

Contexto

Problema: Comprender y modelar modalidad de programas y edad promedio, explorando también estructura

Fecha de generación: 2025-11-21T03:01:33.205973Z

Leakage: Leakage flag=False r2=0.18854441022979995 features=['PROMEDIO EDAD HOMBRE ', 'PRO

Dataset

Matriz de correlación: 200 variables

Fuente: data/raw/*.csv (ver config/config.py para parámetros). Tamaño aproximado post-proceso: ver X_

Metodología

Pipeline: EDA -> Preprocesamiento -> Feature Engineering -> Entrenamiento -> Evaluación -> Interpretaci

Validación temporal aplicada (split por año). HPO: HPO ejecutado (ver hpo_summary.md para top config

Resultados

Clasificación

Métrica: AUC-ROC=0.4644

Figura/Importancias: ver reports/feature_importance_classification.csv (top 5 abajo si disponible).

Regresión

Métrica: pendiente

Figura/Importancias: ver reports/feature_importance_regression.csv.

Clustering

Top clustering (silhouette):

algo|params|silhouette|homogeneity|completeness|runtime_ms

DBSCAN|{'eps': 0.7, 'min_samples': 5}|0.2192823255920242|||743.37

KMeans|{'n_clusters': 7}|0.1822504906134569|||185.74

DBSCAN|{'eps': 0.7, 'min_samples': 10}|0.1820721923050643|||796.53

Figura: clustering_dbSCAN_eps_grid.png

Anomalías

Anomalías (fracción vs target):

algo|anomaly_fraction|contamination_cfg|runtime_ms

IsolationForest|0.05|0.05|1091.82

LocalOutlierFactor|0.05|0.05|874.62

LocalOutlierFactor|0.05|0.05|887.97

EllipticEnvelope|0.05|0.05|2058.34

Figura: anomaly_fraction_by_algo.png

Interpretabilidad

Top importancias clasificación:

feature|importance

CARRERA CLASIFICACIÓN NIVEL 2_Carreras Técnicas|0.2572521855202013

CLASIFICACIÓN INSTITUCIÓN NIVEL 1_Universidades|0.2394819847710022

CARRERA CLASIFICACIÓN NIVEL 1_Técnico de Nivel Superior|0.1896091493932357

NOMBRE INSTITUCIÓN_IP AIEP|0.1038523508856253

CLASIFICACIÓN INSTITUCIÓN NIVEL 2_Universidades CRUCH|0.0632917366810482

Importancias regresión: pendiente

Recomendaciones

- Consolidar almacenamiento de modelos entrenados y versionar.

- Incorporar monitoreo de drift y recalibración anual.

- Extender modelos a boosting y SHAP para interpretabilidad avanzada.
- Integrar panel UI para exploración de anomalías y clusters.