Campus Puebla

Work's name:

Actividad 4.2 Regresión Logística

Course:

Analitica de Datos II

Student:

Ivanna Maldonado Cervantes
Paula Simonetta Madrid Pérez
Ania Diaz Gonzalez
Miranda Eugenia Colorado Arróniz
Omar Alejandro Quinn Toledo

En esta actividad aplicamos **regresión logística** para explorar cómo ciertas características de los proyectos (tipo, alcance, organización, etc.) Se relacionan con resultados operativos como **On-hold**, **State**, **Project size**, **Project Health** y **Percent complete**.

Fuente: dataset interno de proyectos (Forvia).
Limpieza básica: estandarización de formatos (por ejemplo, convertir "Percent complete" de "85%" a **85.0**), manejo prudente de nulos.
Codificación: las columnas categóricas clave —*Project Type, Geographical scope, Project manager, State, Project size, Project organization, BG, Project Health, On-hold*— se codificaron para poder usarlas en el modelo.

1. **Definición X/Y.** Para cada modelo, definimos la **Y** (variable objetivo) y las **X** (predictores) según las preguntas de negocio.
2. **Split 70/30.** train_test_split con 30% para prueba.
3. **Estandarización.** StandardScaler sobre las **X** (la logística es sensible a escala).
4. **Entrenamiento.** LogisticRegression, probando versión **balanceada** (class_weight='balanced') cuando hay desbalance.
5. **Evaluación.** Matriz de confusión + **precisión**, **recall**, **accuracy** y **F1** (reportadas **por clase**, no solo global).
6. **Lectura de umbral.** No nos casamos con 0.5: el **threshold** se ajusta según si priorizamos recall o precisión.

- **Matriz de confusión:** muestra dónde se equivoca el modelo (qué clase confunde con cuál).
- **Precisión (precision):** "cuando digo 'positivo', ¿qué tanto atino?".
- **Sensibilidad (recall):** "de todos los positivos reales, ¿cuántos detecto?".
- **Accuracy:** aciertos totales. (Se busca balancear las clases donde se requiere)
- **F1:** promedio entre precisión y recall..

Modelo 1 — *On-hold* (sí/no)

- **Y:** On-hold
- **X:** Project Type, Geographical scope, Project size
- **Propósito: alerta temprana** de pausas.
- **Lectura de métricas:** da prioridad al **recall** de la clase "sí está en hold" (mejor detectar la mayoría de los casos reales, aunque haya alguno falso positivo).
- **Uso práctico:** si el recall es bueno, convierte la predicción en **trigger** para seguimiento (revisión de bloqueo, escalamiento).

Modelo 2 — *State* (binario)

- **Y:** State (agrupado a 2 niveles)
- **X:** Project size, Project Type, On-hold
- **Propósito:** entender si tamaño/tipo y pausas se reflejan en el **estado**.
- **Lectura de métricas:** mira **F1** por clase (que no gane la mayoría "por inercia").
- **Uso práctico:** diseñar **políticas por tamaño** (p. ej., protocolos más estrictos para LARGE/MEDIUM con historial de *On-hold*).

Modelo 3 — *Project size* (SMALL vs LARGE/MEDIUM)

- **Y:** Project size (binario: SMALL vs LARGE/MEDIUM)
- **X:** Project manager, State, Project organization
- **Propósito: planeación de recursos** (anticipar demanda).
- **Lectura de métricas:** cuida el **recall** de **SMALL** si esa clase es minoritaria (que no se "olvide" del pequeño).
- **Uso práctico:** mejora la **asignación** de equipo/tiempo sin esperar a señales tardías.

Modelo 4 — *Project Health* (binario o multiclase)

- **Y:** Project Health
- **X:** Project organization, State, On-hold
- **Propósito:** relacionar **estructura y estatus** con la **salud** del proyecto.
- **Lectura de métricas:** si es binaria, vigila **F1** por clase; si es multiclase, mira **F1 por clase** y confusiones específicas.
- **Uso práctico:** priorizar **intervenciones** (proyectos con mala salud + On-hold/estado crítico).

Modelo 5 — *Percent complete* (alto vs bajo)

- **Y:** Percent complete (arriba vs abajo del promedio/umbral)
- **X:** Project Health, State, On-hold
- **Propósito:** identificar si salud/estado/pausas **anticipan avance**.
- **Lectura de métricas:** típicamente **F1** puede ser más bajo si faltan features de calendario.
- **Uso práctico:** integrar predicción con **planeación** (adelantar soporte a proyectos que pinta que irán por debajo).

- **Ponderar clases ayuda.** Con class_weight='balanced' suele subir el **recall** de la clase chica (clave para *On-hold*y *SMALL*).
- **Las variables organizacionales sí traen señal.** *Project organization*, *manager* y *state* no son “decoración”: impactan resultados.
- **On-hold ≈ proxy de riesgo.** Cuando se predice bien, funciona como **semáforo** para atención temprana.
- **Para “Percent complete” pide más contexto.** Calendario real (fechas plan vs real, nº de re-planes, últimas WAR) tiende a **elevar F1**.