Tecnológico de Monterrey

Inteligencia artificial avanzada para la ciencia de datos I

Entregable Módulo 2: Análisis y Reporte sobre el desempeño del modelo

Profesor:

Jorge Adolfo Ramírez Uresti

Zaide Islas Montiel | A01751580

Fecha de entrega:
11 de Septiembre de 2023

Entregable Módulo 2: Análisis y Reporte sobre el desempeño del modelo

Introducción

Machine learning representa una faceta de la inteligencia artificial que posibilita que un sistema adquiera conocimientos a partir de datos en lugar de depender de una programación explícita. Sin embargo, el proceso de machine learning no es trivial. A medida que el algoritmo se nutre con datos de entrenamiento, puede desarrollar modelos más precisos basados en dicha información. Un modelo de machine learning se configura como el resultado generado al entrenar un algoritmo de machine learning con datos específicos. Después de este proceso de entrenamiento, al suministrar datos a dicho modelo, éste proporcionará una salida correspondiente. Por ejemplo, un algoritmo predictivo generará un modelo predictivo, y al alimentarse datos, obtendrás una predicción basada en los datos utilizados durante el entrenamiento del modelo.

Machine learning posibilita que los modelos se preparan utilizando conjuntos de datos antes de su implementación. Algunos de estos modelos son de naturaleza continua y operan en línea. Este proceso iterativo de modelos en línea conlleva una mejora en las conexiones establecidas entre los elementos de los datos. Debido a su complejidad y tamaño, estas pautas y relaciones podrían haber pasado inadvertidas para la observación humana. Una vez que un modelo ha sido entrenado, puede utilizarse en tiempo real para aprender de los datos. Las mejoras en la precisión resultan del proceso de entrenamiento y la automatización inherente al machine learning.

Enfoques en el machine learning:

Las técnicas de *machine learning* son esenciales para mejorar la precisión de los modelos predictivos. Dependiendo de la naturaleza del problema empresarial, se emplean diferentes enfoques según el tipo y el volumen de los datos. En esta sección, se describen las categorías del *machine learning*.

- Aprendizaje supervisado: El aprendizaje supervisado comienza generalmente con un conjunto de datos establecido y una comprensión previa de cómo se clasifican esos datos. Su objetivo es encontrar patrones en los datos que puedan aplicarse a procesos analíticos. Estos datos cuentan con etiquetas que definen su significado. Por ejemplo, se puede desarrollar una aplicación de machine learning que, basada en imágenes y descripciones escritas, distinga entre millones de animales.
- Aprendizaje no supervisado: El aprendizaje no supervisado se utiliza cuando se manejan grandes volúmenes de datos sin etiquetar, como en redes sociales como Twitter, Instagram y Snapchat. La comprensión del significado detrás de estos datos requiere algoritmos que clasifiquen los datos según los patrones o clústeres que encuentren. El aprendizaje no supervisado implica un proceso iterativo de análisis de datos sin intervención humana y se aplica en tecnología de detección de spam en correos electrónicos, donde es difícil etiquetar masivamente correos no deseados debido a la diversidad de variables.
- Aprendizaje de refuerzo: El aprendizaje de refuerzo es un modelo de aprendizaje basado en el comportamiento. El algoritmo recibe retroalimentación a medida que analiza los datos, lo que guía al usuario hacia el resultado óptimo. A diferencia de otros tipos de aprendizaje supervisado, el sistema no se entrena con ejemplos de

datos, sino que aprende a través de la prueba y el error. Las secuencias de decisiones acertadas refuerzan el proceso, ya que se consideran las más efectivas para resolver el problema.

- Deep learning: El deep learning es un enfoque específico del machine learning que incorpora redes neuronales con múltiples capas para aprender de los datos de manera iterativa. Es especialmente útil para comprender patrones en datos no estructurados. Las complejas redes neuronales del deep learning imitan el funcionamiento del cerebro humano y permiten que las computadoras traten con abstracciones y problemas vagamente definidos. Este enfoque se emplea frecuentemente en el reconocimiento de imágenes, voz y aplicaciones de visión por computadora.

Dataset utilizado

El *Student Performance Dataset* es un conjunto de datos diseñado para examinar los factores que influyen en el rendimiento académico de los estudiantes. El conjunto de datos consta de 10,000 registros de estudiantes, cada uno de los cuales contiene información sobre diversos predictores y un índice de rendimiento.

Variables:

- Horas estudiadas: El número total de horas dedicadas al estudio por cada estudiante.
- Puntuaciones anteriores: Las puntuaciones obtenidas por los alumnos en pruebas anteriores.
- Actividades Extraescolares: Si el alumno participa en actividades extraescolares (Sí o No).
- Horas de sueño: El número medio de horas de sueño diarias del alumno.
- Cuestionarios de muestra practicados: Número de cuestionarios de muestra que ha practicado el alumno.
- Variable objetivo - Índice de rendimiento: Medida del rendimiento global de cada alumno. El índice de rendimiento representa el rendimiento académico del alumno y se ha redondeado al número entero más próximo. El índice oscila entre 10 y 100, y los valores más altos indican un mejor rendimiento. El objetivo del conjunto de datos es proporcionar información sobre la relación entre las variables predictoras y el índice de rendimiento. Los investigadores y analistas de datos pueden utilizar este conjunto de datos para explorar el impacto de las horas de estudio, las calificaciones anteriores, las actividades extraescolares, las horas de sueño y los modelos de cuestionarios en el rendimiento de los estudiantes.

  El dataset fue elegido debido a las siguientes consideraciones:

- Tamaño adecuado: El dataset debe tener suficientes ejemplos o instancias para que los modelos puedan aprender patrones significativos. Un tamaño de muestra pequeño puede llevar a modelos sobreajustados (overfitting) o ineficaces. En este caso, como se mencionó previamente, el conjunto consta de 10,000 datos.
- Calidad de los datos: Los datos deben estar limpios y ser de alta calidad. Esto significa que no deben contener errores, valores atípicos (outliers) o datos faltantes importantes que puedan afectar la calidad del modelo. Esto se verificó en el primer entregable.

- Etiquetas o categorías claras: En problemas supervisados, es esencial que los datos tengan etiquetas o categorías claras que indiquen cuál es el resultado o la variable objetivo que se está tratando de predecir.
- Características relevantes: Las características o variables incluidas en el dataset deben ser relevantes para el problema que se está abordando. Características irrelevantes pueden introducir ruido y dificultar el aprendizaje del modelo.
- Diversidad: El dataset debe contener una variedad de ejemplos que cubran diferentes escenarios o casos. Esto ayuda a que el modelo generalice mejor a nuevas situaciones.
- Consistencia: Los datos deben ser consistentes en su formato y estructura. Esto facilita el preprocesamiento y la creación de modelos.

Desarrollo

Se realizó la implementación de dos técnicas de aprendizaje máquina, con la diferencia del uso de frameworks. La primera de ellas fue un árbol de decisión, mientras que el segundo fue el uso de *k Nearest Neighbors*. El análisis de este reporte será correspondiente a la segunda técnica implementada, mencionada anteriormente.

1. *Separación y evaluación del modelo con un conjunto de prueba y un conjunto de validación (Train/Test/Validation)*

La separación de los datos en conjuntos de entrenamiento (train), prueba (test), y validación (validation) es una práctica común en el aprendizaje automático. Estos conjuntos se utilizan para diferentes propósitos en el desarrollo de modelos de aprendizaje automático y para evaluar su rendimiento de manera objetiva.

- Conjunto de Entrenamiento (Train):
    - Propósito: El conjunto de entrenamiento se utiliza para entrenar el modelo de aprendizaje automático. Los algoritmos de aprendizaje utilizan estos datos para aprender patrones, relaciones y características en los datos de entrada, lo que les permite hacer predicciones o tomar decisiones.
    - Tamaño: Es el conjunto de datos más grande y generalmente representa el 60-80% del conjunto de datos total. En este caso, se consideró 60%.
- Conjunto de Prueba (Test):
    - Propósito: El conjunto de prueba se utiliza para evaluar el rendimiento del modelo después del entrenamiento. Es independiente del conjunto de entrenamiento y se utiliza para medir qué tan bien el modelo generaliza datos nuevos y no vistos.
    - Tamaño: Por lo general, representa el 20-30% del conjunto de datos total. En este caso, se consideró 20%.
- Conjunto de Validación (Validation):
    - Propósito: El conjunto de validación se utiliza durante el proceso de ajuste de hiperparámetros y selección del modelo. Después de entrenar el modelo en el conjunto de entrenamiento, se evalúa en el conjunto de validación para ajustar los hiperparámetros y evitar el sobreajuste (overfitting) al conjunto de entrenamiento.

- Tamaño: Es más pequeño que el conjunto de entrenamiento y el conjunto de prueba, generalmente representa el 10-20% del conjunto de datos total. Se consideró 20%.

La separación en estos tres conjuntos ayuda a garantizar que el modelo se desarrolle de manera robusta, evitando el sobreajuste y permitiendo una evaluación justa de su rendimiento.

```
# Dividir el conjunto de datos en entrenamiento (60%), validación (20%) y prueba (20%)
x_train, x_temp, y_train, y_temp = train_test_split(X, y, test_size=0.40, random_state=42)
x_val, x_test, y_val, y_test = train_test_split(x_temp, y_temp, test_size=0.50, random_state=42)
```

Figura 1. Separación de datos en entrenamiento, validación y prueba.

```
Tamaño del conjunto de entrenamiento: (6000, 5)
Tamaño del conjunto de validación: (2000, 5)
Tamaño del conjunto de prueba: (2000, 5)
```

Figura 2. Tamaño de los conjuntos

Se realizaron gráficas con el objetivo de observar cómo se comparan las predicciones del modelo con los valores reales en los conjuntos de validación y prueba.

Conjunto de Validación Predictions:

	Actual	Predicted
0	45	42.4
1	79	76.2
2	49	47.8
3	29	33.6
4	79	78.2
5	65	65.2
6	50	47.4
7	74	73.8
8	73	71.0
9	37	35.4

Figura 3. Impresión de un segmento de predicción de datos con el conjunto de Validación

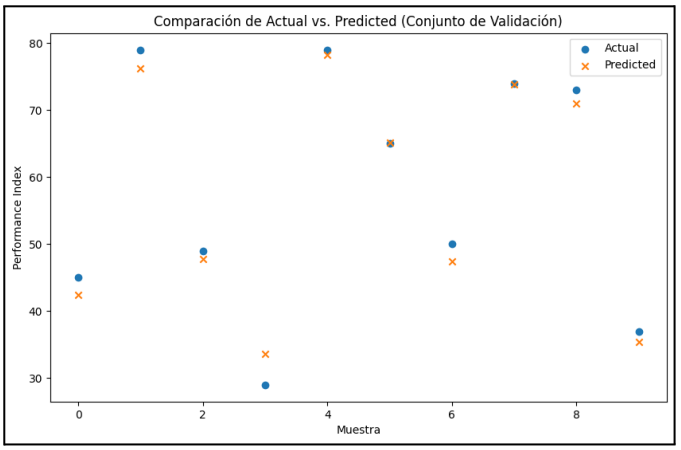


Figura 4. Gráfica de comparación entre valores actuales vs predichos por el conjunto de Validación

Estas gráficas muestran una comparación entre las primeras 10 muestras en el conjunto de validación. Los puntos marcados con *Actual* representan los valores reales del *Performance Index* para esas muestras específicas. Los puntos marcados con Predicted representan las predicciones del modelo para el *Performance Index* de esas mismas muestras.Si las predicciones del modelo son precisas, esperaríamos que los puntos *Predicted* estén cerca de los puntos *Actual*. La distancia entre los puntos *Actual* y *Predicted* en el eje vertical (*Performance Index*) indica cuán cercanas o distantes están las predicciones del modelo con respecto a los valores reales. Cuanto más cerca estén, mejor será el rendimiento del modelo.

Conjunto de Prueba Predictions:

	Actual	Predicted
0	53	49.2
1	66	66.4
2	76	78.2
3	54	49.8
4	61	63.6
5	61	62.0
6	22	21.0
7	26	26.6
8	33	33.2
9	76	73.2

Figura 5. Impresión de un segmento de predicción de datos con el conjunto de Prueba

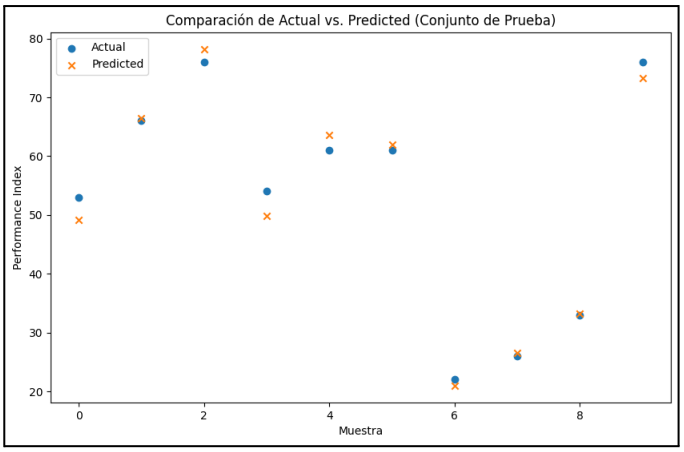


Figura 6. Gráfica de comparación entre valores actuales vs predichos por el conjunto de Prueba

Como se observa, estas gráficas funcionan de la misma manera que las de conjunto de validación pero muestran la comparación entre las primeras 10 muestras en el conjunto de prueba. Al igual que en las gráficas de conjunto de validación, los puntos *Actual* representan los valores reales y los puntos *Predicted* representan las predicciones del modelo.

2. *Diagnóstico y explicación el grado de bias o sesgo: bajo medio alto*

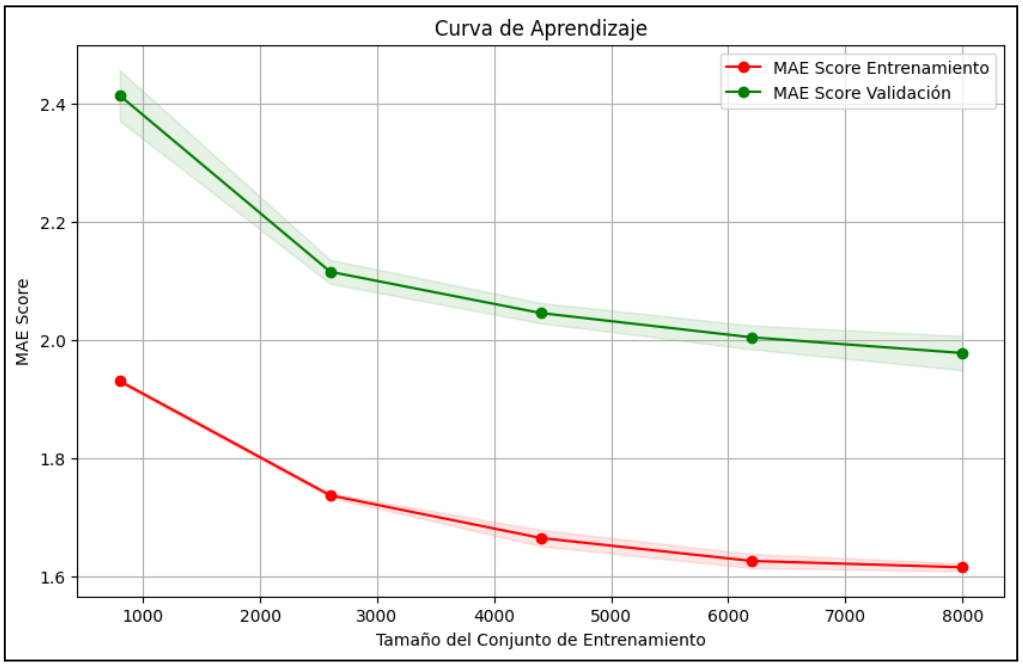


Figura 7. Comparación de curvas de aprendizaje - MAE

La línea roja muestra cómo varía el MAE en el conjunto de entrenamiento a medida que aumenta el tamaño del conjunto de entrenamiento. Idealmente, se espera que el MAE en el conjunto de entrenamiento disminuya a medida que se tienen más datos para entrenar el modelo. Esta línea converge a un valor bajo, lo que demuestra que el modelo está aprendiendo bien del conjunto de entrenamiento.

La línea verde muestra cómo varía el MAE en el conjunto de validación (o prueba) a medida que aumenta el tamaño del conjunto de entrenamiento. Similar al conjunto de entrenamiento, se espera que el MAE en el conjunto de validación disminuya a medida que se

tienen más datos para entrenar el modelo. Sin embargo, hay una brecha significativa entre la línea de entrenamiento y la línea de validación, lo que puede indicar que el modelo está sobreajustando los datos de entrenamiento y no generaliza bien a nuevos datos.

3. *Diagnóstico y explicación el grado de varianza: bajo medio alto*

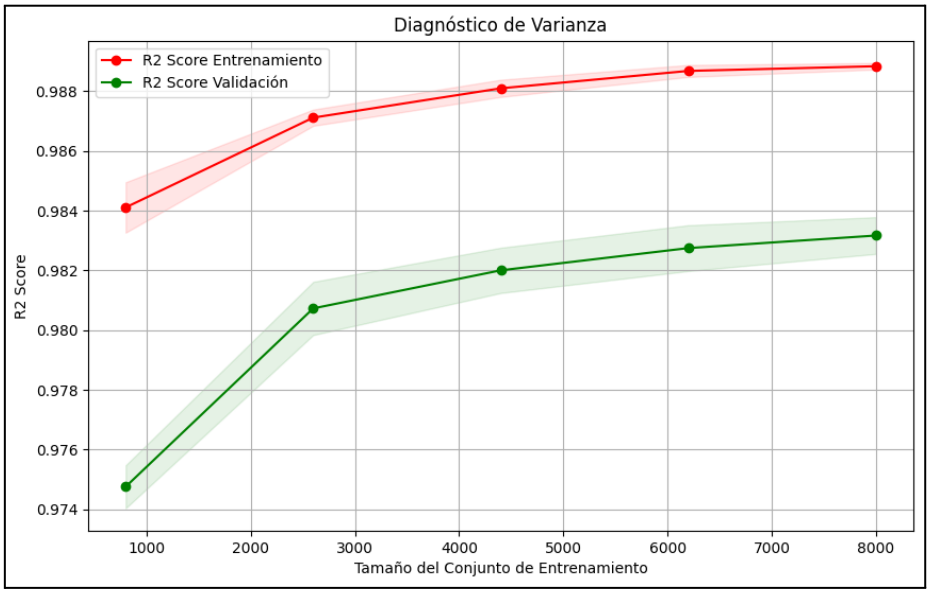


Figura 8. Comparación de curvas de aprendizaje - R2

La línea roja representa el R2 Score en el conjunto de entrenamiento a medida que aumenta el tamaño del conjunto de entrenamiento. Un crecimiento exponencial en esta línea sugiere que a medida que se proporcionan más datos de entrenamiento, el modelo mejora significativamente su rendimiento en el conjunto de entrenamiento.

Por su parte, la línea verde muestra el R2 Score en el conjunto de validación (o prueba) a medida que aumenta el tamaño del conjunto de entrenamiento. El hecho de que esta línea también muestre un crecimiento exponencial sugiere que el modelo es altamente sensible al tamaño de los datos de entrenamiento. A medida que se proporcionan más datos de entrenamiento, el modelo mejora su rendimiento en el conjunto de validación.

En conclusión, el modelo está sobreajustando (overfitting) los datos de entrenamiento. Para abordar este problema de alta varianza, se podría considerar simplificar el modelo, la regularización o validación cruzada.

4. *Diagnóstico y explicación el nivel de ajuste del modelo: underfit, fit, overfit*

Para evaluar el nivel de ajuste, es posible utilizar la gráfica utilizada en el punto 3, sin embargo, se decidió agregar una más para reafirmar la hipótesis, se utilizará MSE.

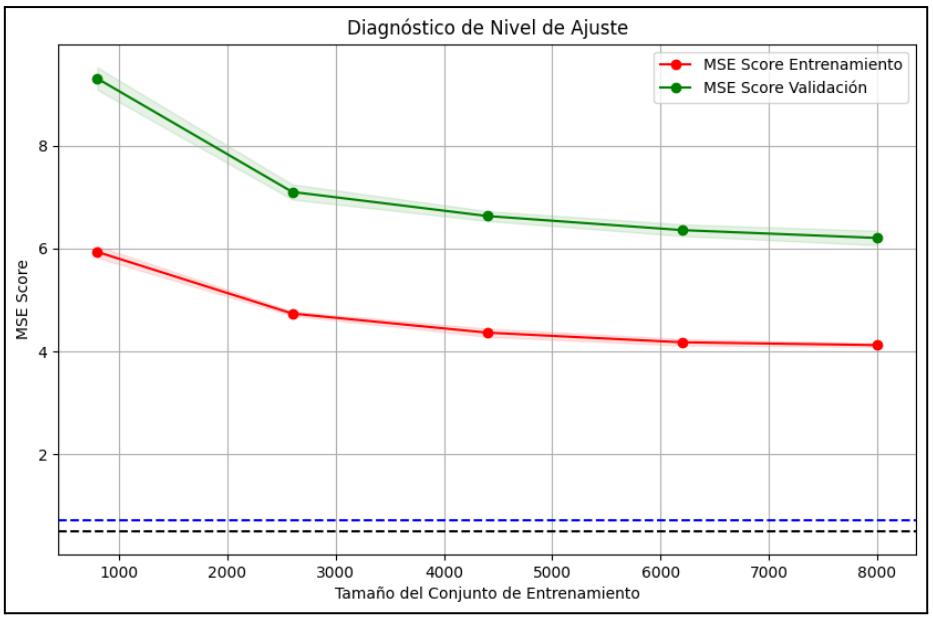


Figura 9. Comparación de curvas de aprendizaje - MSE

La línea roja representa el MSE en el conjunto de entrenamiento a medida que aumenta el tamaño del conjunto de entrenamiento. Idealmente, se espera que el MSE en el conjunto de entrenamiento disminuya a medida que se tienen más datos para entrenar el modelo. Un descenso en la línea de entrenamiento indica que el modelo está aprendiendo y mejorando su ajuste a los datos de entrenamiento.

La línea verde muestra el MSE en el conjunto de validación (o prueba) a medida que aumenta el tamaño del conjunto de entrenamiento. El objetivo es que el MSE en el conjunto de validación disminuya a medida que se tienen más datos de entrenamiento, lo que indicaría que el modelo generaliza bien a nuevos datos.

La línea horizontal negra representa un umbral de MSE. Si las líneas de entrenamiento y validación están por debajo de esta línea, indica que el modelo tiene un ajuste adecuado y es capaz de generalizar bien a nuevos datos. Por el contrario, la línea azul en el gráfico representa un umbral más alto de MSE; las líneas de entrenamiento o validación están por encima de esta línea, lo que sugiere un posible sobreajuste del modelo a los datos de entrenamiento, lo que significa que el modelo se ajusta demasiado a los datos y no generaliza bien.

5. *Uso de técnicas de regularización o ajuste de parámetros para mejorar el desempeño*
   - Búsqueda de hiper parámetros: Se utilizó la técnica de búsqueda de hiper parámetros para encontrar el número óptimo de vecinos (n_neighbors) y la métrica de distancia (p) que maximizan el rendimiento del modelo en el conjunto de validación. Los resultados impresos fueron los siguientes:

```
Validation Set MAE: 1.9465000000000001
Validation Set MSE: 6.092749999999999
Validation Set RMSE: 2.4683496510826823
Validation Set R2: 0.9835288189235224
```

```
Mejores hiperparámetros encontrados: {'n_neighbors': 9, 'p': 2}
Validation Set MAE con mejores hiperparámetros: 1.8595555555555556
Validation Set MSE con mejores hiperparámetros: 5.686802469135802
Validation Set RMSE con mejores hiperparámetros: 2.3847017568525843
Validation Set R2 con mejores hiperparámetros: 0.9846425663125375
```

Figura 11. Métricas del modelo con selección de hiper parámetros

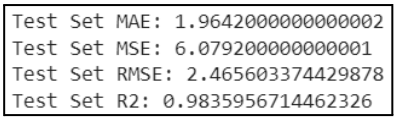
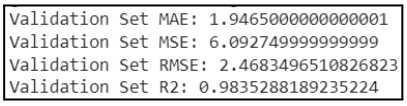


Figura 10. Métricas del modelo

Se observa una ligera mejora en los resultados al correr el modelo con los mejores hiper parámetros.

- Escala de características: Se realizó el escalaimento de las características para asegurar que todas tengan el mismo impacto en el modelo. Esto es especialmente importante en KNN, ya que se basa en distancias entre puntos.



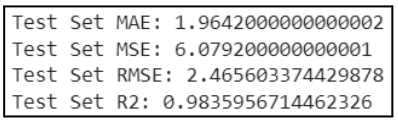


Figura 12. Métricas del model

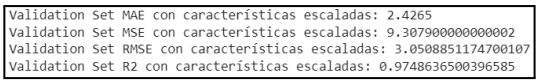
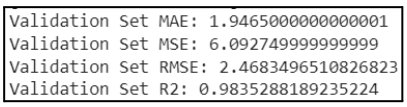


Figura 13. Métricas del modelo con características escaladas

En este caso, a pesar de ser una técnica común de preprocesamiento de datos para realizar mejoras, no se encontró una mejora en el modelo, a excepción de lo medido por R2.

- Selección de características: Se buscó seleccionar las mejores características a través del uso de *SelectKBest* y con eso, entrenar al modelo para realizar predicciones.



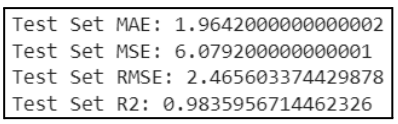


Figura 14. Métricas del model

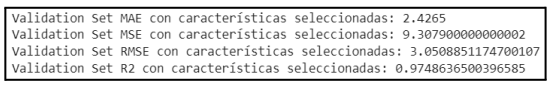


Figura 13. Métricas del modelo con características escaladas

Como se observó en la técnica previa, no se encontró mejora al modelo, por lo cual, debe ser necesario realizar diversos ajustes para seleccionar de mejor forma las características más importantes.

Conclusiones

El presente reporte ha proporcionado una visión detallada del proceso de implementación de técnicas de aprendizaje automático, específicamente el algoritmo k Nearest Neighbors (KNN), utilizando un conjunto de datos de rendimiento estudiantil. Se han abordado aspectos clave, como la separación de datos en conjuntos de entrenamiento, validación y prueba, la evaluación del rendimiento del modelo a través de métricas como MAE, R2 y MSE, y la búsqueda de hiperparámetros para mejorar el desempeño del modelo.

Se identificaron problemas de sobreajuste en el modelo KNN, lo que sugiere la necesidad de tomar medidas para reducir la alta varianza. Se han propuesto soluciones como la simplificación del modelo y la regularización. Además, se exploraron técnicas como el

escalado de características y la selección de características, aunque no se observó una mejora significativa en el rendimiento del modelo en este contexto particular.

Gracias al presente reporte, se encontró una base sólida para continuar refinando el modelo KNN y abordar los desafíos de bias y varianza para lograr un ajuste adecuado y una mejor generalización a nuevos datos.

Fuentes consultadas

- ¿Qué es Machine Learning? (s/f). Ibm.com. Recuperado el 7 de septiembre de 2023, de https://www.ibm.com/mx-es/analytics/machine-learning
- Conjunto de datos obtenido de Narayan, N. (2023). Student Performance [Data set]. En Student Performance (Multiple Linear Regression). https://www.kaggle.com/datasets/nikhil7280/student-performance-multiple-linear-regression