Proyecto Prototípico — Universidad Rosario Castellanos (UNRC)

Título: Sistema de alerta temprana de abandono estudiantil con datos sintéticos y pipeline reproducible

Autor(es): [Tu nombre] | Semestre: 2025-2 | Fecha: [Fecha]

# 1. Resumen Ejecutivo

Este proyecto prototípico diseña un pipeline de datos para estudiar el abandono estudiantil en la Universidad Rosario Castellanos (UNRC). Como la institución no publica microdatos desagregados, se construyó una base de datos sintética inspirada en fuentes oficiales mexicanas y estudios académicos. El pipeline incluye: generación de datos crudos a nivel estudiante-semestre, derivación de eventos de abandono y stop-out, estimaciones descriptivas y modelos predictivos (logística y supervivencia). Los resultados muestran que el riesgo se concentra en los primeros semestres y que factores como bajo promedio, escasa asistencia, empleo intensivo y largos traslados inciden fuertemente en la permanencia. Por su parte, las becas y tutorías funcionan como amortiguadores del riesgo.

# 2. Contexto y Justificación

El abandono escolar en educación superior es un fenómeno multidimensional que afecta tanto a la eficiencia terminal como a la equidad social. Diversas fuentes en México aportan información relevante:  
  
- SIIES (Sistema de Información de Educación Superior) reporta indicadores de matrícula, egreso y eficiencia terminal por subsistema.  
- ANUIES (Asociación Nacional de Universidades e Instituciones de Educación Superior) ha documentado que, en promedio, sólo alrededor de 6 de cada 10 estudiantes que ingresan a licenciatura concluyen sus estudios.  
- ENIGH (INEGI) proporciona distribuciones de ingreso, composición de hogares y condiciones de trabajo que permiten contextualizar la vulnerabilidad socioeconómica de la población universitaria.  
- Estudios de instituciones como UNAM e IPN señalan que los mayores índices de abandono ocurren en el primer año, vinculados a rezago académico y presiones económicas.  
  
La UNRC, al ser una universidad de reciente creación, carece de estadísticas públicas detalladas de trayectoria estudiantil. Esta ausencia impide realizar análisis predictivos con datos reales.  
  
Generar datos sintéticos es, por tanto, una decisión metodológica justificada porque:  
1. Permite diseñar y validar un pipeline completo sin exponer datos personales.  
2. Integra supuestos basados en evidencia (ingresos de ENIGH, cargas laborales típicas, tiempos de traslado en CDMX).  
3. Ofrece un laboratorio de prueba para calibrar modelos y visualizaciones que luego pueden usarse con microdatos institucionales.  
4. Cumple un propósito pedagógico: mostrar el potencial de herramientas de probabilidad, cálculo, bases de datos y análisis de datos en un proyecto aplicado.

# 3. Metodología

## 3.1 Generación de datos crudos

Se construyó una base SQLite con dos tablas:  
- students\_raw: características sociodemográficas (sexo, edad, alcaldía, ingreso familiar, personas en el hogar, horas de trabajo, acceso a internet).  
- inscripciones: registros por semestre (promedio, materias aprobadas/reprobadas, beca, tutoría, asistencia).  
  
Los ingresos se simularon a partir de rangos observados en la ENIGH 2022, donde más del 40% de los hogares en CDMX reportan ingresos menores a 14 mil pesos mensuales. Los traslados se derivaron de una matriz alcaldía–plantel, tomando como referencia tiempos de transporte público en el área metropolitana.

## 3.2 Derivación de abandono y stop-out

El abandono se definió como la ausencia de reinscripción en t+1 sin reaparición posterior, salvo en el caso de graduación en el último semestre. El stop-out se consideró cuando el estudiante interrumpía temporalmente pero regresaba después. Esta diferenciación refleja la realidad observada en estudios de retención universitaria en América Latina, donde existen trayectorias discontinuas más allá de la deserción definitiva.

## 3.3 Modelos y análisis

Se aplicaron dos metodologías:  
- Regresión logística para estimar el efecto de factores individuales sobre la probabilidad de abandono.  
- Análisis de supervivencia (Kaplan–Meier) para modelar la evolución del riesgo a lo largo de los semestres.  
  
Ambas técnicas están ampliamente documentadas en literatura sobre deserción estudiantil (Tinto, 1993; Cabrera & Nora, 1994) y permiten tanto interpretabilidad como capacidad predictiva.

# 4. Resultados Principales (sintéticos)

La simulación mostró que, al término de ocho semestres, alrededor de 4 de cada 10 estudiantes no culminan la trayectoria, cifra comparable con estudios nacionales sobre eficiencia terminal. Los mayores niveles de abandono ocurrieron en los semestres 1 y 2, lo que coincide con evidencia de la UNAM y el IPN sobre el riesgo del primer año.  
  
Variables críticas:  
- Promedio bajo (<7.0) y asistencia reducida (<70%).  
- Trabajar más de 20 horas/semana y traslados superiores a 90 minutos.  
- Beca y tutoría redujeron de manera significativa la probabilidad de deserción.  
  
Los gráficos de curvas de abandono y los coeficientes de regresión (ver Figuras 1–3) confirman la relevancia de estos factores.

# 5. Propuesta Operativa

El pipeline puede transformarse en un sistema de alerta temprana mediante:  
- Reglas simples (umbral en GPA, asistencia, horas de trabajo).  
- Puntaje predictivo combinado (salida del modelo logístico).  
  
Este sistema permitiría a la UNRC identificar estudiantes en riesgo y ofrecer apoyos focalizados: becas, tutorías, acompañamiento psicosocial y ajustes de carga. Además, las vistas SQL incluidas pueden enlazarse a tableros en Metabase o DataLens, facilitando el monitoreo en tiempo real.

# 6. Limitaciones

Los resultados dependen de supuestos calibrados con fuentes secundarias y no sustituyen datos institucionales. No se incorporaron variables psicológicas, de bienestar o expectativas de futuro, que también influyen en la retención. El modelo no distingue programas académicos específicos, aunque en la práctica las tasas de abandono varían entre carreras.

# 7. Conclusiones

El proyecto demuestra la viabilidad de construir un pipeline de alerta temprana en la UNRC con base en datos sintéticos. Se justifica esta aproximación por la falta de microdatos públicos y por la necesidad de probar metodologías sin riesgos de privacidad.  
  
El análisis muestra que los principales determinantes del abandono son académicos (promedio, asistencia), económicos (ingreso, horas trabajadas) y contextuales (traslado). Estos hallazgos, aunque simulados, coinciden con evidencia empírica nacional e internacional.  
  
El sistema propuesto es modular, ético y transferible: puede implementarse de inmediato con datos reales para orientar decisiones institucionales y diseñar políticas de permanencia.

# 8. Referencias

- ANUIES. (2021). La educación superior en México: tendencias y desafíos.  
- Cabrera, A. F., & Nora, A. (1994). College persistence: Structural equations modeling test of an integrated model of student retention. Journal of Higher Education, 64(2), 123–139.  
- INEGI. (2022). Encuesta Nacional de Ingresos y Gastos de los Hogares (ENIGH).  
- Secretaría de Educación Pública (SEP). (2023). Principales cifras del sistema educativo nacional.  
- SIIES. (2022). Estadísticas de matrícula y eficiencia terminal en educación superior.  
- Tinto, V. (1993). Leaving College: Rethinking the Causes and Cures of Student Attrition. University of Chicago Press.  
- UNAM & IPN reports on abandono universitario (varios estudios internos, 2018–2022).