# **A document with text and images AI-generated content may be incorrect.**

**Tema: Predicción de deserción universitaria en los primeros semestres**

**1. Problemática**

La deserción universitaria representa uno de los principales retos en la educación superior. En México y en otros países, un porcentaje significativo de estudiantes abandona sus estudios en los primeros semestres debido a factores académicos, personales y socioeconómicos. Este fenómeno afecta tanto a las instituciones, que pierden matrícula, como a los propios estudiantes, que ven interrumpido su desarrollo profesional.

El problema radica en que muchas universidades carecen de un sistema preventivo que les permita identificar a los alumnos en riesgo de abandonar. Sin un diagnóstico temprano, resulta difícil implementar apoyos específicos como tutorías, becas o programas de acompañamiento.

**2. Uso de datos para resolver la problemática**

La deserción no depende de un único factor, sino de una combinación de variables. Con un enfoque de ciencia de datos, es posible recolectar, analizar y modelar información para encontrar patrones que anticipen qué estudiantes tienen más probabilidades de abandonar sus estudios.

Al contar con un modelo predictivo, las universidades podrían generar alertas tempranas y tomar decisiones basadas en evidencia, en lugar de suposiciones.

**3. Pregunta de investigación**

¿Se pueden predecir los estudiantes con mayor riesgo de deserción universitaria en los primeros semestres utilizando información académica y socioeconómica?

**4. Posibles variables (sin dataset todavía)**

Para un futuro dataset, se consideran los siguientes campos:

* **Edad**: factor asociado a la adaptación a la vida universitaria.
* **Promedio académico**: indicador del rendimiento del estudiante.
* **Asistencia a clases**: refleja el compromiso y la continuidad.
* **Horas de estudio por semana**: tiempo dedicado a reforzar aprendizajes.
* **Horas trabajadas por semana**: carga laboral que puede interferir con el estudio.
* **Nivel socioeconómico**: influencia de recursos disponibles en la permanencia.
* **Estado de abandono (sí/no)**: variable objetivo para el modelo.

**5. Futuras etapas del proyecto**

* **Obtención del dataset**: ya sea de fuentes abiertas (ej. Kaggle) o mediante encuestas internas.
* **Explicación de cada característica**: documentar qué representa cada campo y su importancia.
* **Limpieza de datos**: detección de duplicados, manejo de valores nulos, normalización y justificación de qué variables se mantienen o eliminan.
* **Modelado**: aplicar algoritmos de clasificación para predecir el riesgo de abandono.