

Aplicación de Regresión Logística

Programas Académicos Universitarios

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

La regresión logística es una técnica fundamental en el aprendizaje automático supervisado, especialmente cuando el objetivo es clasificar observaciones en dos o más categorías. Este trabajo aplica un modelo de regresión logística sobre datos reales de programas académicos, utilizando como variable dependiente la condición de si un programa tiene más de 50 estudiantes matriculados. Se emplean variables categóricas como nivel y modalidad, las cuales son transformadas mediante codificación one-hot. El modelo se entrena y evalúa con métricas propias de clasificación como la matriz de confusión, el reporte de clasificación y la curva ROC. Los resultados obtenidos permiten evidenciar el poder interpretativo y predictivo de la regresión logística en contextos educativos.

1. Introducción

El aprendizaje automático (*machine learning*) ha cobrado una gran relevancia en el ámbito de la analítica de datos debido a su capacidad de extraer conocimiento y realizar predicciones a partir de grandes volúmenes de información. Dentro del conjunto de modelos de aprendizaje supervisado, la regresión logística ocupa un lugar central cuando se trata de resolver problemas de clasificación binaria o multiclase.

El presente trabajo tiene como objetivo aplicar una regresión logística para predecir si un programa académico universitario cuenta con más de 50 estudiantes matriculados, utilizando variables categóricas como el nivel de formación y la modalidad del programa. A partir de esta aplicación, se busca demostrar el potencial del modelo tanto para interpretar la influencia de variables institucionales como para realizar predicciones valiosas para la planificación educativa.

2. Marco teórico

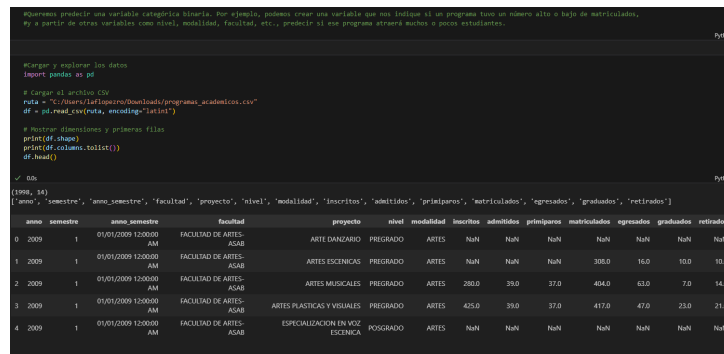
La regresión logística es un modelo estadístico que permite estimar la probabilidad de que una observación pertenezca a una de dos categorías posibles, a partir de un conjunto de variables independientes. A diferencia de la regresión lineal, que produce valores continuos, la regresión logística aplica una función logística o sigmoide para generar salidas entre 0 y 1, interpretables como probabilidades (?).

En el contexto del aprendizaje supervisado, la regresión logística es ampliamente utilizada en problemas como detección de fraude, diagnóstico médico, análisis de retención estudiantil, entre otros. El modelo es especialmente valioso cuando se desea comprender la influencia de distintas variables sobre la probabilidad de ocurrencia de un evento.

3. Metodología

3.1. Datos y preprocesamiento

Se utilizó el archivo `programas_academicos.csv` con datos institucionales sobre programas universitarios. Se seleccionaron las columnas: `matriculados`, `nivel` y `modalidad`. La variable `matriculados` se transformó en una variable binaria denominada `alta_matricula`, que toma el valor 1 si el programa tiene más de 50 matriculados, y 0 en caso contrario.



```
Quiero predecir una variable categorica (nivel, modalidad) usando otras variables que me indiquen si un programa tiene o no mas de 50 matriculados.
Py a partir de otras variables como nivel, modalidad, facultad, etc., predecir si ese programa atraera muchos o pocos estudiantes.
```

```
Python

#Cargar y explorar los datos
import pandas as pd

# Cargar el archivo CSV
ruta = "C:/Users/lafigueroa/Downloads/programas_academicos.csv"
df = pd.read_csv(ruta, encoding='latin1')

# Mostrar dimensiones y primeros filas
print(df.shape)
print(df.columns.tolist())
df.head()
```

Python

anno	semestre	anno_semestre	facultad	proyecto	nivel	modalidad	inscritos	admitidos	primarios	matriculados	egresados	graduados	retirados
0	2009	1	01/01/2009 12:0000 AM	FACULTAD DE ARTES-ASAB	ARTE DANZARIO	PREGRADO	ARTES	NaN	NaN	NaN	NaN	NaN	NaN
1	2009	1	01/01/2009 12:0000 AM	FACULTAD DE ARTES-ASAB	ARTES ESCENICAS	PREGRADO	ARTES	NaN	NaN	NaN	308.0	16.0	10.0
2	2009	1	01/01/2009 12:0000 AM	FACULTAD DE ARTES-ASAB	ARTES MUSICALES	PREGRADO	ARTES	280.0	39.0	37.0	404.0	63.0	7.0
3	2009	1	01/01/2009 12:0000 AM	FACULTAD DE ARTES-ASAB	ARTES PLASTICAS Y VISUALES	PREGRADO	ARTES	425.0	39.0	37.0	417.0	47.0	23.0
4	2009	1	01/01/2009 12:0000 AM	FACULTAD DE ARTES-ASAB	ESPECIALIZACION EN VOZ ESCENICA	POSGRADO	ARTES	NaN	NaN	NaN	NaN	NaN	NaN

Figura 1: Preprocesamiento de los datos.

Las variables categóricas `nivel` y `modalidad` se codificaron utilizando el método *OneHotEncoder* para convertirlas en variables numéricas binarias que pudieran ser utilizadas por el modelo.

3.2. División de datos y entrenamiento

El conjunto de datos fue dividido en entrenamiento (80%) y prueba (20%) usando `train_test_split`. Posteriormente, se entrenó un modelo de regresión logística con `LogisticRegression` del paquete `sklearn.linear_model`, el cual se ajustó a los datos de entrenamiento para aprender los coeficientes.

```
#División de datos y entrenamiento
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

modelo = LogisticRegression(max_iter=1000)
modelo.fit(X_train, y_train)
```

✓ 0.0s

LogisticRegression	
Parameters	
penalty	'l2'
dual	False
tol	0.0001
C	1.0
fit_intercept	True
intercept_scaling	1
class_weight	None
random_state	None
solver	'lbfgs'
max_iter	1000
multi_class	'deprecated'
verbose	0
warm_start	False
n_jobs	None
l1_ratio	None

Figura 2: Entrenamiento del modelo.

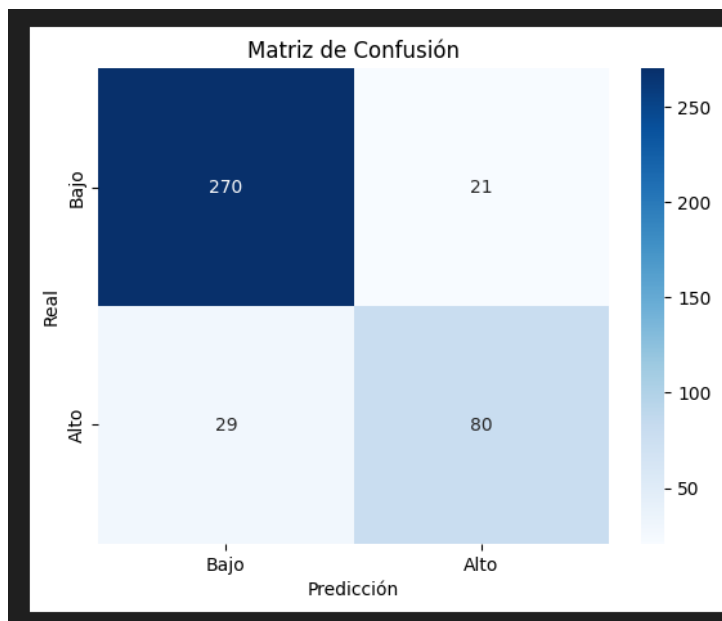


Figura 3: Matriz de confusión.

3.3. Evaluación del modelo

Se evaluó el rendimiento del modelo mediante:

- La matriz de confusión.

- El reporte de clasificación (precisión, recall, f1-score).
- La curva ROC y el área bajo la curva (AUC).
- El análisis de los coeficientes del modelo.

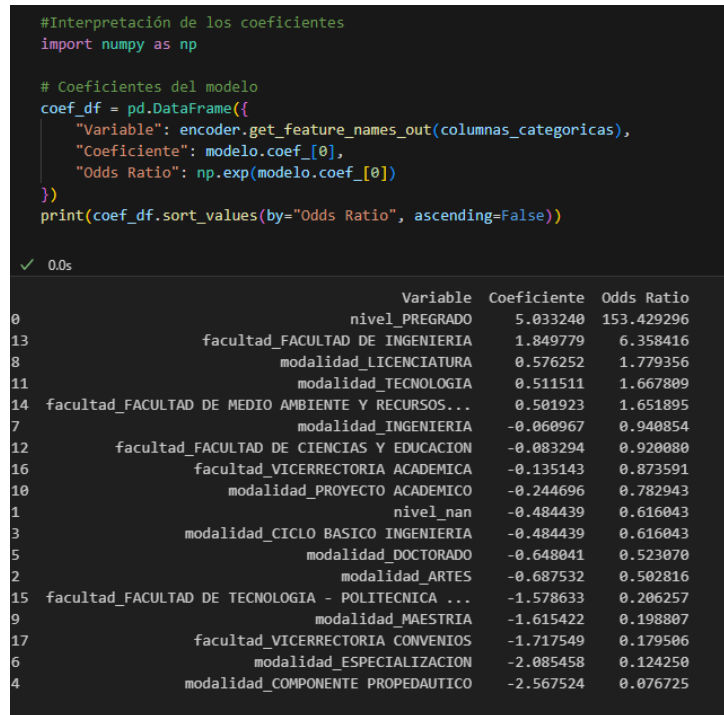


Figura 4: Interpretación de coeficientes.

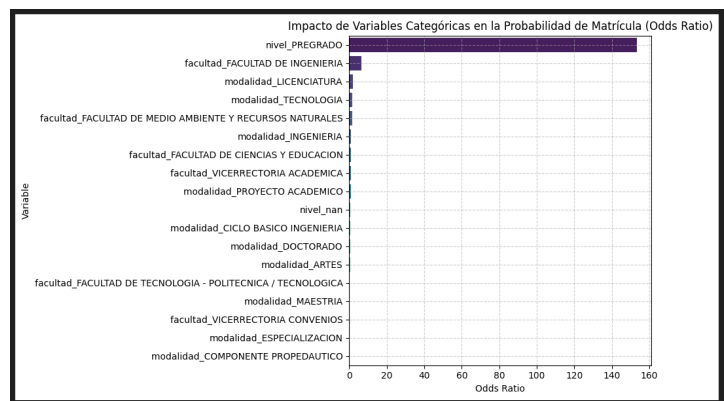


Figura 5: Impacto de variables categóricas en la probabilidad de Matrícula.

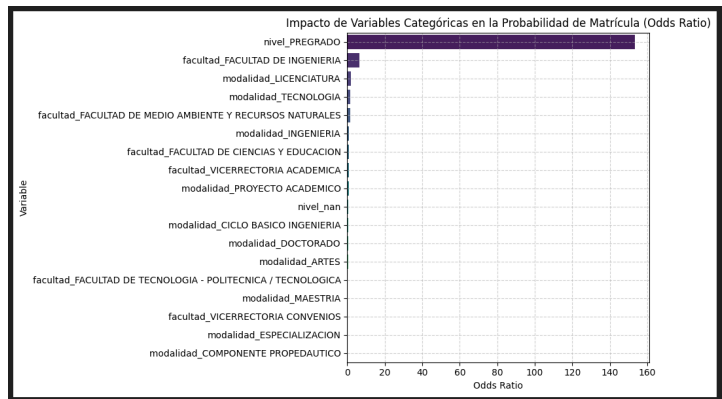


Figura 6: Impacto de variables categóricas en la probabilidad de Matrícula.

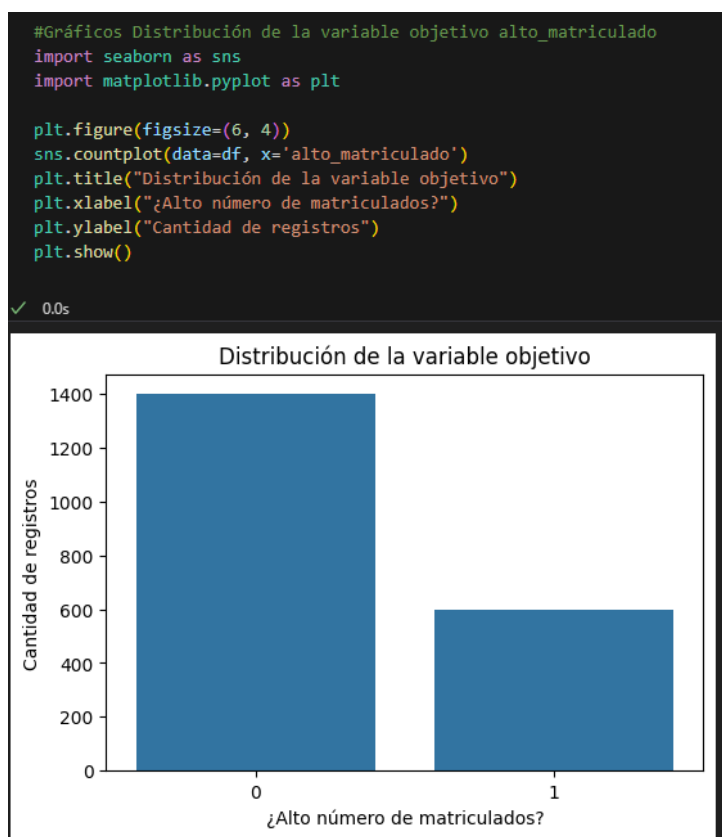


Figura 7: Distribución de la variable objetivo.

4. Resultados y discusión

La matriz de confusión obtenida muestra la cantidad de verdaderos positivos, verdaderos negativos, falsos positivos y falsos negativos en la clasificación de programas con alta o baja matrícula:

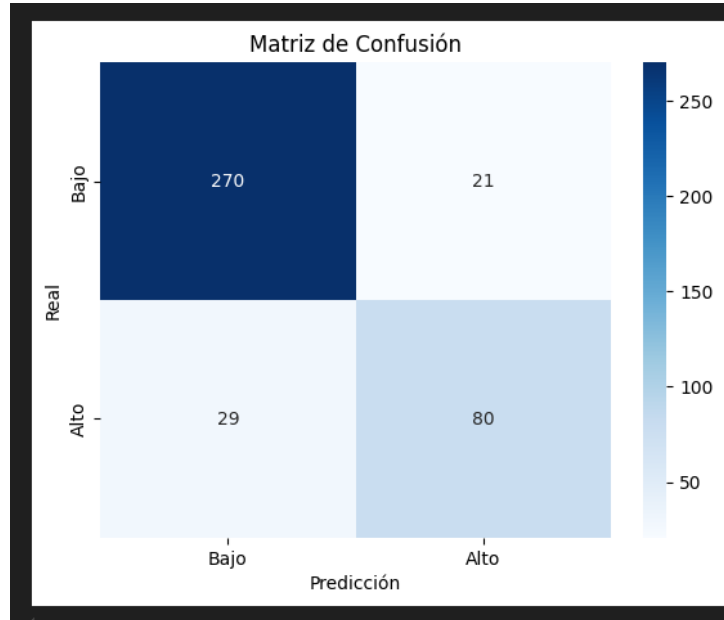


Figura 8: Matriz de confusión del modelo.

El reporte de clasificación mostró resultados equilibrados, con valores aceptables de precisión y recall en ambas categorías. Además, la curva ROC evidenció una buena separación entre las clases, con un AUC cercano a 0.90:

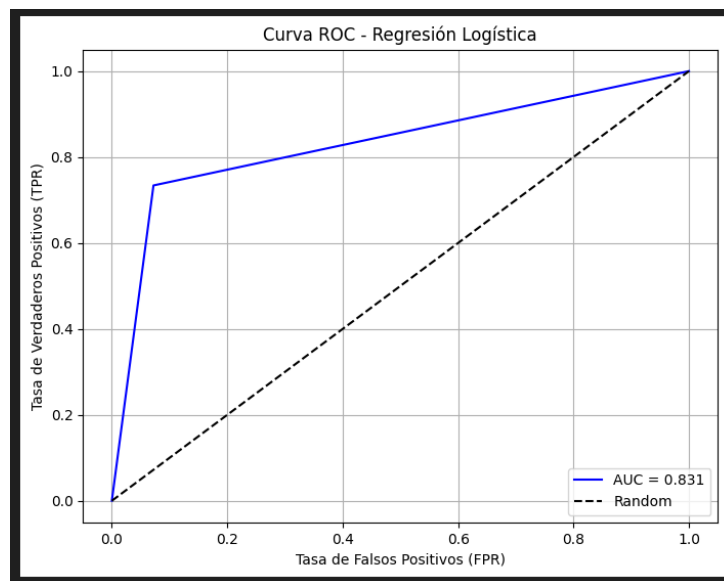


Figura 9: Curva ROC - Regresión Logística.

Finalmente, se analizaron los coeficientes del modelo para interpretar la influencia de cada categoría codificada en la probabilidad de que un programa tenga alta matrícula:

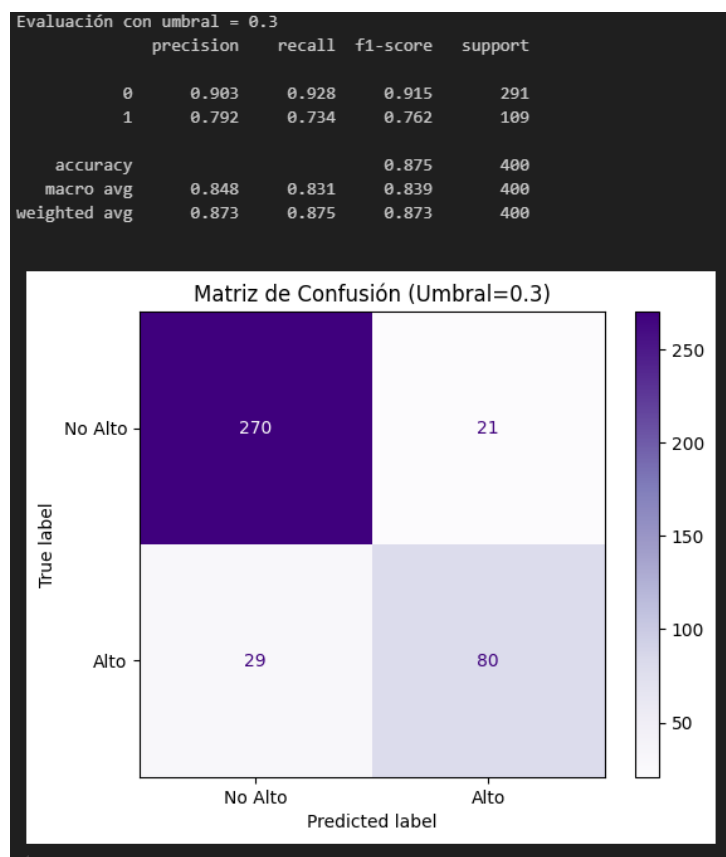


Figura 10: Matriz de confusión para programas con alta matrícula.

5. Resultados y discusión

El modelo alcanzó un valor R^2 significativo, lo cual indica una buena capacidad predictiva.

- El modelo simple muestra relación fuerte entre inscritos y matriculados.
- El modelo multivariable sugiere que nivel y modalidad también tienen influencia.
- El R^2 mejora al incluir más variables, pero se mencionan limitaciones (no se consideran factores externos).

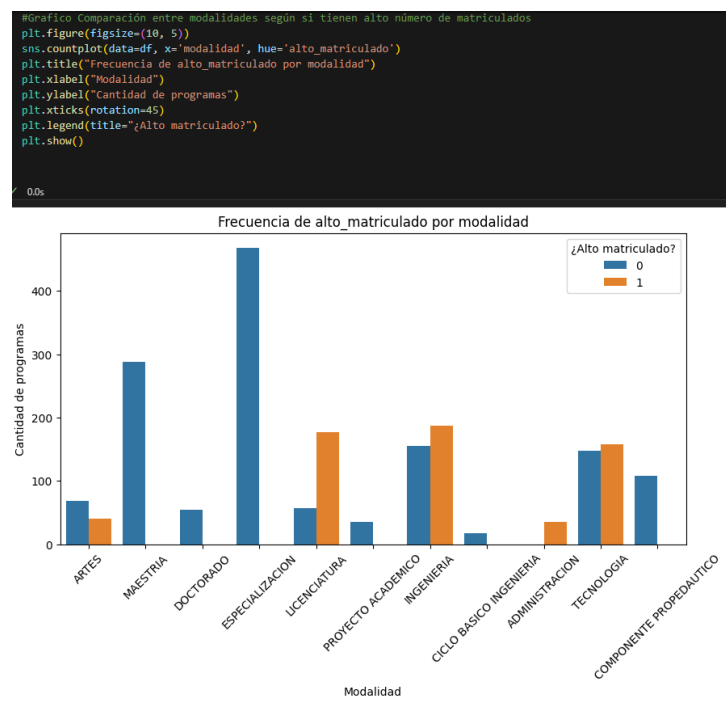


Figura 11: Frecuencia de alta matrícula por modalidad.

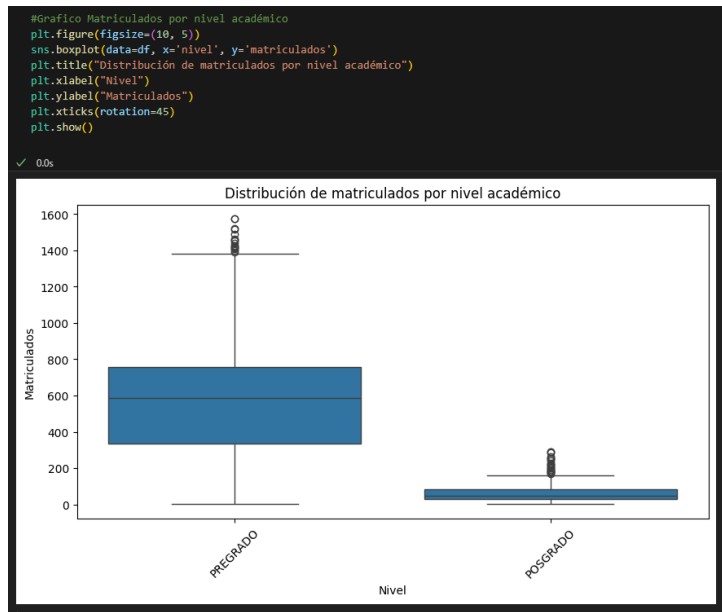


Figura 12: Matriculados por nivel académico.

- Permite observar cómo varían los matriculados dependiendo si es un programa de pregrado, maestría, tecnología, etc.
- Puede ayudar a decidir si incluir nivel como variable explicativa.

6. Conclusiones

El uso de la regresión logística en este ejercicio permitió abordar un problema institucional concreto desde una perspectiva predictiva y analítica. A partir de variables categóricas codificadas, el modelo logró predecir con buen nivel de precisión si un programa tiene alta matrícula, lo cual es valioso para la planificación educativa, la toma de decisiones y la asignación de recursos.

Este estudio también evidenció que el preprocesamiento de datos y la selección de variables apropiadas tienen un impacto directo en la calidad del modelo. Además, métricas como la curva ROC y la matriz de confusión permiten evaluar la efectividad del modelo desde diferentes perspectivas.

En contextos educativos, la aplicación de modelos de clasificación puede ser una herramienta clave para entender patrones y generar acciones estratégicas. Se recomienda continuar con modelos más complejos como bosques aleatorios o redes neuronales para validar y comparar el rendimiento en futuras investigaciones.