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

**Actividad INDIVIDUAL**

Luis David Maza Ramírez A01747890

Inteligencia artificial avanzada para la ciencia de datos I
TC3006C
Grupo 101
11 de septiembre del 2024

***Implementación de Machine Learning a analizar:***

Random Forest.

***Información del Dataset:***

El dataset utilizado contiene información acerca de los componentes del cemento, con el objetivo de predecir la dureza del cemento dependiendo de los materiales usados, tiempo de producción, etc. Algunos de los componentes que utiliza el modelo para predecir la dureza del cemento son; la cantidad de mezcla de cemento utilizada, el tiempo de horneado de la mezcla, la cantidad de agua utilizada, cantidad de superplastico utilizado, cuantos días se dejó secar la mezcla y por último el Target, que representa la fuerza o la dureza del mismo.

***Justificación del uso del Dataset:***

Uno de los motivos principales por los cuáles se decidió usar este dataset, es porque al hacer un análisis profundo de las diferentes variables y columnas, se puede encontrar que en las mismas existe una relación bastante profunda, lo cual facilita la interpretación se los resultados, dando como resultado un modelo más eficiente a la hora de hacer sus predicciones.

Sumado a ello, aunque podemos encontrar que los datos del dataset en su mayoría son numéricos, se puede asumir que a la hora de obtener la variable Target, se le pude asignar una categorización a la hora de sacar su predicción, es decir determinar si la dureza de un cemento es buena o mala con respecto a los componentes de la mezcla. El hacer esta categorización facilita el proceso de medir y sacar nuestras métricas del modelo.

Por último, es importante mencionar que este dataset cuenta con una gran cantidad de datos con los cuales se puede trabajar de una manera muy eficiente, sobre todo a la hora de dividir los datos en componentes de entrenamiento, prueba y validación.

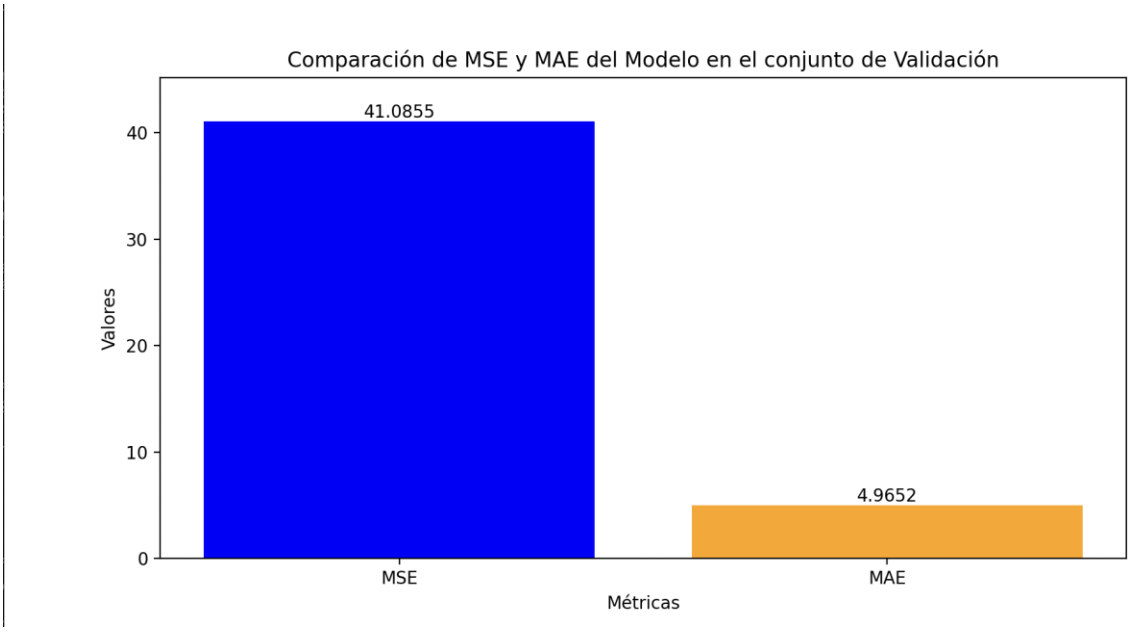
***Separación y evaluación del modelo con un conjunto de prueba y un conjunto de validación (Train/Test/Validation).***

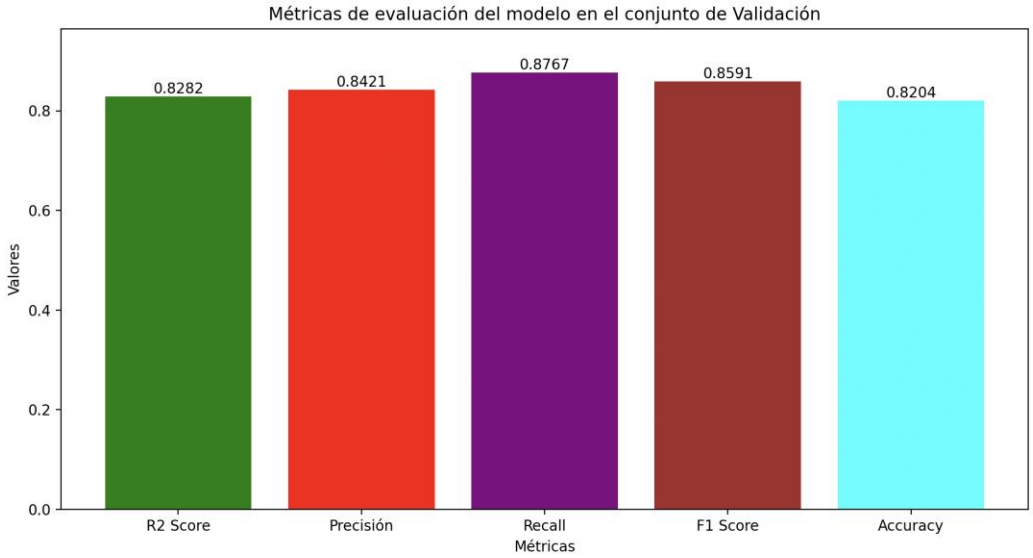
Hacer la separación respectiva de datos entre conjuntos de entrenamiento, pruebas y validación, fue un proceso de suma importancia a la hora de implementar el modelo de machine learning, esto debido a que gracias a dichos conjuntos, evitamos que el modelo se ajustara a un solo conjunto de datos, aprendiendose muy bien todos los patrones y perdiendo la capacidad de generalización de nuevos datos. Sumado a ello, gracias al conjunto de prueba pudimos generar una evaluación objetiva final del modelo, sobre datos que no había visto antes.

Para la implementación de este modelo, la parte de la separación de datos representó algo muy estructurado, dividiendo el 60% de los datos para el conjunto de entrenamiento, 20% para el conjunto de validación y el último 20% restante para el conjunto de pruebas.

```
Dataset completo:
   CementComponent   BlastFurnaceSlag  FlyAshComponent  WaterComponent  SuperplasticizerComponent  CoarseAggregateComponent  FineAggregateComponent  AgeInDays  Strength
0            540.0               0.0              0.0           162.0                        2.5                    1040.0                   676.0         28     79.99
1            540.0               0.0              0.0           162.0                        2.5                    1055.0                   676.0         28     61.89
2            332.5             142.5              0.0           228.0                        0.0                     932.0                   594.0        270     40.27
3            332.5             142.5              0.0           228.0                        0.0                     932.0                   594.0        365     41.05
4            198.6             132.4              0.0           192.0                        0.0                     978.4                   825.5        360     44.30
Tamaño del dataset:
(1030, 9)

Conjunto de entrenamiento:
     CementComponent   BlastFurnaceSlag  FlyAshComponent  WaterComponent  SuperplasticizerComponent  CoarseAggregateComponent  FineAggregateComponent  AgeInDays
546            333.0               0.0              0.0           192.0                        0.0                     931.2                   842.6         28
21             139.6             209.4              0.0           192.0                        0.0                    1047.0                   806.9         28
669            288.0             192.0              0.0           192.0                        0.0                     932.0                   717.8          3
742            480.0               0.0              0.0           192.0                        0.0                     936.0                   721.0         90
545            255.0               0.0              0.0           192.0                        0.0                     889.8                   945.0         28
Tamaño del conjunto de entrenamiento:
(618, 8)

Conjunto de validación
     CementComponent   BlastFurnaceSlag  FlyAshComponent  WaterComponent  SuperplasticizerComponent  CoarseAggregateComponent  FineAggregateComponent  AgeInDays
634            275.0               0.0              0.0           183.0                        0.0                    1088.0                   808.0         28
101            388.6              97.1              0.0           157.9                       12.1                     852.1                   925.7          7
813            310.0               0.0              0.0           192.0                        0.0                     970.0                   850.0        180
345            213.7               0.0            174.7           154.8                       10.2                    1053.5                   776.4         14
263            212.6               0.0            100.4           159.4                       10.4                    1003.8                   903.8        100
Tamaño del conjunto de validación:
(206, 8)

Conjunto de prueba
     CementComponent   BlastFurnaceSlag  FlyAshComponent  WaterComponent  SuperplasticizerComponent  CoarseAggregateComponent  FineAggregateComponent  AgeInDays
325            252.3               0.0             98.8           146.3                       14.2                     987.8                   889.0         14
836            304.0             140.0              0.0           214.0                        6.0                     895.0                   722.0         28
538            480.0               0.0              0.0           192.0                        0.0                     936.2                   712.2          7
20             427.5              47.5              0.0           228.0                        0.0                     932.0                   594.0        180
71             313.3             262.2              0.0           175.5                        8.6                    1046.9                   611.8          3
Tamaño del conjunto de prueba:
(206, 8)
(ml_env) (base) %n@%m %1~ %#
```

Sumado a ello, gracias al conjunto de validación, se pudieron obtener las diferentes métricas del modelo en base al conjunto de validación, lo cual nos permitió observar y analizar de una mejor manera el desempeño del modelo frente a diferentes datos.

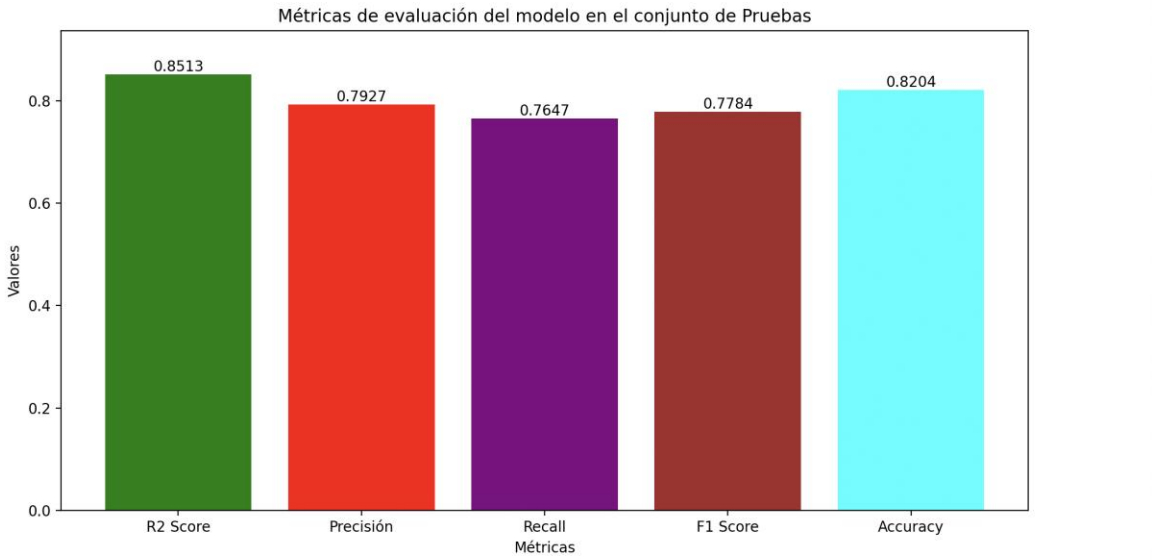





***Diagnóstico y explicación el grado de bias o sesgo: bajo medio alto:***

Para la implementación de este modelo y después de analizar detenidamente las métricas de nuestro modelo, se puede entender que el bias de este modelo es bajo, esto debido a que los diferentes indicadores como el recall, el f1 score, la precisión y el accuracy, se encuentran muy parecidos entre sí, si bien, es importante mencionar que aunque los datos si se encuentran muy cercanos, estos mismos se encuentran en el límite entre el bias bajo y el bias medio, ya que datos como el R2 score y el accuracy del modelo, si tienen datos un poco mayores, sin embargo, todavía se encuentran en lo considerable para hablar de un bias bajo.

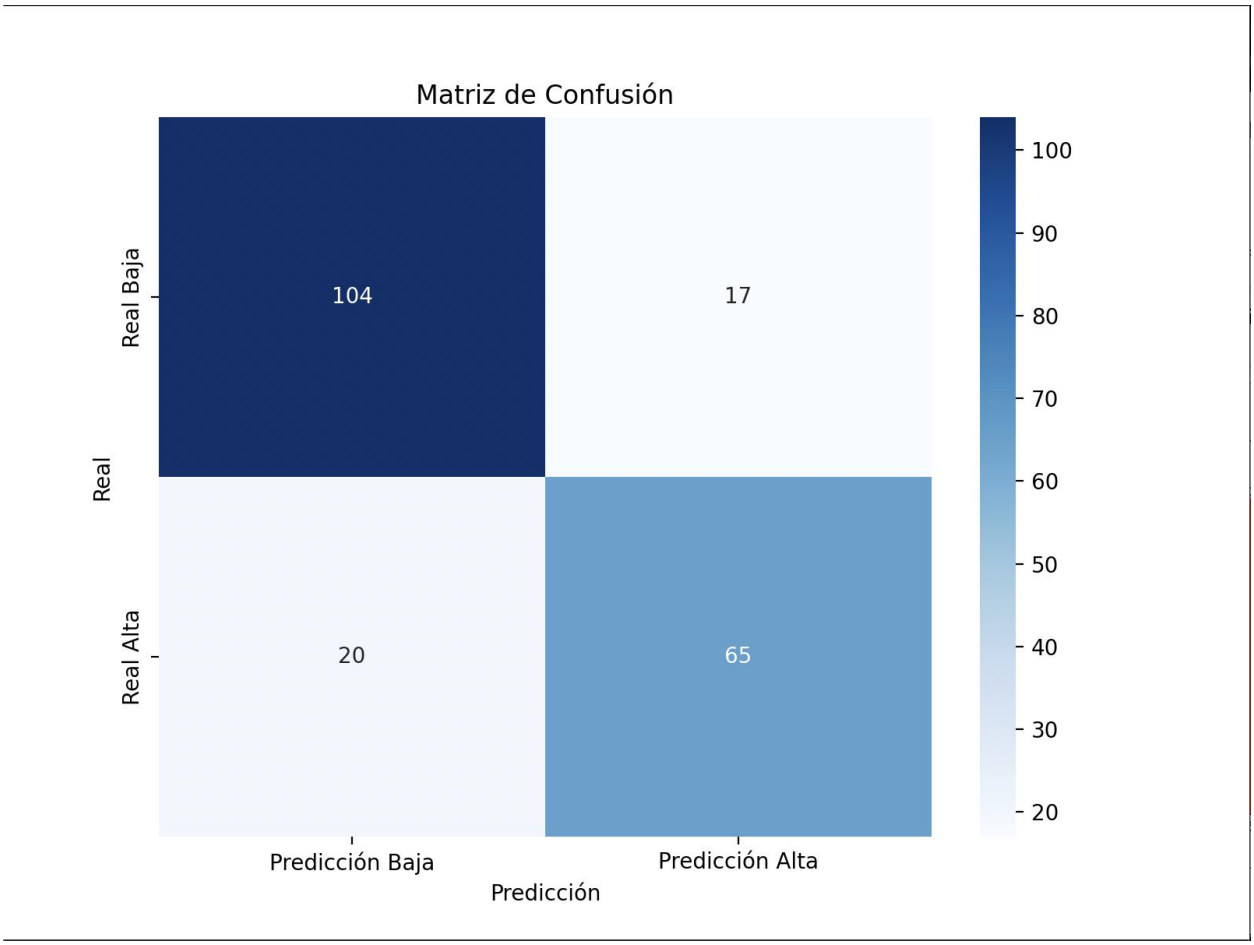
```
Test MSE: 40.8955
Test MAE: 5.0751
Test R2 Score: 0.8513
Accuracy: 0.8204
Test Precisión: 0.7927
Test Recall: 0.7647
Test F1 Score: 0.7784
```




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

Gracias al análisis de nuestra matriz de confusión, podemos detectar que nuestro modelo si bien tiene un buen número de predicciones, en algunas ocasiones, a la hora de predecir los datos tiene una que otra predicción equivocada.

Dicho eso y comprendiendo que la cantidad de verdaderos positivos y falsos negativos es bastante alta, se puede entender que el grado de varianza del modelo es medio, ya que, si tiene muchos aciertos, es probable que, con un conjunto de prueba con mayor cantidad de datos, el modelo tienda se equivoca un poco más, pero en esta ocasión el rendimiento del modelo con los datos usados, son consistentes y con una buena predicción.




***Diagnóstico y explicación el nivel de ajuste del modelo: underfitt fitt overfitt:***

Finalmente, gracias a las diferentes métricas de evaluación del modelo calculadas previamente, junto con sus gráficas comparativas, se puede concluir que el nivel de ajuste del modelo es adecuado y está relacionado con un fitting correcto. Esto se evidencia en que las métricas de rendimiento, como el MSE, MAE y R2 score, son consistentes en los conjuntos de entrenamiento y en los de prueba y validación.

Sumado a ello, gracias al análisis previo, se puede entender que contamos con un nivel de bias bajo y un nivel de varianza medio, lo cual da como consecuencia un modelo con bastante coherencia en el rendimiento de varios subconjuntos de datos. Si bien, el conjunto tiene muchas áreas en las cuales mejorar para expresar mejores predicciones y generalizaciones, el modelo presenta un ajuste preciso en la mayoría de sus predicciones.

***Técnicas de regularización o ajuste de parámetros para mejorar el desempeño del modelo:***

**AJUSTE DE HIPERPARÁMETROS. (n_estimators)**

Para aplicar este modelo, se pudo aplicar la técnica de ajustar los parámetros de nuestro modelo, más específicamente en el número de estimadores, es decir, aumentaremos el número de árboles de nuestro bosque para tener una mejor precisión en el modelo.

ANTES:

```
(n_estimators=19,
```

```
Accuracy: 0.8204
Test Precisión: 0.7927
Test Recall: 0.7647
Test F1 Score: 0.7784
```

DESPUÉS:

```
(n_estimators=50,
```

```
Accuracy: 0.8447
Test Precisión: 0.8354
Test Recall: 0.7765
Test F1 Score: 0.8049
```

**AJUSTE DE HIPERPARÁMETROS. (min_samples_leaf)**

De la misma manera, podemos ajustar el parámetro de nuestro árbol de nuestro min_smples_leaf, es decir ajustar el número mínimo de muestras que debe tener una hoja de cada árbol. Este parámetro es de suma importancia ya que nos ayuda a controlar la profundidad del árbol y ayuda a evitar el sobreajuste.

ANTES:

```
min_samples_leaf=8)
```

```
Accuracy: 0.8204
Test Precisión: 0.7927
Test Recall: 0.7647
Test F1 Score: 0.7784
```

DESPUÉS:

```
min_samples_leaf=10)
```

```
Accuracy: 0.8447
Test Precisión: 0.8354
Test Recall: 0.7765
Test F1 Score: 0.8049
```

**Cross Validation:**

Una técnica también implementada para mejorar el rendimiento del modelo es la validación cruzada por medio de KFold, esto con el objetivo de no solo usar los conjuntos previamente establecidos de entrenamiento, validación y prueba, si no, que por medio de esta técnica se divide el dataset en diferentes particiones de igual tamaño, proporcionando una validación más robusta y menos sesgada del rendimiento del modelo que una única división en entrenamiento y prueba.

```
# Implementamos la validación cruzada con K-Folds (k=5 en este caso)
kf = KFold(n_splits=5, shuffle=True, random_state=0)
```