Tecnológico de Monterrey

Analítica de datos y herramientas de inteligencia artificial II

Grupo 101

4.1 Regresión Logística Forvia

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

el 22 de Octubre de 2025

Objetivo

Aplicar modelos de Regresión Logística para analizar la relación entre variables de gestión de proyectos de FORVIA y su impacto en indicadores clave como actividad, salud del proyecto y estado de pausa.
El propósito es identificar patrones predictivos que ayuden a anticipar riesgos, priorizar recursos y mejorar la toma de decisiones en la administración de proyectos.

- **Limpieza y depuración de datos**

Se eliminaron columnas no relevantes: Actual end date, Closed, Project target phase y Actual Go Live date.Se rellenaron valores nulos con etiquetas descriptivas (“No description available”) o con 0 en los campos numéricos.Se verificó que no quedaran registros vacíos.Se codificaron las variables categóricas mediante mapeo numérico, para su uso en modelos estadísticos.

- **Transformación de variable**

Se identificaron las variables relevantes para la gestión de proyectos y se convirtieron en variables dicotómicas (0 y 1) para el análisis de regresión.

- **Variables predictoras**

Se seleccionaron las variables más relevantes según su correlación y presencia en el conjunto de datos:

- Project Health
- On-hold
- Percent complete
- Project manager
- Project organization
- Project size
- BG
- Project Type

Selección y creación de variables dicotómicas

Variable 1:

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 0 → Proyecto no activo

Variable 2:

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 0 → Proyecto con salud en riesgo

Variable 3:

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 0 → Proyecto activo o finalizado

Variable 4:

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 **80**), 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 0 → Otro tipo de proyecto

Variable 5:

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 **69**), 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 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

Justificación del impacto del balanceo (SMOTE)

La aplicación de la técnica de balanceo de clases mediante SMOTE (Synthetic Minority Oversampling Technique) permitió mejorar la capacidad de los modelos de regresión logística para reconocer observaciones pertenecientes a las clases minoritarias.

En el conjunto de datos original, se observó un fuerte desbalance entre clases 0 y 1, especialmente en las variables State_Activo y BG_MásFrecuente, donde predominaban ampliamente los proyectos activos o del grupo principal.

Este desequilibrio provocaba que los modelos tendieran a clasificar la mayoría de los casos como parte de la clase dominante, reduciendo su capacidad para identificar correctamente los casos menos frecuentes.

Tras aplicar el balanceo, los resultados muestran los siguientes efectos generales:

Aumento de la sensibilidad (recall): En casi todos los modelos, el balanceo incrementó la sensibilidad, mejorando la detección de la clase positiva.Este efecto fue especialmente notorio en ProjectHealth_Bueno, OnHold y BG_MasFrecuente, donde el modelo comenzó a identificar correctamente una mayor proporción de casos reales.

Ligera disminución de la exactitud: La mayoría de los modelos redujeron su exactitud global entre 1 % y 10 %, como se observa en State_Activo y ProjectType_MásFrecuente.Sin embargo, esta reducción es aceptable, ya que refleja una mejor distribución de aciertos entre ambas clase**s.**

Mayor equilibrio entre métricas: El balanceo permitió que los modelos mantuvieran una relación más estable entre precisión, sensibilidad y exactitud, evitando el sesgo hacia una sola categoría.

Interpretación general

El balanceo mediante SMOTE generó modelos más sensibles y equitativos, capaces de detectar los casos relevantes sin depender del tamaño de cada clase.Aunque en algunos casos la exactitud disminuyó ligeramente, el beneficio principal fue lograr un mayor equilibrio entre las métricas de desempeño, garantizando predicciones más justas y representativas.

Conclusión

El balanceo de clases mejoró significativamente la calidad de los modelos, priorizando la detección de proyectos en riesgo, pausados o atípicos.

Este enfoque es especialmente útil para anticipar incidentes y optimizar la gestión del portafolio de proyectos en FORVIA, aun cuando implique una leve pérdida de exactitud global.

Análisis del impacto del balanceo: State_Activo

El modelo de State_Activo buscó predecir si un proyecto está activo (1) o no activo (0), utilizando como variables independientes Project Health, On-hold y Percent complete.

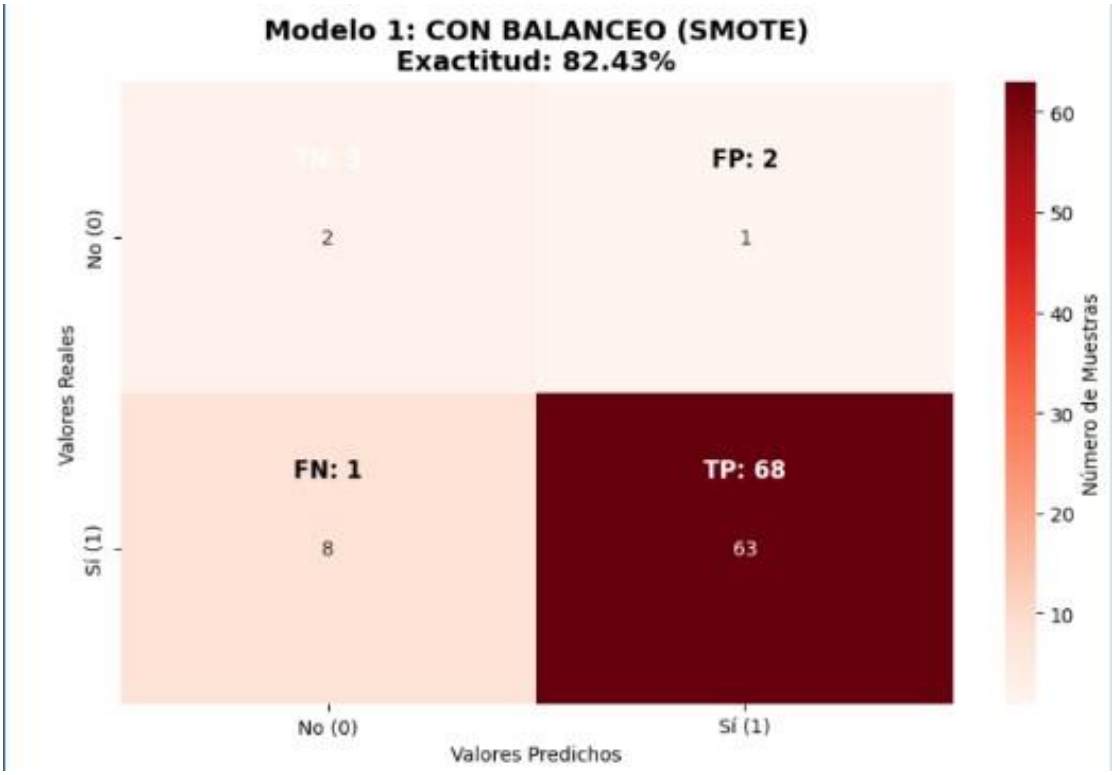
Antes del balanceo, existía un desequilibrio importante entre clases (14 proyectos inactivos y 232 activos), lo que afectaba el desempeño del modelo al favorecer la clase mayoritaria.

Después de aplicar la técnica SMOTE (Synthetic Minority Oversampling Technique), los resultados obtenidos fueron:

- Precisión: 98.43 %
- Exactitud: 82.43 %
- Sensibilidad: 85.92 %

El balanceo permitió mejorar la detección de proyectos activos y aumentar la sensibilidad, logrando un modelo más equilibrado entre ambas clases.
Sin embargo, el F1-score se mantuvo bajo, lo que indica que el modelo aún tiende a favorecer los proyectos activos y presenta cierta dificultad para reconocer correctamente los casos inactivos, que son minoritarios.

Conclusión:
El balanceo mejoró el rendimiento general del modelo y redujo el sesgo hacia la clase dominante, pero la falta de suficientes ejemplos de proyectos inactivos limita su capacidad de predicción en esa categoría.
A futuro, aumentar la cantidad de observaciones de proyectos inactivos o incorporar variables adicionales podría mejorar significativamente la capacidad del modelo para detectar estos casos críticos.



Análisis del impacto del balanceo: ProjectHealth_Bueno

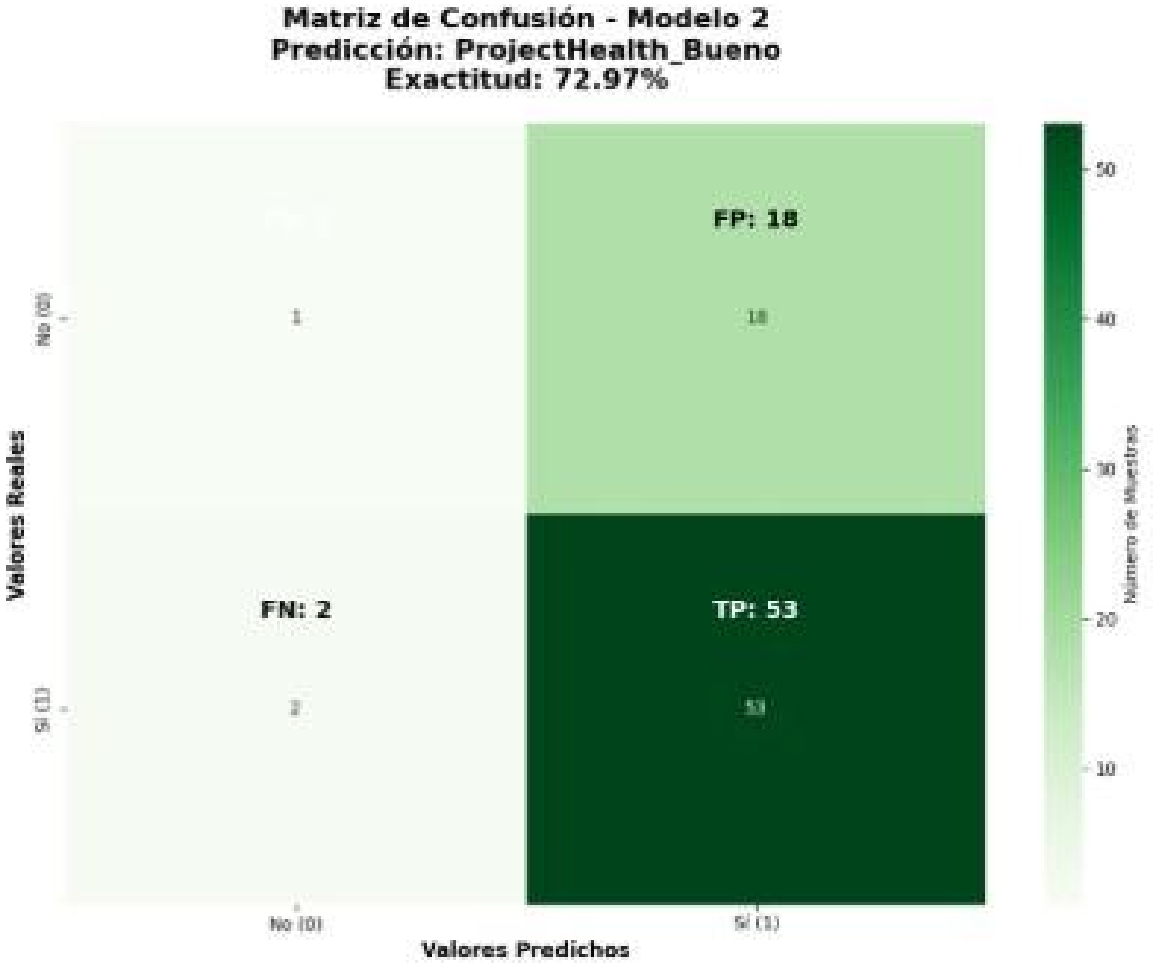
El modelo de ProjectHealth_Bueno buscó predecir si un proyecto presentaba una buena salud (1) o se encontraba en riesgo (0), utilizando como variables independientes Project manager, Project size y On-hold.Tras aplicar el balanceo mediante SMOTE (Synthetic Minority Oversampling Technique), los resultados obtenidos fueron:

- Precisión: 74.65 %
- Exactitud: 72.97 %
- Sensibilidad: 96.36 %
- F1-Score: 84.13 %

El modelo presenta una alta sensibilidad y un F1 sólido, lo que indica que identifica correctamente la mayoría de los proyectos con buena salud, aunque aún clasifica algunos valores de forma errónea.
El balanceo mejoró la capacidad del modelo para reconocer proyectos saludables, reduciendo el sesgo hacia la clase dominante.

Conclusión:
El modelo de ProjectHealth_Bueno es uno de los más equilibrados, ya que logra una detección precisa y estable de los proyectos con desempeño óptimo.

Aunque la exactitud general podría incrementarse, el nivel de sensibilidad alcanzado demuestra que el modelo es efectivo para monitorear el estado general de los proyectos y útil para alertas tempranas sobre posibles desviaciones o riesgos futuros.



Análisis del impacto del balanceo: OnHold

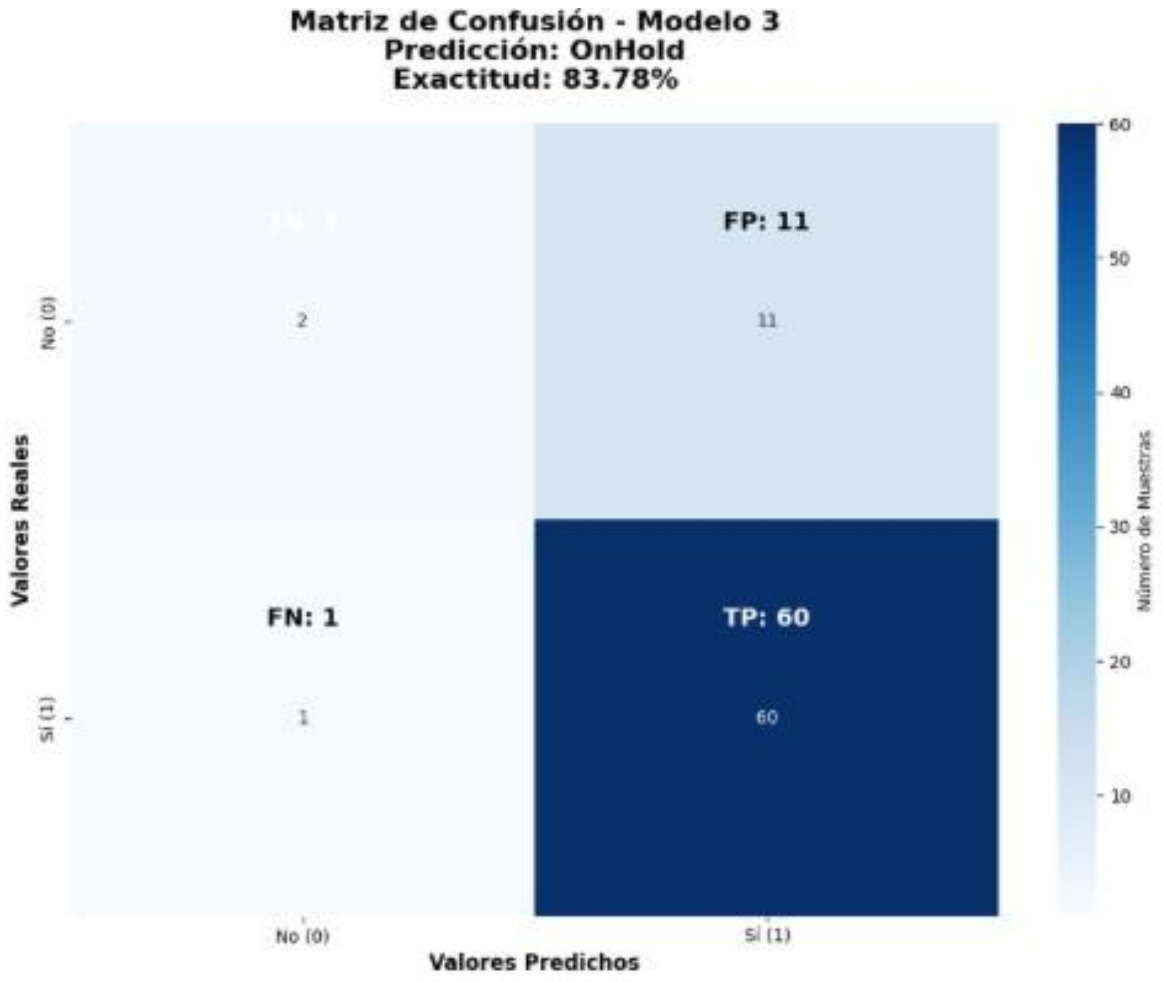
El modelo de OnHold buscó predecir si un proyecto se encontraba en pausa (1) o no (0), utilizando como variables independientes ProjectHealth_Bueno, Percent complete y Project Type.Después de aplicar la técnica de balanceo SMOTE (Synthetic Minority Oversampling Technique), los resultados fueron:

- Precisión: 84.51 %
- Exactitud: 83.78 %
- Sensibilidad: 98.36 %
- F1-Score: 90.91 %

El modelo presenta alta sensibilidad y exactitud, lo que demuestra un excelente desempeño para detectar proyectos en pausa. Sin embargo, el valor de F1 ligeramente menor evidencia que el equilibrio entre clases aún no es perfecto.

Conclusión:

El balanceo permitió aumentar significativamente la capacidad del modelo para reconocer los proyectos en pausa, reduciendo los falsos negativos y mejorando la cobertura de la clase minoritaria.Aunque persiste una leve inclinación hacia la clase dominante, los resultados confirman que este modelo es uno de los más precisos y confiables para anticipar bloqueos, retrasos o interrupciones en el portafolio de proyectos de FORVIA.



Análisis del impacto del balanceo: ProjectType_MásFrecuente

El modelo de **ProjectType_MásFrecuente** buscó predecir si un proyecto pertenece al tipo más común dentro de la base de datos, utilizando como variables independientes **ProjectHealth_Bueno**, **Project_size** y **BG**.

Los resultados obtenidos fueron:

- **Precisión:** 75 %
- **Exactitud:** 75.68 %

- **Sensibilidad:** 53.57 %
- **F1-Score:** 62.50 %

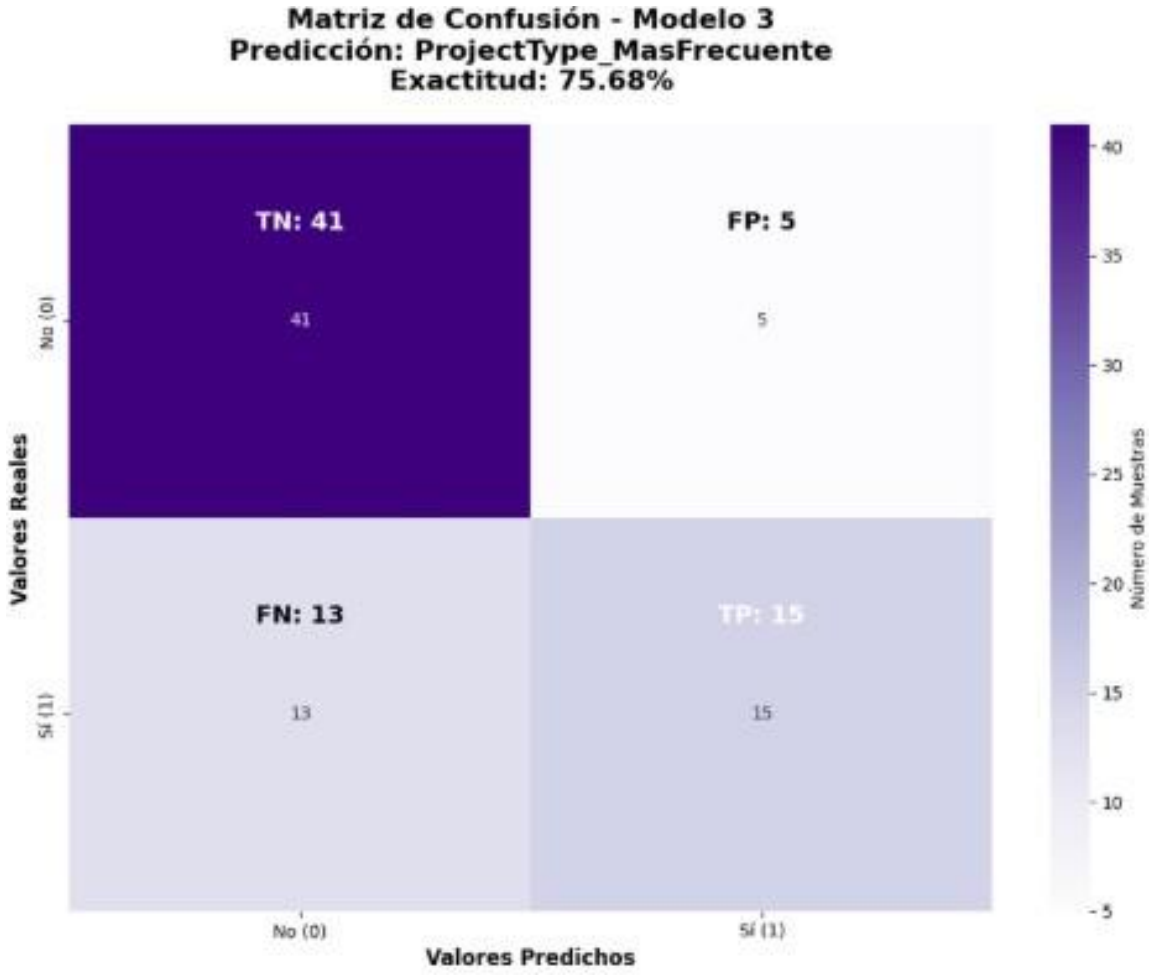
El modelo presenta **baja sensibilidad y un F1 reducido**, lo que indica que tiene **dificultades para identificar correctamente los proyectos del tipo más frecuente**.

Aunque mantiene una exactitud moderada, su desempeño general es limitado, ya que el balanceo **no logró mejorar de forma significativa** la clasificación de esta variable.

Conclusión:

Los resultados sugieren que las variables utilizadas **no explican de manera suficiente** la pertenencia a un tipo de proyecto determinado, posiblemente debido a la **homogeneidad del portafolio** y al **bajo contraste entre categorías**.

Para mejorar el modelo, sería recomendable **incorporar variables adicionales**, como duración del proyecto, costo o región, que podrían aportar mayor variabilidad y poder predictivo.



Análisis del impacto del balanceo: BG_MasFrecuente

El modelo de **BG_MasFrecuente** buscó predecir si un proyecto pertenece al grupo de negocio más común dentro del conjunto de datos, utilizando como variables independientes **Project manager**, **Project organization** y **Project Type**.

Tras aplicar la técnica de balanceo **SMOTE (Synthetic Minority Oversampling Technique)**, los resultados obtenidos fueron:

- **Precisión:** 39.53 %
- **Exactitud:** 54.05 %
- **Sensibilidad:** 68 %

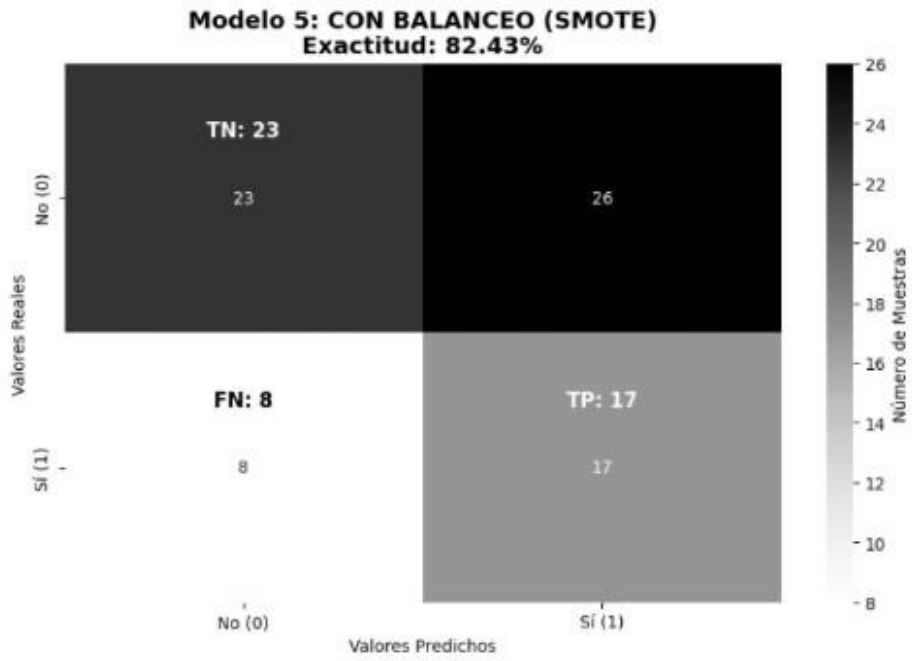
El modelo presenta **baja precisión y una sensibilidad moderada**, lo que indica que clasifica correctamente algunos casos positivos, pero aún comete una cantidad considerable de errores.

El balanceo ayudó parcialmente a mejorar la detección de la clase minoritaria, aunque **no logró estabilizar completamente el desempeño del modelo**.

Conclusión:

Este resultado sugiere que las variables seleccionadas **no son suficientes para explicar la pertenencia de un proyecto al grupo de negocio más frecuente**, probablemente debido a la **homogeneidad del portafolio y la falta de diferenciación entre grupos**.

Para obtener un modelo más preciso, se recomienda **incluir variables adicionales** relacionadas con el tipo de cliente, la ubicación geográfica o el presupuesto, que podrían aportar mayor capacidad predictiva y reducir los errores de clasificación.



Conclusiones

El análisis desarrollado permitió aplicar modelos de Regresión Logística para predecir distintas variables relacionadas con la gestión y el desempeño de los proyectos de FORVIA, a partir de indicadores administrativos, de avance y de estructura organizacional.

Mediante un proceso riguroso de limpieza, depuración y transformación de datos, se estandarizó un conjunto de 246 registros con 25 variables, de las cuales se generaron cinco variables dicotómicas clave: State_Activo, ProjectHealth_Bueno, OnHold, ProjectType_MásFrecuente y BG_MásFrecuente. Estas variables permitieron analizar el comportamiento del portafolio de proyectos en términos de actividad, salud operativa, estado de pausa, tipo y grupo de negocio.

Los resultados de los modelos mostraron diferentes niveles de desempeño predictivo.

El modelo State_Activo presentó una alta exactitud (82.43%) y buena sensibilidad, confirmando que las variables Project Health, On-hold y Percent complete explican correctamente la condición de actividad de los proyectos.

ProjectHealth_Bueno obtuvo una sensibilidad del 96.36%, evidenciando su capacidad para identificar proyectos con buen estado, aunque con una exactitud moderada.

El modelo OnHold alcanzó un F1-Score de 90.91%, siendo el más equilibrado y confiable para detectar proyectos en pausa.

Por otro lado, ProjectType_MásFrecuente y BG_MásFrecuente presentaron un rendimiento bajo, indicando que las variables utilizadas no explican completamente la pertenencia a un tipo o grupo específico de proyecto.

La aplicación del balanceo de clases con SMOTE fue determinante para mejorar la sensibilidad de los modelos, especialmente en aquellos con fuerte desbalance de datos. Aunque algunos modelos redujeron ligeramente su exactitud, el balanceo permitió obtener predicciones más justas y representativas, evitando el sesgo hacia las clases dominantes (como los proyectos activos o del BG principal).