Data for Education 2025

Universidad Tecnológica de Pereira

Identificación temprana y explicable de estudiantes en riesgo: Ingeniería de características longitudinales con interpretabilidad para intervención universitaria efectiva

Steven Galindo Noreña
Cristian Alejandro Blanco Martínez

Universidad Tecnológica de Pereira

Conocer más

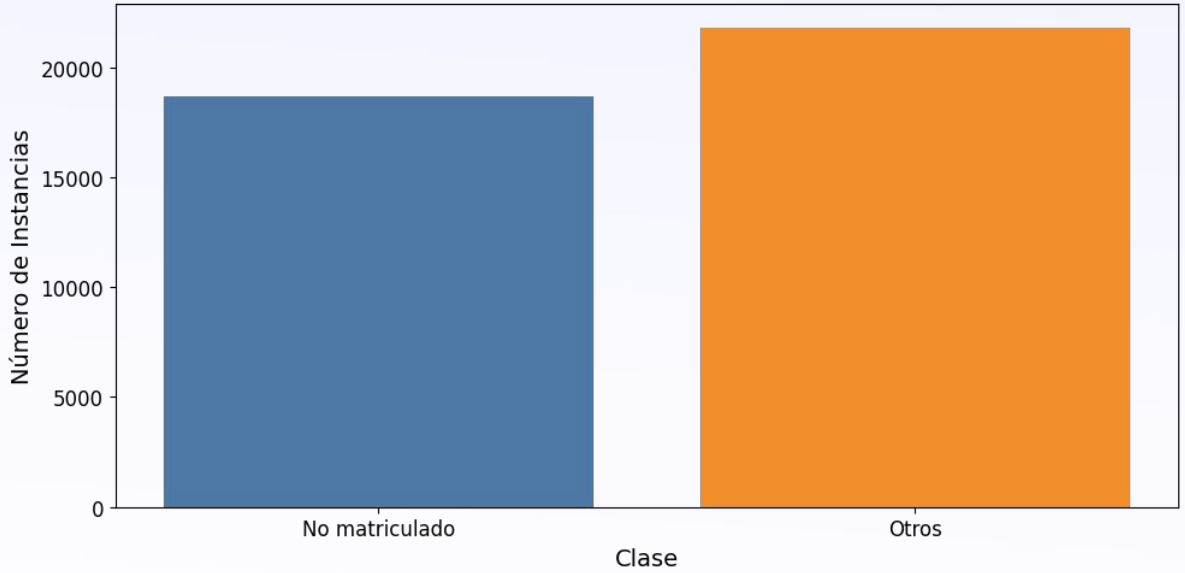
Inteligencia Artificial para la Transformación Educativa

Agenda

1. Problema y objetivo
2. Base de datos
3. Metodología
4. Modelo usado
5. Resultados
6. Interpretabilidad

Problema y objetivo

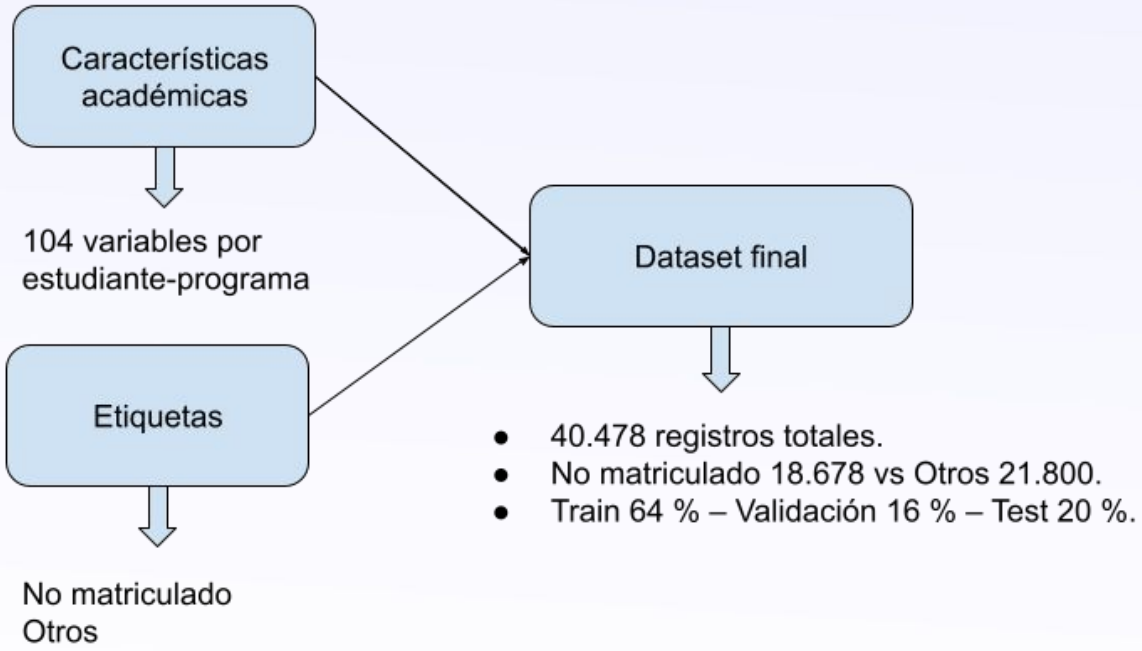
La deserción universitaria es un reto crítico: en nuestro histórico encontramos casi 18 mil casos de estudiantes que no se matricularon nuevamente, lo cual representa un impacto fuerte en la trayectoria educativa y en la planeación institucional.



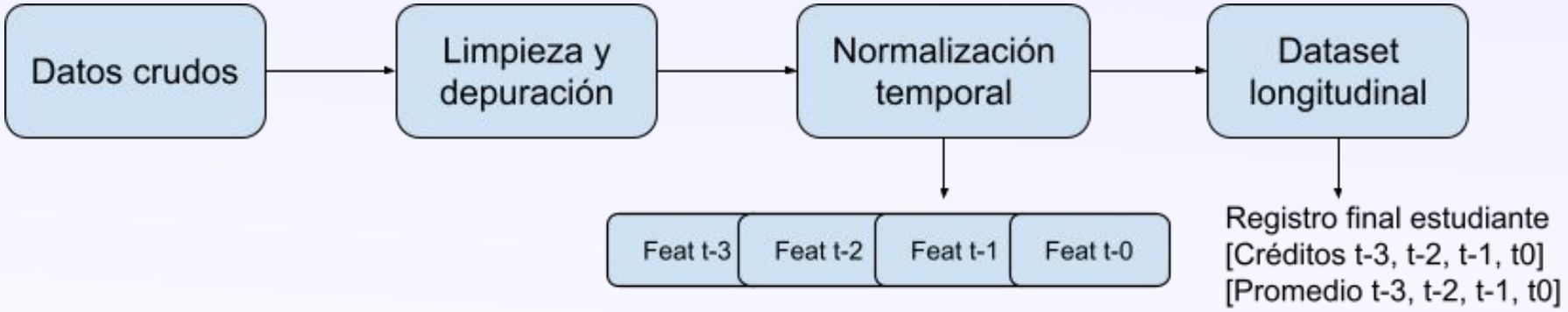
Objetivo: Diseñar una metodología reproducible y explicable para identificar tempranamente estudiantes en riesgo de no matricularse, usando señales académicas longitudinales.

Datos

Las características son un conjunto de 104 variables que incluyen, entre otras, créditos matriculados, aprobados, reprobados y cancelados en ventanas temporales, promedios semestrales, número de materias y estados administrativos.



Metodología



El reto radica en los datos académicos, los cuales son **transaccionales y fragmentados**: cada estudiante tiene múltiples registros por periodo y programa.

Nuestra solución fue **normalizarlos en el tiempo** y resumirlos en **ventanas relativas t-3 a t-0**, para que cada registro capture no solo un instante, sino el comportamiento completo.

Metodología

Variable	t-3	t-2	t-1	t0
Créditos aprobados	12	10	8	6
Promedio semestre	4.1	3.9	3.2	2.8
Materias reprobadas	1	2	3	4

En este ejemplo vemos un estudiante cuya carga y rendimiento van cayendo: menos créditos aprobados, promedios decrecientes, más materias reprobadas.

Modelo usado

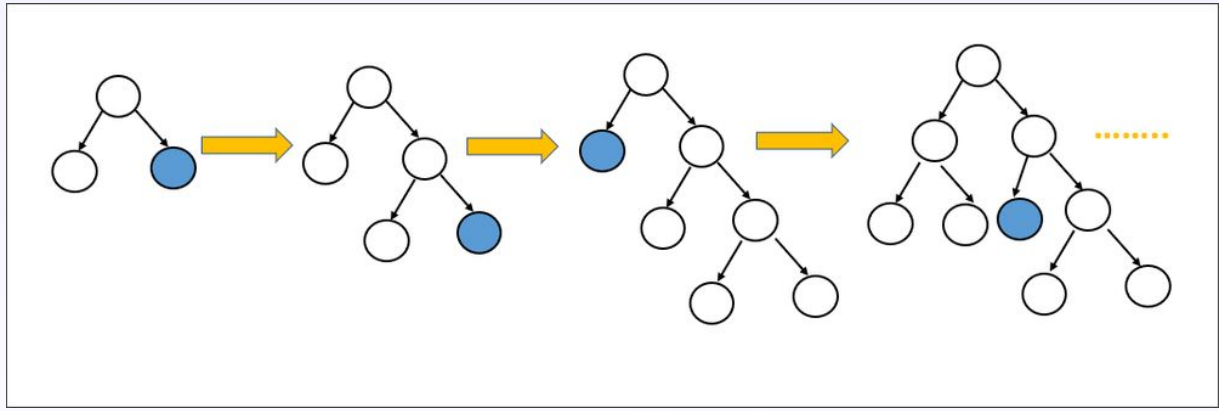
Búsqueda de hiperparámetros (RandomizedSearchCV):

Iteraciones: 30

Validación cruzada: 3 folds

Métrica de optimización: F1-score

LightGBM:



Rangos explorados:

Número de árboles: entre 100 y 300

Tasa de aprendizaje: entre 0.01 y 0.2

Profundidad máxima de los árboles: entre 3 y 10

Mejores valores encontrados:

Número de árboles: 152

Tasa de aprendizaje: 0.137

Profundidad máxima: 8

Resultados

Métricas de desempeño:

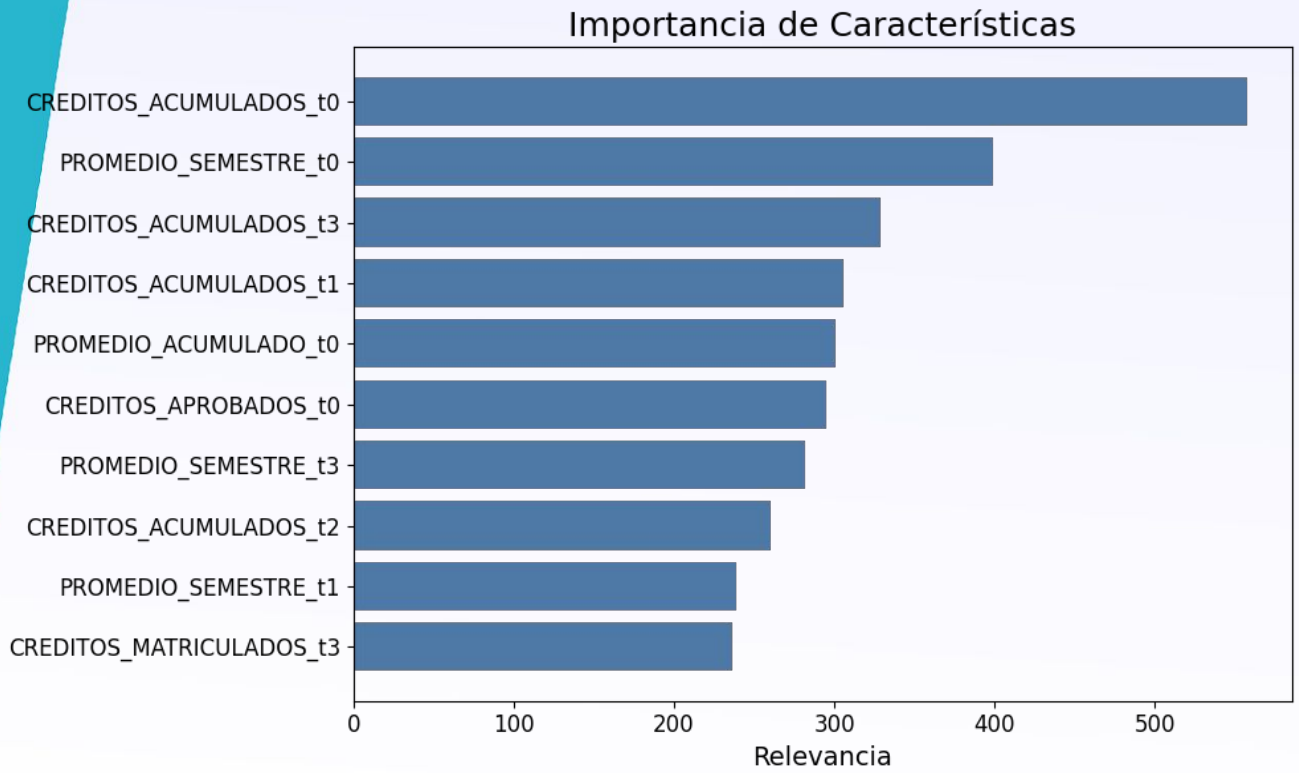
Conjunto	Accuracy	F1-macro
Validación	0.946	0.945
Test	0.944	0.944

Clase	Precisión	Recall	F1
No matriculado	0.95	0.93	0.94
Otros	0.94	0.96	0.95

Gracias a LightGBM pudimos confirmar la importancia relativa de las características, destacando créditos acumulados y promedio reciente.

- Desempeño alto y balanceado entre clases.
- Consistencia entre validación y test (F1 ~0.94).
- Señales más predictivas: créditos acumulados y promedio reciente (t0).

Interpretabilidad (LightGBM)



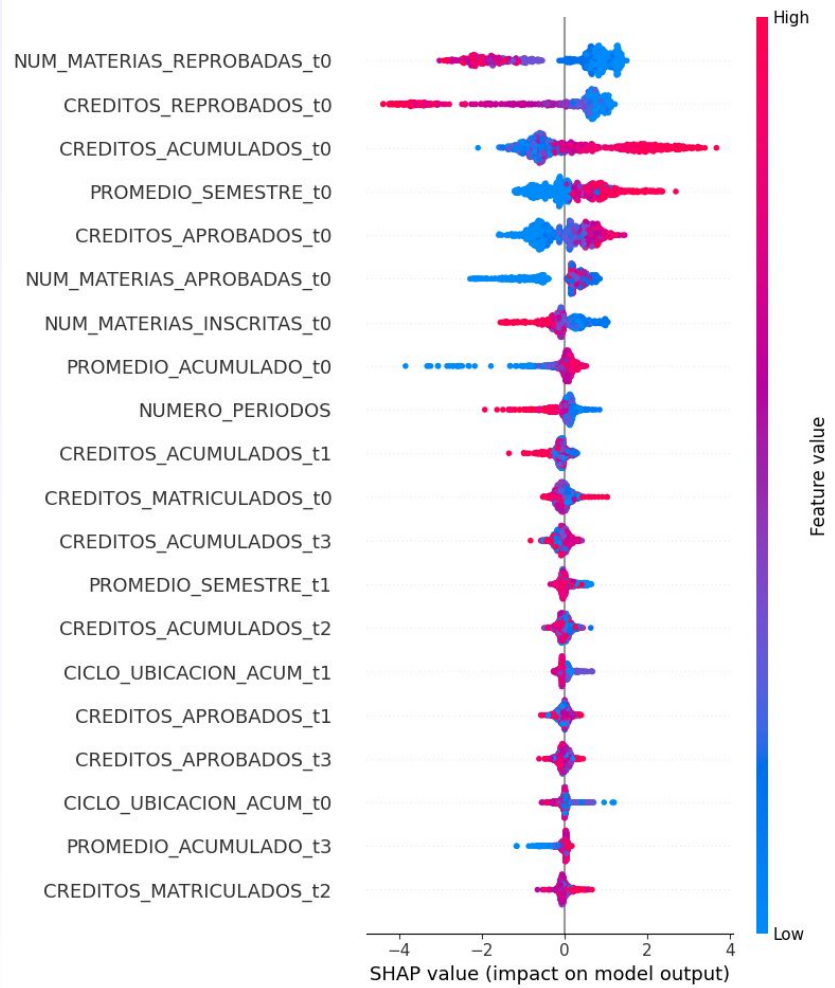
- El avance académico reciente es el factor más determinante: menos créditos acumulados en el último periodo (t0) se asocian a mayor riesgo de deserción.
- El rendimiento inmediato también pesa: un promedio de semestre bajo en t0 incrementa la probabilidad de deserción.
- La historia académica completa importa: variables de periodos previos (t1, t2, t3) aportan contexto, pero la ventana más reciente domina.
- Señales de trayectoria como créditos aprobados o reprobados reflejan la persistencia o dificultades del estudiante.

Interpretabilidad (SHAP)

El análisis SHAP confirma que las variables más influyentes son los créditos acumulados y el desempeño académico reciente.

- En el gráfico, cada punto es un estudiante:
  - El eje horizontal indica cómo esa variable empuja la predicción hacia deserción (izquierda) o permanencia (derecha).
  - Los colores marcan si el valor de la variable es bajo (azul) o alto (rojo).

Así, por ejemplo, tener muchos créditos reprobados en t0 (puntos rojos a la izquierda) incrementa fuertemente el riesgo de deserción, mientras que altos créditos acumulados en t0 (puntos rojos a la derecha) favorecen la permanencia.



Conclusiones

- Señales predictivas claras: el avance académico reciente (créditos acumulados en t0) y el promedio del último semestre son los principales indicadores de permanencia o riesgo de deserción.

- Potencial de alertas tempranas: la metodología permite identificar trayectorias de deterioro antes de que ocurra la deserción, facilitando intervenciones oportunas como tutorías o acompañamiento académico.

- Hacia un modelo más integral: incorporar variables socioeconómicas podría enriquecer la capacidad predictiva y orientar políticas más inclusivas.

- Robustez futura: aplicar validación temporal garantizará que el modelo mantenga su desempeño al predecir cohortes futuras, no solo históricas.

Lecciones aprendidas

- La ingeniería de datos es decisiva: la calidad del preprocesamiento y la construcción de variables tuvo más impacto que la elección del algoritmo.

- Enfoque en señales accionables: priorizar variables que la institución ya recolecta y puede monitorear facilita la implementación de alertas tempranas.

- Interpretabilidad como requisito: un modelo útil no solo debe predecir, sino explicar. La transparencia en las variables refuerza la confianza de docentes y directivos.

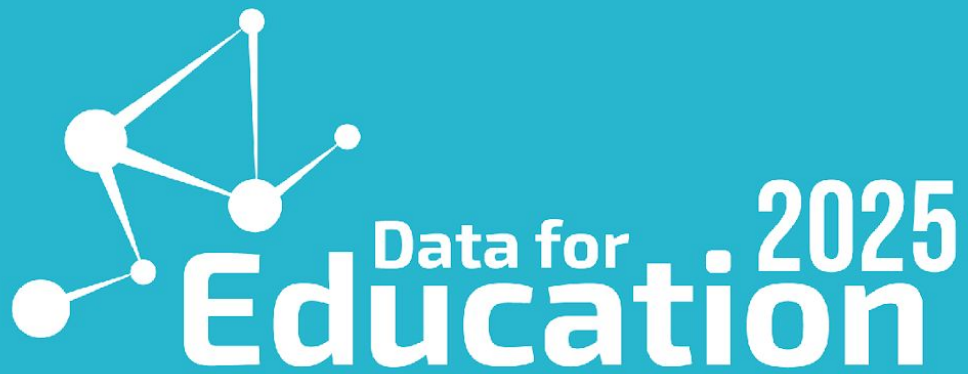
Data for Education 2025

Universidad Tecnológica de Pereira

Muchas gracias

Steven Galindo Noreña
sgalindo@utp.edu.co

Cristian Alejandro Blanco Martínez
cristian.blanco@utp.edu.co


Data for
Education
2025


Universidad
Tecnológica
de Pereira