Instituto Tecnológico y de Estudios Superiores de Monterrey

Campus Estado de México

Fecha de entrega: 11 de septiembre del 2023

Análisis y Reporte sobre el desempeño del modelo

Inteligencia artificial avanzada para la ciencia de datos I (Gpo. 101)

Profesor:

Jorge Adolfo Ramírez Uresti

Alumno:

José Ángel García Gómez A01745865

Justificación de la selección del dataset:

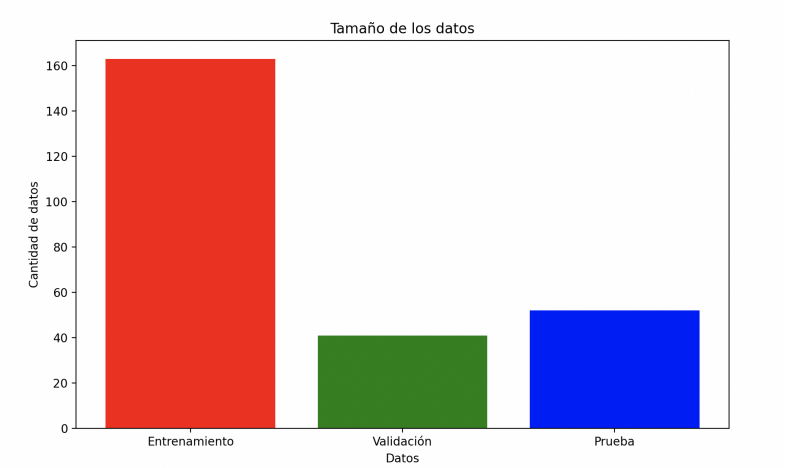
En primer lugar, es importante destacar que la elección de esta fuente de datos surge de un interés personal. El tema de los automóviles es de gran interés y pasión, lo que proporciona una motivación adicional para llevar a cabo esta investigación.

Además, la fuente de datos que se ha elegido se caracteriza por ser bastante completa y contar con una gran cantidad de información. Esta abundancia de datos es esencial cuando se trata de entrenar modelos de aprendizaje automático, como Random Forest. Cuantos más datos tengamos, nuestros resultados serán más precisos y estables, lo que facilitará una mejor generalización. La capacidad del modelo para generalizar, gracias a la fuente de datos proporcionada, se demuestra en el siguiente gráfico que representa la distribución de las variables objetivo:

Otro aspecto crucial es la estructura de la fuente de datos en sí. Esta fuente de datos de automóviles es especialmente adecuada para la implementación de un modelo de Random Forest de clasificación debido a su composición de variables. En particular, la fuente de datos incluye una variable objetivo categórica: la nacionalidad del automóvil. Por otro lado, se disponen de variables predictoras numéricas que representan las características de los automóviles. Esta combinación de variables es esencial, ya que se ajusta perfectamente a la metodología de Random Forest, que se destaca en la clasificación de datos con múltiples variables predictoras y una variable objetivo categórica.

Separación y evaluación del modelo con un conjunto de prueba y un conjunto de validación:

La estrategia de dividir la fuente de datos en conjuntos de entrenamiento, validación y prueba es esencial para garantizar la solidez y la capacidad de generalización de un modelo de Random Forest. En este enfoque, inicialmente, se asigna el 80% de los datos al conjunto de entrenamiento, reservando el 20% restante para el conjunto de prueba, asegurando que el modelo no vea estos datos durante el entrenamiento. Luego, se realiza una segunda división en el conjunto de entrenamiento, separando un 20% adicional para fines de validación. Esta división se puede apreciar mejor en el siguiente gráfico:



Asimismo, a continuación, se muestra una selección de registros de cada subconjunto para asegurar que no se repitan los datos:

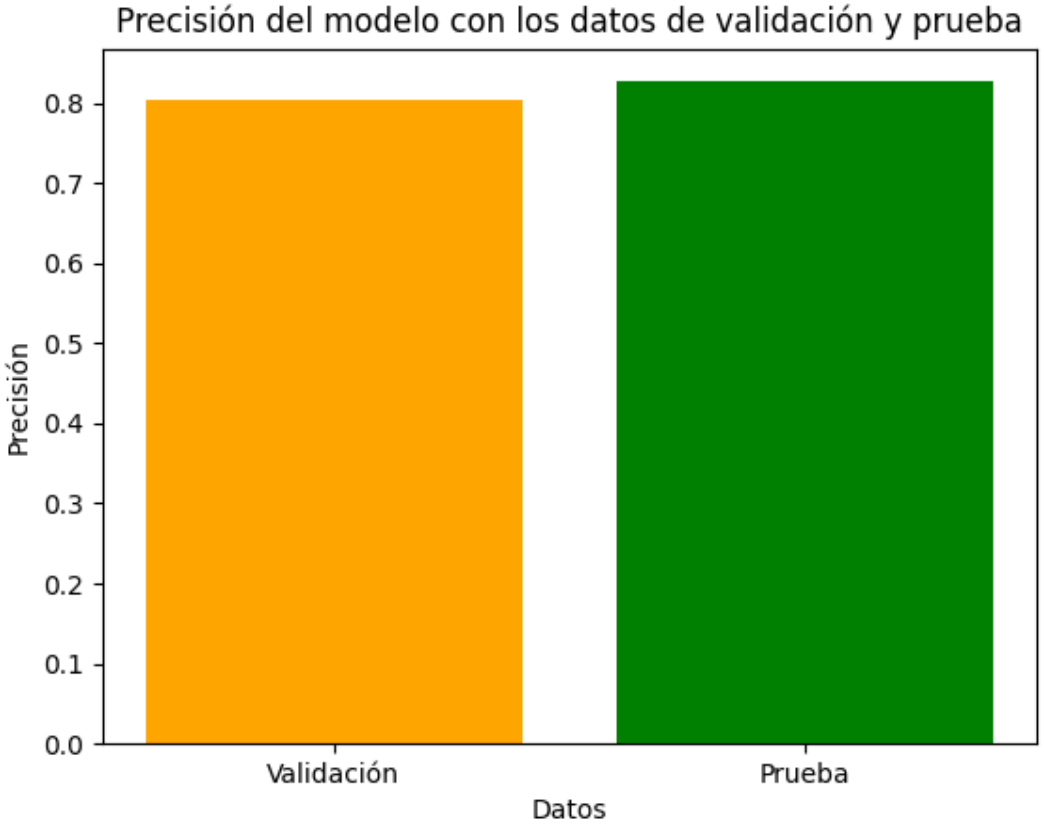
```
Datos de entrenamiento:
      mpg  cylinders  cubicinches   hp  weightlbs  time-to-60  year
109  18.0          8        307.0  130     3504.0          12  1971
141  29.0          4         97.0   75     2171.0          16  1976
88   28.0          4         90.0   75     2125.0          15  1975
116  24.5          4        151.0   88     2740.0          16  1978
131  28.0          4        151.0   90     2678.0          17  1981

Datos de validación:
      mpg  cylinders  cubicinches   hp  weightlbs  time-to-60  year
183  18.0          8        318.0  150     3436.0          11  1971
176  14.0          8        400.0  175     4385.0          12  1973
97   18.0          6        250.0   88     3139.0          15  1972
101  12.0          8        400.0  167     4906.0          13  1974
121  26.0          4         79.0   67     1963.0          16  1975

Datos de prueba:
      mpg  cylinders  cubicinches   hp  weightlbs  time-to-60  year
233  16.0          8        318.0  150     4498.0          15  1976
6    13.0          8        351.0  158     4363.0          13  1974
82   19.2          8        305.0  145     3425.0          13  1979
211  17.7          6        231.0  165     3445.0          13  1979
120  27.0          4        101.0   83     2202.0          15  1977
```

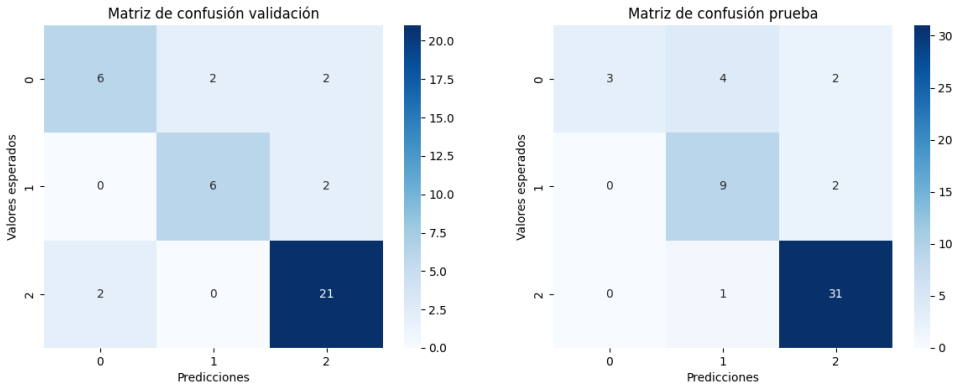
Esta división en tres conjuntos tiene como objetivo principal prevenir el sobreajuste del modelo, donde este se ajustaría excesivamente a los datos de entrenamiento y no sería capaz de generalizar correctamente a datos desconocidos. Asimismo, el conjunto de prueba, que se mantiene completamente aparte durante el proceso de entrenamiento y validación, sirve como una evaluación imparcial del desempeño real del modelo en datos no observados previamente. En resumen, esta estrategia asegura que el modelo Random Forest esté ajustado

de manera adecuada, optimizado y evaluado de manera objetiva, lo que es esencial para su eficacia en la clasificación y predicción de nuevos datos.

Diagnóstico y explicación el grado de bias o sesgo

Para identificar posibles sesgos en el modelo, es esencial comparar las precisiones alcanzadas con diferentes tipos de datos, como los conjuntos de validación y prueba. Esta comparación nos permite evaluar tanto las diferencias entre las precisiones como la capacidad del modelo para clasificar los datos de manera efectiva. Para llevar a cabo este análisis, se representaron gráficamente los resultados obtenidos para cada tipo de dato. Este enfoque nos permite visualizar de manera clara y efectiva cómo el modelo se desempeña en diferentes conjuntos de datos y, por lo tanto, identificar posibles sesgos o discrepancias en su rendimiento.



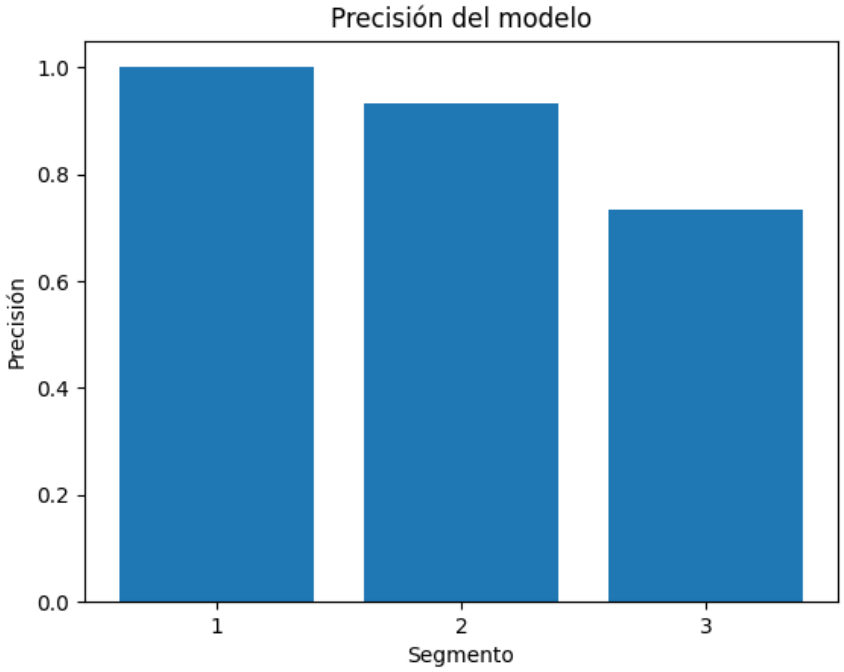


Después de analizar detenidamente las gráficas resultantes, se evidencia que el rendimiento del modelo es altamente preciso tanto en los datos de validación como en los datos de prueba. Este hallazgo sugiere que existe un **bajo sesgo** en nuestro modelo, lo que significa que está realizando predicciones con gran precisión en diferentes conjuntos de datos. Sin embargo, es importante destacar que, al examinar las matrices de confusión, se observa que el modelo también comete errores de clasificación. Estos errores indican que, a pesar de su precisión general, el modelo no es infalible y ocasionalmente falla al predecir correctamente algunas instancias. Esto subraya la importancia de una evaluación exhaustiva y de comprender las

áreas en las que el modelo puede necesitar mejoras adicionales para lograr un rendimiento aún más sólido.

Diagnóstico y explicación el grado de varianza

Para evaluar el grado de variabilidad en el rendimiento del modelo, es necesario validar su desempeño utilizando diferentes conjuntos de datos. Esto nos permite cuantificar cuánto varía el rendimiento en función de las diferentes muestras de datos utilizadas para entrenar y evaluar el modelo. Esta evaluación de la variabilidad es fundamental para comprender cómo el modelo se comporta en diversas situaciones y si su rendimiento es consistente o fluctúa significativamente. En última instancia, nos proporciona información valiosa sobre la estabilidad y la robustez del modelo en un contexto más amplio.



Luego de evaluar el rendimiento del modelo con distintos conjuntos de datos, se observa que existe una **variabilidad media** en su desempeño. Esta variabilidad se manifiesta en los primeros conjuntos de datos, donde las diferencias en el rendimiento son relativamente pequeñas. Sin embargo, es importante destacar que se encuentra una variación significativa en el rendimiento al considerar el tercer segmento de datos. Este hallazgo sugiere que el modelo es generalmente estable y coherente en su rendimiento, pero existen circunstancias o características particulares en el tercer conjunto de datos que pueden influir en su capacidad

predictiva de manera más notable. Por lo tanto, es fundamental prestar atención a estos escenarios específicos para comprender mejor las condiciones en las que el modelo puede beneficiarse de mejoras adicionales o ajustes específicos.

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

Para la validación del nivel de ajuste de un modelo de clasificación de Random Forest se basa en una evaluación exhaustiva del rendimiento del modelo en tres subconjuntos de datos clave: entrenamiento, validación y prueba. Al observar las métricas de evaluación en cada uno de estos conjuntos, se puede concluir que el modelo exhibe un buen nivel de ajuste.

```
***** Datos de validación ******
Precisión de los datos de validación:  0.7317073170731707
Reporte de clasificación de los datos de validación:
              precision    recall  f1-score   support

     Europe.       0.71      0.50      0.59        10
      Japan.       0.50      0.38      0.43         8
         US.       0.79      0.96      0.86        23

    accuracy                           0.73        41
   macro avg       0.67      0.61      0.63        41
weighted avg       0.71      0.73      0.71        41

***** Datos de prueba ******
Precisión de los datos de prueba:  0.8653846153846154
Reporte de clasificación de los datos de prueba:
              precision    recall  f1-score   support

     Europe.       1.00      0.56      0.71         9
      Japan.       0.69      0.82      0.75        11
         US.       0.91      0.97      0.94        32

    accuracy                           0.87        52
   macro avg       0.87      0.78      0.80        52
weighted avg       0.88      0.87      0.86        52

***** Datos de entrenamiento ******
Precisión de los datos de entrenamiento:  0.9202453987730062
Reporte de clasificación de los datos de entrenamiento:
              precision    recall  f1-score   support

     Europe.       0.84      0.96      0.90        28
      Japan.       0.81      0.91      0.85        32
         US.       0.99      0.91      0.95       103

    accuracy                           0.92       163
   macro avg       0.88      0.93      0.90       163
weighted avg       0.93      0.92      0.92       163
```

Se destaca que las métricas de evaluación en estos subconjuntos son comparables, lo que indica que el modelo no sufre de sobreajuste (donde el rendimiento es alto en entrenamiento pero bajo en prueba) ni subajuste (donde el rendimiento es bajo en todos los conjuntos), sino que logra un equilibrio adecuado. Además, se resalta la precisión alcanzada en cada conjunto: 73% en validación, 87% en prueba y 92% en entrenamiento. Estas cifras reflejan un rendimiento razonablemente alto y coherente del modelo en todos los subconjuntos.

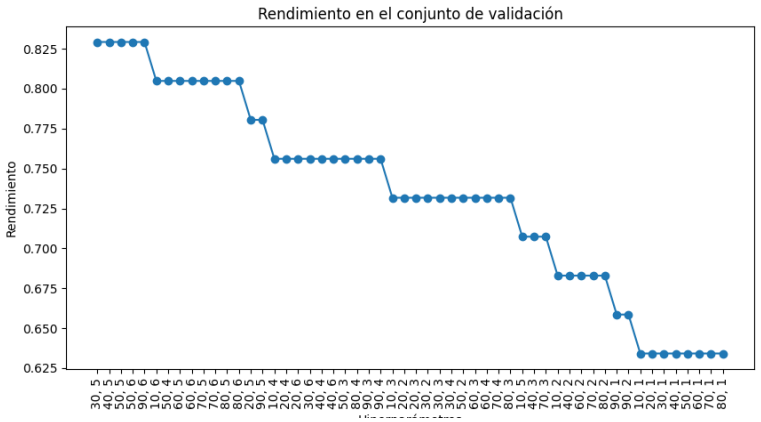
Además, al utilizar el método de 'Cross Validation' para la evaluación, se confirmó que el modelo presenta un rendimiento consistente y aceptable en todas las situaciones, como se observa en la captura de pantalla a continuación:

```
Precisiones obtenidas con cross validation:  [0.78846154 0.70588235 0.78431373 0.82352941 0.78431373]
Precisión promedio del modelo:  0.7773001508295625
```

Mejoramiento del rendimiento del modelo

Dado que los Random Forests se construyen con una base sólida para evitar el sobreajuste, la aplicación de técnicas de regularización L1 o L2, diseñadas principalmente para controlar el sobreajuste en modelos lineales, puede no producir mejoras significativas en su desempeño. Los Random Forests ya cuentan con mecanismos internos que limitan la profundidad de los árboles y evitan el ajuste excesivo a los datos, lo que disminuye la necesidad de una regularización adicional.

En lugar de centrarse en L1 o L2, se optó por explorar la optimización de otros hiper parámetros específicos de Random Forest, como el número de árboles en el ensemble así como la profundidad máxima de los árboles. Estos ajustes suelen tener un impacto más significativo en el rendimiento de un Random Forest que la aplicación de regularización L1 o L2, que están diseñadas para abordar desafíos específicos de los modelos lineales. El rendimiento de diferentes combinaciones de hiper parámetros se muestra en el siguiente gráfico comparativo:



El gráfico sugiere que la configuración de hiper parámetros óptima sería seleccionar 30 estimadores (árboles generados) con una profundidad máxima (max_depth) de 5, con los cuales se obtiene un rendimiento del 0.83%.

Para apreciar de manera más efectiva las mejoras logradas mediante la modificación de los hiper parámetros, se lleva a cabo una comparación de la precisión mediante la técnica de validación cruzada para cada uno de los modelos. Esta comparación nos proporciona resultados que indican claramente una mejora en la precisión. La validación cruzada nos permite evaluar el rendimiento de los modelos de manera más robusta al considerar múltiples divisiones de los datos. Al analizar estos resultados, se confirma de manera concluyente que las modificaciones en los hiper parámetros han conducido a una mejora en la precisión del modelo, lo que respalda la efectividad de los ajustes realizados. Esta validación cruzada refuerza la confianza en las mejoras implementadas y demuestra su impacto positivo en la calidad del modelo.