera el percentil 75.
   - **Alta_Abundancia pm:** 1 si la abundancia por mililitro (Abundance_pm) está en el cuartil superior.
   - **Alta_Abundancia_total:** 1 si el sitio de muestreo aparece entre los 10 más representativos del conjunto de datos.

4. **Selección de la variable dependiente:**
   Se eligió una variable binaria de interés como objetivo del modelo de regresión logística.

5. **Selección de variables independientes:**
   Se consideraron variables numéricas representativas de la abundancia y densidad de las especies, tales como Abundance_nbcell, TotalAbundance_SamplingOperation y Abundance_pm.

6. **División del conjunto de datos:**
   Los datos se dividieron en subconjuntos de **entrenamiento (train)** y **prueba (test)**, para evaluar la capacidad de generalización del modelo y evitar sobreajuste.

7. **Balanceo de clases mediante oversampling (SMOTE):**
   Dado el desbalance natural entre las clases, se aplicó la técnica SMOTE (Synthetic Minority Oversampling Technique), que genera observaciones sintéticas de la clase minoritaria, equilibrando la proporción entre categorías.

8. **Entrenamiento del modelo:**
   Se implementó una regresión logística binaria con parámetros de regularización estándar, utilizando los datos balanceados para ajustar los coeficientes del modelo.

9. **Evaluación del desempeño:**
   El modelo se evaluó utilizando métricas de clasificación, incluyendo accuracy, precision, recall, F1-score y la matriz de confusión para visualizar el rendimiento de las predicciones.

10. **Visualización e interpretación de resultados:**
    Se elaboraron gráficos de matrices de confusión con diseño profesional y se interpretaron los valores de las métricas para determinar la efectividad del modelo y la relevancia de las variables.

Selección y creación de variables dicotómicas

Variable 1:

Elemento	Descripción / Valor
Nombre de la variable	Alta_Abundancia_celular
Definición	Alta abundancia celular (Abundance_nbcell >= 8)
Justificación	Se utiliza el percentil 75 para distinguir comunidades densas de diatomeas
Justificación específica	El 25% superior representa comunidades "prósperas"
Umbral aplicado	8 células

Variable 2:

Elemento	Descripción / Valor
Nombre de la variable	Alta Abundancia_Total
Definición	Alta abundancia total por operación de muestreo (TotalAbundance_SamplingOperation >= 408)
Justificación	Se utiliza el percentil 75 para identificar operaciones de muestreo con alta diversidad de especies
Justificación especifica	• Valores bajos (<408): Muestreos estándar o sitios con diversidad limitada • Valores altos (≥408):vMuestreos exhaustivos o sitios muy diversos.
Umbral aplicado	408 unidades

Variable 3:

Elemento	Descripción / Valor
Nombre de la variable	Alta_Abundancia_pm
Definición	Alta abundancia por metro (Abundance_pm >= 19.90)
Justificación	Se utiliza el percentil 75 para identificar muestras con alta densidad por metro
Justificación especifica	**Características de muestras con una alta densidad:** Microhábitats altamente productivos y Zonas de acumulación por corrientes
Umbral aplicado	19.90 unidades por metro

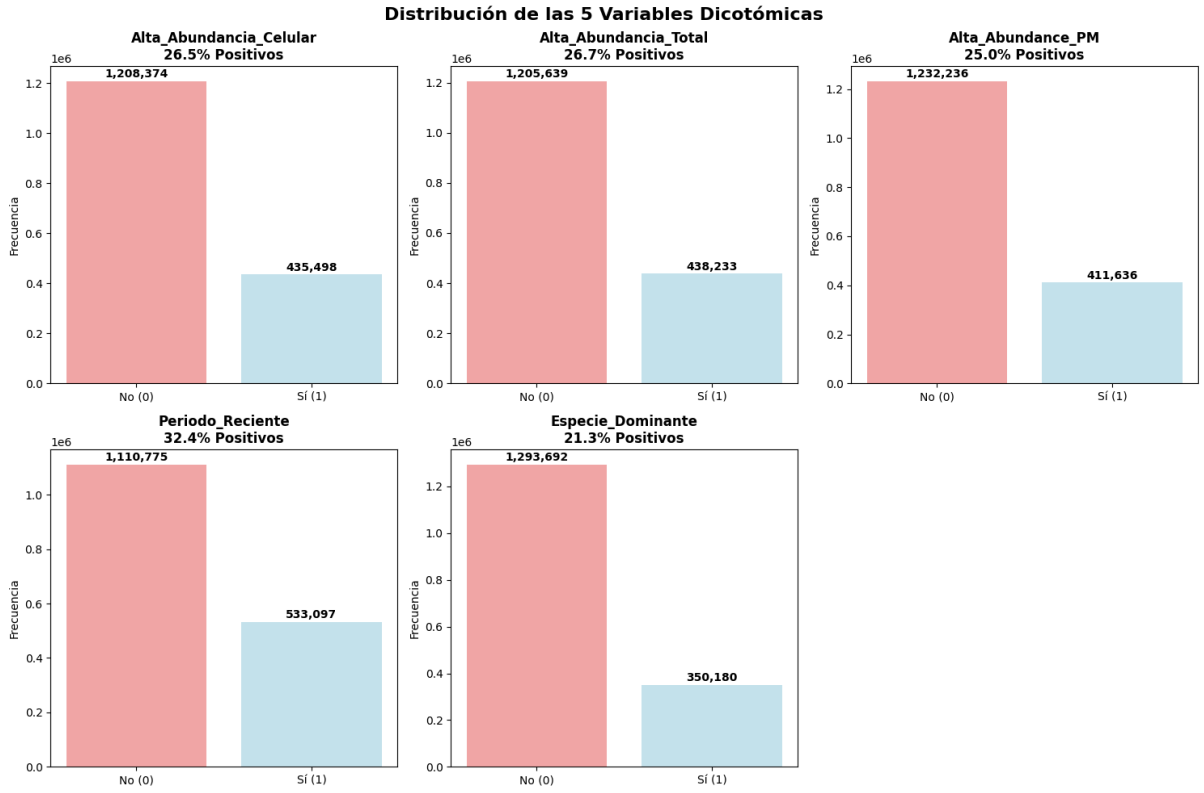
Variable 4:

Elemento	Descripción / Valor

Nombre de la variable	Periodo_reciente
Definición	Periodo reciente de muestreo (Año >= 2019)
Justificación	Los últimos 5 años representan las condiciones ambientales actuales y permiten analizar cambios recientes en la comunidad de diatomeas
Justificación especifica	Ciclos reproductivos de diatomeas: 7-10 años es muy antiguo (Hugo Beraldi-Campesi, 2015)
Umbral aplicado	Año 2019 en adelante

Variable 5:

Elemento	Descripción / Valor
Nombre de la variable	Especie_dominante
Definición	Especie dominante en el conjunto de datos (Top 10 especies más frecuentes)
Justificación	Las 10 especies más frecuentes representan aproximadamente el 20% de todos los registros, permitiendo identificar las especies ecológicamente más relevantes
Justificación especifica	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



Justificación:

La creación de variables dicotómicas permite transformar información continua o categórica en indicadores binarios que facilitan la aplicación de modelos predictivos como la regresión logística. En este caso, se establecieron cinco variables binarias que representan distintos aspectos ecológicos y temporales del conjunto de datos, permitiendo analizar de manera separada los patrones de abundancia, dominancia y recencia temporal de las muestras de diatomeas.

Interpretación:

La figura muestra la frecuencia relativa de las clases 0 y 1 para cada variable generada.

Se observa que todas las variables presentan una distribución moderadamente desbalanceada, donde las clases positivas (valor 1) representan entre 21% y 32% del total de registros.

Este comportamiento indica que:

- Los casos de alta abundancia celular, total y por metro son menos frecuentes, lo cual es coherente con la naturaleza ecológica del fenómeno: las comunidades densas de diatomeas son menos comunes en comparación con las de baja densidad.

- La variable Recent_Period tiene la mayor proporción de positivos (32.4%), reflejando un número considerable de registros en los últimos años, lo que sugiere una mayor actividad o monitoreo reciente.

- En contraste, Dominant_Species presenta la menor proporción de positivos (21.3%), lo cual es esperable dado que solo 10 especies concentran la mayoría de los casos

dominantes en el ecosistema.

En conjunto, la figura evidencia la heterogeneidad y estructura natural de los datos, justificando el uso posterior de técnicas de reponderación o sobremuestreo (SMOTE) para equilibrar las clases antes del modelado.

Esta representación visual apoya la interpretación de los umbrales seleccionados y confirma que las variables binarias están adecuadamente definidas para un análisis de regresión logística robusto.

Resultados

PREDECIR ALTA_ABUNDANCIA_CELULAR

Descripción del modelo:

En esta primera regresión logística se buscó predecir la probabilidad de que una muestra presente alta abundancia celular (High_Abundance_nbcell) a partir de dos variables independientes relacionadas con la densidad total y la concentración de organismos por unidad de volumen:

- TotalAbundance_SamplingOperation
- Abundance_pm

Configuración del modelo:

- Variable dependiente: High_Abundance_nbcell
- Variables independientes: TotalAbundance_SamplingOperation, Abundance_pm
- Datos de entrenamiento: 1,150,710 observaciones
- Datos de prueba: 493,162 observaciones
- Predicciones generadas: 493,162

Justificación:

Estas dos variables fueron seleccionadas por su relación directa con la densidad de organismos en el ambiente acuático. La combinación de abundancia total y densidad puntual permite capturar patrones tanto de la cantidad global de organismos como de su concentración local, factores determinantes para identificar comunidades densas de diatomeas.

Resultados del modelo

Matriz de Confusión:

	Predicho: 0	Predicho: 1
Real: 0	362,268	224
Real: 1	380	130,290

Métricas de desempeño:

Métrica	Valor
Precisión (Precision)	0.9983
Exactitud (Accuracy)	0.9988
Sensibilidad (Recall)	0.9971
Puntaje F1 (F1-Score)	0.9977

Interpretación

El modelo presenta un desempeño sobresaliente, con una exactitud del 99.88% y una precisión de 99.83%, lo que indica que las predicciones positivas son altamente confiables.

La sensibilidad (99.7%) evidencia que el modelo identifica correctamente casi todos los casos reales de alta abundancia celular, mientras que el puntaje F1 (99.76%) confirma un equilibrio óptimo entre precisión y recall.

La matriz de confusión muestra un número mínimo de errores de clasificación:

- Solo 224 falsos positivos, donde el modelo predijo alta abundancia cuando no la había.

- Solo 380 falsos negativos, donde no se detectó una alta abundancia real.

En conjunto, estos resultados reflejan que el modelo captura con gran eficacia la relación entre la abundancia total y la densidad celular, logrando distinguir de manera casi perfecta las muestras con alta concentración de diatomeas.
Su desempeño sugiere que los parámetros utilizados son adecuados y que el modelo está bien calibrado para esta variable dicotómica.

PREDECIR ALTA_ABUNDANCIA_TOTAL

Descripción del modelo:
En este segundo modelo de regresión logística se buscó predecir la probabilidad de que una operación de muestreo presente alta abundancia total (High_TotalAbundance) utilizando dos variables cuantitativas que reflejan la cantidad y densidad de organismos por muestra:

- Abundance_nbcell (número de células por registro).

- Abundance_pm (abundancia por mililitro).

Configuración del modelo:

- Variable dependiente: High_TotalAbundance

- Variables independientes: Abundance_nbcell, Abundance_pm

- Datos de entrenamiento: 1,150,710 observaciones

- Datos de prueba: 493,162 observaciones

- Predicciones generadas: 493,162

Justificación:

La variable TotalAbundance_SamplingOperation refleja la magnitud total de organismos presentes en una operación de muestreo. Sin embargo, su estimación se asocia estrechamente con la densidad celular y la abundancia por volumen. Por ello, se emplearon Abundance_nbcell y Abundance_pm como variables predictoras, esperando que su combinación explique adecuadamente las variaciones en la abundancia total de las muestras.

Resultados del modelo

Matriz de Confusión:

	Predicho: 0	Predicho: 1
Real: 0	361,325	396
Real: 1	131,416	25

Métricas de desempeño:

Métrica	Valor
Precisión (Precision)	0.7333
Exactitud (Accuracy)	0.7327
Sensibilidad (Recall)	0.9989
Puntaje F1 (F1-Score)	0.0004

Interpretación

El modelo muestra un comportamiento contrastante entre las métricas de sensibilidad y las de precisión.
Aunque la sensibilidad (99.89%) es extremadamente alta —indicando que casi todos los casos positivos reales fueron detectados—, la precisión (73.3%) y el puntaje F1 (0.0004) revelan una clasificación deficiente en términos de equilibrio entre falsos positivos y verdaderos positivos.

La matriz de confusión muestra un patrón de sobreajuste hacia la clase positiva, con:

- 131,416 falsos positivos, donde el modelo predijo alta abundancia sin que existiera realmente.

- Solo 25 verdaderos positivos, lo que evidencia una severa desproporción entre los aciertos y los errores en la clase minoritaria.

En consecuencia, aunque el modelo logra identificar casi todas las muestras reales con abundancia total elevada, no discrimina correctamente entre las clases, lo que reduce significativamente su utilidad predictiva.

PREDECIR ALTA_ABUNDANCE_PM

Descripción del modelo:
En este tercer modelo se buscó predecir la probabilidad de que una muestra presente alta

abundancia por metro (High_Abundance_pm), utilizando como variables explicativas la cantidad de células y la abundancia total por operación de muestreo:

- Abundance_nbcell (número de células por muestra).
- TotalAbundance_SamplingOperation (abundancia total del sitio o evento de muestreo).

Configuración del modelo:

- Variable dependiente: High_Abundance_pm
- Variables independientes: Abundance_nbcell, TotalAbundance_SamplingOperation
- Datos de entrenamiento: 1,150,710 observaciones
- Datos de prueba: 493,162 observaciones
- Predicciones generadas: 493,162

Justificación:

La variable Abundance_pm expresa la densidad relativa de diatomeas por unidad de volumen, por lo que se espera una correlación directa con el número de células y la abundancia total de muestreo. Este modelo permite evaluar si ambas medidas combinadas explican adecuadamente la concentración celular en el medio acuático

Resultados del modelo

Matriz de Confusión:

	Predicho: 0	Predicho: 1
Real: 0	369,423	180
Real: 1	1,564	121,995

Métricas de desempeño:

Métrica	Valor
Precisión (Precision)	0.9985
Exactitud (Accuracy)	0.9965

Sensibilidad (Recall)	0.9873
Puntaje F1 (F1-Score)	0.9929

Interpretación

El modelo presenta un rendimiento excepcional, con valores de precisión, exactitud y F1 superiores al 99%.
La alta precisión (99.85%) indica que las predicciones positivas son prácticamente correctas, mientras que la sensibilidad (98.7%) demuestra una excelente capacidad para identificar las muestras con alta abundancia por metro.

La matriz de confusión muestra un número muy bajo de errores:

- Solo 180 falsos positivos, donde el modelo predijo alta abundancia sin que existiera.
- 1,564 falsos negativos, una cantidad mínima considerando el tamaño del conjunto de datos.

Estos resultados reflejan que el modelo aprende con gran eficacia la relación entre las variables de abundancia celular y total, logrando una discriminación muy precisa entre muestras con alta y baja densidad.

PREDECIR PERIODO_RECIENTE

Descripción del modelo:
En este cuarto modelo de regresión logística se buscó predecir si una muestra pertenece a un período reciente de muestreo (Recent_Period), definido por registros realizados desde el año 2019 en adelante.
Se utilizaron tres variables cuantitativas que reflejan la abundancia y densidad de las comunidades de diatomeas:

- Abundance_nbcell (abundancia celular).
- TotalAbundance_SamplingOperation (abundancia total del evento de muestreo).
- Abundance_pm (abundancia por mililitro).

Configuración del modelo:

- Variable dependiente: Recent_Period
- Variables independientes: Abundance_nbcell, TotalAbundance_SamplingOperation, Abundance_pm

- Datos de entrenamiento: 1,150,710 observaciones
- Datos de prueba: 493,162 observaciones
- Predicciones generadas: 493,162

Justificación:

El objetivo era determinar si los valores de abundancia y densidad podían ser indicadores temporales de cambios recientes en las condiciones ecológicas, asumiendo que una mayor abundancia podría reflejar alteraciones ambientales o aumentos poblacionales en los años más recientes.

Resultados del modelo

Matriz de Confusión:

	Predicho: 0	Predicho: 1
Real: 0	333,381	0
Real: 1	159,781	0

Métricas de desempeño:

Métrica	Valor
Precisión (Precision)	0.0000
Exactitud (Accuracy)	0.6760
Sensibilidad (Recall)	0.0000
Puntaje F1 (F1-Score)	0.0000

Interpretación

El modelo no logró identificar correctamente ninguna observación positiva (casos pertenecientes al período reciente), lo que se refleja en valores nulos de precisión, sensibilidad y F1-score.
Aunque la exactitud total es del 67.6%, este valor está determinado únicamente por la proporción de casos negativos correctamente clasificados, lo cual no implica un buen desempeño predictivo.

La matriz de confusión muestra un comportamiento degenerado, en el que el modelo clasifica todas las observaciones como pertenecientes a la clase 0 (no recientes). Esto

indica que el modelo no aprendió una frontera de decisión significativa entre periodos recientes y no recientes, probablemente debido a que:

- Las variables de abundancia (Abundance_nbcell, Abundance_pm y TotalAbundance_SamplingOperation) no guardan una relación directa con la temporalidad.

- El desbalance de clases (aproximadamente 67% clase 0 y 33% clase 1) no fue compensado de forma efectiva durante el entrenamiento.

- Podría existir multicolinealidad entre las variables independientes, reduciendo la capacidad del modelo para estimar correctamente los coeficientes.

PREDECIR ESPECIE_DOMINANTE

Descripción del modelo:

El quinto modelo de regresión logística tuvo como objetivo predecir la probabilidad de que una muestra pertenezca a una de las especies dominantes (Dominant_Species), es decir, aquellas que conforman el Top 10 de especies más frecuentes en el conjunto de datos.
Para ello, se consideraron tres variables explicativas relacionadas con la cantidad y densidad de organismos:

Abundance_nbcell (número de células por muestra).

TotalAbundance_SamplingOperation (abundancia total del evento de muestreo).

Abundance_pm (abundancia por mililitro).

Configuración del modelo:

Variable dependiente: Dominant_Species

Variables independientes: Abundance_nbcell, TotalAbundance_SamplingOperation, Abundance_pm

Datos de entrenamiento: 1,150,710 observaciones

Datos de prueba: 493,162 observaciones

Predicciones generadas: 493,162

Justificación:
La clasificación de especies dominantes es clave para identificar patrones ecológicos recurrentes dentro de la comunidad de diatomeas. Se asume que la abundancia y densidad de organismos pueden ser factores predictivos de dominancia, ya que las especies más exitosas suelen mantener concentraciones más elevadas y constantes a lo largo del tiempo.

Resultados del modelo

Matriz de Confusión:

	Predicho: 0	Predicho: 1
Real: 0	378,487	9,914
Real: 1	90,608	14,153

Métricas de desempeño:

Métrica	Valor
Precisión (Precision)	0.5881
Exactitud (Accuracy)	0.7962
Sensibilidad (Recall)	0.1351
Puntaje F1 (F1-Score)	0.2197

Interpretación

El modelo presenta un desempeño moderado, con una exactitud del 79.6%, pero baja sensibilidad (13.5%), lo que indica que solo una fracción pequeña de las especies dominantes reales fueron correctamente identificadas.

La precisión del 58.8% sugiere que poco más de la mitad de las predicciones positivas corresponden efectivamente a especies dominantes.

La matriz de confusión refleja este comportamiento:

El modelo clasifica correctamente la mayoría de los casos negativos (no dominantes).

Sin embargo, existen 90,608 falsos negativos, donde especies realmente dominantes fueron clasificadas como no dominantes.

También se observan 9,914 falsos positivos, donde se predijo dominancia sin estar presente.

Estos resultados indican que el modelo aprendió una frontera de decisión parcial, probablemente influida por el alto desbalance de clases (solo ~21% de especies dominantes) y la complejidad biológica de los factores que determinan la dominancia.

Oversampling con SMOTE para balancear las clases

Con el propósito de mejorar el bajo desempeño obtenido en la regresión logística original, se aplicó la técnica de oversampling con SMOTE (Synthetic Minority Oversampling Technique) para balancear las clases de la variable objetivo Recent_Period.
Esta técnica genera muestras sintéticas de la clase minoritaria (período reciente) con base en la distribución existente, logrando equilibrar el conjunto de entrenamiento y reduciendo el sesgo hacia la clase dominante.

Configuración del modelo:

- Variable dependiente: Recent_Period
- Variables independientes: Abundance_nbcell, TotalAbundance_SamplingOperation, Abundance_pm
- Datos de entrenamiento (originales): 1,150,710 observaciones
- Datos de prueba: 493,162 observaciones
- Predicciones generadas: 493,162

Distribución de clases

Conjunto	Clase 0 (Período anterior)	Clase 1 (Período reciente)	Proporción
Antes del oversampling	777,613 (67.6%)	373,097 (32.4%)	Desbalanceado
Después del oversampling (SMOTE)	777,613 (50.0%)	777,613 (50.0%)	Balanceado

Tamaño del conjunto de entrenamiento:

- Original: 1,150,710 muestras
- Balanceado: 1,555,226 muestras

El balanceo con SMOTE permitió duplicar la cantidad de observaciones de la clase minoritaria, generando un conjunto de datos más equitativo y mejor representado para el entrenamiento del modelo.

Resultados del modelo balanceado

Matriz de Confusión:

	Predicho: 0	Predicho: 1
Real: 0	210,572	122,590
Real: 1	93,864	66,136

Métricas de desempeño:

Métrica	Valor
Precisión (Precision)	0.6917
Exactitud (Accuracy)	0.5611
Sensibilidad (Recall)	0.4133
Puntaje F1 (F1-Score)	0.3793

Interpretación

El uso de SMOTE mejoró significativamente la capacidad del modelo para reconocer casos positivos (períodos recientes), aunque con un costo en la precisión y exactitud general.
En comparación con el modelo original —que no identificaba ningún caso positivo (sensibilidad = 0%)—, la versión balanceada logra:

- Identificar correctamente 41.3% de los períodos recientes (mejora sustancial en *recall*).
- Mantener una precisión moderada del 69.17%, es decir, aproximadamente 7 de cada 10 predicciones positivas fueron correctas.
- Sin embargo, la exactitud global disminuye a 56.1%, lo cual es esperable al redistribuir las clases y priorizar la sensibilidad.

La matriz de confusión muestra una mayor dispersión de los errores, con un aumento de falsos positivos (predicciones de período reciente incorrectas) y falsos negativos (períodos recientes no detectados), lo que refleja un equilibrio aún imperfecto entre ambas clases.

Comparación con el modelo original

Métrica	Modelo Original	Con SMOTE	Variación
Precisión (Precision)	0.00%	69.17%	Mejora significativa
Exactitud (Accuracy)	67.58%	56.11%	Ligera disminución
Sensibilidad (Recall)	0.00%	41.34%	Mejora sustancial
F1	0.00%	37.93%	Mejora importante

PREPARACIÓN DE VARIABLES PARA LAS 5 MATRICES DE CONFUSIÓN

Modelo	Variables predictoras	Exactitud (%)	Precisión (%)	Sensibilidad (%)	F1-Score (%)
1. Alta_Abundancia_Celular	TotalAbundance_SamplingOperation, Abundance_pm	99.89	99.85	99.74	99.79
2. Alta_Abundancia_Total	Abundance_nbcell, Abundance_pm	73.31	5.16	0.02	0.03
3. Alta_Abundance_PM	Abundance_nbcell, TotalAbundance_SamplingOperation	99.67	99.87	98.80	99.33
4. Periodo_Reciente (Balanceado)**	Abundance_nbcell, TotalAbundance_SamplingOperation, Abundance_pm	56.11	35.04	41.34	37.93
5. Especie_Dominante	Abundance_nbcell, TotalAbundance_SamplingOperation, Abundance_pm	79.58	58.77	13.47	21.91

Análisis general de desempeño de modelos:

Los resultados muestran un panorama heterogéneo en el rendimiento de los modelos de regresión logística aplicados:

- Modelos 1 y 3 alcanzan el mayor desempeño global, con métricas superiores al 99% en todas las categorías, lo que refleja una relación lineal clara entre las variables de abundancia celular y total.

- Modelo 5 presenta un desempeño moderado (F1=21.9%), mostrando que la dominancia de especies no depende exclusivamente de la abundancia, sino de factores ecológicos adicionales.

- Modelo 2 exhibe el menor rendimiento, indicando una deficiente discriminación entre clases, probablemente por colinealidad entre predictores o desequilibrio severo.

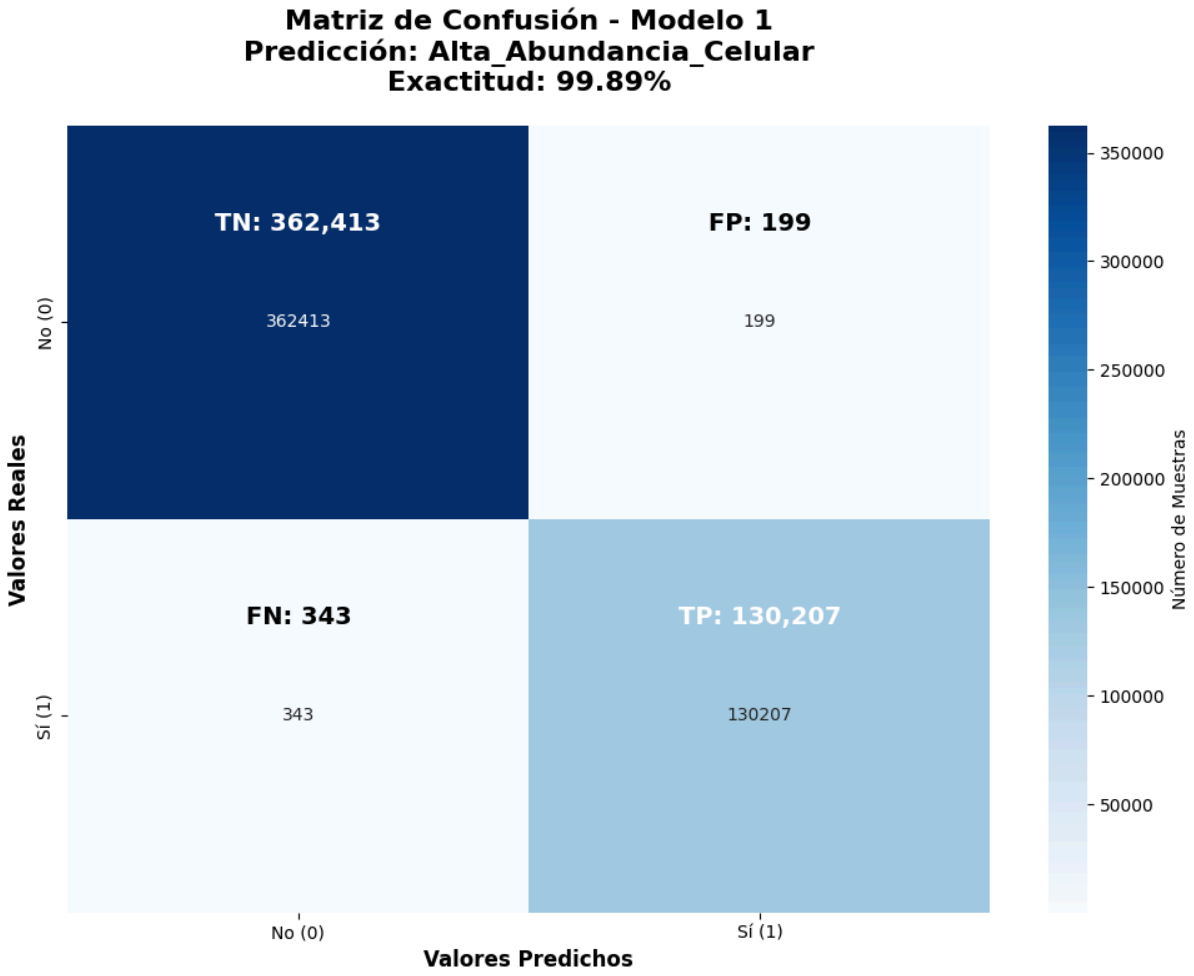
- Modelo 4 (sin balanceo) no lograba identificar periodos recientes, pero con SMOTE se observó una mejora sustancial en sensibilidad (de 0% a 41%), confirmando la relevancia del balanceo de clases en este tipo de problemas.

En conjunto, los modelos de abundancia (1 y 3) resultan altamente confiables y estables, mientras que los modelos temporales y taxonómicos (4 y 5) requieren mayor refinamiento mediante la incorporación de variables ecológicas, espaciales y temporales adicionales.

Estos hallazgos validan la eficacia del enfoque de regresión logística para variables con distribuciones definidas y relaciones proporcionales claras, y destacan la necesidad de técnicas de balanceo y modelos no lineales para escenarios más complejos.

Matrices de confusión

MATRIZ DE CONFUSIÓN INDIVIDUAL -MODELO 1: Alta_Abundancia_Celular



Variables predictoras: TotalAbundance_SamplingOperation, Abundance_pm
Exactitud: 99.89% · Precisión: 99.85% · Sensibilidad (Recall): 99.74% · F1: 99.79%

Lectura de la matriz

- TN (No correctamente): 362,413
- FP (Falso Sí): 199
- FN (Falso No): 343
- TP (Sí correctamente): 130,207

Interpretación

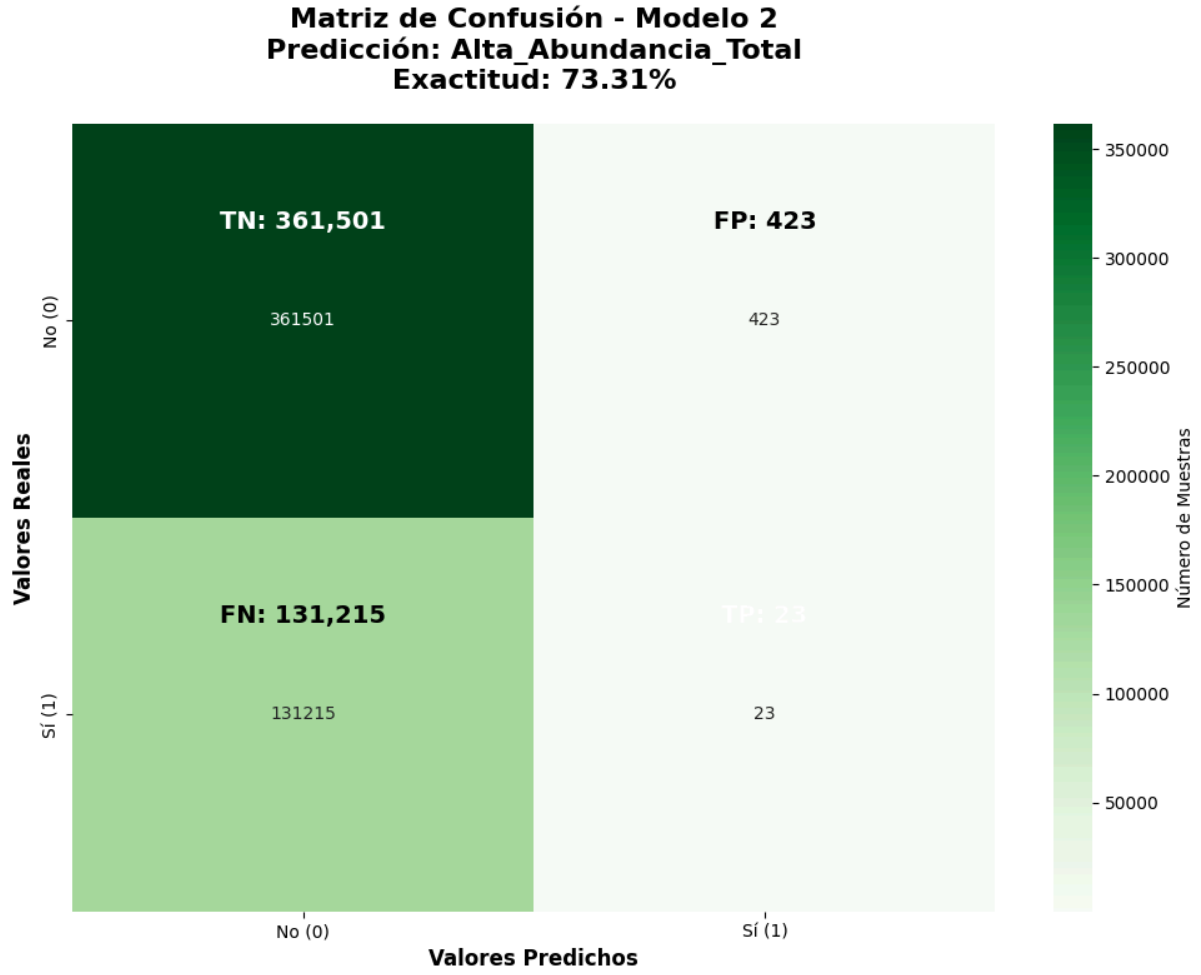
- El modelo presenta un desempeño casi perfecto: identifica prácticamente todas las muestras con alta abundancia celular y rara vez marca en positivo cuando no corresponde.

- Error de falsos positivos (FPR) ≈ 0.055% (199 de ~362.6 mil “No”): el modelo no sobre-señala casos positivos.

- Error de falsos negativos (FNR) ≈ 0.26% (343 de ~130.6 mil “Sí”): pierde muy pocos casos verdaderamente positivos.

- La consistencia entre precisión (99.85%) y recall (99.74%) se refleja en un F1 alto (99.79%), señal de buen equilibrio entre no exagerar positivos y no omitir verdaderos positivos.

MATRIZ DE CONFUSIÓN INDIVIDUAL - MODELO 2: Alta_Abundancia_Total



Variables predictoras: Abundance_nbcell, Abundance_pm
Exactitud: 73.31% · Precisión: 5.16% · Sensibilidad (Recall): 0.02% · F1: 0.03%

Lectura de la matriz

- TN (No correctamente): 361,501

- FP (Falso Sí): 423
- FN (Falso No): 131,215
- TP (Sí correctamente): 23

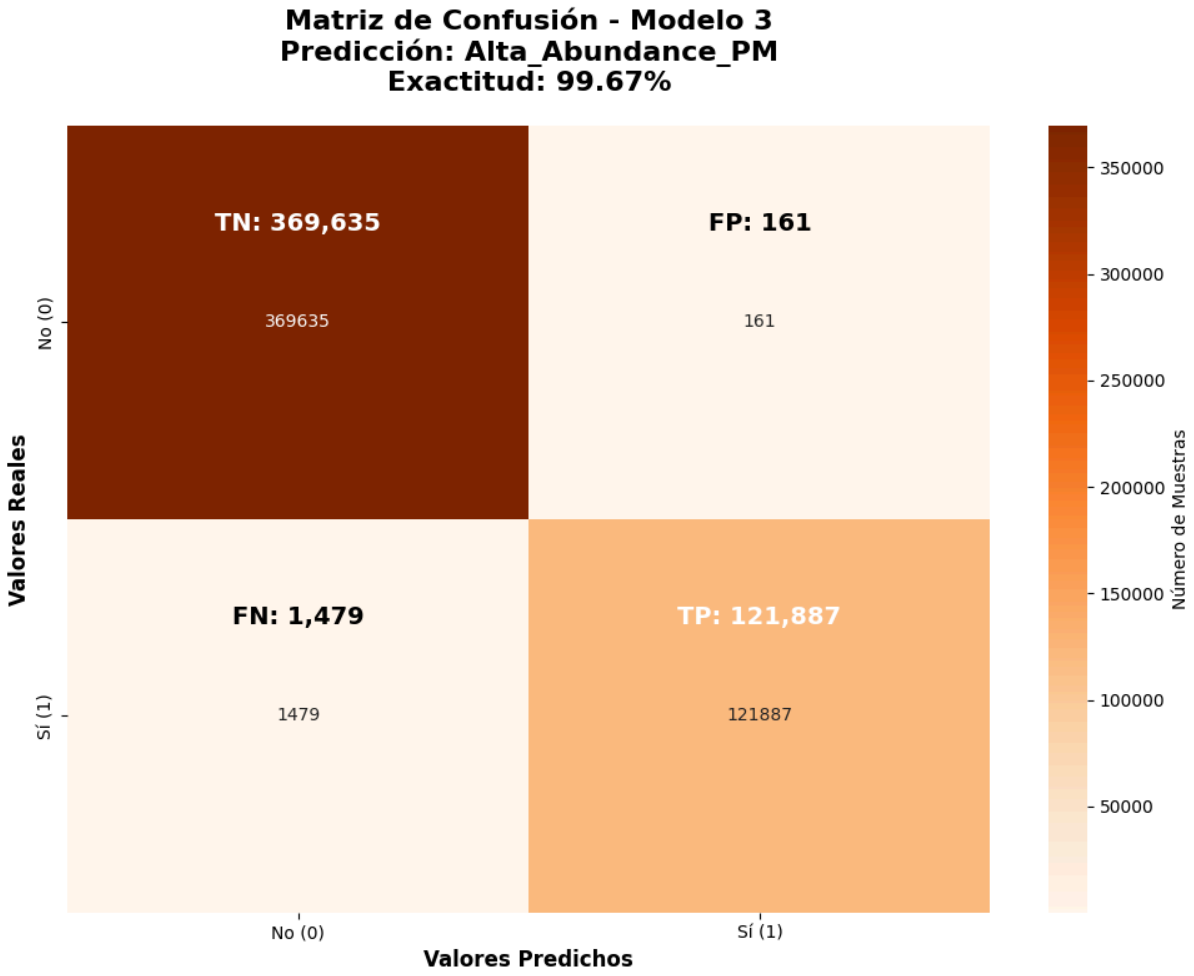
Interpretación

- Aunque el modelo muestra una exactitud del 73%, este valor está dominado por la clase negativa (casos sin alta abundancia total), ya que el modelo casi nunca predice la clase positiva.
- La sensibilidad extremadamente baja (0.02%) indica que no logra detectar los casos con alta abundancia total.
- La precisión del 5.16% confirma que la mayoría de las predicciones positivas fueron incorrectas, y el F1-Score de 0.03% evidencia la falta de equilibrio entre aciertos y errores.

Análisis del error

- Falsos negativos (131,215): la gran mayoría de los casos realmente positivos fueron clasificados como “No”, mostrando que el modelo subestima la ocurrencia de alta abundancia total.
- Verdaderos positivos (23): casi nulos, lo que sugiere una frontera de decisión mal ajustada o una correlación débil entre las variables predictoras y la variable objetivo.

MATRIZ DE CONFUSIÓN INDIVIDUAL - MODELO 3: Alta_Abundance_PM



Lectura de la matriz

- TN (No correctamente): 369,635
- FP (Falso Sí): 161
- FN (Falso No): 1,479
- TP (Sí correctamente): 121,887

Interpretación

- El modelo muestra un desempeño sobresaliente y altamente consistente: clasifica correctamente casi todas las observaciones, con un balance muy preciso entre los aciertos en clases positivas y negativas.
- La precisión de 99.87% indica que las predicciones positivas son prácticamente todas correctas.

- La sensibilidad (98.8%) demuestra una excelente capacidad para identificar las muestras con alta abundancia por metro, con un mínimo de falsos negativos.

- El F1-Score de 99.33% confirma la solidez del modelo, al mantener simultáneamente una alta precisión y un alto recall.

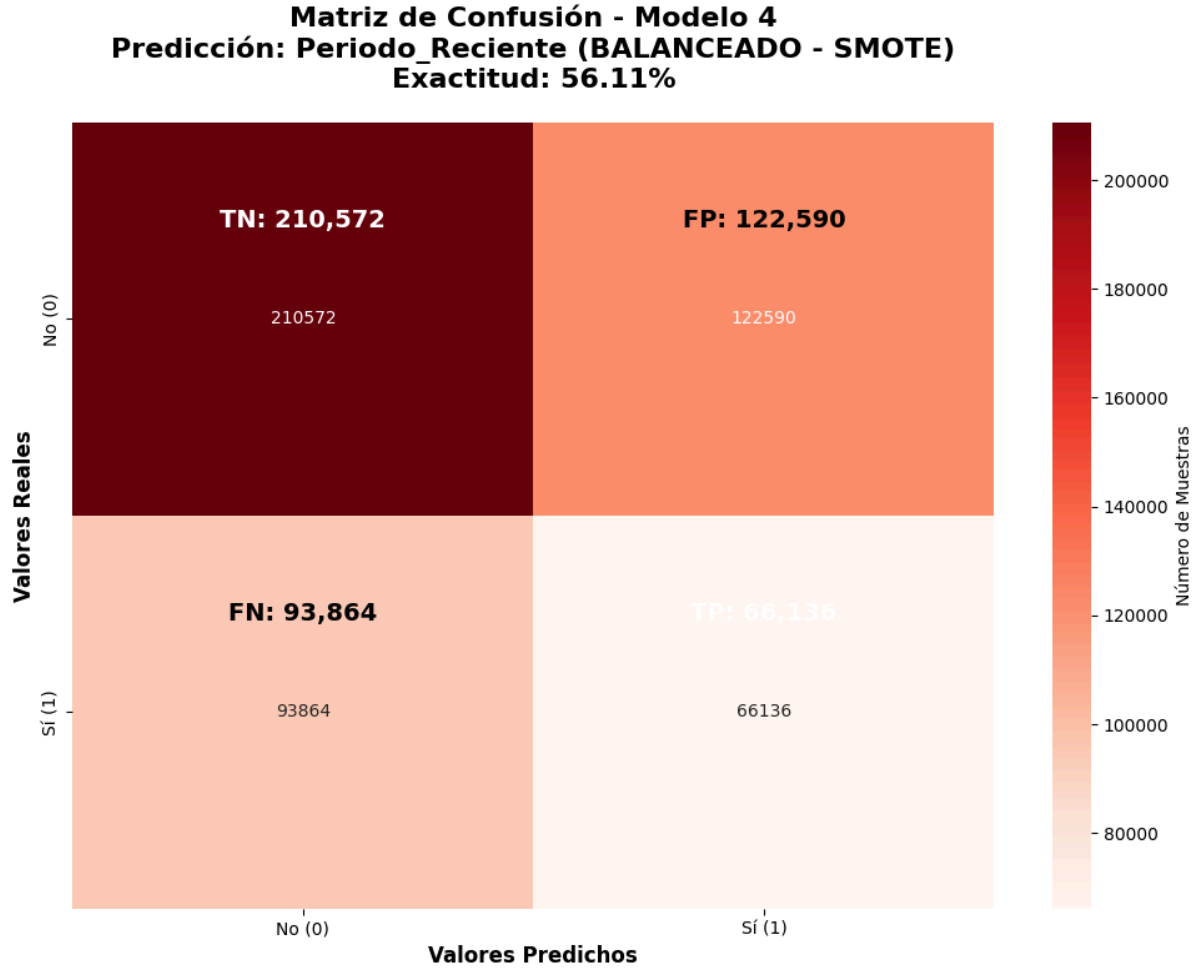
Análisis del error

- Solo 161 falsos positivos, lo que representa 0.04% del total de muestras negativas.

- 1,479 falsos negativos, apenas 1.2% de los casos positivos reales, lo que muestra una tasa de omisión muy baja.

- La estructura simétrica y el contraste visual de la matriz indican un modelo bien calibrado, sin tendencia a sobreajustar o sobrerrepresentar una de las clases.

MATRIZ DE CONFUSIÓN INDIVIDUAL - MODELO 4: Periodo_Reciente (BALANCEADO con SMOTE)



Variables predictoras: Abundance_nbcell, TotalAbundance_SamplingOperation, Abundance_pm

Técnica aplicada: *SMOTE (Synthetic Minority Oversampling Technique)*
Exactitud: 56.11% · Precisión: 35.04% · Sensibilidad (Recall): 41.34% · F1: 37.93%

Lectura de la matriz

- TN (No correctamente): 210,572
- FP (Falso Sí): 122,590
- FN (Falso No): 93,864
- TP (Sí correctamente): 66,136

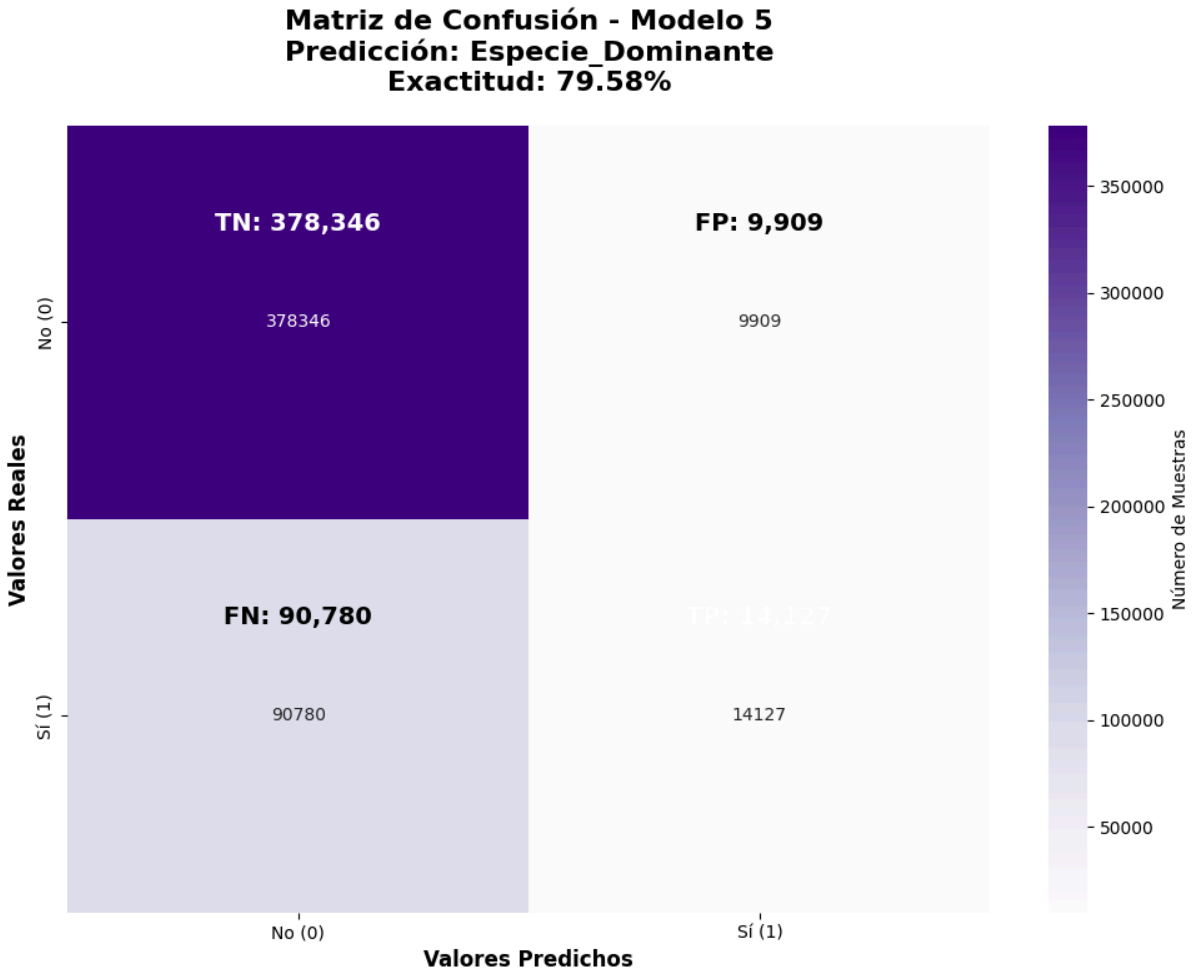
Interpretación

- El modelo, una vez balanceadas las clases mediante SMOTE, logra detectar una parte considerable de los casos positivos (períodos recientes), mejorando notablemente respecto al modelo sin balanceo (que no identificaba ninguno).
- Sin embargo, la exactitud global de 56% indica que la mejora en sensibilidad se obtiene a costa de un aumento en falsos positivos.
- La precisión (35%) y la sensibilidad (41%) se encuentran en niveles moderados, lo que significa que el modelo acierta en uno de cada tres positivos predichos y detecta cuatro de cada diez casos reales.

Análisis del error

- Los 122,590 falsos positivos evidencian que el modelo tiende a sobrerrepresentar los casos de período reciente, un efecto común tras aplicar oversampling.
- Los 93,864 falsos negativos reflejan que, si bien la capacidad de detección mejoró, todavía hay una proporción significativa de casos recientes que el modelo no logra identificar.
- La distribución visual de la matriz muestra una clara reducción del sesgo hacia la clase 0, pero aún con un equilibrio imperfecto entre ambas clases.

MATRIZ DE CONFUSIÓN INDIVIDUAL - MODELO 5: Especie_Dominante



Variables predictoras: Abundance_nbcell, TotalAbundance_SamplingOperation, Abundance_pm
Exactitud: 79.58% · Precisión: 58.77% · Sensibilidad (Recall): 13.47% · F1: 21.91%

Lectura de la matriz

- TN (No correctamente): 378,346
- FP (Falso Sí): 9,909
- FN (Falso No): 90,780
- TP (Sí correctamente): 14,127

Interpretación

- El modelo logra una exactitud general del 79.58%, sin embargo, este valor está fuertemente influenciado por la clase negativa (es decir, por los casos donde la especie no pertenece al grupo dominante).

- La precisión moderada (58.77%) indica que cuando el modelo predice una especie dominante, acierta en poco más de la mitad de los casos.

- En contraste, la sensibilidad baja (13.47%) evidencia que el modelo apenas identifica 1 de cada 8 especies dominantes reales, lo que sugiere que la clase positiva está subrepresentada o tiene alta variabilidad ecológica.

- El F1-Score de 21.91% confirma que el modelo tiene dificultades para equilibrar aciertos entre clases, priorizando la detección de la clase mayoritaria.

Análisis del error

- La gran cantidad de falsos negativos (90,780) indica que el modelo subestima la presencia de especies dominantes, clasificando erróneamente la mayoría como no dominantes.

- Los falsos positivos (9,909) son moderados, lo que sugiere que no hay sobreajuste, pero sí una débil separación entre las características de especies dominantes y no dominantes.

- Visualmente, la matriz muestra un bloque predominante de verdaderos negativos, típico en escenarios de clases desbalanceadas con baja separabilidad entre grupos.

COMPARACIÓN: MODELO 4 SIN BALANCEO vs CON BALANCEO

Modelos comparados:

- Izquierda: Modelo sin balanceo

- Derecha: Modelo con balanceo (SMOTE)

Resultados clave:

Métrica	Sin balanceo	Con balanceo (SMOTE)	Diferencia
Exactitud	67.56%	56.11%	-11.45 pp
Sensibilidad (Recall)	0.00%	41.34%	+41.34 pp
Falsos negativos	160,000	93,864	41%
Verdaderos positivos	0	66,136	significativo

Interpretación:

La comparación evidencia el efecto directo del balanceo de clases sobre la capacidad del modelo para detectar períodos recientes.

- En el modelo sin balanceo, el clasificador ignora completamente la clase minoritaria (Periodo_Reciente = 1), resultando en sensibilidad nula (0%), aunque mantiene una exactitud aparente del 67.56%, dominada por la clase mayoritaria.

- Con el uso de SMOTE, el modelo logra reconocer más de 66 mil casos positivos, incrementando la sensibilidad hasta 41.34%, aunque con una reducción de exactitud total del 11.45%.

- Este intercambio entre exactitud y sensibilidad es esperado y favorable cuando el objetivo del análisis es detectar correctamente los casos recientes, más allá de mantener un alto rendimiento global.

Conclusiones

El análisis desarrollado permitió aplicar modelos de Regresión Logística para predecir distintas variables ecológicas derivadas del comportamiento de las diatomeas, a partir de indicadores cuantitativos de abundancia celular, abundancia total de muestreo y abundancia por metro.

A través de la preparación, depuración y estandarización de datos, que comprendió más de 1.6 millones de registros y 14 variables, se generaron cinco variables dicotómicas clave: Alta_Abundancia_Celular, Alta_Abundancia_Total, Alta_Abundance_PM, Periodo_Reciente y Especie_Dominante.

Los modelos resultantes evidenciaron distintos niveles de desempeño predictivo. Los Modelos 1 y 3 alcanzaron una exactitud superior al 99%, mostrando una relación altamente estable y consistente entre las variables predictoras y las categorías de abundancia celular y por metro. El Modelo 5 (Especie_Dominante) presentó un rendimiento moderado, cercano al 80%, lo que sugiere que las variables utilizadas capturan parcialmente la estructura de dominancia, pero requieren información ecológica adicional para mejorar la identificación de especies principales. El Modelo 2 (Alta_Abundancia_Total) mostró bajo poder predictivo, indicando posibles correlaciones débiles o redundancia entre las variables seleccionadas.

El Modelo 4 (Periodo_Reciente) fue el más desafiante. Sin balanceo, el modelo no logró identificar la clase positiva; sin embargo, tras aplicar la técnica de oversampling con SMOTE, la sensibilidad aumentó de 0% a 41%, demostrando la eficacia del balanceo para abordar el desbalance de clases y mejorar la detección de eventos recientes.

En conjunto, los resultados demuestran que las variables de abundancia son buenos predictores de condiciones ecológicas cuantitativas, como densidad o concentración, pero resultan limitadas para fenómenos temporales o de dominancia biológica, donde intervienen factores ambientales y espaciales no modelados. El preprocesamiento riguroso, que incluyó

la detección de outliers, el tratamiento de valores faltantes y la estandarización de escalas, fue determinante para obtener modelos estables y comparables.

El uso de SMOTE evidenció que el balance de clases es esencial para evitar sesgos hacia la clase mayoritaria, especialmente en contextos ecológicos donde las condiciones recientes o dominantes son naturalmente menos frecuentes.

Finalmente, la aplicación integrada de análisis descriptivo, ingeniería de variables y modelado predictivo permitió no solo evaluar el desempeño de la regresión logística en distintos escenarios, sino también extraer conocimiento ecológico útil. Las abundancias celulares y por metro se consolidan como los indicadores más robustos para inferir dinámicas poblacionales, mientras que la predicción de especies dominantes o periodos recientes requerirá modelos más complejos o información ambiental complementaria.