Universidad Católica San Pablo

Laboratorio: Regresión Logística

Computer Science — Tópicos en Inteligencia Artificial

Harold Alejandro Villanueva Borda

1. Introducción

La regresión logística es una técnica de clasificación ampliamente utilizada en problemas de clasificación binaria. Este informe detalla los experimentos realizados con un modelo de regresión logística, aplicado a un conjunto de datos para predecir una variable objetivo binaria. Los datos fueron revisados y preparados para garantizar la calidad del modelado. Se llevaron a cabo múltiples pruebas con diferentes enfoques de preprocesamiento y modelado, incluyendo la normalización de datos, selección de características, y evaluación de métricas de rendimiento, como la curva ROC y el análisis de la importancia de características.

Este informe documenta el proceso completo, desde la revisión inicial de los datos hasta la implementación y evaluación del modelo de regresión logística. A través de una cuidadosa inspección y ajustes en el modelado, se busca determinar qué enfoques ofrecen los mejores resultados y cómo la selección de características afecta el desempeño del modelo.

2. Experimento

2.1. Revisión de Datos

Los datos cargados inicialmente fueron revisados para identificar cualquier anomalía que pudiera impactar en el rendimiento del modelo. El conjunto de datos cuenta con un total de N ejemplos distribuidos en varias características predictivas. A continuación, se verificó si había ejemplos con valores faltantes.

Se detectaron algunos ejemplos con valores faltantes en ciertas características. Para manejar estos valores, se tomaron los siguientes enfoques:

- Eliminación de ejemplos con múltiples características faltantes, ya que estos podrían introducir ruido o sesgo en el modelo.

- Imputación de valores faltantes utilizando la mediana para características numéricas, con el fin de preservar la mayor cantidad de datos posible.

2.2. Inspección de la Frecuencia de Clases

Una de las primeras inspecciones realizadas fue verificar la distribución de la variable objetivo. En la Figura 1 podemos que la distribución entre las dos clases fue evaluada visualmente y numéricamente, lo que nos permitió determinar si el conjunto de datos estaba balanceado o no. Los resultados indicaron que las clases no estaban completamente equilibradas:

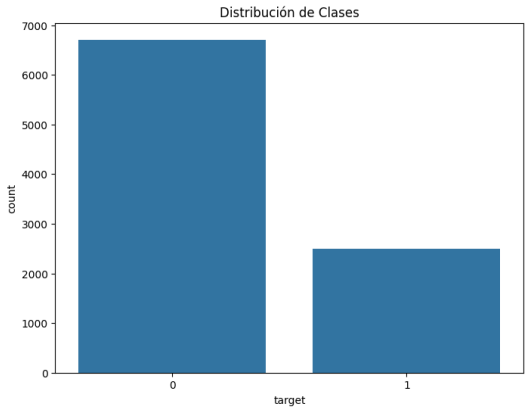

Figura 1: Inspección de la frecuencia de clases

- Clase 0: Representa el 72.9 % de los ejemplos.
- Clase 1: Representa el 27.1 % de los ejemplos.

Este desbalance de clases puede afectar negativamente el rendimiento del modelo, favoreciendo las predicciones de la clase mayoritaria. Por lo tanto, se evaluaron técnicas para manejar este desbalance, como el sobremuestreo de la clase minoritaria o el uso de ponderación de clases en el modelo.

2.3. Normalización de los Datos

Para mejorar el rendimiento del modelo de regresión logística, se aplicó una normalización de características. Esto es particularmente útil cuando las características tienen diferentes escalas. La normalización convierte las características a un rango común, eliminando sesgos que podrían provenir de escalas muy diferentes.

2.4. Prueba 1: Modelo Base

El primer experimento fue el entrenamiento de un modelo de regresión logística sin realizar una selección previa de características. El modelo base se entrenó con todas las características disponibles, sin ningún preprocesamiento adicional. Los resultados iniciales se evaluaron en términos de precisión y la métrica AUC (Área Bajo la Curva ROC).

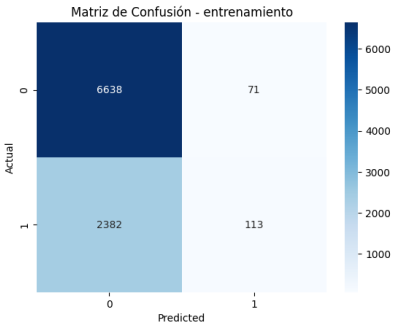

Figura 2: Matrix vs Train

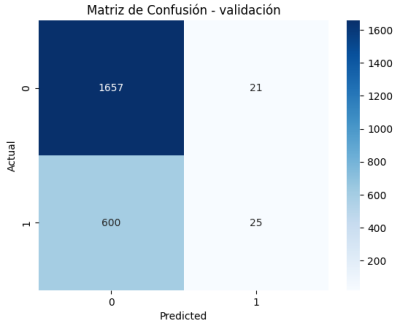

Figura 3: Matrix vs Validation

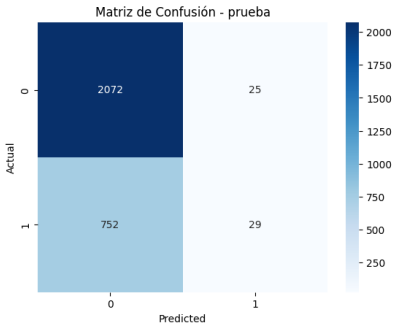

Figura 4: Matrix vs Test

2.4.1. Conjunto de Validación:

Matriz de Confusión:

$$\begin{bmatrix} 6638 & 71 \\ 2382 & 113 \end{bmatrix}$$

Métricas:

- Accuracy: 0.7335
- Tasa de Verdaderos Positivos (TPR): 0.0453
- Tasa de Verdaderos Negativos (TNR): 0.9894

La matriz muestra que:

- 8639 son verdaderos negativos (clase 0 correctamente clasificada).
- 71 son falsos positivos (clase 0 clasificada incorrectamente como 1).
- 2382 son falsos negativos (clase 1 clasificada incorrectamente como 0).
- 113 son verdaderos positivos (clase 1 correctamente clasificada).

Reporte de Clasificación:

	Precision	Recall	F1-Score	Support
0	0.74	0.99	0.84	6709
1	0.61	0.05	0.08	2495
Accuracy			**0.73**	9204
Macro Avg	0.68	0.52	0.46	9204
Weighted Avg	0.70	0.73	0.64	9204

Cuadro 1: Reporte de clasificación del modelo base

- Accuracy: 0.7835. El modelo tiene una precisión general del 78.35 % en el conjunto de entrenamiento.
- Tasa de Verdaderos Positivos (TPR): 0.0453, lo que indica que el modelo detecta el 4.53 % de los positivos reales (clase 1).
- Tasa de Verdaderos Negativos (TNR): 0.9918, el modelo detecta correctamente el 99.18 % de los negativos (clase 0).
- Precisión: 0.6144. Solo el 61.44 % de las predicciones positivas son correctas.
- Recall (sensibilidad): 0.0453. Muy bajo, el modelo no detecta bien los positivos.
- F1-Score: 0.0846. La media armónica entre precisión y recall muestra que el modelo tiene un rendimiento bajo para la clase positiva.

En la figura 15 el modelo en entrenamiento tiene un buen rendimiento para detectar la clase negativa (TNR alto), pero es pobre para detectar la clase positiva (TPR bajo y F1-score bajo). Esto indica que el modelo está desbalanceado y se inclina hacia la clase negativa (clase mayoritaria).

2.4.2. Conjunto de Validación:

Matriz de Confusión:

$$\begin{bmatrix} 1657 & 21 \\ 600 & 25 \end{bmatrix}$$

La matriz muestra que:

- 1637 son verdaderos negativos.
- 21 son falsos positivos.
- 600 son falsos negativos.
- 25 son verdaderos positivos.

Métricas:

- Accuracy: 0.7304
- Tasa de Verdaderos Positivos (TPR): 0.0400
- Tasa de Verdaderos Negativos (TNR): 0.9875

Reporte de Clasificación:

	Precision	Recall	F1-Score	Support
0	0.73	0.99	0.84	1678
1	0.54	0.04	0.07	625
Accuracy			**0.73**	2303
Macro Avg	0.64	0.51	0.46	2303
Weighted Avg	0.68	0.73	0.63	2303

Cuadro 2: Reporte de clasificación del conjunto de validación

Métricas:

- Accuracy: 0.7285. El modelo tiene una precisión general del 72.85 % en el conjunto de validación.
- TPR: 0.04. El modelo detecta el 4 % de los positivos reales.
- TNR: 0.9873. Detecta correctamente el 98.73 % de los negativos.
- Precisión: 0.5435. El 54.35 % de las predicciones positivas son correctas.
- Recall: 0.04. Muy bajo, el modelo sigue sin detectar bien los positivos.
- F1-Score: 0.0741. Refleja un bajo equilibrio entre precisión y recall.
- Soporte: 25 positivos reales en el conjunto de validación.

En la figura 16 el modelo muestra un comportamiento similar al de entrenamiento, con un fuerte sesgo hacia la clase negativa. La tasa de verdaderos positivos sigue siendo baja, lo que significa que el modelo no generaliza bien para la clase positiva.

2.4.3. Conjunto de Prueba:

Matriz de Confusión:

$$\begin{bmatrix} 2072 & 25 \\ 752 & 29 \end{bmatrix}$$

La matriz muestra que:

- 2072 son verdaderos negativos.

- 25 son falsos positivos.
- 752 son falsos negativos.
- 29 son verdaderos positivos.

Métricas:

- Accuracy: 0.7300
- Tasa de Verdaderos Positivos (TPR): 0.0371
- Tasa de Verdaderos Negativos (TNR): 0.9881

Reporte de Clasificación:

	Precision	Recall	F1-Score	Support
0	0.73	0.99	0.84	2097
1	0.54	0.04	0.07	781
Accuracy			**0.73**	2878
Macro Avg	0.64	0.51	0.46	2878
Weighted Avg	0.68	0.73	0.63	2878

Cuadro 3: Reporte de clasificación del conjunto de prueba

Métricas:

- Accuracy: 0.73. El modelo tiene una precisión general del 73 % en el conjunto de prueba.
- TPR: 0.0371. Detecta solo el 3.71 % de los positivos reales.
- TNR: 0.9881. Muy alta, el modelo detecta correctamente el 98.81 % de los negativos.
- Precisión: 0.5370. El 53.7 % de las predicciones positivas son correctas.
- Recall: 0.0371. Muy bajo, con un pobre rendimiento para la clase positiva.
- F1-Score: 0.0691. Refleja un bajo equilibrio entre precisión y recall.
- Soporte: 29 positivos reales en el conjunto de prueba.

En la figura 13 nuevamente, el modelo muestra un sesgo hacia la clase negativa en el conjunto de prueba. El TPR sigue siendo muy bajo (3.71 %), lo que indica que el modelo tiene dificultades para detectar la clase positiva en el conjunto de prueba.

Los tres gráficos muestran curvas ROC (Receiver Operating Characteristic) para diferentes conjuntos de datos: entrenamiento, validación y prueba. Estas curvas evalúan el rendimiento de un modelo de clasificación binaria.

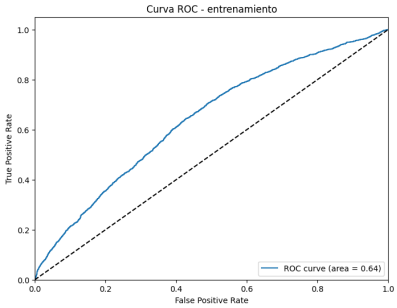

Figura 5: Entrenamiento

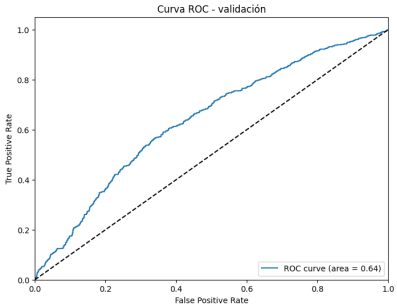

Figura 6: Validación

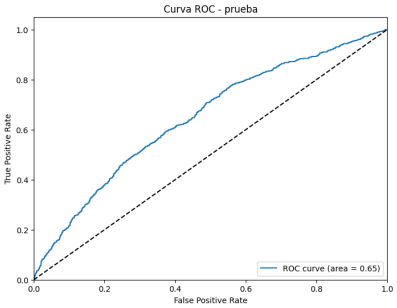

Figura 7: Prueba

1. Curva ROC - entrenamiento:
   - Área bajo la curva (AUC) = 0.64
   - Muestra un rendimiento moderado del modelo en los datos de entrenamiento.
2. Curva ROC - validación:
   - AUC = 0.64
   - El rendimiento en el conjunto de validación es similar al de entrenamiento, lo que sugiere que no hay sobreajuste significativo.
3. Curva ROC - prueba:
   - AUC = 0.65
   - El rendimiento en el conjunto de prueba es ligeramente mejor, pero consistente con los resultados anteriores.

Las tres curvas indican un rendimiento del modelo por encima del azar (línea diagonal punteada) pero con margen de mejora, ya que un AUC de 0.64-0.65 sugiere una capacidad discriminativa moderada del modelo en todos los conjuntos de datos.

3. Prueba 2: Modelo con Selección de Características

En este experimento, se seleccionaron las características más representativas utilizando técnicas de análisis de importancia de características. Se utilizó una selección basada en el impacto de cada característica en el modelo para filtrar aquellas menos relevantes. Luego, el modelo fue entrenado de nuevo con este subconjunto de características.

Conjunto de Entrenamiento:

- **Matriz de confusión:**
  - Verdaderos negativos (TN): 6672

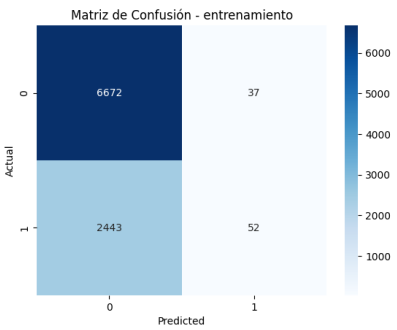

Figura 8: Matrix vs Train

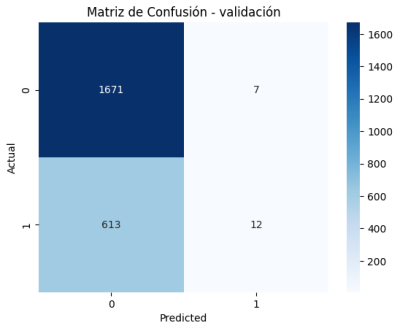

Figura 9: Matrix vs Validation

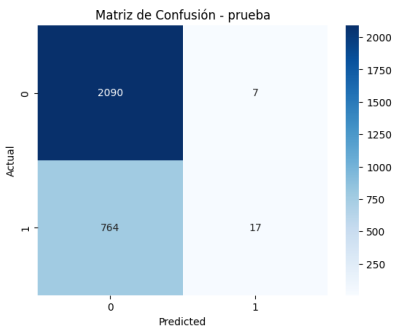

Figura 10: Matrix vs Test

  - Falsos negativos (FN): 2443
  - Falsos positivos (FP): 37
  - Verdaderos positivos (TP): 52
- **Accuracy**: 0.7306 (73.06 %)
  - El modelo clasifica correctamente el 73 % de los datos.
- **Tasa de Verdaderos Positivos (TPR o recall)**: 0.0208
  - Solo el 2.08 % de los positivos reales son correctamente clasificados, lo que indica un **muy bajo rendimiento** para la clase positiva.
- **Tasa de Verdaderos Negativos (TNR)**: 0.9945
  - El 99.45 % de los negativos reales son clasificados correctamente, mostrando un **muy buen rendimiento** para la clase negativa.
- **Métricas adicionales:**
  - La clase 0 (negativa) tiene una precisión y recall altos, pero la clase 1 (positiva) tiene un **f1-score bajo** (0.04) debido a su bajo recall (0.02), lo que sugiere que el modelo tiene dificultad en detectar los casos positivos.

Conjunto de Validación;

- **Matriz de confusión:**
  - TN: 1671, FN: 613, FP: 7, TP: 12
- **Accuracy**: 0.7308 (73.08 %)
- **TPR**: 0.0192 (1.92 %)

- **TNR**: 0.9958 (99.58 %)
  - Similar al conjunto de entrenamiento, el modelo muestra un **bajo rendimiento** en detectar la clase positiva, pero sigue siendo muy preciso para la clase negativa.
- **Métricas adicionales:**
  - La clase 1 (positiva) sigue teniendo un f1-score bajo de 0.04, con una alta precisión pero un recall muy bajo (0.02), lo que significa que el modelo no detecta bien los verdaderos positivos.

Conjunto de Prueba:

- **Matriz de confusión:**
  - TN: 2090, FN: 764, FP: 7, TP: 17
- **Accuracy**: 0.7321 (73.21 %)
- **TPR**: 0.0218 (2.18 %)
- **TNR**: 0.9967 (99.67 %)
  - Nuevamente, el modelo sigue mostrando un buen rendimiento en clasificar correctamente los negativos, pero tiene **problemas graves** en identificar la clase positiva (bajo TPR).
- **Métricas adicionales:**
  - En la clase 1 (positiva), el f1-score sigue siendo bajo (0.04), debido a un recall pobre (0.02), aunque la precisión es relativamente alta.

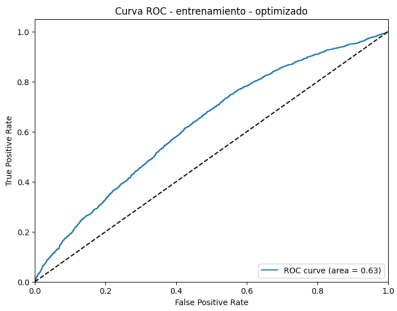

Figura 11: Entrenamiento

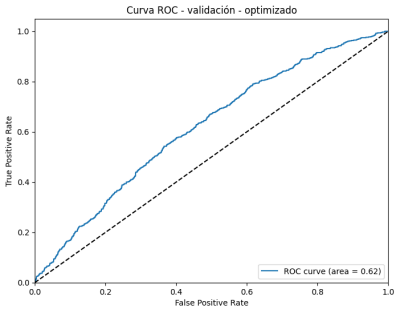

Figura 12: Validación

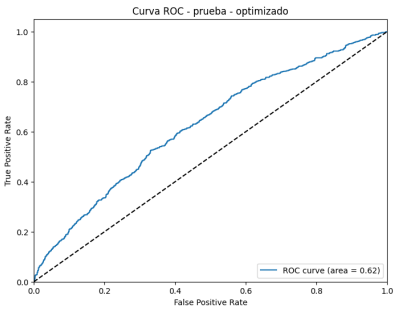

Figura 13: Prueba

Los tres gráficos muestran curvas ROC (Receiver Operating Characteristic) para diferentes conjuntos de datos: entrenamiento, prueba y validación, todos etiquetados como .ºptimizado".

1. Curva ROC - entrenamiento:

- Área bajo la curva (AUC) = 0.63
- Muestra el mejor rendimiento entre los tres conjuntos

2. Curva ROC - validación:
   - AUC = 0.62
   - Rendimiento ligeramente inferior al conjunto de entrenamiento
3. Curva ROC - prueba:
   - AUC = 0.62
   - Rendimiento similar al conjunto de prueba

Las tres curvas indican un rendimiento del modelo por encima del azar (línea diagonal punteada) pero con margen de mejora, ya que un AUC de 0.64-0.65 sugiere una capacidad discriminativa moderada del modelo en todos los conjuntos de datos.

4. Prueba 2: Modelo con Selección de Características

Las tres curvas están por encima de la línea diagonal (que representa el azar), lo que indica que el modelo tiene cierta capacidad predictiva. Sin embargo, con AUC alrededor de 0.62-0.63, el rendimiento del modelo es modesto. La similitud entre las curvas de prueba y validación sugiere que el modelo generaliza de manera consistente, pero hay margen de mejora en su capacidad predictiva global.

5. Comparación de Acuracia de los Modelos

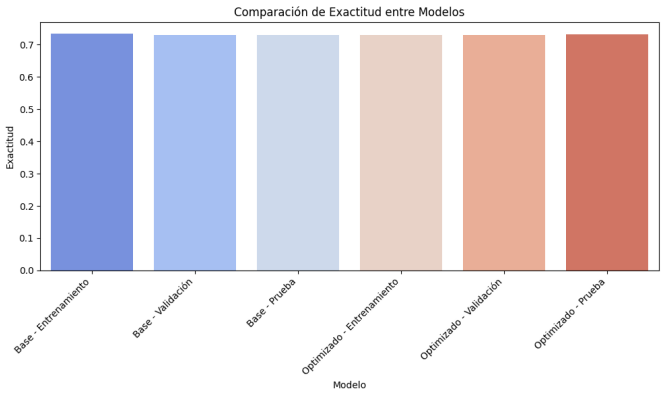

Figura 14: Entrenamiento

Modelos Base: La exactitud en el conjunto de entrenamiento, validación y prueba es bastante similar, alrededor de 0.7, lo que sugiere que el modelo base es consistente, pero su capacidad predictiva es limitada.

Modelos Optimizados: La exactitud de los modelos optimizados muestra una mejora ligera en comparación con el modelo base, aunque la diferencia no es significativa. Esto indica que el proceso de optimización ha logrado un ajuste algo mejor, pero aún hay margen de mejora.

6. Frontera de Decisión

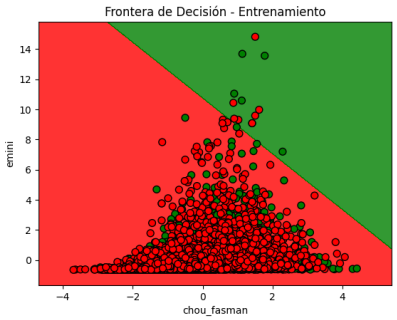

Figura 15: Entrenamiento

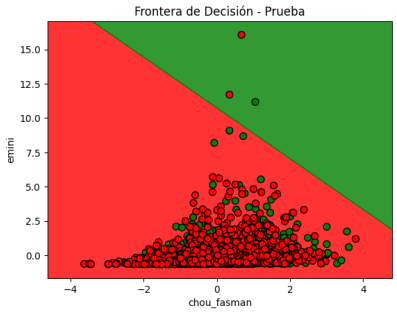

Figura 16: Prueba

Frontera de Decisión - Entrenamiento: El modelo ha separado los datos de entrenamiento en dos clases (rojo y verde) con una línea divisoria clara. Sin embargo, algunos puntos rojos están en la zona verde y viceversa, lo que indica que el modelo tiene cierta dificultad para separar completamente las clases.

Frontera de Decisión - Prueba: En los datos de prueba, la frontera es similar a la de entrenamiento, pero hay una mayor cantidad de puntos en la región incorrecta. Esto sugiere que el modelo tiene problemas de generalización y podría estar sobreajustado.

7. Conclusión

Este informe analizó dos modelos de regresión logística para un problema de clasificación binaria con datos desbalanceados. Los hallazgos clave son:

1. Ambos modelos (base y optimizado) mostraron un sesgo significativo hacia la clase mayoritaria, con alta precisión para la clase negativa (TNR ¿98 %) pero muy baja para la positiva (TPR ¡5 %).

2. La precisión general de ambos modelos fue de aproximadamente 73 %, pero esta métrica no refleja adecuadamente el rendimiento en clases desbalanceadas.

3. Los modelos demostraron una capacidad muy limitada para detectar la clase positiva (minoritaria), con valores F1 extremadamente bajos (alrededor de 0.07).

4. Las curvas ROC y los valores AUC (cerca de 0.64) indican un rendimiento moderado, apenas por encima del azar.

5. La visualización de las fronteras de decisión reveló una separación clara entre clases, pero con una cantidad significativa de puntos mal clasificados.

8. Repositorio

La implementación se encuentra en el siguiente respositorio: GitHub.