Instituto Tecnológico y de Estudios Superiores de Monterrey Campus Estado de México

Materia:

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

Análisis y reporte sobre el desempeño del modelo

Santiago Espinosa Domínguez

A01747478

Jorge Adolfo Ramírez Uresti

11 de septiembre del 2024

La implementación en la que se va a hacer el análisis va a ser del modelo de predicción de dígitos escritos a mano. Los resultados se dividen en 2, la primera parte es la prueba con el conjunto de validación y la segunda es con el conjunto de prueba.

Los resultados del desempeño del modelo de clasificación evaluado en conjuntos de validación muestran las métricas de precisión, recall, puntaje F1 y el soporte para cada dígito. La precisión mide la proporción de identificaciones correctas para cada clase, donde tenemos un promedio de 97% el cual llega a ser muy alto, donde todos los dígitos excepto el 0, 1, 8 y 9 alcanzan una precisión perfecta del 100%. El recall refleja la proporción de positivos verdaderos identificados correctamente, destacando solo los dígitos 0, 2, 6 y 9 alcanzan un 100%, lo que indica que fueron correctamente reconocidos. El puntaje F1, que es la media armónica de la precisión y el recall, muestra un equilibrio perfecto para 2 dígitos, que son el 2 y el 6, con un puntaje de 1.00. Finalmente, el soporte indica el número de muestras reales de cada clase en el conjunto de prueba, ayudando a contextualizar las otras métricas según el volumen de datos evaluados por clase, como podemos observar hay un poco de varianza entre las clases ya que va desde 29 muestras del digito 8 hasta 46 muestras del digito 3.

En conclusión, el modelo con los datos de validación muestra ser altamente efectivo y equilibrado en su clasificación, con métricas de rendimiento impresionantemente altas tanto a nivel global como por clase individual.

Precisión en validación: 0.975
Reporte de Clasificación para el Conjunto de Validación:

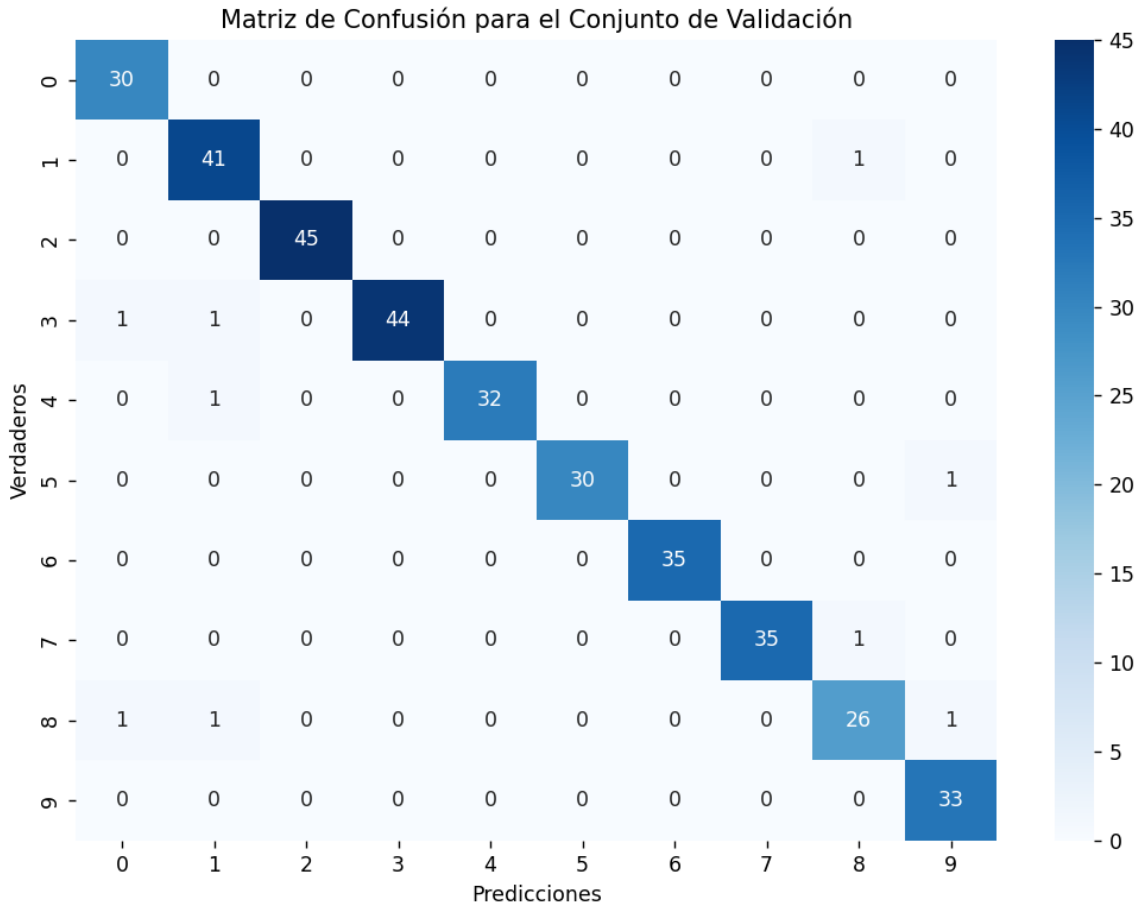
	precision	recall	f1-score	support
0	0.94	1.00	0.97	30
1	0.93	0.98	0.95	42
2	1.00	1.00	1.00	45
3	1.00	0.96	0.98	46
4	1.00	0.97	0.98	33
5	1.00	0.97	0.98	31
6	1.00	1.00	1.00	35
7	1.00	0.97	0.99	36
8	0.93	0.90	0.91	29
9	0.94	1.00	0.97	33
accuracy			0.97	360
macro avg	0.97	0.97	0.97	360
weighted avg	0.98	0.97	0.98	360

Reporte de clasificación para el conjunto de validación

Como de igual manera se puede observar la matriz de confusión reafirman los resultados anteriores, la matriz muestra los dígitos predichos por el algoritmo y los dígitos verdaderos en la cual podemos ver que llega a cometer solo 9 errores los cuales son:

1. El dígito era un 3 pero el modelo predijo un 0
2. El dígito era un 3 pero el modelo predijo un 1
3. El dígito era un 4 pero el modelo predijo un 1
4. El dígito era un 8 pero el modelo predijo un 0
5. El dígito era un 8 pero el modelo predijo un 1
6. El dígito era un 1 pero el modelo predijo un 8
7. El dígito era un 7 pero el modelo predijo un 8
8. El dígito era un 5 pero el modelo predijo un 9
9. El dígito era un 8 pero el modelo predijo un 9

Los errores presentados son muy pocos y están muy variados entre las clases, lo cual demuestra que no se está atorando en predecir algún dígito en específico.



Matriz de confusión para el conjunto de validación

Después de ajustar el modelo de acuerdo con los datos de validación se probó con los datos de prueba, datos que jamás había visto y el resultado fue el siguiente.

Los resultados muestran que en la precisión tenemos un promedio de 97% el cual llega a ser muy alto, donde solo los dígitos 0,2, 3 y 8 alcanzan una precisión perfecta del 100%. El recall del modelo muestra que solo los dígitos 1,2 y 4 alcanzan un 100%, lo que indica que fueron correctamente reconocidos. El resultado del puntaje F1, muestra un equilibrio perfecto para el dígito 2 con un puntaje de 1.00. Finalmente, la varianza del soporte es más grande que la del conjunto de validación ya que va desde 28 muestras del digito 1 hasta 47 muestras del digito 5.

Precisión en prueba: 0.9694444444444444
Reporte de Clasificación para el Conjunto de Prueba:

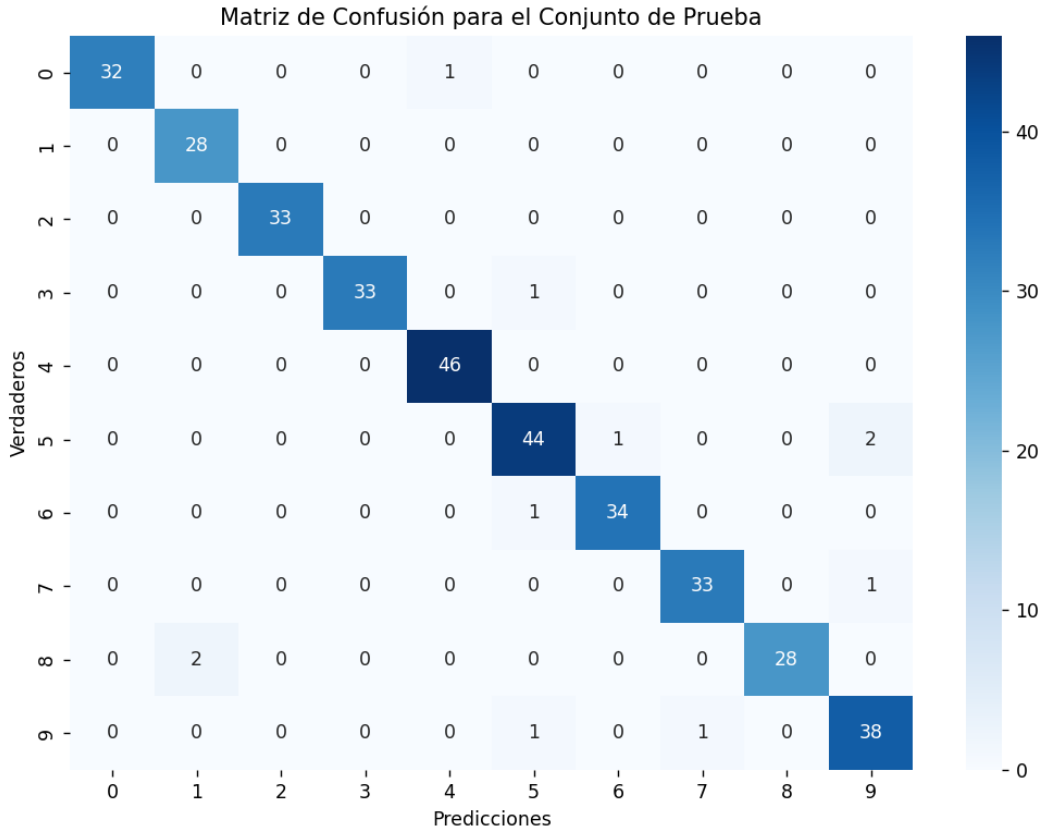
	precision	recall	f1-score	support
0	1.00	0.97	0.98	33
1	0.93	1.00	0.97	28
2	1.00	1.00	1.00	33
3	1.00	0.97	0.99	34
4	0.98	1.00	0.99	46
5	0.94	0.94	0.94	47
6	0.97	0.97	0.97	35
7	0.97	0.97	0.97	34
8	1.00	0.93	0.97	30
9	0.93	0.95	0.94	40
accuracy			0.97	360
macro avg	0.97	0.97	0.97	360
weighted avg	0.97	0.97	0.97	360

Reporte de clasificación para el conjunto de prueba

Para reafirmar los datos anteriores, acudimos a la matriz de confusión la cual muestra los dígitos predichos por el algoritmo con el conjunto de prueba y los dígitos verdaderos en la cual podemos ver que llega a cometer solo 10 errores los cuales son:

1. El dígito era un 8 pero el modelo predijo un 1 dos veces
2. El dígito era un 0 pero el modelo predijo un 4
3. El dígito era un 3 pero el modelo predijo un 5
4. El dígito era un 6 pero el modelo predijo un 5
5. El dígito era un 9 pero el modelo predijo un 5
6. El dígito era un 5 pero el modelo predijo un 6
7. El dígito era un 5 pero el modelo predijo un 9 dos veces
8. El dígito era un 7 pero el modelo predijo un 9

Los errores presentados son muy pocos, pero llegan a ser mayores que los del conjunto de validación, pero solo por uno, lo bueno es que están muy variados entre las clases, lo cual demuestra que no se le está dificultando predecir algún digito en específico lo cual muestra un gran rendimiento del modelo.



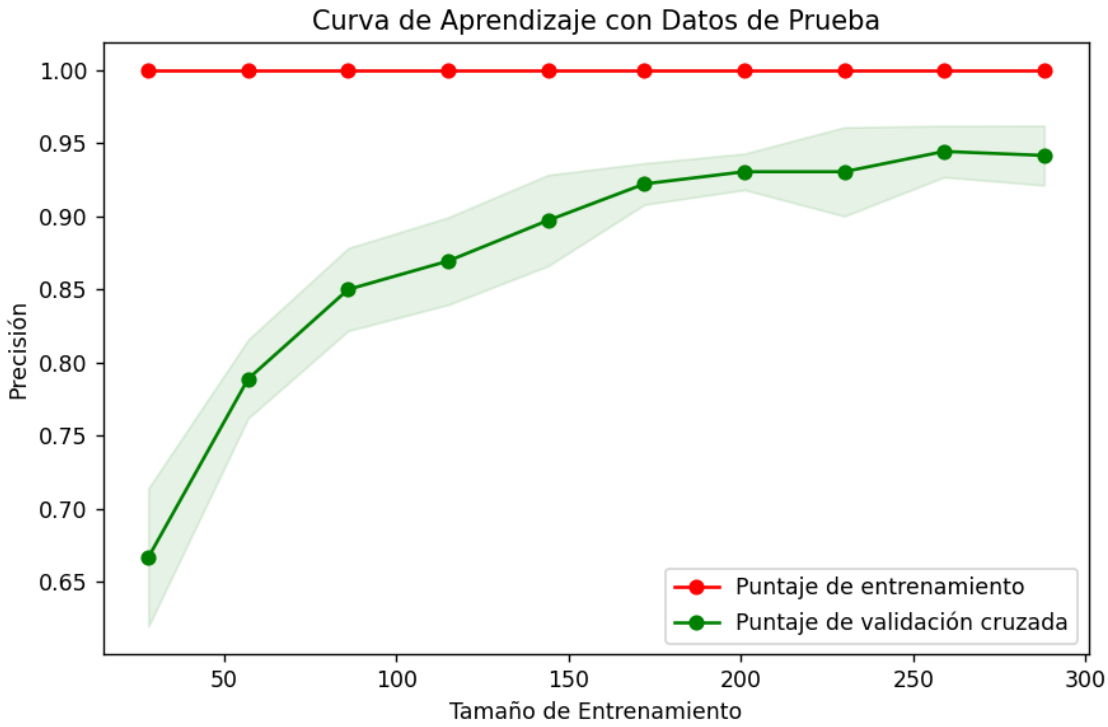
Matriz de confusión para el conjunto de prueba

Para poder hacer un análisis mas profundo del modelo se requirió usar una curva de aprendizaje la cual nos permite detectar problemas de sesgo o varianza, evaluar si el modelo se beneficia de más datos de entrenamiento, y determinar cuándo el modelo ha alcanzado su máximo rendimiento. Este análisis es fundamental para decidir si se requiere ajustar el modelo o modificar los datos.

En cuanto al puntaje de entrenamiento se observa que se mantiene en 1.0, lo que indica que el modelo ajusta perfectamente los datos o puede ser un signo de que sufre un overfitting. Para el puntaje de los datos de la validación cruzada comienza en un valor bajo (~0.65) cuando son pequeños los datos de entrenamientos, pero a medida que aumentan, el puntaje de validación cruzada mejora significativamente, alcanzando un máximo alrededor de 0.95, y no llega cambiar mucho alrededor de los 200 datos por lo que aumentar datos no mostraría una mejora en el rendimiento del modelo.

En cuanto al sesgo no parece haber un problema, ya que la precisión del modelo en el conjunto de validación alcanza un valor alto (~0.95), lo que indica que el modelo es capaz de generalizar bien cuando se le dan suficientes datos de entrenamiento.

Pero en cuanto a la varianza el hecho de que la precisión del conjunto de entrenamiento sea tan alta (1.0) mientras que la precisión de los resultados de la validación cruzada es más baja, lo que sugiere que el modelo puede tener alta varianza, lo que significa que está sobre ajustando (overfitting) los datos de entrenamiento.



Curva de aprendizaje con datos de prueba

Para reducir el overfitting se ajustaron los siguientes hiperparámetros.

```python
param_grid = {
    'n_estimators': [50, 100],              # Aumentar el número de estimadores
    'criterion': ['gini'],
    'max_depth': [3, 5],                    # Limitar más la profundidad del árbol
    'min_samples_split': [10, 15],          # Aumentar el número mínimo de muestras por división
    'min_samples_leaf': [4, 6],             # Aumentar el número mínimo de muestras en las hojas
    'max_features': ['sqrt', 'log2', 0.5],  # Limitar más las características por split
    'max_samples': [0.7, 0.8]               # Utilizar una fracción de muestras para cada árbol
}
```

Modificación de hiperparámetros

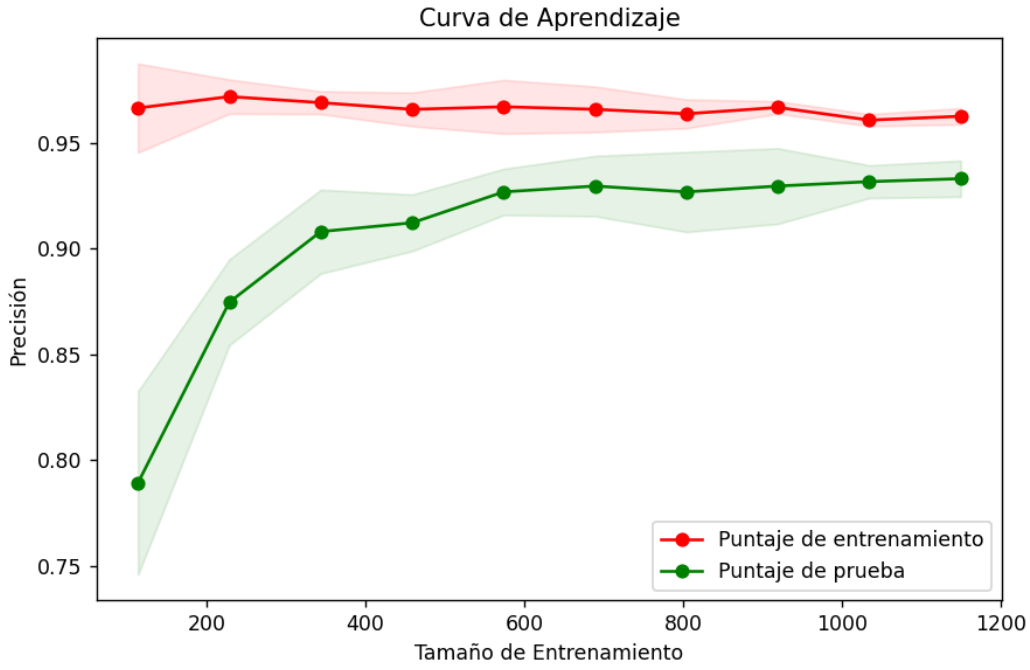
Para el “n_estimator” al aumentar el número de árboles ayuda a mejorar la capacidad del modelo de generalizar al reducir el impacto de árboles sobre ajustados.

En cuanto al hiper parámetro de “max_depth” al reducir la profundidad máxima de los árboles a valores menores como 3 o 5 fuerza al modelo a aprender reglas más simples, lo que son menos propensas a sobre ajustarse.

Para los hiperparámetros de “min_samples_split” y “min_samples_leaf” aumentarlos obliga a los árboles a no dividirse con pequeñas cantidades de datos, haciendo que los modelos sean menos específicos.

Con “max_features” se limita aún más el número de características consideradas en cada división reduce la capacidad del modelo para ajustarse demasiado a las características particulares del conjunto de datos de entrenamiento.

Finalmente, con “max_samples” entrena cada árbol con solo una fracción de las muestras lo que puede ayudar a crear diversidad entre los árboles, mejorando la generalización.



Curva de aprendizaje modificada con datos de prueba

Reporte de Clasificación para el Conjunto de Validación:

	precision	recall	f1-score	support
0	1.00	1.00	1.00	30
1	0.93	0.95	0.94	42
2	0.96	0.96	0.96	45
3	0.95	0.89	0.92	46
4	1.00	0.91	0.95	33
5	0.96	0.87	0.92	31
6	1.00	1.00	1.00	35
7	0.92	1.00	0.96	36
8	0.83	0.86	0.85	29
9	0.84	0.94	0.89	33
accuracy			0.94	360
macro avg	0.94	0.94	0.94	360
weighted avg	0.94	0.94	0.94	360

Reporte de clasificación para el conjunto de validación después de modificar los hiperparámetros

Reporte de Clasificación para el Conjunto de Prueba:

	precision	recall	f1-score	support
0	0.97	0.97	0.97	33
1	0.90	0.96	0.93	28
2	1.00	1.00	1.00	33
3	1.00	0.91	0.95	34
4	0.96	0.96	0.96	46
5	0.93	0.89	0.91	47
6	0.97	0.97	0.97	35
7	0.89	0.97	0.93	34
8	0.93	0.87	0.90	30
9	0.86	0.90	0.88	40
accuracy			0.94	360
macro avg	0.94	0.94	0.94	360
weighted avg	0.94	0.94	0.94	360

Reporte de clasificación para el conjunto de prueba después de modificar los hiperparámetros

En conclusión, los resultados obtenidos después de ajustar los hiperparámetros muestran que el modelo ha logrado reducir el sobreajuste sin comprometer significativamente el rendimiento. Los cambios aplicados a los hiperparámetros, como la reducción de la profundidad de los árboles, el aumento de las muestras mínimas por hoja y por división, y la limitación en las características evaluadas en cada división, han permitido que el modelo generalice mejor los datos. Esto se refleja en una menor diferencia entre los puntajes de precisión en los conjuntos de entrenamiento y prueba, lo que indica que el modelo no está capturando ruido o patrones específicos del conjunto de entrenamiento. De este modo, el modelo tiene una mejor capacidad de predicción en datos nuevos.