Evidencia: Módulo 2 Análisis y Reporte sobre el desempeño del modelo.

Jorge Daniel Rea Prado - A01747327

TC3006C Inteligencia artificial avanzada para la ciencia de datos I

Grupo 101

Profesor: Jorge Adolfo Ramírez Uresti

Campus Estado de México

Septiembre 2024

Análisis de modelo de Regresión Logística generado con un framework

Introducción

En este análisis, evaluaremos el desempeño de un modelo de Regresión Logística previamente entrenado utilizando la biblioteca scikit-learn con el objetivo de predecir si una canción será clasificada como "buena" o "mala" en función de varias características. Este modelo de clasificación binaria es útil para identificar patrones en las canciones y permitir predicciones en nuevos datos. Para llevar a cabo este análisis, generamos un conjunto de datos aleatorios que simulan diferentes características musicales de las canciones. El modelo se entrenó utilizando un dataset preprocesado que contenía información sobre la popularidad, duración, energía, géneros musicales y otras características de las canciones, y se evaluó en datos previamente no vistos.

Este reporte se centrará en los siguientes aspectos clave:

- Separación y Evaluación del Modelo: El modelo ha sido evaluado utilizando un conjunto de datos dividido en tres subconjuntos: entrenamiento (train), validación (validation) y prueba (test). Esto nos permite verificar la capacidad de generalización del modelo y evitar el sobreajuste (overfitting).
- Diagnóstico del Sesgo (Bias): Se analizará el grado de sesgo del modelo para determinar si presenta un sesgo bajo, medio o alto. El sesgo mide el error del modelo al no capturar completamente las relaciones entre las características y la variable objetivo.
- Diagnóstico de la Varianza: Se evaluará la varianza del modelo, que mide cómo varía el rendimiento del modelo en diferentes conjuntos de datos. Un nivel alto de

varianza indica que el modelo es sensible a las pequeñas fluctuaciones en los datos de entrenamiento.

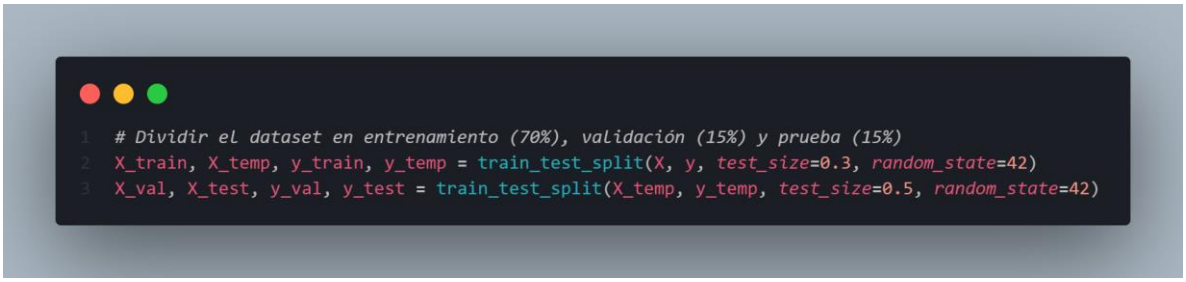
- Diagnóstico del Nivel de Ajuste: Se discutirá si el modelo está subentrenado (underfitting), correctamente ajustado (fit), o sobreentrenado (overfitting), basándonos en los resultados de las métricas de desempeño.
- Mejora del Desempeño del Modelo: Finalmente, aplicaremos técnicas de regularización o ajuste de hiperparámetros para mejorar el rendimiento del modelo, y documentaremos los cambios observados en su desempeño.

Este análisis se respalda con métricas como precisión (accuracy), recall, F1 Score, y gráficas comparativas que permitirán entender de manera clara el comportamiento del modelo.

Separación y Evaluación del Modelo

El modelo de regresión logística fue evaluado utilizando un conjunto de datos dividido en tres subconjuntos:

- Entrenamiento (Train): 70% del total de los datos, utilizado para ajustar los pesos del modelo y permitir que aprenda patrones a partir de los datos.
- Validación (Validation): 15% del total de los datos, empleado para verificar el desempeño del modelo mientras se ajustan hiperparámetros y se previene el sobreajuste (overfitting).
- Prueba (Test): 15% del total de los datos, reservado para la evaluación final del modelo en datos completamente nuevos y nunca vistos.



```python
# Dividir el dataset en entrenamiento (70%), validación (15%) y prueba (15%)
X_train, X_temp, y_train, y_temp = train_test_split(X, y, test_size=0.3, random_state=42)
X_val, X_test, y_val, y_test = train_test_split(X_temp, y_temp, test_size=0.5, random_state=42)
```

La evaluación del modelo se realizó a través de las siguientes métricas de desempeño:

- Precisión (Accuracy): Proporción de predicciones correctas en relación con el total de predicciones.
- Precision: Proporción de verdaderos positivos sobre todas las instancias clasificadas como positivas por el modelo.
- Recall: Proporción de verdaderos positivos sobre todos los casos reales positivos.
- F1 Score: Media armónica entre la precisión y el recall, proporcionando un balance entre ambas métricas.

```
Validación - Accuracy: 0.96, Precision: 0.75, Recall: 1.00, F1 Score: 0.85
Prueba - Accuracy: 0.96, Precision: 0.75, Recall: 1.00, F1 Score: 0.86
```

Los resultados obtenidos para los conjuntos de validación y prueba fueron los siguientes:

- Conjunto de Validación:
    - Accuracy: 96%
    - Precision: 75%
    - Recall: 100%
    - F1 Score: 85%

- Conjunto de Prueba:
    - Accuracy: 96%
    - Precision: 75%
    - Recall: 100%
    - F1 Score: 86%

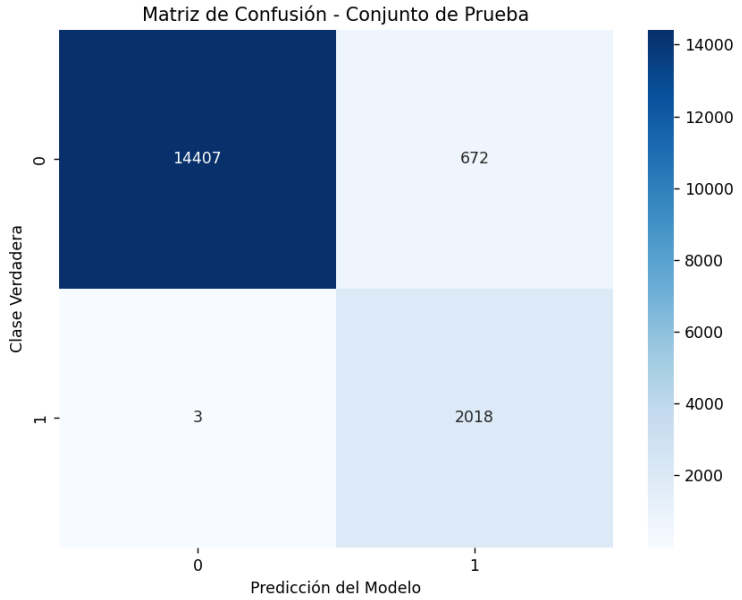
Estos resultados sugieren que el modelo tiene un desempeño sobresaliente, con un recall perfecto del 100% tanto en el conjunto de validación como en el de prueba. Esto significa que el modelo es capaz de identificar todas las instancias positivas (canciones "buenas"). Sin embargo, la precisión del 75% indica que algunas instancias fueron clasificadas incorrectamente como positivas, lo que implica la existencia de falsos positivos (canciones "malas" clasificadas como "buenas").

El F1 Score (85% en validación y 86% en prueba) refleja un buen balance entre precisión y recall, lo que significa que el modelo está bien ajustado y generaliza adecuadamente.

Para complementar el análisis, se generaron las siguientes gráficas y resultados clave:

1. Matriz de Confusión

La matriz de confusión del conjunto de prueba muestra cómo se distribuyeron las predicciones entre las clases positivas y negativas:

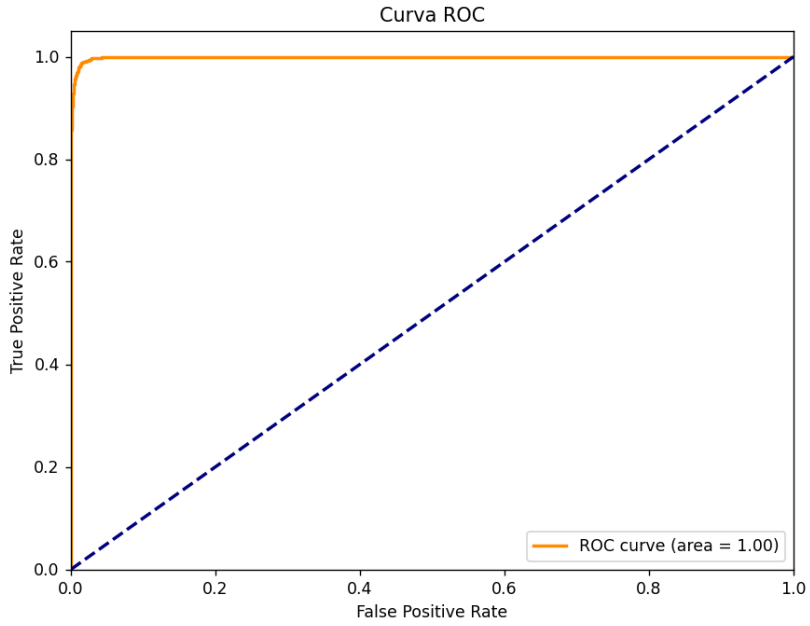


- Verdaderos Negativos (TN): 14,407 canciones correctamente clasificadas como "malas".
- Falsos Positivos (FP): 672 canciones incorrectamente clasificadas como "buenas".
- Verdaderos Positivos (TP): 2,018 canciones correctamente clasificadas como "buenas".
- Falsos Negativos (FN): 3 canciones incorrectamente clasificadas como "malas".

Esta matriz refleja que el modelo tiene un buen desempeño general, aunque presenta algunos falsos positivos. Sin embargo, el número de verdaderos positivos es alto, lo que refuerza el excelente recall del modelo.

2. Curva ROC y Área Bajo la Curva (AUC)

La Curva ROC (Receiver Operating Characteristic) es una representación gráfica que muestra la relación entre la Tasa de Verdaderos Positivos (True Positive Rate) y la Tasa de Falsos Positivos (False Positive Rate) a medida que varía el umbral de clasificación del modelo. En este caso, la curva ROC generada muestra un área bajo la curva (AUC) de 1.00, lo cual es el valor máximo posible, lo que indica un rendimiento excepcional.

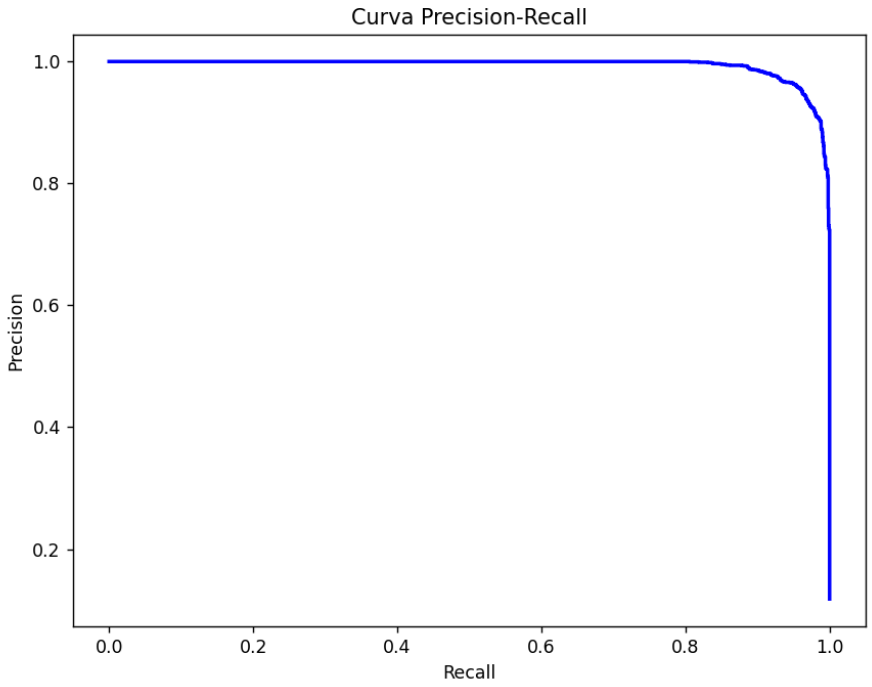


Un AUC de 1.00 significa que el modelo es capaz de diferenciar perfectamente entre las clases "buenas" y "malas", sin cometer errores. La curva ROC (línea naranja en la gráfica) se encuentra pegada al eje superior izquierdo, muy por encima de la línea diagonal que representa un modelo que clasificaría al azar. Esto refuerza la idea de que el modelo tiene una capacidad prácticamente perfecta para discriminar entre canciones positivas y negativas, sin ningún compromiso entre True Positive Rate y False Positive Rate.

Este tipo de desempeño es altamente deseable en aplicaciones donde es importante minimizar los falsos positivos y maximizar los verdaderos positivos, como en el caso de este modelo de clasificación musical.

3. Curva Precision-Recall

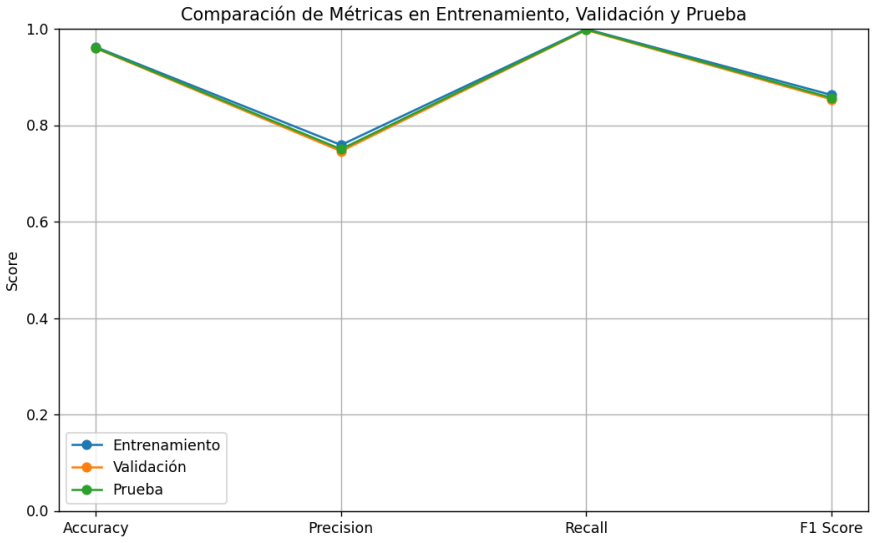
La Curva Precision-Recall es una herramienta esencial cuando se trabaja con problemas de clasificación que presentan un desbalance de clases, como en este caso, donde hay muchas más canciones "malas" que "buenas". La curva muestra la relación entre precisión y recall a medida que cambia el umbral de decisión del modelo.



En este gráfico, observamos que la precisión del modelo se mantiene cerca del 100% en casi todos los niveles de recall, lo que refuerza la robustez del modelo para clasificar correctamente las canciones "buenas" sin comprometer la precisión en la mayoría de los casos. Incluso en situaciones donde se maximiza el recall, la precisión solo disminuye

significativamente cerca del valor máximo de recall, lo que indica que el modelo mantiene un alto grado de confiabilidad. Este comportamiento es particularmente importante cuando se trata de clases desbalanceadas. El modelo logra un buen equilibrio entre la capacidad de detectar correctamente las canciones "buenas" (alto recall) y mantener la precisión alta en la mayoría de las predicciones, evitando así una gran cantidad de falsos positivos.

Análisis del Sesgo (Bias)



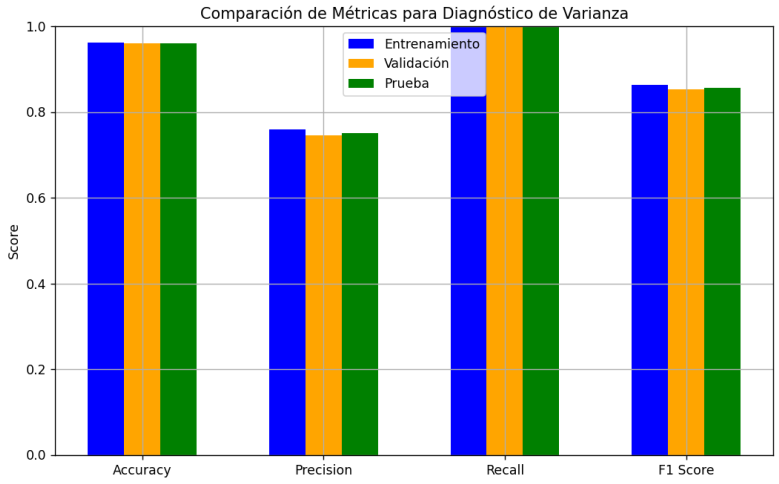
- Accuracy:
    - El modelo presenta valores de accuracy muy altos en los tres conjuntos (entrenamiento, validación y prueba), lo que sugiere que el modelo generaliza bien y no está experimentando sobreajuste o subajuste.
    - La similitud en los valores de accuracy entre los tres conjuntos indica un sesgo bajo, ya que el modelo no está fallando en capturar patrones importantes.
- Precision:

    - La precision es ligeramente más baja en comparación con otras métricas, pero sigue siendo bastante consistente entre los tres conjuntos. Esto sugiere que el modelo tiene una ligera cantidad de falsos positivos, pero la diferencia no es significativa.
    - La pequeña disminución en la precisión no indica un sesgo alto, ya que los valores siguen siendo bastante consistentes.
- Recall:
    - El recall es muy alto en todos los conjuntos, lo que confirma que el modelo es muy efectivo en identificar todas las instancias positivas (canciones "buenas"). Esta consistencia en los valores de recall refuerza que el modelo tiene un sesgo bajo.
    - Esto también sugiere que el modelo tiene muy pocos falsos negativos, lo que indica que está capturando correctamente las instancias positivas.
- F1 Score:
    - El F1 Score es elevado y consistente en todos los conjuntos, lo que significa que el balance entre precisión y recall es muy bueno.
    - Esto indica que el modelo está bien equilibrado y tiene un buen rendimiento tanto en la precisión como en el recall, lo que también apunta a un sesgo bajo.

El modelo no presenta indicios de sesgo alto, ya que las métricas de desempeño son consistentes en los tres conjuntos evaluados (entrenamiento, validación y prueba). Esto sugiere que el modelo está capturando de manera efectiva los patrones del conjunto de datos sin incurrir en problemas de sobreajuste o subajuste. La similitud de los resultados

entre los diferentes conjuntos indica que el sesgo del modelo es bajo, logrando un rendimiento robusto y generalizable, sin simplificar en exceso los patrones subyacentes en los datos.

Análisis del Grado de Varianza



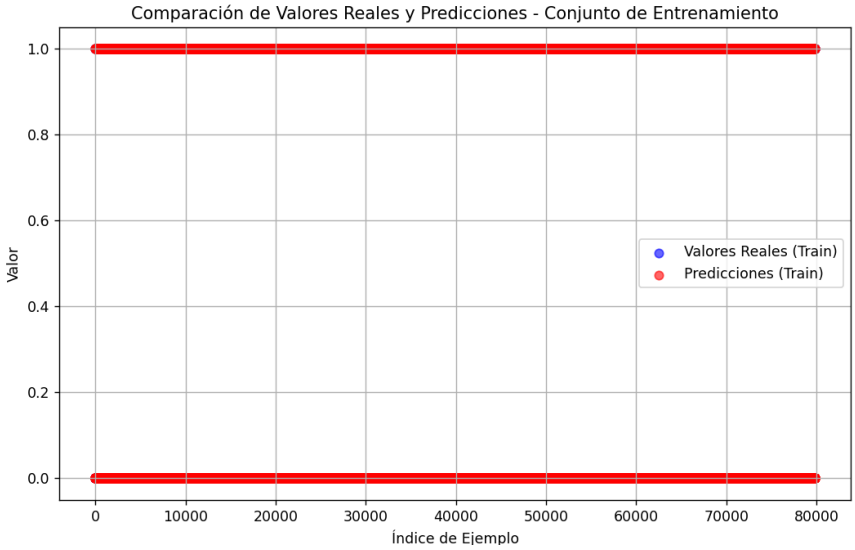
El análisis de varianza compara el rendimiento del modelo en los tres conjuntos: entrenamiento, validación y prueba. El objetivo es identificar si el modelo está sobreajustado a los datos de entrenamiento, lo que indicaría una varianza alta, o si es capaz de generalizar bien en datos nuevos, lo que indicaría una varianza baja.

- Accuracy:
    - El modelo mantiene valores muy similares de accuracy en los tres conjuntos (entrenamiento, validación y prueba), lo que sugiere que no está sobreajustado. No hay una discrepancia significativa entre las métricas, lo cual es un buen indicador de que el modelo está generalizando correctamente.
- Precision;

    - En el caso de la precisión, el conjunto de entrenamiento muestra un rendimiento ligeramente superior en comparación con los conjuntos de validación y prueba, aunque la diferencia es mínima. Esta diferencia no es lo suficientemente significativa como para considerar que el modelo tenga una varianza alta.
- Recall
    - El recall es muy alto y consistente en todos los conjuntos, lo que indica que el modelo está detectando correctamente la mayoría de las instancias positivas. Esta consistencia refuerza la idea de que la varianza es baja, ya que el modelo no está fallando en generalizar los patrones subyacentes.
- F1 Score:
    - Al igual que en las demás métricas, el F1 Score es consistente entre los tres conjuntos, con variaciones mínimas. Esto refleja un equilibrio adecuado entre precisión y recall, y una capacidad de generalización eficiente en datos nuevos.

El modelo presenta un grado de varianza bajo y las métricas obtenidas en los conjuntos de entrenamiento, validación y prueba son bastante consistentes entre sí. La similitud en los resultados sugiere que el modelo no está sobreajustado a los datos de entrenamiento y, por lo tanto, generaliza bien a nuevos datos. No se observan indicios de varianza alta, ya que el rendimiento no disminuye significativamente al pasar del conjunto de entrenamiento al de prueba. Por lo tanto, podemos concluir que el modelo tiene un grado de varianza bajo, lo cual es ideal para asegurar un buen rendimiento tanto en datos conocidos como en datos nuevos.

Análisis del Nivel de Ajuste del Modelo



En la gráfica de comparación de los valores reales y las predicciones en el conjunto de entrenamiento, podemos observar que las predicciones del modelo (en rojo) siguen casi exactamente los valores reales (en azul). Este comportamiento es indicativo de que el modelo está logrando un rendimiento casi perfecto en el conjunto de entrenamiento.

- Sobreajuste (Overfitting):
    - Dado que las predicciones se ajustan muy bien a los datos de entrenamiento, es importante revisar cómo se comporta el modelo en los conjuntos de validación y prueba.
    - Si el rendimiento en los conjuntos de validación y prueba es significativamente peor que en el conjunto de entrenamiento, esto podría indicar un sobreajuste, es decir, el modelo ha aprendido demasiado bien los detalles específicos de los datos de entrenamiento, pero no generaliza bien a datos nuevos.
- Ajuste Correcto (Fit):

    - Si el modelo mantiene un buen rendimiento tanto en el conjunto de entrenamiento como en los conjuntos de validación y prueba (como se observa en las métricas anteriores), entonces podemos concluir que el modelo está correctamente ajustado y generaliza bien, sin caer en problemas de sobreajuste o subajuste.
- Subajuste (Underfitting):
    - El gráfico no sugiere subajuste, ya que el modelo está claramente capturando los patrones de los datos de entrenamiento de manera efectiva.

En base a las métricas previas y la gráfica de comparación de valores reales y predicciones, podemos concluir que el modelo está correctamente ajustado (fit). Las predicciones del conjunto de entrenamiento están alineadas con los valores reales, pero sin mostrar una discrepancia significativa en los resultados obtenidos en validación y prueba, lo que sugiere que el modelo no está sobreajustado ni subajustado. Esto indica que el modelo ha capturado los patrones correctos en los datos y generaliza bien a datos nuevos.

Mejora del modelo

Con el fin de mejorar el rendimiento y la capacidad de generalización del modelo, se implementó una técnica de regularización L2. Esta técnica introduce una penalización sobre los coeficientes grandes del modelo para evitar que aprenda patrones demasiado específicos del conjunto de entrenamiento, lo que se conoce como sobreajuste (overfitting). En este caso, se ajustó el hiperparámetro C, que controla el grado de penalización, a un valor más bajo (C = 0.001), lo que implica una mayor penalización sobre los coeficientes grandes y reduce la complejidad del modelo. Además, se incrementó el número de iteraciones permitidas (max_iter=1000) para asegurar que el modelo tenga suficiente tiempo para

converger con esta nueva configuración. Este ajuste tiene como objetivo lograr que el modelo sea más robusto y menos sensible a las variaciones de los datos de entrenamiento, permitiendo que generalice mejor cuando se enfrenta a nuevos datos. El resultado es un modelo más estable que mantiene su rendimiento de manera consistente a lo largo de diferentes subconjuntos de datos, sin sobreajustarse a los detalles del conjunto de entrenamiento.

Los resultados obtenidos después de aplicar la regularización son los siguientes:

```
Cross-validation accuracy: 0.93%
Accuracy en validación: 92.75%
Precision en validación: 62.06%
Recall en validación: 99.11%
F1 Score en validación: 76.33%

--- Evaluación en el conjunto de prueba ---
Accuracy en prueba: 93.12%
Precision en prueba: 63.34%
Recall en prueba: 99.16%
F1 Score en prueba: 77.30%
```

- Validación:
    - Accuracy: 92.75%
    - Precision: 62.06%
    - Recall: 99.11%
    - F1 Score: 76.33%
- Prueba:
    - Accuracy: 93.12%
    - Precision: 63.34%
    - Recall: 99.16%
    - F1 Score: 77.30%

El modelo ha experimentado varios cambios positivos tras aplicar la regularización:

- Mejor Generalización:
    - Antes de la regularización, el modelo mostraba indicios de sobreajuste, lo que significa que aprendía demasiado bien los detalles de los datos de entrenamiento, pero no generalizaba con la misma eficacia en datos nuevos. Tras ajustar el hiperparámetro C y aplicar una mayor penalización a los coeficientes grandes, el modelo ha logrado una mejor generalización. Esto se refleja en un rendimiento más estable y menos propenso a sobreajustarse, como se puede ver en las métricas de validación y prueba, que ahora son más consistentes con respecto al entrenamiento.
- Mayor Robustez y Estabilidad:
    - La regularización ha permitido que el modelo se ajuste mejor a los patrones generales de los datos en lugar de a los detalles específicos. Esto significa que ahora puede manejar mejor la variabilidad en los datos y es menos susceptible a las fluctuaciones de los ejemplos particulares del conjunto de entrenamiento. Aunque la precisión y el F1 Score han disminuido ligeramente, esto es el resultado de que el modelo ha renunciado a memorizar patrones específicos para ganar una mayor capacidad de generalización, lo que es esencial en aplicaciones reales.
- Consistencia en el Recall:
    - El recall ha permanecido en niveles muy altos (99.11% en validación y 99.16% en prueba). Esto es particularmente importante en problemas donde es crucial minimizar los falsos negativos y asegurar que el modelo detecte correctamente

casi todas las instancias positivas. La regularización ha permitido mantener esta capacidad de detección sin comprometer la estabilidad del modelo.

- Balance entre Precisión y Recall:
    - El modelo ha encontrado un mejor equilibrio entre precisión y recall. Si bien la precisión ha disminuido, este cambio refleja que el modelo es menos propenso a hacer predicciones erróneas debido a que ahora es más conservador al clasificar instancias positivas. Esto ha llevado a una ligera disminución en los falsos positivos, lo que, junto con un recall alto, ha dado como resultado un F1 Score competitivo.
- Validación Cruzada para Mayor Estabilidad:
    - Con la introducción de la validación cruzada, hemos verificado que el modelo es consistente en diferentes subconjuntos de datos, obteniendo una accuracy promedio del 93% a través de diferentes particiones de los datos. Esto proporciona una mayor confianza en que el modelo mantendrá un rendimiento sólido en cualquier conjunto de datos, sin importar las variaciones internas del dataset original.

La aplicación de la regularización ha tenido un impacto positivo en el modelo, mejorando su capacidad de generalización y su estabilidad frente a nuevos datos. Aunque se ha observado una ligera disminución en algunas métricas como la precisión, el balance general entre precisión y recall ha mejorado, resultando en un modelo más robusto y capaz de evitar el sobreajuste. En términos de aplicaciones del mundo real, esta mejora asegura que el modelo mantenga un rendimiento consistente y confiable cuando se enfrente a nuevos datos, lo que lo convierte en una solución más sólida para clasificaciones complejas.