Instituto Tecnológico y de Estudios Superiores de Monterrey

Campus Estado de México

Momento de Retroalimentación: Módulo 2 Análisis y Reporte sobre el desempeño del modelo. (Portafolio Implementación)

Inteligencia artificial avanzada para la ciencia de datos I

Daniel Alejandro Martínez Cienfuegos - A01745412

Grupo 101

11/09/2024

Análisis de Desempeño del Modelo de Árbol de Decisión con el Dataset del Titanic

El conjunto de datos del Titanic se hizo uso debido a que este conjunto de datos proporciona información clara y concisa sobre pasajeros, como su edad, sexo, clase de boleto, número de hermanos/esposos a bordo, número de padres/hijos a bordo, tarifa del boleto, y puerto de embarque. La variable objetivo es "Survived", que indica si un pasajero sobrevivió o no al hundimiento del Titanic. Este conjunto de datos es ideal para los algoritmos de aprendizaje automático ya que permite demostrar la capacidad de generalización de un modelo. Además, la estructura del conjunto de datos permite la implementación de técnicas de preprocesamiento como la gestión de valores nulos y la codificación de variables categóricas, lo que facilita la preparación de datos para los modelos de clasificación.

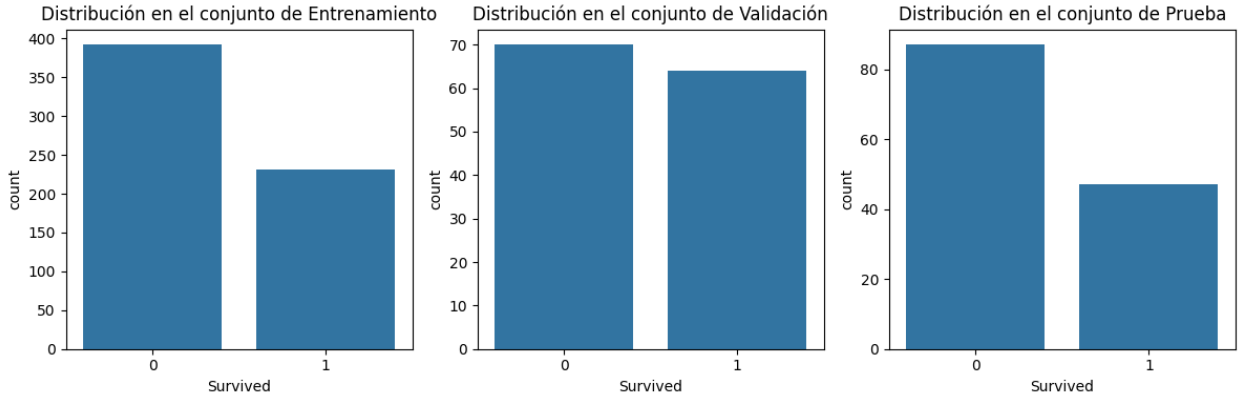
	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

(Fig. 1. Dataset)

Separación y Evaluación del Modelo con Conjuntos de Entrenamiento, Prueba y Validación

Para garantizar una evaluación justa y efectiva del modelo, el conjunto de datos se dividió en tres partes: un conjunto de entrenamiento, un conjunto de validación y un conjunto de prueba. El conjunto de entrenamiento se utilizó para entrenar el modelo de árbol de decisión; el conjunto de validación ayudó a ajustar los hiperparámetros del modelo y prevenir el sobreajuste (overfitting); y el conjunto de prueba se utilizó para evaluar el rendimiento del modelo final de manera imparcial. La división se realizó de la siguiente manera: el 70% de los datos se asignaron al conjunto de entrenamiento, mientras que el 30% restante se dividió equitativamente en los conjuntos de validación y prueba, cada uno conteniendo el 15% de los datos originales.

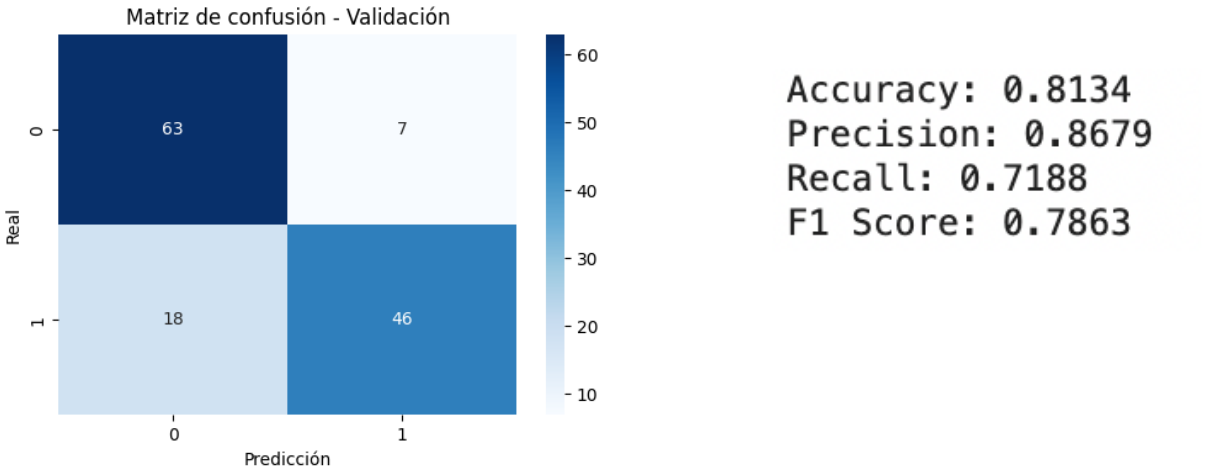
Esta separación asegura que el modelo tenga suficiente información para aprender los patrones y, al mismo tiempo, evalúe su capacidad para generalizar a datos no vistos. Las características del conjunto de datos fueron adecuadamente preparadas mediante la eliminación de columnas irrelevantes, la imputación de valores nulos y la codificación de variables categóricas a formato numérico.



(Fig. 2. Evaluación Train, Test and Validation)

Diagnóstico y Grado de Sesgo (Bias)

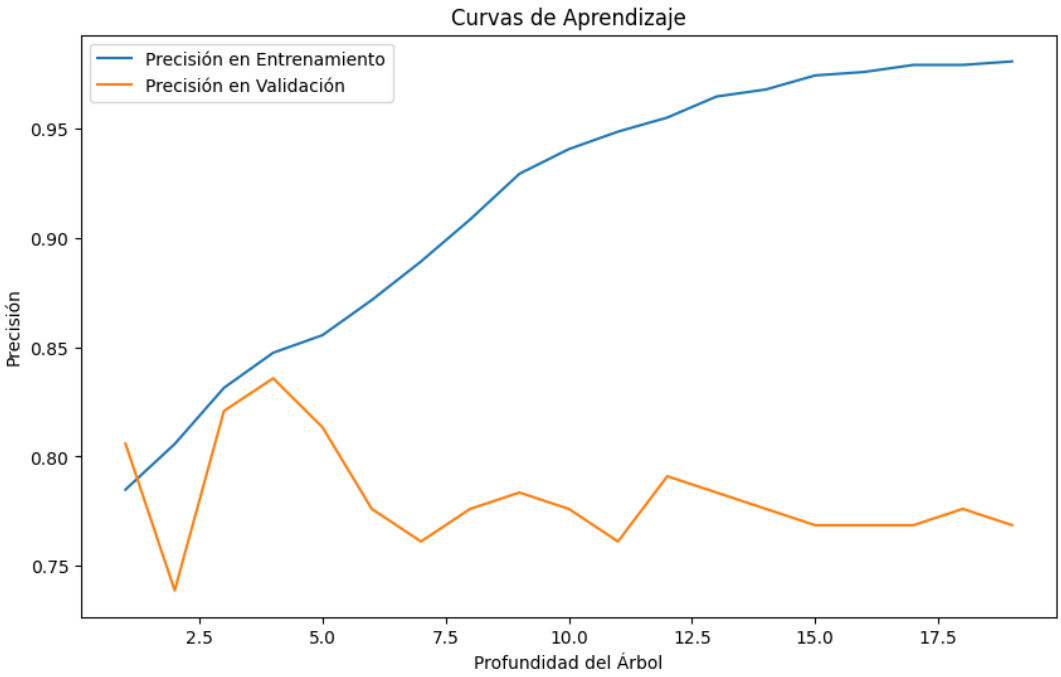
El modelo entrenado sin técnicas de regularización mostró un sesgo bajo. Esto se evidenció por la alta precisión del modelo en el conjunto de entrenamiento, indicando que el modelo se ajustaba bien a los datos de entrenamiento. Sin embargo, es importante notar que un sesgo demasiado bajo puede llevar al problema de la varianza alta, ya que el modelo puede captar demasiado los detalles de los datos de entrenamiento y no generalizar bien a nuevos datos. En el caso del modelo sin regularización, el sesgo bajo significaba que estaba muy bien ajustado a los datos de entrenamiento pero podría no desempeñarse tan bien con datos de prueba debido al sobreajuste.



(Fig. 3. Matriz de Confusión)

Diagnóstico y Explicación del Grado de Varianza

El diagnóstico del modelo sin regularización reveló un alto grado de varianza, lo que se conoce como "overfitting". Esto quedó claro cuando se evaluó el modelo en el conjunto de validación y prueba, donde la precisión disminuyó significativamente en comparación con el conjunto de entrenamiento. El alto grado de varianza indica que el modelo es demasiado complejo y se ha ajustado demasiado a los detalles específicos del conjunto de entrenamiento, capturando incluso el ruido de los datos. Como resultado, el modelo no generaliza bien a los datos nuevos y muestra un desempeño inferior en los conjuntos de validación y prueba.



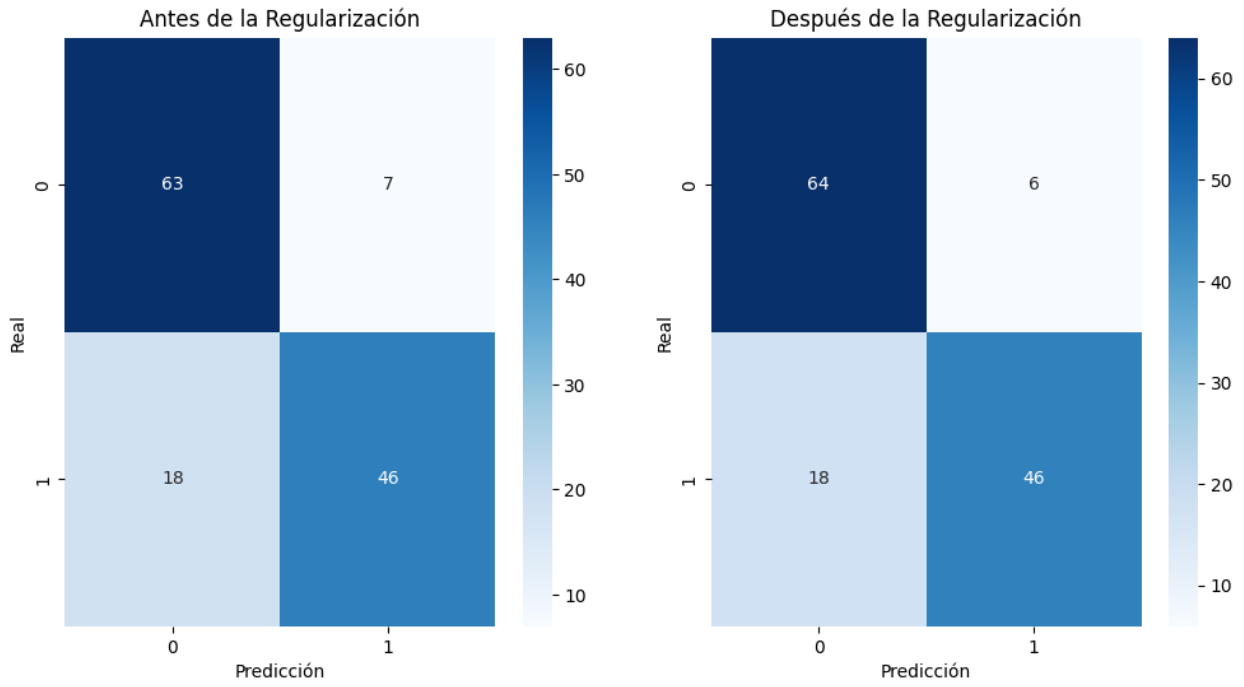
Diagnóstico y Explicación del Nivel de Ajuste del Modelo

El análisis inicial del nivel de ajuste del modelo mostró que, sin regularización, el modelo estaba claramente sobreajustado (overfit). Esto es común en modelos de árboles de decisión que no tienen restricciones en su crecimiento, ya que pueden crear nodos y ramas que se ajustan excesivamente a las observaciones del conjunto de entrenamiento. Para mitigar este problema, se implementaron técnicas de regularización que limitaron la profundidad del árbol y controlaron el número mínimo de muestras necesarias para dividir un nodo. Al limitar la complejidad del modelo, se redujo el sobreajuste, mejorando así la capacidad del modelo para generalizar a nuevos datos.

Mejora del Desempeño del Modelo Mediante Técnicas de Regularización

Se implementaron específicamente tres técnicas de regularización para mejorar el rendimiento del modelo y reducir el sobreajuste. Primero, se ajustó la profundidad máxima del árbol de decisión max_depth, para limitar cuán profundo podría crecer el árbol. Esto redujo la complejidad del modelo, disminuyendo la probabilidad de que se ajuste demasiado a los datos de entrenamiento. Además, se establecieron los hiperparámetros min_samples_split y min_samples_leaf, para controlar el número mínimo de muestras requeridas para realizar una división en un nodo y el número mínimo de muestras necesarias en una hoja. Estas técnicas aseguraron que el árbol no se volviera demasiado específico en las divisiones, lo cual mejoró la capacidad de generalización.

```
Accuracy (Regularización): 0.8209
Precision (Regularización): 0.8846
Recall (Regularización): 0.7188
F1 Score (Regularización): 0.7931
```



(Fig. 5. Comparación de Desempeño Antes y Después de la Regularización)

Comparación de Desempeño Antes y Después de la Regularización

Para evaluar la efectividad de la regularización, se realizaron comparaciones del rendimiento del modelo antes y después de aplicar las técnicas de regularización. Las métricas de rendimiento clave utilizadas incluyeron precisión (accuracy), precisión positiva (precision), recall y puntaje F1. Antes de la regularización, el modelo presentó alta precisión en el conjunto de entrenamiento, pero mostró una disminución significativa en el conjunto de prueba, lo cual indica sobreajuste. Después de aplicar las técnicas de regularización, el modelo mostró una precisión más balanceada entre los conjuntos de entrenamiento, validación y prueba. La precisión en el conjunto de prueba mejoró, indicando que el modelo generaliza mejor a datos no vistos. Las gráficas de las matrices de confusión antes y después de la regularización también destacaron esta mejora. Antes de la regularización, la matriz de confusión mostraba que el

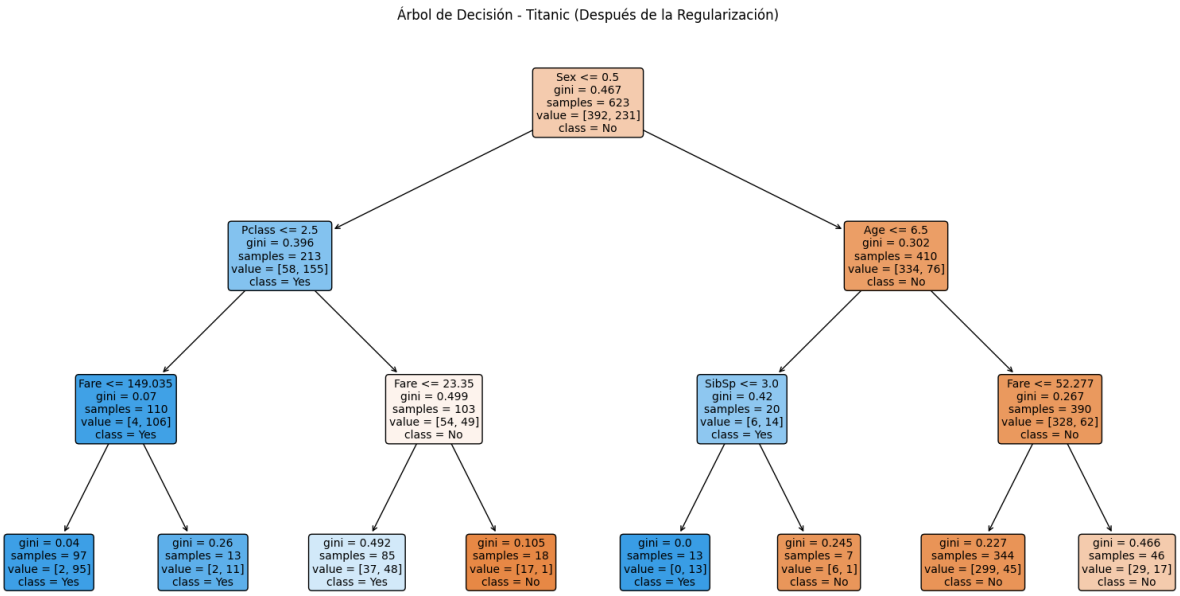
modelo cometía más errores al clasificar correctamente los casos de supervivencia y no supervivencia en los datos de prueba. Después de la regularización, el número de errores disminuyó, lo que confirma una mejora en el rendimiento del modelo. En general, la implementación de estas técnicas de regularización ayudó a reducir tanto el grado de varianza como el sobreajuste, logrando un modelo más robusto que se desempeña de manera consistente y precisa en diferentes conjuntos de datos.

Árbol de Decisión Regularizado

La visualización del árbol de decisión después de la aplicación de técnicas de regularización muestra un modelo más sencillo y menos profundo en comparación con el árbol inicial. Esta simplificación del modelo es resultado directo de la aplicación de técnicas de regularización como la restricción de la profundidad máxima del árbol, el ajuste del número mínimo de muestras requeridas para dividir un nodo y el número mínimo de muestras necesarias en un nodo hoja. Inicialmente, el modelo de árbol de decisión se entrenó sin restricciones significativas, lo que permitió que el árbol creciera en profundidad y generara múltiples ramas para ajustar los datos de entrenamiento. Este proceso, si bien optimiza el ajuste a los datos de entrenamiento, también puede llevar a un sobreajuste. Un modelo sobreajustado capta demasiadas variaciones y ruido de los datos de entrenamiento, lo que reduce su capacidad para generalizar a datos nuevos y no vistos. Para mejorar la capacidad de generalización del modelo y evitar el sobreajuste, se aplicaron técnicas de regularización.

Después de aplicar las técnicas de regularización, la visualización del árbol de decisión muestra un modelo con menos niveles de profundidad y menos nodos totales. Las ramas del árbol son más cortas y las decisiones en cada nodo son más claras y están mejor definidas. Este cambio indica que el modelo ahora está capturando patrones más generales y significativos de los datos en lugar de detalles específicos del conjunto de entrenamiento. Al comparar el árbol antes y después de la regularización, se observa que el árbol original era más profundo y más ramificado, lo que indicaba una alta complejidad. Sin embargo, el modelo ajustado con

regularización es más compacto y menos profundo, reflejando un mejor balance entre bias y varianza. Este nuevo árbol es más probable que generalice bien en datos nuevos, mostrando una reducción en el riesgo de sobreajuste.



(Fig. 6. Árbol de Decisión después de la Regularización)

Conclusiones

El análisis realizado con el conjunto de datos del Titanic demuestra que, al aplicar técnicas de regularización adecuadas y ajustar correctamente los hiperparámetros del modelo de árbol de decisión, se puede mejorar significativamente el rendimiento del modelo, logrando un equilibrio entre el sesgo y la varianza. Esto permite que el modelo generalice de manera más efectiva a datos no vistos, aumentando su precisión y su capacidad predictiva. Además, el conjunto de datos del Titanic es adecuado para este tipo de análisis, ya que proporciona una variedad de características que son representativas y fácilmente manejables para los algoritmos de aprendizaje automático, demostrando así su utilidad en el desarrollo de modelos de clasificación robustos.