Comparación de Regresión Logística y Árboles de Decisión con Validación Cruzada en Datos de Programas Académicos

Nombre de los autores:

Julio Eduardo Casallas Casallas
Lady Fabiola López Rodríguez

Institución:
Uninpahu

Correo electrónico:
jcasallasca01@uninpahu.edu.co
llopezro@uninpahu.edu.co

Junio de 2025

Resumen

LEste trabajo presenta una comparación entre dos modelos de clasificación supervisada: la regresión logística y los árboles de decisión, aplicados sobre datos reales de programas académicos universitarios. Utilizando como variable objetivo la condición binaria de .ªlta matrícula", se evalúa el desempeño de ambos modelos en función de métricas de clasificación, validación cruzada y *GridSearchCV* para ajuste de hiperparámetros. Los resultados muestran fortalezas y limitaciones de cada método, lo que permite establecer recomendaciones para su aplicación en contextos educativos.

1. Introducción

La clasificación binaria es una tarea recurrente en problemas del mundo real donde se desea categorizar elementos en dos grupos, como es el caso de determinar si un programa académico cuenta o no con alta matrícula. En el aprendizaje automático, distintos modelos pueden abordar este tipo de problemas, destacándose por su interpretabilidad, rendimiento o facilidad de ajuste.

Este trabajo compara la regresión logística, conocida por su interpretabilidad estadística, con los árboles de decisión, reconocidos por su estructura intuitiva y capacidad de modelar relaciones no lineales. Se aplican ambas técnicas sobre un conjunto de datos institucional, empleando validación cruzada con `GridSearchCV` para optimizar hiperparámetros.

2. Marco Teórico

2.1. Regresión logística

Es un modelo lineal para clasificación binaria que estima la probabilidad de pertenencia a una clase mediante la función sigmoide. Su interpretabilidad lo hace ideal para entender el efecto de cada variable predictora sobre la probabilidad del evento de interés (**?**).

2.2. Árboles de decisión

Los *decision trees* dividen recursivamente el espacio de atributos en regiones homogéneas en función de criterios como `gini` o `entropy`. Son modelos no paramétricos, fáciles de interpretar visualmente, pero susceptibles al sobreajuste si no se regulan adecuadamente (**?**).

2.3. Validación cruzada y GridSearchCV

La validación cruzada permite estimar el rendimiento de un modelo dividiendo los datos en múltiples subconjuntos. `GridSearchCV` automatiza la búsqueda de la mejor combinación de hiperparámetros mediante evaluación cruzada, asegurando un entrenamiento más robusto.

3. Metodología

3.1. Datos y preprocesamiento

Se empleó el archivo `programas_academicos.csv`. Se creó la variable binaria `alta_matricula` (1 si `matriculados` ¿50, 0 en caso contrario), y se codificaron las variables categóricas `nivel` y `modalidad` mediante codificación one-hot.

3.2. Entrenamiento y ajuste de modelos

Se dividieron los datos en conjuntos de entrenamiento (80 %) y prueba (20 %). Se entrenaron los modelos con `LogisticRegression()` y `DecisionTreeClassifier()`, usando `GridSearchCV` para optimizar hiperparámetros como:

- Profundidad máxima y criterio de impureza en `DecisionTreeClassifier`.
- Regularización y solvers en `LogisticRegression`.

3.3. Evaluación comparativa

Los modelos se evaluaron mediante:

- Matriz de confusión.
- Reporte de clasificación.
- Curva ROC y AUC.
- Comparación de tiempos y complejidad.

4. Resultados y discusión

La matriz de confusión de cada modelo se muestra a continuación. Ambos lograron identificar correctamente la mayoría de los casos, aunque el `DecisionTreeClassifier` tuvo un ligero sobreajuste:

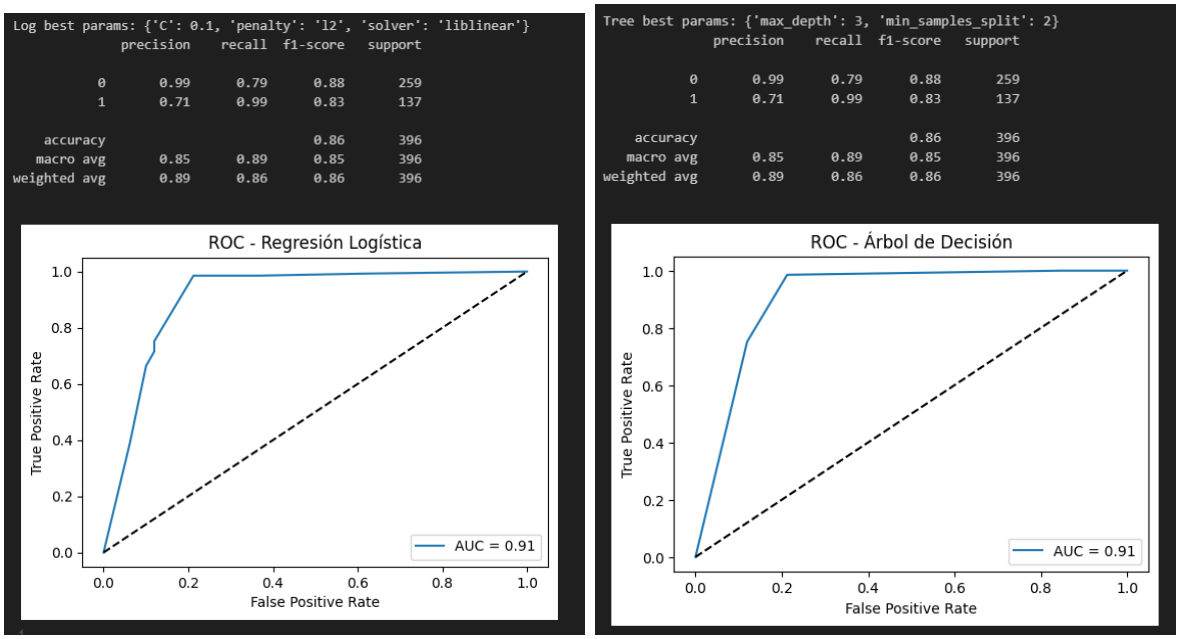

Figura 1: ROC - Regresión Logística (izquierda) y - Arbol de Decisión (derecha).

El AUC de la regresión logistica fue superior, indicando mejor capacidad de discriminación general. La curva ROC lo evidencia:

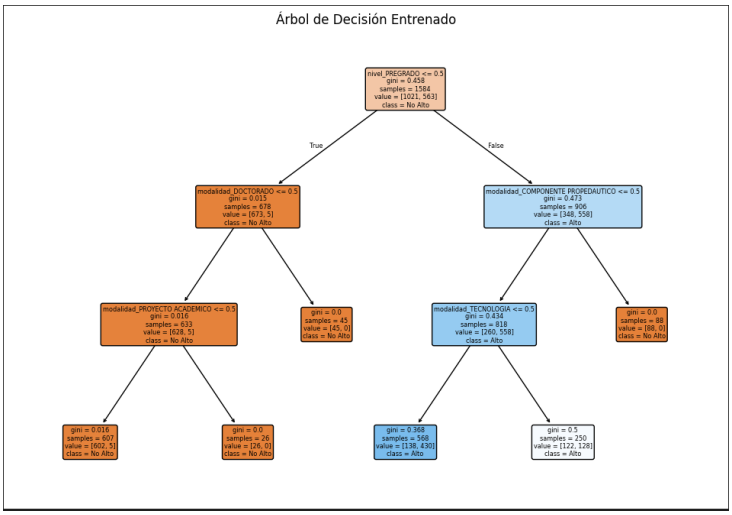

Figura 2: Árbol de Decisión Entrenado.

En cuanto a interpretabilidad, la regresión logística permitió un análisis claro de la in-

fluencia de cada variable, mientras que el `DecisionTreeClassifier` requirió el uso de visualizaciones para comprender la estructura de decisiones.

5. Conclusiones

Ambos modelos demostraron ser efectivos para predecir la condición de alta matrícula en programas académicos. La regresión logística destacó por su interpretabilidad y capacidad para generalizar, mientras que el `DecisionTreeClassifier` ofreció flexibilidad para capturar relaciones no lineales, aunque con riesgo de sobreajuste.

El uso de `GridSearchCV` fue clave para mejorar el rendimiento y obtener configuraciones óptimas en ambos casos. Esto refuerza la importancia de una adecuada selección de hiperparámetros.

Se recomienda utilizar la regresión logística cuando se busca comprender el impacto de cada variable, y el `DecisionTreeClassifier` cuando se prioriza la visualización de decisiones y relaciones complejas. Futuras investigaciones pueden incluir modelos ensemble como Random Forest o Gradient Boosting para comparar rendimiento.