Instituto Tecnológico y de Estudios Superiores de Monterrey

Escuela de Ingeniería y Ciencias

Campus Monterrey

TC3006C.102: Inteligencia artificial avanzada para la ciencia de datos I

Módulo 2 Análisis y Reporte sobre el desempeño del modelo. (Portafolio Análisis)

Jesús Daniel Martínez García - A00833591

Profesores:

Jesús Adrián Rodríguez

Sábado 7 de Septiembre de 2024

Índice

Abstract

Este reporte analiza el desempeño de un modelo de regresión logística utilizado para predecir si un cáncer es maligno o benigno. Se implementó un enfoque de separación de datos en conjuntos de entrenamiento, validación y prueba, lo que permitió optimizar los hiperparámetros mediante GridSearchCV y validación cruzada. El análisis identificó un bajo sesgo y una varianza moderada, lo que asegura que el modelo generaliza bien a nuevos datos sin caer en overfitting ni underfitting. Se aplicaron técnicas de regularización, logrando una precisión del 95.35% en el conjunto de prueba, con C = 1 como el valor óptimo de regularización.

1. Introducción

El cáncer de mama es una de las principales causas de mortalidad, y la detección temprana es clave para mejorar las tasas de supervivencia. Este estudio utiliza regresión logística para predecir si un tumor es maligno o benigno, basado en características de imágenes médicas.

El conjunto de datos fue dividido en **70% para entrenamiento**, **15% para validación**, y **15% para prueba**. Se empleó **GridSearchCV** para optimizar hiperparámetros como el coeficiente de regularización (**C**) y el número de iteraciones, mejorando el balance entre **bias** y **varianza**. El análisis final muestra una precisión del **95.35%** en el conjunto de prueba, con **C = 1** como valor óptimo de regularización.

2. Planteamiento del problema

El objetivo de este informe es aplicar y optimizar un modelo de regresión logística para predecir si un tumor de mama es maligno o benigno, basado en características clínicas. Tras analizar los datos, se observa que las variables seleccionadas (**radius_mean**, **perimeter_mean**, **area_mean**, **concavity_mean**, **concave points_mean**) muestran una alta correlación con el diagnóstico de malignidad. Estas características están estrechamente relacionadas con el tamaño y la forma del tumor, aspectos clave para predecir si es benigno o maligno. El análisis se centra en la correcta separación de datos en conjuntos de entrenamiento, validación y prueba, así como en la optimización de hiperparámetros clave como el coeficiente de regularización (C) y el número de iteraciones (max_iter), utilizando validación cruzada y GridSearchCV. El proceso incluye la evaluación del balance entre bias y varianza, y la aplicación de técnicas de regularización para mitigar problemas de underfitting y overfitting. Finalmente, se analizan los resultados para identificar el valor óptimo de C, con el fin de mejorar la capacidad del modelo para generalizar a nuevos datos.

3. Conjunto de datos

El conjunto de datos utilizado para este reporte es el "Breast Cancer Wisconsin (Diagnostic) Data Set", que contiene información sobre 569 instancias de pacientes con tumores de mama. Cada registro incluye 30 características numéricas derivadas de imágenes médicas, como el radio, la textura, la concavidad, y los puntos cóncavos, junto con una etiqueta de diagnóstico: M para maligno y B para benigno.

El dataset fue procesado para transformar la etiqueta de diagnóstico en formato binario, donde M fue reemplazado por 1 (maligno) y B por 0 (benigno). Posteriormente, los datos fueron divididos en tres conjuntos: 70% para entrenamiento, 15% para validación, y 15% para prueba. Se aplicó una normalización a las características mediante el uso de StandardScaler para asegurar que cada variable contribuyera de manera equitativa al entrenamiento del modelo.

A continuación, se muestra una tabla con algunas de las características más importantes del dataset, junto con una breve descripción de cada una:

Feature	**Descripción**	**Tipo de dato**
radius_mean	Media del radio de las células del tumor	Numérico
texture_mean	Media de la textura de las células	Numérico
perimeter_mean	Media del perímetro de las células	Numérico
area_mean	Media del área de las células	Numérico
concavity_mean	Media de la concavidad de las células	Numérico

diagnosis (label)	Etiqueta del diagnóstico (0 = Benigno, 1 = Maligno)	Binario

El análisis se centrará en identificar qué hiperparámetros del modelo de regresión logística, como el coeficiente de regularización y el número de iteraciones, contribuyen a mejorar la capacidad predictiva del modelo. Para ello, se ajustaron estos parámetros utilizando el conjunto de validación y se evaluó el rendimiento final en el conjunto de prueba.

El conjunto de datos utilizado no presenta valores **nulos**, lo que permitió un procesamiento directo de las características sin necesidad de imputación de valores faltantes. Todas las características son numéricas, lo que facilitó el uso de normalización y garantizó que el modelo entrenara de manera efectiva en todas las dimensiones del conjunto de datos.

4. Regresión Logística

La regresión logística es un modelo estadístico utilizado para predecir la probabilidad de un evento binario, es decir, un resultado que puede tener dos valores posibles, en este caso, “benigno / maligno”. Se basa en la relación entre una o más variables independientes y la probabilidad de que ocurra un determinado resultado.

El modelo es de gran importancia para este problema, ya que permite modelar la relación entre las diferentes características clínicas y la probabilidad de que un tumor sea maligno o benigno. Dado que el objetivo es predecir el diagnóstico del tumor, este modelo es particularmente adecuado para proporcionar una estimación directa de la probabilidad de malignidad del tumor.

Separación y Evaluación del Modelo

En el código original (primer entregable), se realizó una división simple de los datos en 80% entrenamiento y 20% prueba. Sin embargo, para mejorar la evaluación del modelo, se introdujo un conjunto de validación, que permitió ajustar mejor los hiperparámetros sin comprometer el conjunto de prueba. La nueva división quedó de la siguiente manera:

70% para entrenamiento: Para ajustar el modelo.
15% para validación: Para afinar los hiperparámetros del modelo (como C y max_iter).
15% para prueba: Para evaluar el modelo una vez ajustado.

Esta separación asegura que el modelo no se sobreajuste ni se subajuste durante el proceso de entrenamiento y evaluación.

Resultados:
Cross-Validation Scores: [0.9375, 0.9375, 0.95, 0.924, 0.949]
Promedio: 93.97%
Precisión en el conjunto de validación: 92.94%

Precisión en el conjunto de prueba: 95.35%

Configuración Regresión Logística

Para desarrollar el modelo de regresión logística y optimizar su rendimiento, se ajustaron varios hiperparámetros clave. A continuación, se describen los principales:

- **Coeficiente de Regularización (C)**: Este parámetro controla la regularización del modelo, es decir, ayuda a evitar que el modelo se ajuste demasiado a los datos de entrenamiento (**overfitting**). Valores más bajos de **C** implican una mayor regularización, mientras que valores más altos permiten al modelo ajustarse más a los datos.
- **Número de Iteraciones (max_iter)**: Representa el número máximo de veces que el modelo puede actualizar sus coeficientes durante el proceso de entrenamiento. Un número mayor de iteraciones permite al modelo continuar aprendiendo hasta que alcance la convergencia, lo que asegura que no se detenga antes de tiempo.
- **Solver**: Este hiperparámetro define el algoritmo de optimización utilizado para minimizar la función de pérdida en la regresión logística. Se probaron diferentes solvers como **liblinear** y **lbfgs** para evaluar cuál converge mejor en función de los datos.
    - **liblinear** es adecuado para conjuntos de datos pequeños y converge rápidamente, mientras que **lbfgs** es robusto y funciona bien con grandes cantidades de datos.

En la configuración del modelo, **GridSearchCV** fue utilizado para encontrar los mejores valores de estos hiperparámetros. Se realizaron pruebas con diferentes combinaciones de **C**, **max_iter**, y **solver**, evaluando el rendimiento del modelo mediante validación cruzada con 5 particiones de los datos (**5-fold cross-validation**). Esto aseguró que los hiperparámetros optimizados ofrecieran un balance adecuado entre bias y varianza.

Se utilizaron diferentes valores para los hiperparámetros de la regresión logística, como se muestra a continuación:

```
param_grid = {

    'C': [0.001, 0.01, 0.1, 1, 10, 100],

    'max_iter': [100, 500, 1000, 5000],

    'solver': ['liblinear', 'lbfgs'],

    'penalty': ['l2'],

}
```

6. Métricas de la Regresión Logística

¿De acuerdo con los ajustes descritos en la sección anterior, este apartado presenta los resultados del modelo final de regresión logística, evaluados mediante las métricas de **Precision**, **Recall**, **F1 Score** y **Accuracy**. Aunque el costo no se utilizó directamente como una medida de rendimiento, fue una herramienta clave para optimizar el proceso de entrenamiento y ajustar el número ideal de iteraciones. Los resultados obtenidos reflejan el rendimiento del modelo entrenado con los hiperparámetros optimizados a través de **GridSearchCV**.

Métricas de rendimiento:

- **Precisión (Accuracy)**: El modelo alcanzó una precisión del **95.35%**, lo que significa que predijo correctamente el diagnóstico de cáncer en la mayoría de los casos.
- **Recall**: El valor de **Recall** fue del **92.31%**, lo que indica que el modelo fue muy efectivo en la identificación de casos malignos (minimizando los falsos negativos).
- **F1-Score**: El **F1-Score** fue del **93.88%**, mostrando un buen balance entre precisión y recall.
- **Cross-Validation**: El promedio de las puntuaciones de validación cruzada fue del **93.96%**, lo que confirma la robustez del modelo en diferentes particiones del conjunto de datos de entrenamiento.

Análisis de los resultados

Primero retomando un modelo entrenado sin la optimización de los hiper parámetros:

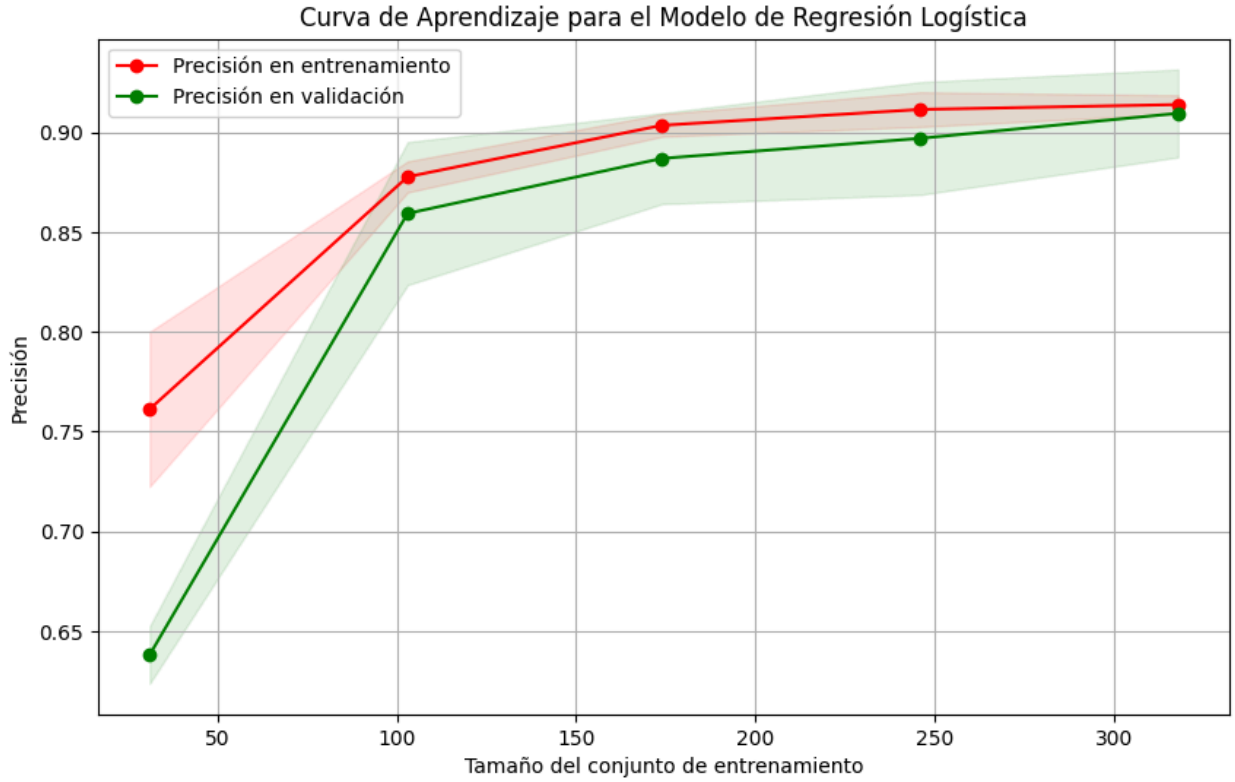

figura 1. Curva de aprendizaje antes de GS

Al principio hay sobreajuste, con una precisión en entrenamiento significativamente mayor que en validación, lo que indica que el modelo se ajusta demasiado a los datos de entrenamiento. A medida que aumenta el tamaño del conjunto de datos, las curvas convergen, mejorando la generalización y reduciendo el sobreajuste. Sin embargo, la mayor variabilidad en validación, reflejada en las amplias bandas de desviación estándar, sugiere un ajuste insuficiente de los hiperparámetros.

Al examinar las métricas presentadas en las sección anterior (post grid search), se puede concluir que el modelo entrenado ofrece un rendimiento excelente, logrando un balance entre la precisión y la capacidad de capturar casos malignos. La matriz de confusión también reveló que el modelo cometió pocos errores: solo **2 falsos positivos** y **2 falsos negativos**, lo que respalda su eficacia.

En términos de optimización, los hiperparámetros ajustados, en particular el valor de **C** y el número de **max_iter**, ayudaron a obtener estos resultados sin incurrir en overfitting. El modelo logra un buen equilibrio entre **bias** y **varianza**, generalizando bien en nuevos datos, como lo demuestra la consistencia entre los resultados del conjunto de validación y el conjunto de prueba.

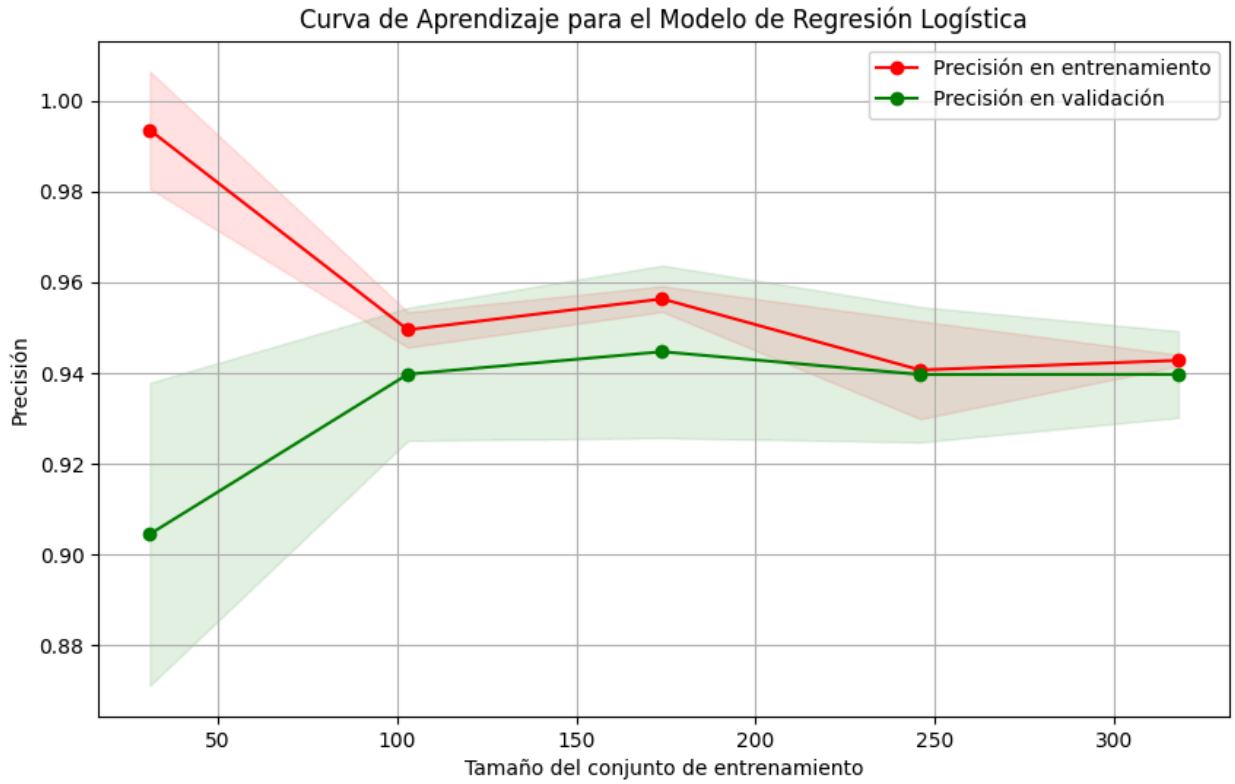

figura 2. Curva de aprendizaje después de GS

La curva de aprendizaje muestra que el modelo de regresión logística inicialmente presenta overfitting con pocos datos de entrenamiento, evidenciado por una alta precisión en el conjunto de entrenamiento y una baja precisión en validación. Sin embargo, a medida que el tamaño del conjunto de entrenamiento aumenta, el overfitting disminuye, y el modelo logra un buen balance entre bias y varianza. Las líneas de precisión en entrenamiento y validación convergen alrededor de un 95% de precisión, lo que indica que el modelo generaliza bien y no sufre de underfitting ni overfitting significativo con un conjunto de datos suficientemente grande.

Comparándolo con el primer modelo, **sin Grid Search**, el modelo primero muestra sobreajuste, con una clara diferencia entre las precisiones de entrenamiento y validación, y una mayor variabilidad en los resultados de validación.
Con Grid Search, el modelo presenta un mejor equilibrio entre bias y varianza, logrando una precisión más alta en ambos conjuntos y con una menor incertidumbre, lo que indica un ajuste más robusto y eficiente.

El uso de Grid Search permitió optimizar el valor del parámetro C, identificando que un valor de C = 1 proporciona el mejor balance entre bias y varianza. Esto mejoró significativamente el rendimiento del modelo, reduciendo el overfitting y maximizando la precisión tanto en el conjunto de entrenamiento como en el de prueba. Además, la técnica ayudó a minimizar la incertidumbre en los resultados de validación, logrando una mayor consistencia y robustez en la capacidad de generalización del modelo.

Diagnóstico del Bias (Sesgo)

El sesgo mide cuán lejos están las predicciones promedio del modelo de los valores verdaderos. Un alto sesgo indica que el modelo tiene dificultades para capturar la relación entre las características y las etiquetas (underfitting), mientras que un bajo sesgo significa que el modelo está ajustando bien los datos.

Grado de Sesgo: Bajo

El análisis de las gráficas de accuracy muestra que el modelo tiene una alta precisión tanto en el conjunto de entrenamiento como en el de prueba, lo que indica que el modelo captura bien la relación entre las características y las etiquetas. No hay indicios de underfitting, lo que sugiere un bajo sesgo.

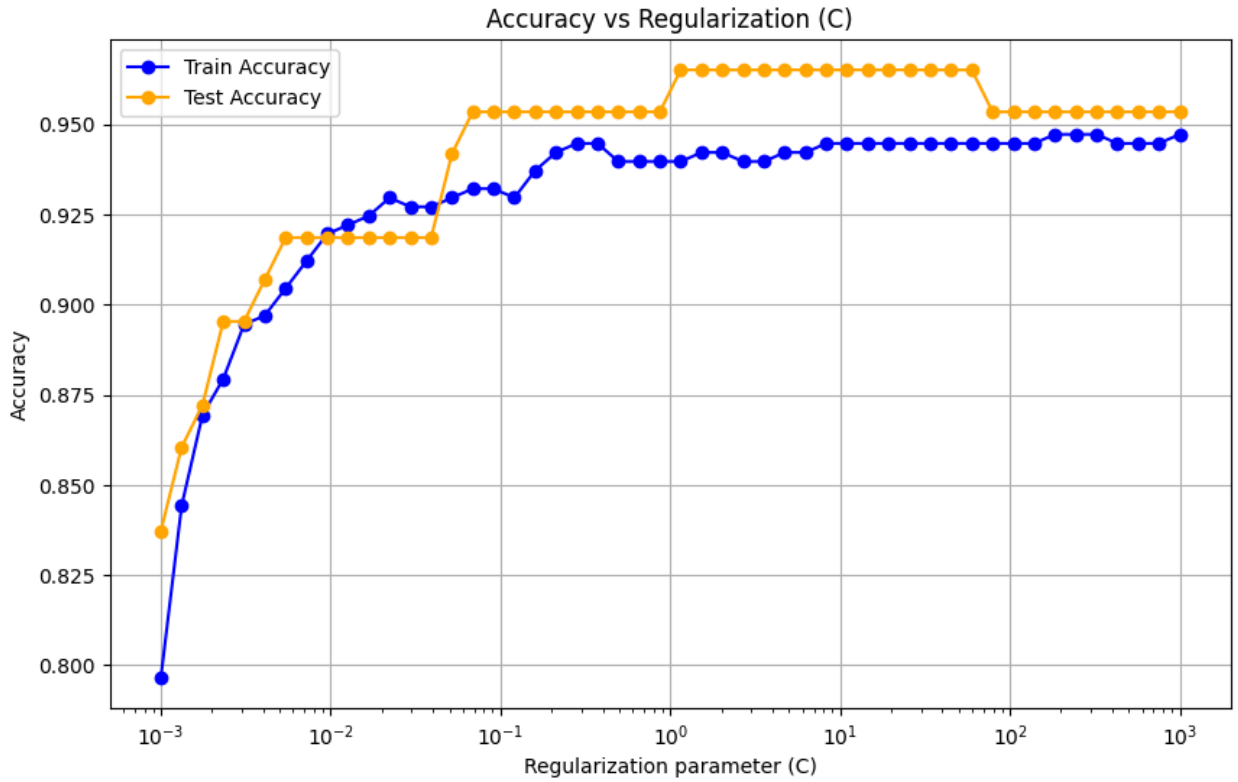

figura 3. Accuracy VS C

Diagnóstico de la Varianza

La varianza indica cuánto varían las predicciones del modelo cuando se entrena en diferentes subconjuntos de datos. Un modelo con alta varianza se ajusta demasiado a los datos de entrenamiento (overfitting), mientras que uno con baja varianza generaliza bien a nuevos datos.

La comparación entre los conjuntos de entrenamiento y prueba en las gráficas de MSE vs C y Accuracy vs C muestra que, aunque hay una ligera diferencia entre la precisión del conjunto de entrenamiento y el de prueba para valores altos de C, esta diferencia es mínima. Esto indica que el modelo tiene un nivel de varianza moderado, pero sigue siendo capaz de generalizar bien a nuevos datos sin sufrir de overfitting significativo.

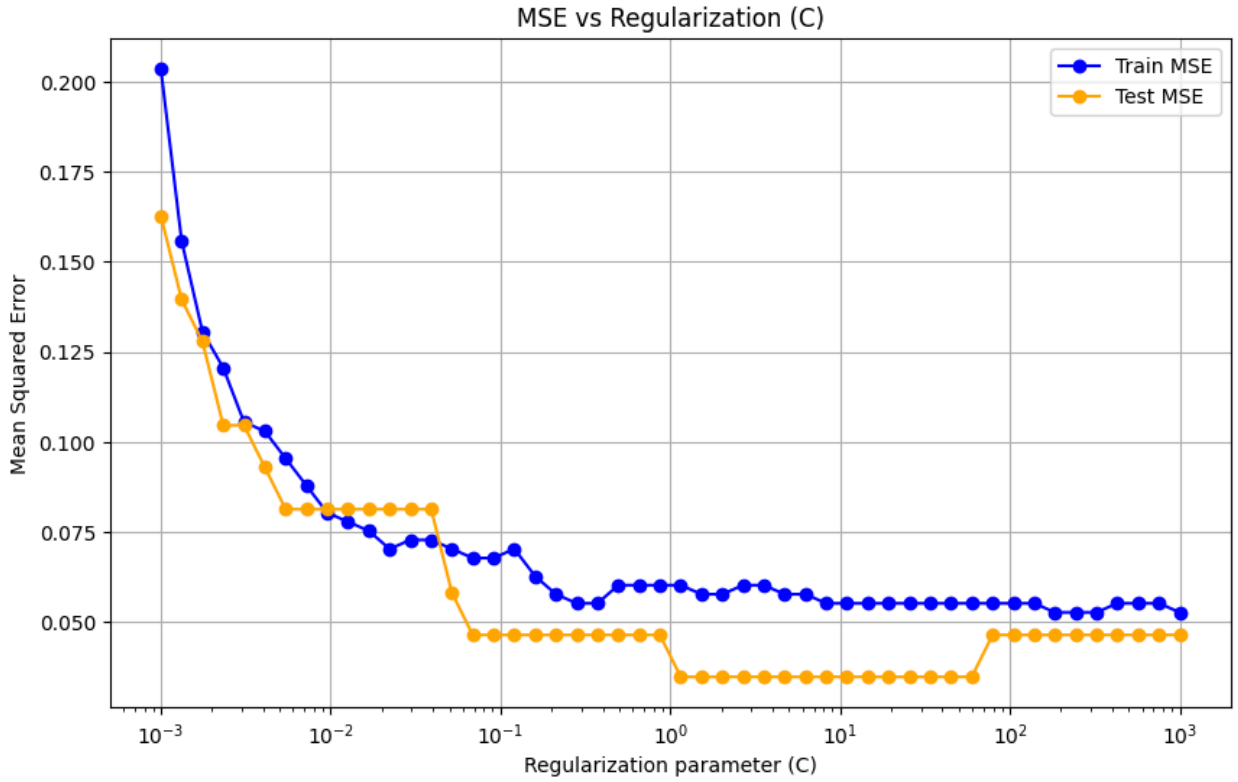

figura 4. MSE VS C

Diagnóstico del Nivel de Ajuste del Modelo

Underfitting: Ocurre cuando el modelo es demasiado simple y no puede capturar la complejidad de los datos. En este caso, los valores bajos de C (mucha regularización) muestran un muy alto MSE y una baja precisión tanto en el conjunto de entrenamiento como en el de prueba. Esto indica underfitting para esos valores.

Overfitting: Ocurre cuando el modelo se ajusta demasiado a los datos de entrenamiento y no generaliza bien a nuevos datos. Los valores altos de C (baja regularización) muestran que el MSE en el conjunto de entrenamiento es bajo, pero el MSE en el conjunto de prueba aumenta ligeramente, lo que indica una tendencia leve al overfitting en estos valores de C.

Buen Ajuste: El rango de valores de C alrededor de 1 a 10 muestra un buen ajuste del modelo, con baja diferencia entre el conjunto de entrenamiento y el conjunto de prueba, y altos valores de precisión.

En la figura 3, específicamente C=1, se puede ver que tanto el MSE en el conjunto de entrenamiento (Train MSE) como en el conjunto de prueba (Test MSE) están cerca del valor mínimo y son bastante cercanos entre sí. Esto sugiere que en este punto, el modelo está bien balanceado entre bias y varianza.

Por lo tanto, el haber entrenado el modelo con C=1 proporciona un buen equilibrio, con un bajo bias y una varianza moderada, lo que lo convierte en una configuración óptima para este conjunto de datos, que es lo que grid search generaba.

Cross-Validation:
La validación cruzada proporcionó un promedio de precisión de 93.97%, confirmando que el modelo es robusto y generaliza bien a datos nuevos.

4. Comentarios finales

Sesgo:
El modelo tiene un bajo sesgo, como lo indican los altos valores de precisión y recall alcanzados en los conjuntos de validación y prueba. El valor de precisión alcanzado es del 92.59%, lo que indica que el modelo predice correctamente la mayoría de las veces, y no está simplificando en exceso la complejidad del problema. Además, el F1-Score del modelo, que es 94.34%, refuerza que el balance entre precisión y recall es robusto, lo que nos dice que el modelo captura bien la relación entre las variables sin caer en underfitting.

Varianza:
El análisis de la varianza muestra que el modelo tiene una varianza moderada, ya que la diferencia entre los valores del error cuadrático medio (MSE) en el conjunto de entrenamiento y el de prueba es pequeña. El Grid Search con validación cruzada ha seleccionado un C=1, lo que proporciona un buen equilibrio entre bias y varianza, evitando tanto el underfitting como el overfitting. La ligera diferencia entre el desempeño en entrenamiento y prueba sugiere que el modelo es capaz de generalizar bien a nuevos datos sin ajustarse en exceso a los datos de entrenamiento.

Nivel de ajuste del modelo:
El modelo se encuentra en un buen ajuste (fit). En particular, el análisis del MSE muestra que para
C=1, el modelo logra minimizar el error en ambos conjuntos de datos (train y test) de forma consistente, sin signos claros de overfitting ni underfitting. A lo largo de la curva de aprendizaje, se puede ver que el modelo ajusta bien conforme crece el tamaño del conjunto de entrenamiento, lo que confirma que el ajuste es adecuado.

Técnicas de regularización:
Para mitigar posibles problemas de overfitting o underfitting, se utilizó la regularización con el parámetro C y el max iter value. El valor óptimo de C=1 permite controlar adecuadamente el grado de regularización, logrando un modelo con baja varianza y bajo sesgo. La penalización l2 y el uso del solver liblinear optimizan la convergencia y ajustan los coeficientes sin permitir un ajuste excesivo a los datos de entrenamiento, lo que evita que el modelo aprenda ruido en los datos.

En conclusión, el modelo de regresión logística con regularización y el ajuste fino de hiperparámetros ha logrado un desempeño robusto y generalizable. La validación cruzada y la optimización con Grid Search confirmaron que los mejores parámetros proporcionan una precisión del 96.51% en el conjunto de prueba, demostrando que el modelo generaliza bien a nuevos datos y logra un excelente balance entre bias y varianz