# UNIVERSIDAD DE LA COSTA

Facultad de Ingeniería de sistemas

## INFORME DE LABORATORIO – CONTROL AUTOMÁTICOS II

**Actividad 2:** Entrenamiento de modelos – Student Dropout Classification

Instructor: José Escorcia-Gutiérrez, Ph.D.

Estudiantes: Carlos Antonio Ardila Ruiz y Hernando Luis Calvo Ochoa

## Resumen

Esta práctica tuvo como objetivo desarrollar y comparar dos modelos de aprendizaje automático, Regresión Logística y Árbol de Decisión, para predecir la deserción estudiantil en programas de pregrado. Se utilizó el conjunto de datos 'student_dropout.csv', que contiene información demográfica, académica y financiera de 4424 estudiantes. El trabajo incluyó limpieza de datos, análisis exploratorio, preprocesamiento con ColumnTransformer, validación cruzada y evaluación de métricas en el conjunto de prueba.

## Objetivo General

Implementar y comparar modelos de Regresión Logística y Árbol de Decisión para predecir la probabilidad de abandono estudiantil, identificando qué modelo ofrece mejor rendimiento para un sistema de alerta temprana.

## Metodología

**Carga y limpieza de datos:** Se cargó el dataset de 37 variables. Se detectó que la columna'Target' tenía tres categorías: Graduate, Dropout y Enrolled. Para el análisis se transformó en binaria (Dropout=1, otros=0).
**Análisis exploratorio:** No se hallaron valores nulos. Se identificaron 35 variables numéricas y 2 categóricas.
**Preprocesamiento:** Se aplicó imputación por mediana para variables numéricas, estandarizacióncon StandardScaler y codificación OneHot para las categóricas mediante ColumnTransformer.

**Entrenamiento de modelos:** Se construyeron dos pipelines: uno con Regresión Logística y otrocon Árbol de Decisión. Se dividieron los datos 80/20 para entrenamiento y prueba. 5. **Validación cruzada:** Se realizó validación 5-fold usando F1 como métrica principal.
**Evaluación:** Se calcularon Accuracy, Precision, Recall, F1-score, matriz de confusión y AUC-ROC.

## Resultados

| Modelo | Accuracy | Precision | Recall | F1-score | ROC AUC |
|---|---|---|---|---|---|
| Regresión Logística | 0.888 | 0.897 | 0.736 | 0.809 | 0.927 |
| Árbol de Decisión | 0.795 | 0.666 | 0.729 | 0.696 | 0.778 |

## Explicación de los modelos

**Regresión Logística:** Es un modelo estadístico que estima la probabilidad de pertenecer a unaclase (abandono o no) en función de variables predictoras. Genera una frontera de decisión lineal y es interpretativo, lo que facilita comprender el peso de cada variable en el resultado.

**Árbol de Decisión:** Divide los datos en ramas según las condiciones que mejor separan las clases.Aunque es fácil de visualizar y entender, puede sobreajustarse si no se controla la profundidad. Es útil para detectar interacciones no lineales entre variables.

## Conclusiones

La Regresión Logística obtuvo el mejor rendimiento general (F1=0.81 y AUC=0.93), lo que la convierte en el modelo más adecuado para implementar un sistema de alerta temprana sobre riesgo de deserción. El Árbol de Decisión, aunque menos preciso, puede resultar útil por su facilidad de interpretación. Como trabajo futuro se propone aplicar técnicas de balanceo de clases y explorar modelos más robustos como Random Forest o XGBoost.

"El verdadero valor de los datos está en las decisiones que permiten tomar."