

PROGRAMA

Machine Learning y Modelado Predictivo

Objetivos: Este curso brinda una formación integral en aprendizaje automático y modelado predictivo, abarcando todas las etapas del desarrollo de modelos, desde el pre-procesamiento de datos hasta su validación e interpretación. Los estudiantes aprenderán a implementar pipelines reproducibles, comparar y optimizar modelos mediante métricas apropiadas, diagnosticar problemas como sesgo, sobreajuste y desbalance, interpretar predicciones utilizando técnicas modernas como SHAP y aplicar métodos avanzados de regularización, optimización bayesiana y modelos de ensamble para construir soluciones predictivas robustas y confiables.

La asignatura tiene como objetivo que los estudiantes aprendan a:

- Implementar pipelines ML desde limpieza hasta validación
- Comparar modelos con métricas teóricamente justificadas
- Diagnosticar sesgo, sobreajuste, desbalance
- Interpretar predicciones con SHAP
- Aplicar regularización y tuning bayesiano
- Usar ensemble methods efectivamente
- Escribir código reproducible y defendible

Herramientas

Se usará como herramientas scikit-learn, XGBoost, LightGBM, SHAP, Optuna - GitHub, Google Colab, DVC y datasets provenientes de Kaggle, UCI Repository, Our World in Data Datos.gob.cl, INDEC Argentina.

Contenidos:

Unidad 1. Introducción y problemas clásicos:

Fundamentos ML & Herramientas para Gestión de Proyectos DS

Conceptos: - Taxonomía ML (supervisado/unsupervised, regresión/clasificación) -

Generalización: train/val/test, validación cruzada - Sesgo vs varianza

Herramientas prácticas: - Estructura de carpetas reproducible - Jupyter Notebooks

+

Colab best practices - Git básico: commits, branches, pull requests - Versionado de datos

(DVC intro) - Documentación con README y docstrings - Requirements.txt y reproducibilidad.

Unidad 2. EDA & Limpieza de Datos:

- *Análisis exploratorio*
- *Valores faltantes: imputación y validación de impacto*
- *Outliers: detección y decisión*

- *Encoding de categóricas*
- *Escalado y normalización*
- *Feature engineering: lógica y validación*

Unidad 3. Clasificación Binaria & Multiclase

- *Regresión logística: odds, probabilidades calibradas*
- *Árboles de decisión: teoría y diagnósticos*
- *Desbalance: SMOTE, class weights, métricas apropiadas*
- *Threshold tuning: ROC, Precision-Recall*
- *Métricas: Precision, Recall, F1, AUC-ROC (nunca Accuracy sola)*

MINI PROYECTO 1: Análisis Exploratorio + Limpieza de Dataset (Semana 6) — 25%

Unidad 4. Ensemble Methods Fundamentales:

- *Random Forests: teoría e importancia de features*
- *Gradient Boosting, XGBoost, LightGBM*
- *Stacking y blending*
- *Tuning de hiperparámetros: grid/random/Bayesian*

Semana 9-10: Interpretabilidad & Explicabilidad

- *Feature importance: tree-based, permutation*
- *SHAP: Shapley values, force plots, dependence plots, interaction*
- *LIME: explicaciones locales*
- *Análisis sistemático de errores*
- *Fairness y sesgo*

Unidad 5. Validación Avanzada & Gotchas

- *Data leakage: tipos y prevención*
- *Temporal data: walk-forward validation, drift*
- *Producción: persistencia, monitoreo, retraining*
- *Reproducibilidad garantizada*
- *Tests estadísticos de diferencia*

Bibliografía

Géron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems (2da ed.). O'Reilly Media.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). An Introduction to Statistical Learning: with Applications in R (2da. ed.). Springer.

Molnar, C. (2022). Interpretable Machine Learning (2da. ed.). Leanpub. Disponible gratuitamente en: <https://christophm.github.io/interpretable-ml/>