



# Base de datos de rendimiento académico en secundaria

Esta base de datos contiene información detallada sobre el rendimiento académico de 1.000 estudiantes de secundaria en tres áreas clave: matemáticas, lectura y escritura. Además, incluye variables sociodemográficas y académicas esenciales para comprender los factores que influyen en el desempeño educativo.

Con estos datos, es posible realizar análisis que identifiquen patrones y determinantes del éxito académico, facilitando la toma de decisiones basadas en evidencia para mejorar las políticas educativas.

Por: Simón Posada, Esteban Ramirez y Samuel Garcia



# Análisis de variables y clasificación

## Género

Variable categórica que clasifica a los estudiantes en femenino (0) y masculino (1), importante para estudiar posibles brechas de rendimiento académico.

## Nivel educativo de los padres

Clasificación ordinal desde "some high school" hasta "master's degree", que ayuda a analizar la influencia del entorno socioeducativo en el rendimiento.

- some high school - 0
- high school - 1
- some college - 2
- associate degree - 3
- bachelor's degree - 4
- master's degree - 5

## Curso de preparación

Indica si el alumno completó un curso preparatorio. Variable clave para evaluar el impacto de intervenciones educativas.

- 0 ninguno
- 1 completado

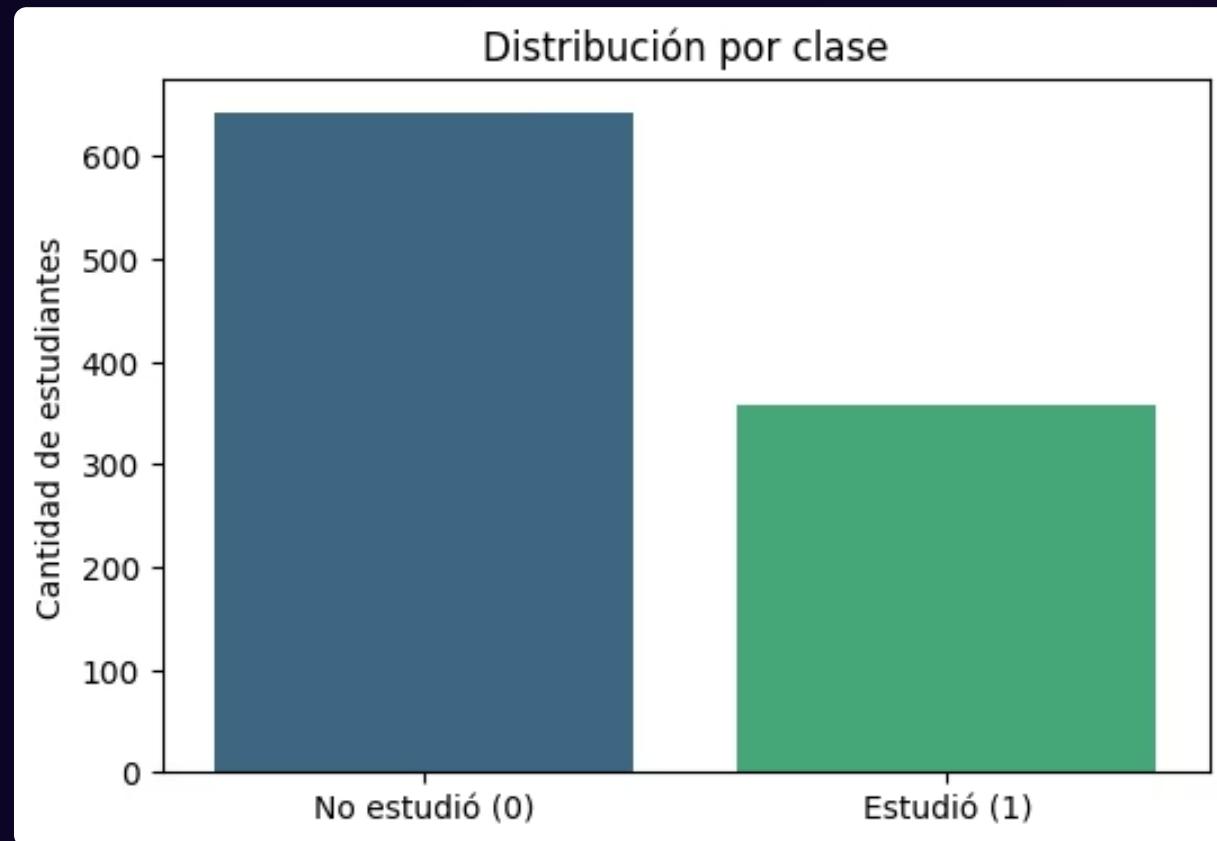
## Puntajes académicos

- Matemáticas (0-100)
- Lectura (0-100)
- Escritura (0-100)

Variables numéricas continuas que permiten realizar análisis predictivos y comparativos.

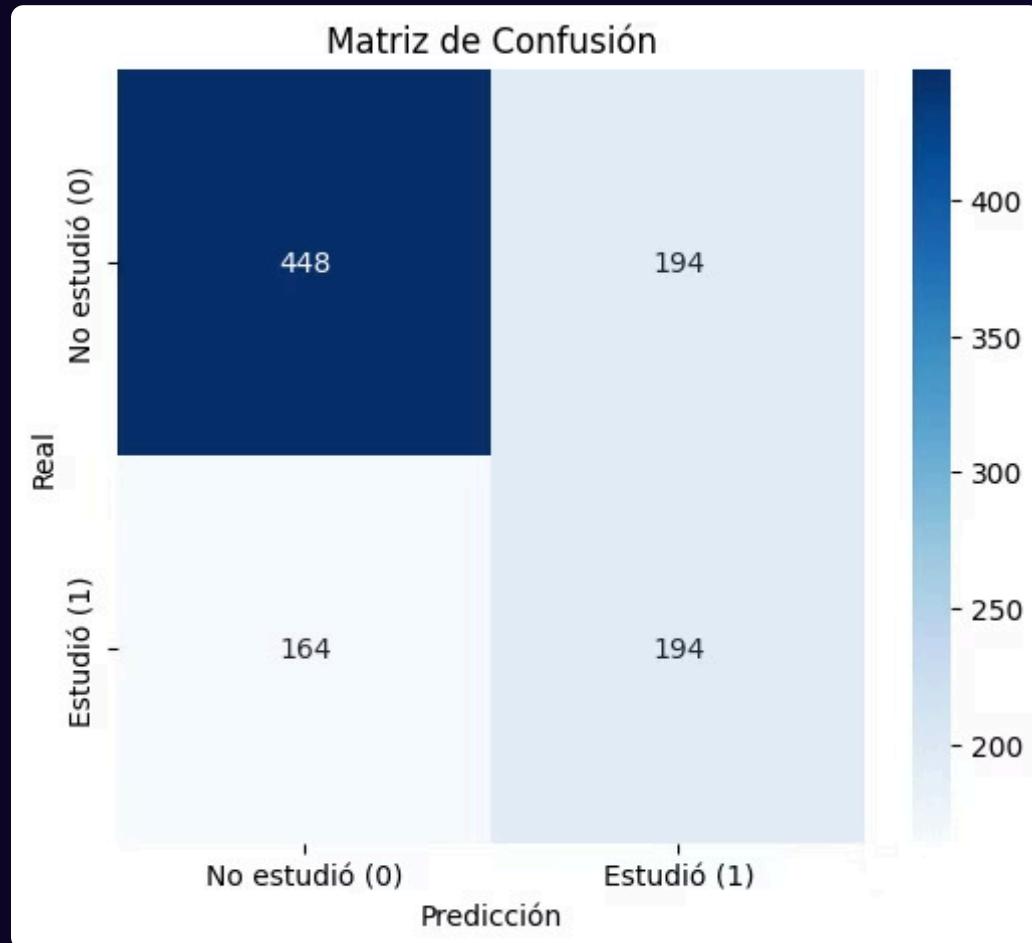
# Modelo 1 - Naive Bayes

El Modelo Naive Bayes es un modelo Machine Learning (ML) por aprendizaje supervisado para la clasificación de los datos en términos de una variable de referencia.



- **Variables usadas:**
- Entrada (X): género, nivel educativo de padres, notas en matemáticas, lectura y escritura.
- Salida (y): curso de preparación (Estudió o No estudió).
- **Distribución de clases:**
  - Total estudiantes: 1000
  - No estudió: 642 (64.2%)
  - Estudió: 358 (35.8%)

# Evaluación del Modelo y Predicción Final



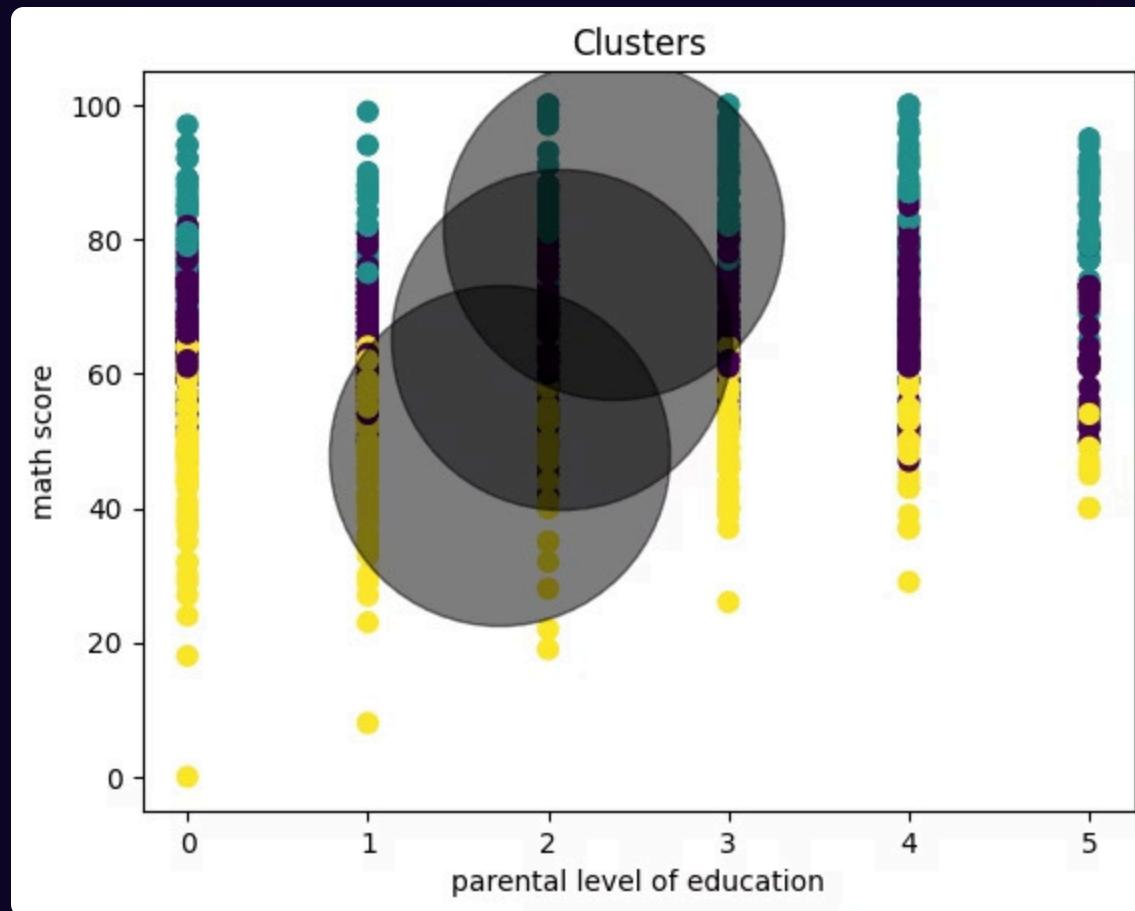
## Métricas del modelo:

- Exactitud: **64.2%**
- Tasa de error: 35.8%
- Sensibilidad (Recall): 54.2%
- Especificidad: 69.8%
- Precisión: 50.0%

Aunque el modelo tiene una precisión moderada, tiende a predecir mejor los casos negativos. La sensibilidad indica que apenas detecta un poco más de la mitad de quienes sí estudiaron.

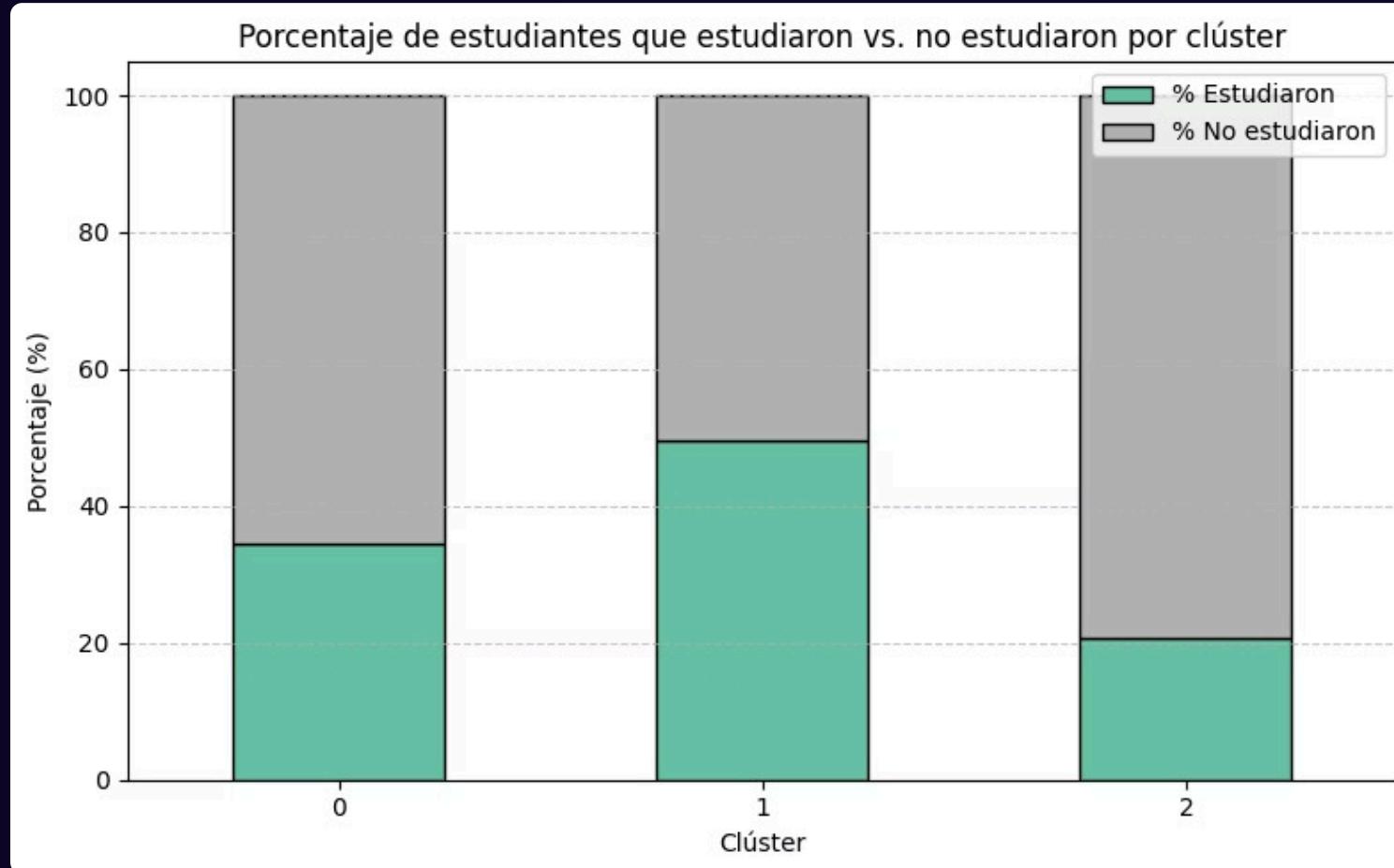
# Modelo Cluster K-means

El modelo de clúster K-Means es un algoritmo de aprendizaje no supervisado que agrupa datos en **k** subconjuntos (clústeres) según similitud, minimizando la distancia entre los datos y el centroide de su grupo.



- Se utilizó el modelo K-Means con **3 clústeres** para segmentar estudiantes según características académicas y sociodemográficas.
- El gráfico muestra la **distribución de estudiantes** por nivel educativo de los padres y puntaje en matemáticas.
- Los **centroídes (puntos negros)** representan el promedio de cada grupo.
- El modelo agrupa estudiantes con **patrones similares de rendimiento**.

# Resultados y Evaluación

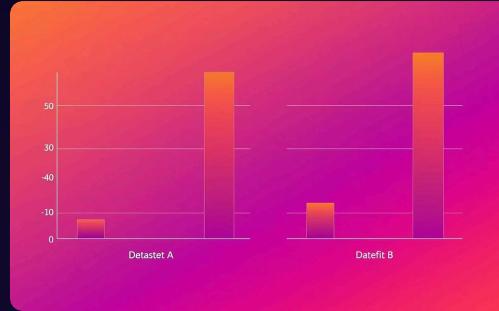
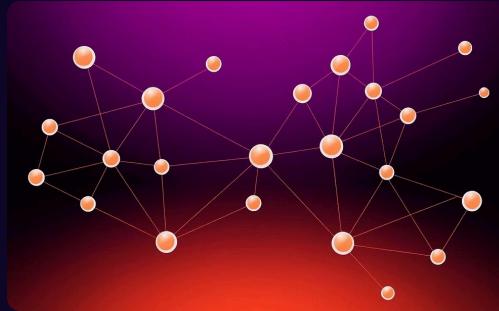


## Distribución por clúster:

- **Clúster 0:** 440 estudiantes → 34.6% estudiaron
- **Clúster 1:** 313 estudiantes → 49.5% estudiaron
- **Clúster 2:** 247 estudiantes → 20.7% estudiaron

El modelo acierta en 6 de cada 10 casos aproximadamente. Mayor precisión al identificar estudiantes que no estudiaron.

# Conclusión



## Precisión Moderada y Sesgo

Ambos modelos (Naive Bayes y clustering) lograron una precisión del 64.2%, mostrando una mayor efectividad para predecir a quienes "no" completaron el curso de preparación, y menos para quienes "sí" lo hicieron.

## Limitaciones del Naive Bayes

El bajo rendimiento del modelo Naive Bayes se relaciona con su supuesto de independencia de variables, el cual podría no cumplirse por completo en los datos, afectando su capacidad predictiva.

## Desequilibrio de Datos

Un problema común en ambos modelos es el desequilibrio de datos, donde hay más ejemplos de personas que "no estudiaron", dificultando que los modelos aprendan a reconocer con precisión los casos positivos.

## Recomendaciones y Mejoras

Para futuras mejoras, se sugiere explorar modelos más sofisticados (ej. árboles de decisión), aplicar técnicas de balanceo de clases e incorporar más variables que aporten información útil.