Taller: Clasificación con *Diabetes Health Indicators*

Fundación Universitaria los Libertadores

10 de noviembre de 2025

Objetivo general

Desarrollar, evaluar e interpretar un **clasificador reproducible** para predecir diabetes usando el conjunto de datos *Diabetes Health Indicators*, cumpliendo el ciclo **CRISP-DM** de extremo a extremo, aplicando **técnicas de balanceo** cuando sea necesario y **priorizando una métrica clave** con justificación técnica.

Resultados de aprendizaje

Al finalizar, cada participante/equipo será capaz de:

1. **Planear y ejecutar CRISP-DM** extremo a extremo (Business Understanding → Deployment simulado).
2. Diseñar **pipelines** de preprocesamiento y modelado en `scikit-learn` evitando *data leakage*.
3. **Diagnosticar desbalanceo** y aplicar **class_weight/SMOTE/undersampling** dentro de la validación cruzada cuando corresponda.
4. **Seleccionar la métrica prioritaria** (p. ej., **PR-AUC** o **Recall** si hay desbalanceo; o **ROC-AUC**/**Balanced Accuracy** si no) y **defenderla** con criterios de costo–error.
5. Ajustar **umbrales** y **calibrar probabilidades** (Platt/Isotónica) según la métrica prioritaria.
6. Comparar modelos (Logística, Árbol/RandomForest; opcional XGBoost/LightGBM), **explicar** resultados (coeficientes/SHAP) y documentar **limitaciones y sesgos**.
7. Mantener **trazabilidad** en GitHub: estructura clara, commits significativos, `README` reproducible y `requirements`.

Requisitos técnicos

- **IDE**: PyCharm *o* Visual Studio Code.
- **Control de versiones**: Git + GitHub (un único repositorio por equipo).

- **Python** 3.10/3.11 con los siguientes paquetes mínimos:
  - `pandas`, `numpy`, `scikit-learn`, `imbalanced-learn`, `matplotlib`, `shap`, `joblib`.
  - (Opcional) `xgboost` o `lightgbm`; `ydata-profiling` para EDA.
- **Gestión de entorno**: `venv`, `conda` o `poetry` (a elección, documentado en el `README`).