Informe Académico: Predicción del Abandono Estudiantil

El presente informe describe el proceso de análisis y modelado realizado para predecir el abandono estudiantil utilizando un algoritmo de Árbol de Decisión. El objetivo fue identificar los factores que más influyen en la continuidad académica de los estudiantes, a partir de un conjunto de datos institucionales. El trabajo se dividió en dos etapas principales: entrenamiento del modelo con datos históricos y aplicación del mismo a un conjunto de estudiantes nuevos.

1. Metodología

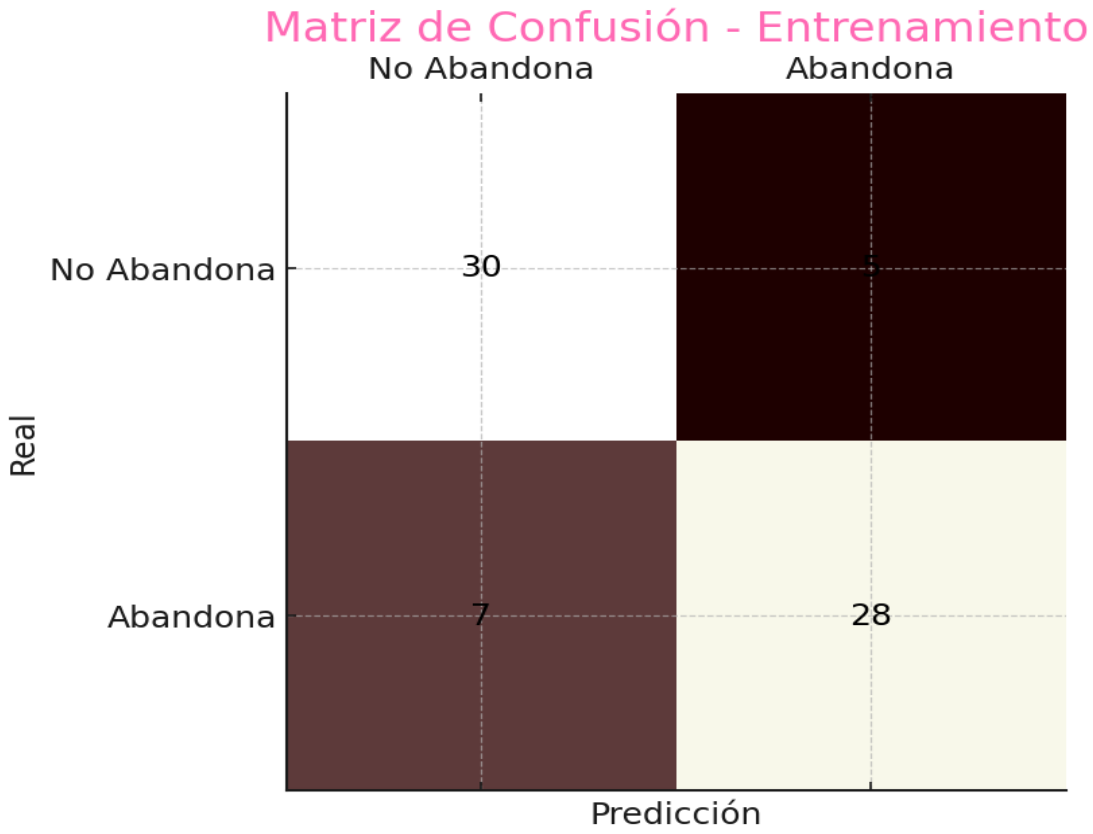
El procedimiento metodológico siguió varias fases:
1. **Carga y limpieza de datos**: se eliminaron valores nulos y se homogeneizó la variable estado_final.
2. **Codificación de variables categóricas**: conversión a valores numéricos mediante códigos de categoría.
3. **División de datos**: 70% para entrenamiento y 30% para prueba, con semilla aleatoria fija para replicabilidad.
4. **Entrenamiento del modelo**: se utilizó un Árbol de Decisión con criterio de entropía, profundidad máxima de 6 y un mínimo de 5 muestras por hoja.
5. **Evaluación**: se midió el rendimiento con matrices de confusión y precisión global.
6. **Predicción**: el modelo se aplicó a un dataset final, generando una columna de predicciones que permiten anticipar el abandono.

2. Resultados del modelo

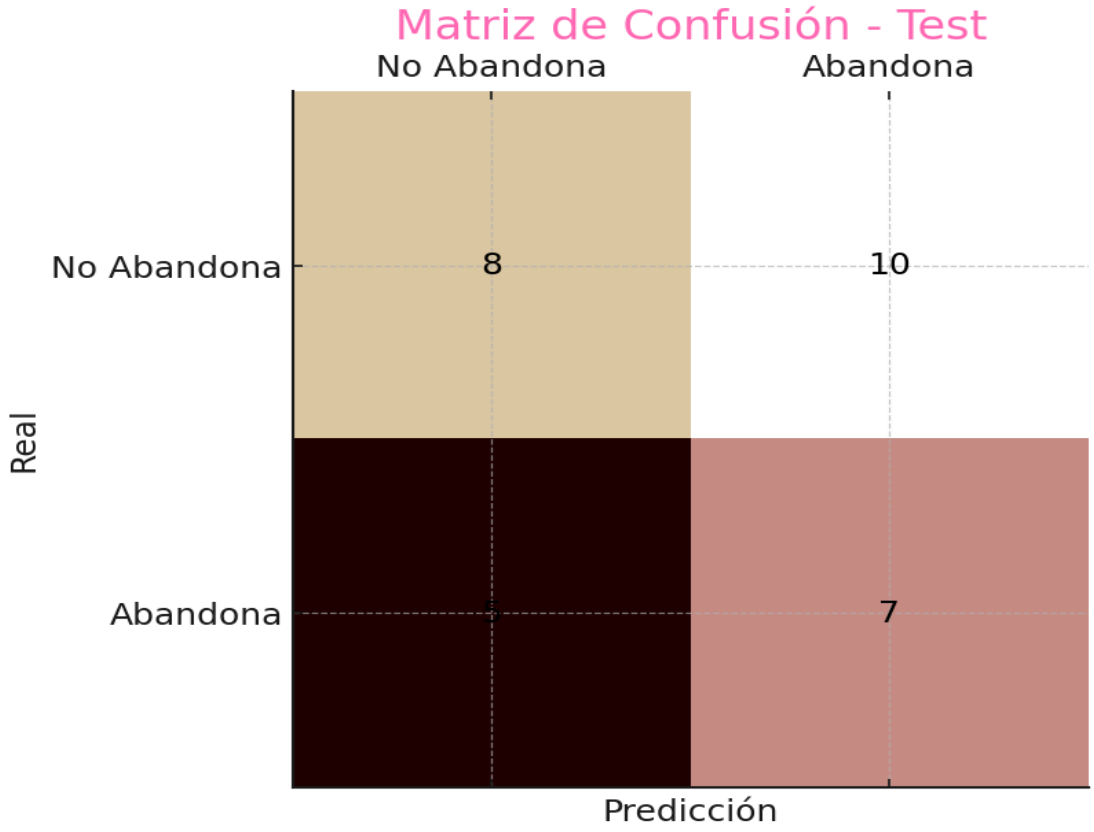
En términos de desempeño, el modelo alcanzó una precisión de **82.9%** en entrenamiento y **50.0%** en prueba. Esto muestra un fenómeno de sobreajuste: el árbol logra aprender muy bien los datos conocidos, pero pierde capacidad de generalización ante información nueva. Las matrices de confusión muestran que el modelo clasifica con mayor acierto a quienes no abandonan, aunque también identifica correctamente a una parte significativa de quienes sí lo hacen.

Conjunto	Precisión
Entrenamiento	82.9%
Prueba	50.0%

Matriz de confusión – Entrenamiento



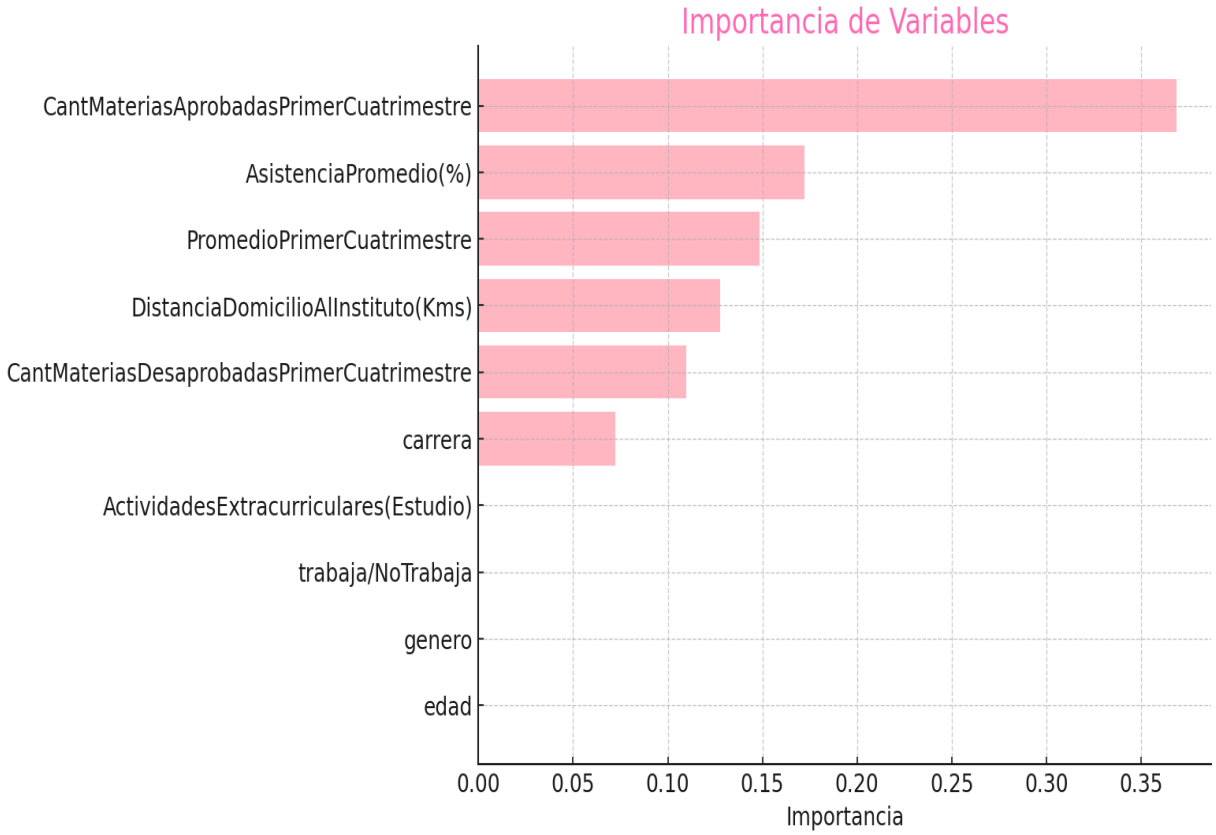
Matriz de confusión – Prueba



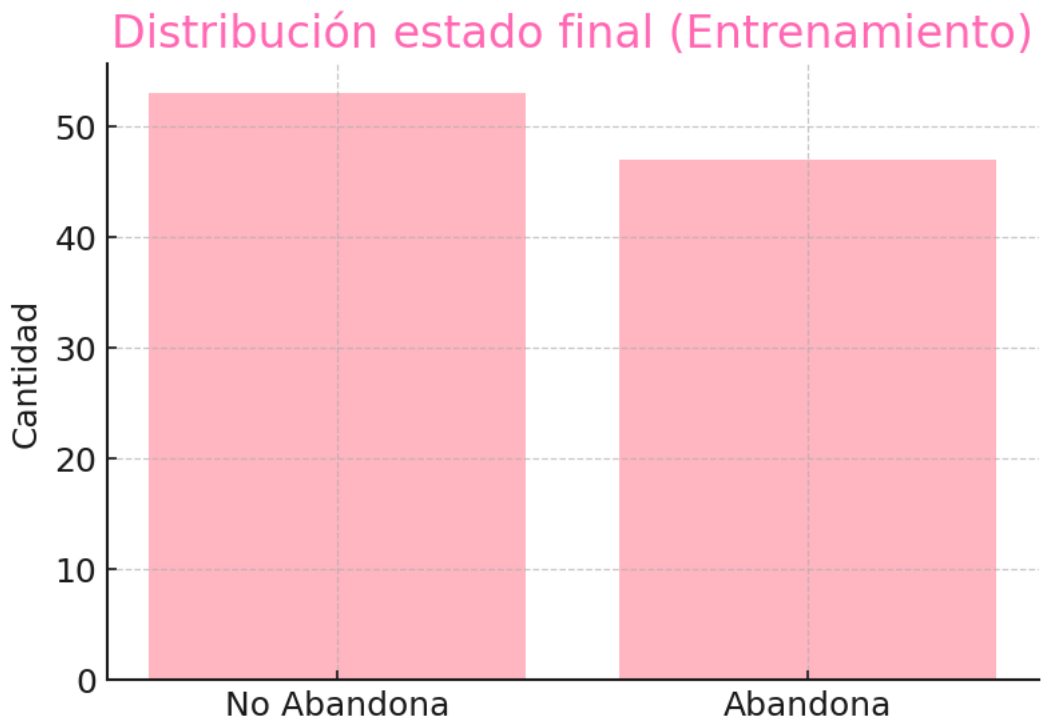
3. Importancia de las variables

Respecto a la importancia de las variables, se observa que la asistencia promedio y el rendimiento académico constituyen los factores más relevantes. Otros elementos influyentes son la cantidad de materias

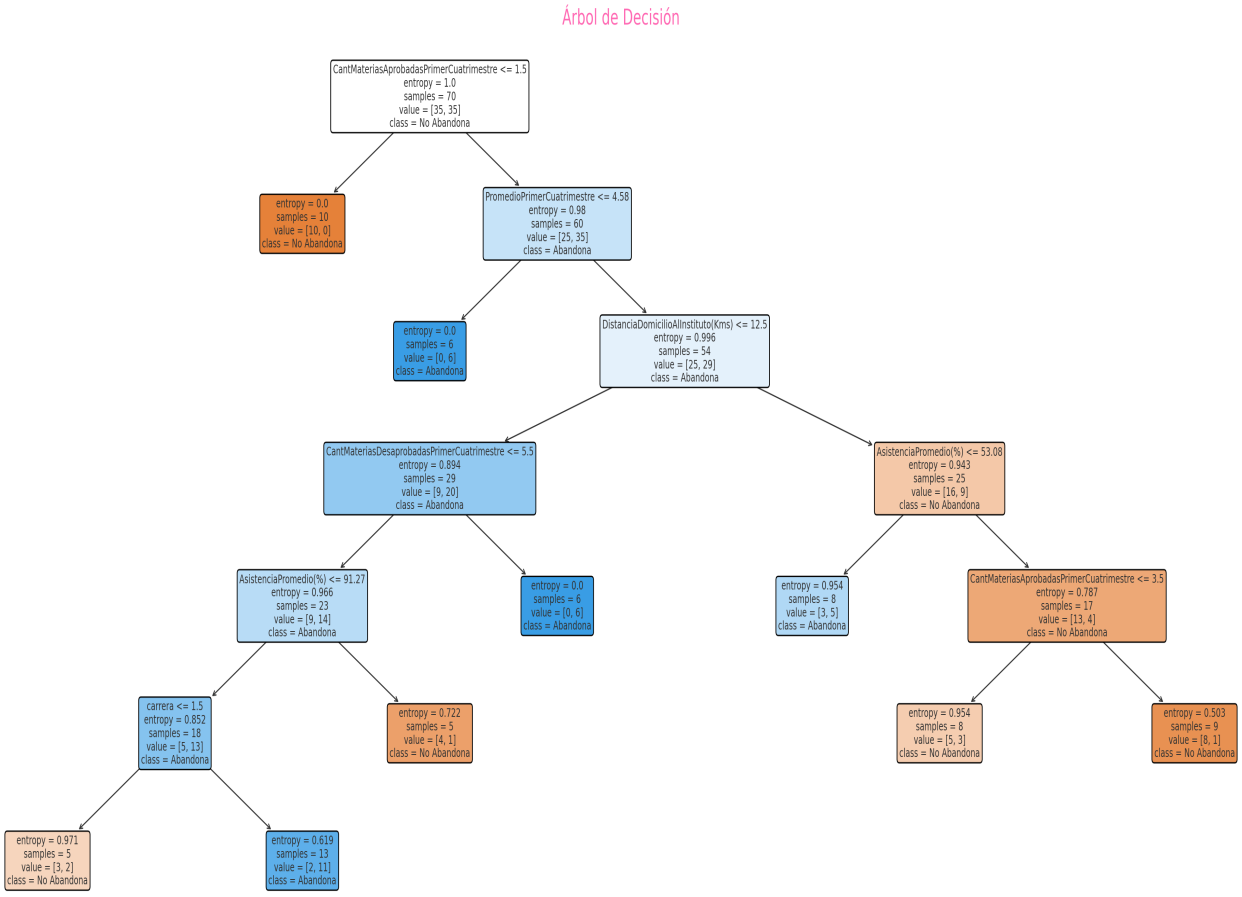
aprobadas/desaprobadas y la situación laboral de los estudiantes. Estos hallazgos permiten orientar políticas de acompañamiento y apoyo.



4. Distribución del estado final



5. Árbol de decisión completo



6. Aplicación a datos finales

Se aplicó el modelo a 100 estudiantes. A continuación se muestran las primeras 10 predicciones:

Estudiante	Predicción Abandono
1	Sí
2	Sí
3	Sí
4	Sí
5	No
6	No
7	Sí
8	Sí
9	Sí
10	No

7. Conclusión y recomendaciones

El modelo de Árbol de Decisión constituye una herramienta valiosa para comprender y anticipar el abandono académico, aunque presenta limitaciones de generalización. Se recomienda avanzar hacia modelos más complejos como Random Forest o Gradient Boosting, además de ampliar la base de datos histórica. La institución puede utilizar estos resultados para diseñar estrategias preventivas tales como tutorías personalizadas, programas de acompañamiento para estudiantes trabajadores y seguimiento digital de la asistencia. En conclusión, este estudio demuestra la potencialidad de la analítica de datos para mejorar la gestión educativa y reducir el abandono estudiantil.