Trabajo Práctico 2

Análisis de datos y clasificación de imágenes

Noviembre 2024

Materia: LABORATORIO DE DATOS
Grupo: 12

Cuatrimestre: 2

Integrante: Sabor, Solana *(449/23)* - solanasaborschvartz@gmail.com
Integrante: Basaldua, Santos *(1118/23)* - santosbasaldua@gmail.com
Integrante: Segovia, Emiliano *(805/23)* - segoviaemiliano769@gmail.com

1 Análisis Exploratorio

1.1 Ejercicio I

Dentro del dataset propuesto por la materia, notamos una cantidad de 10.000 datos, siendo estos números del 0 al 9, ambos incluidos, lo que corresponde a 10 clases diferentes. Cada valor tiene asignada una imagen diferente de 28x28 píxeles. Estas imágenes muestran el número correspondiente en una escala de color gris. Los atributos de cada píxel son sus valores, que varían de 0 a 255, siendo 0 = negro y 255 = blanco, lo que define la intensidad del color.
Teniendo esto en cuenta, los píxeles más relevantes son los que tienen valores mayores a 90, ya que se empiezan a distinguir mejor para formar la imagen del número. Por otro lado, la gran mayoría de los píxeles de los bordes toman valor 0, siendo completamente negros, y no aportan información relevante para la clasificación de las diferentes clases. Dicho esto, podríamos descartar estos píxeles.
Para estar seguros de que este razonamiento, nacido a partir de la visualización de diferentes imágenes de números, es válido, decidimos realizar el promedio de todos los valores por cada píxel, teniendo en cuenta todos los dígitos del 0 al 9. Luego, hicimos un gráfico de heatmap para visualizar el resultado obtenido.

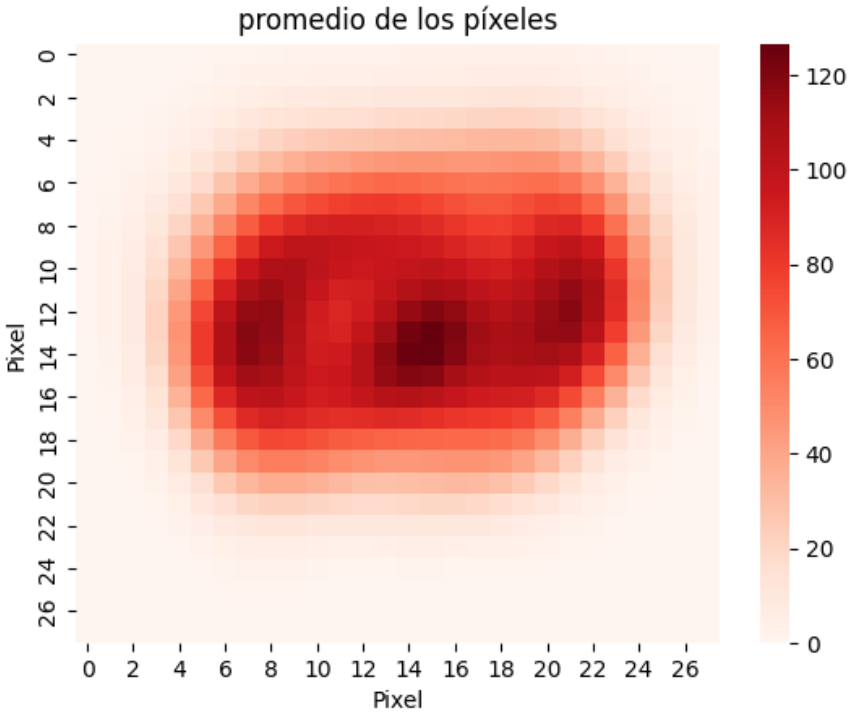

Figure 1: Promedio De Los Pixeles

Tomando como ejemplo este heatmap, vemos cómo varía el color de cada píxel según su valor. Estamos observando una imagen que representa el valor promedio de cada píxel, teniendo en cuenta las 10.000 imágenes diferentes.
Notamos que las imágenes de todos los números están concentradas en el centro y que, a partir de esta imagen, podemos definir los bordes como atributos descartables, ya que no aportan información relevante para la clasificación de las clases.

1.2 Ejercicio II

Analizando las diferentes imágenes de cada dígito, vemos que los dígitos que son muy parecidos entre sí son los siguientes pares: 5 y 6, 3 y 8. Luego, todos los demás mantienen diferencias claras. Los más fáciles de diferenciar son aquellos que se comparen con el 7 y el 1.
Para lograr comparaciones interesantes, elegimos analizar los pares 5

y 6, ya que son números similares en las imágenes disponibles, y, por otro lado, el 1 y el 0, que son muy distinguibles uno del otro.
Procedimos obteniendo, para cada dígito, el promedio de todos los valores de cada píxel de las imágenes que tienen asociadas. Una vez obtenidos estos datos, las imágenes de cada dígito serán una representación general de todas sus imágenes, lo que nos permitirá comparar de manera completa la similitud/diferencia de cada par.

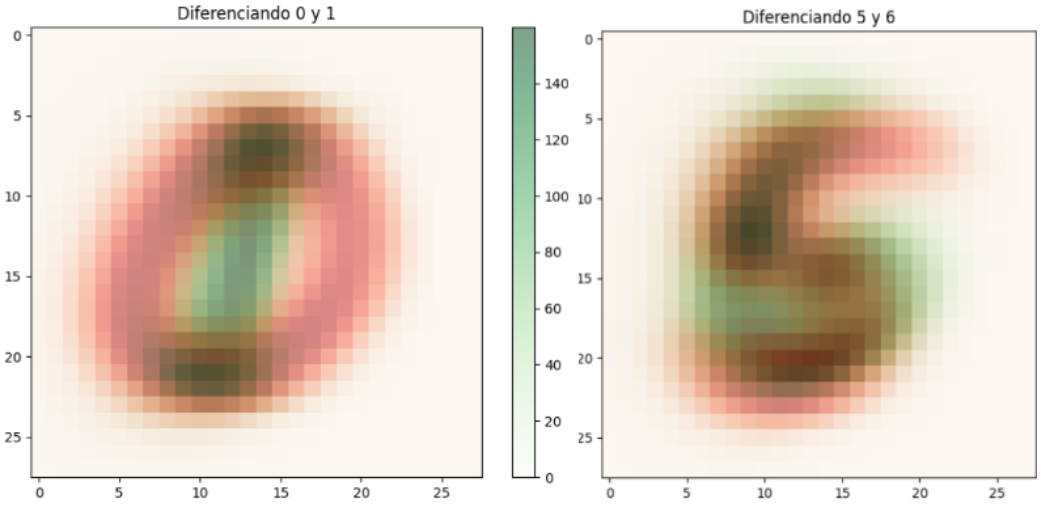

Figure 2: Diferencia Entre Digitos

Para hacer la comparación, decidimos trasponer las imágenes del promedio de cada dígito para así tener una comparación visual más directa y fácil de interpretar.
En el gráfico de la izquierda se observa la imagen del promedio del 0 y encima la imagen del promedio del 1. Vemos claramente que son muy diferentes entre sí, también que son dos grupos de píxeles muy apartados donde coinciden, y que en el resto de los píxeles no se encuentran. De esta manera, encontramos la forma de visualizar la gran diferencia entre este par.
Por otro lado, en el gráfico de la derecha tenemos la imagen del promedio del 6 junto con la del 5. A diferencia del par 0 y 1, notamos que la cantidad de píxeles en los que coinciden es mucho mayor. Visualmente, el promedio de las imágenes de 5 se asemeja mucho al de las de 6, llegando a la conclusión de que, en efecto, los dígitos más similares son el par 5 y 6.

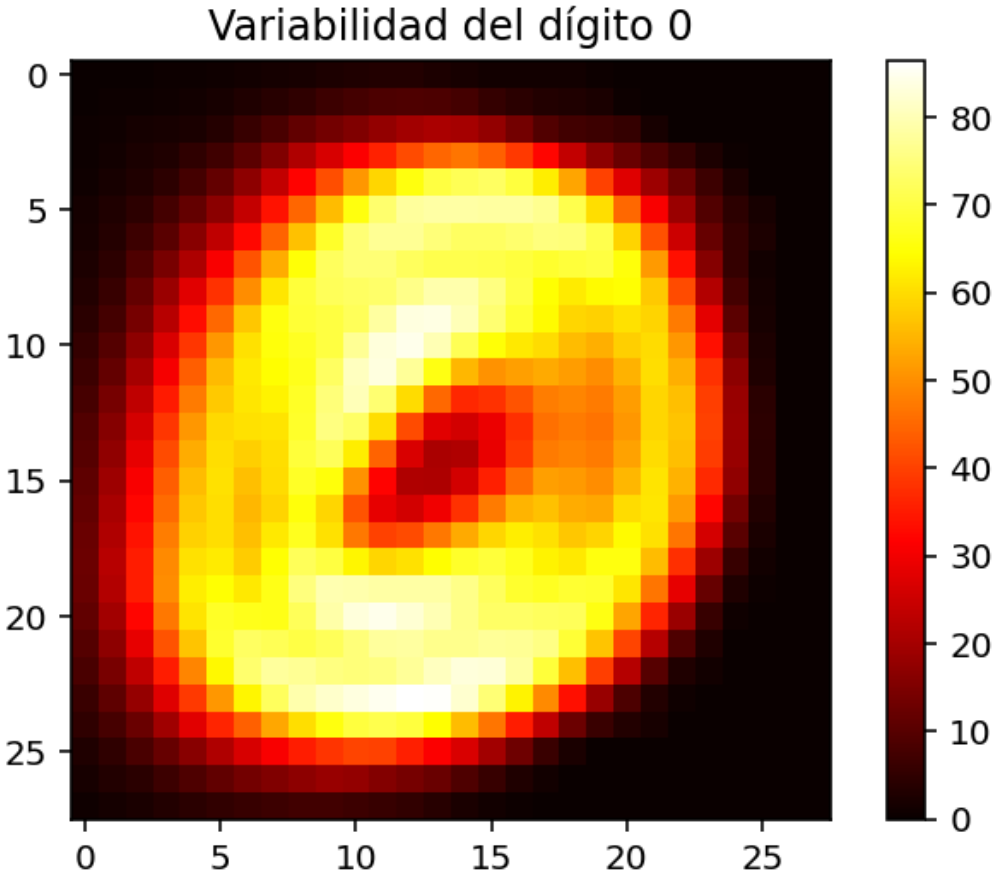

Figure 3: Variabilidad

1.3 Ejercicio III

Tomando la clase del 0 para analizar las similitudes entre sus distintas imágenes, notamos que las imágenes son muy similares entre sí, salvo por algunos casos en los que la inclinación es un poco diferente a la de la mayoría.

Con el objetivo de analizar la dispersión de los valores de los píxeles correspondientes a las imágenes de la clase 0, decidimos calcular la desviación estándar de estos valores, con el fin de realizar un análisis más detallado de su variabilidad.

Este es un gráfico heatmap que representa, para cada píxel de la clase 0, la desviación estándar. La desviación estándar de los valores de los píxeles en las imágenes de la clase 0 nos indica cuán consistentes o variables son los valores de los píxeles en esas imágenes. Lógicamente, sigu-

iendo el planteamiento realizado al inicio, observamos que los píxeles con menor desviación estándar corresponden a los bordes. Posteriormente, notamos tres aspectos adicionales:

1. En primer lugar, se puede observar un borde en forma de cero, representado en rojo. En la escala de valores que estamos utilizando, este color indica una desviación estándar baja, lo que sugiere que estos píxeles presentan una alta consistencia en las diferentes imágenes de la clase.
2. En segundo lugar, ocurre algo similar con un grupo de píxeles ubicados en el centro de la imagen. Interpretamos que este grupo muestra una buena consistencia, ya que corresponde a los valores dentro del "0", especialmente aquellos más cercanos al centro.
3. En tercer lugar, notamos una alta desviación estándar en los píxeles situados dentro de los bordes rojos. Interpretamos que esto se debe a las diferentes inclinaciones que pueden tener las imágenes dentro de esta clase.

2 Clasificación Multiclase

2.1 Ejercicio I

Con el objetivo de separar nuestros datos de interés (información sobre los dígitos 1, 2, 3, 7, 9) del dataset. Implementamos una función para crear 2 array, uno que contenga la información de nuestras clases, es decir, que estén todas las repeticiones de los dígitos mencionados y otro que para cada dato tenga asociada la imagen que le corresponde. De esta manera ahora contamos con el conjunto de datos necesarios para avanzar. Una vez que tenemos nuestro conjunto de datos de interés,

procedemos a separar el conjunto en desarrollo (dev), siendo este el 80% de los datos disponibles y validación (hold-out), siendo este el 20% restante. Para utilizar en los siguientes dos incisos los datos de desarrollo y en el último los de validación.

2.2 Ejercicio II

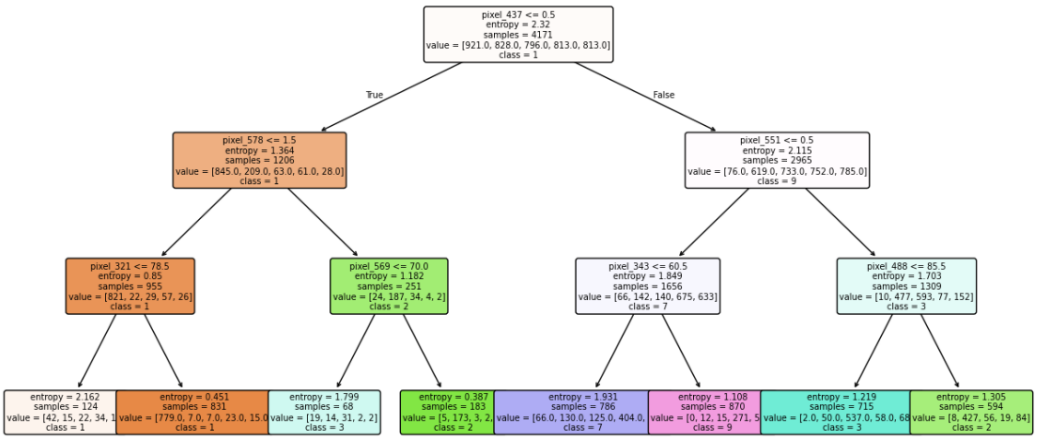

Figure 4: Entropia Longitud 3

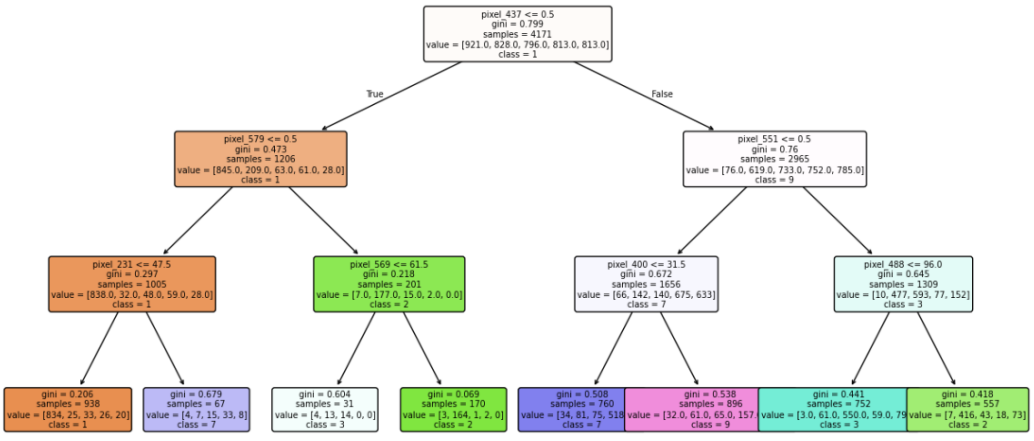

Figure 5: Gini Longitud 3

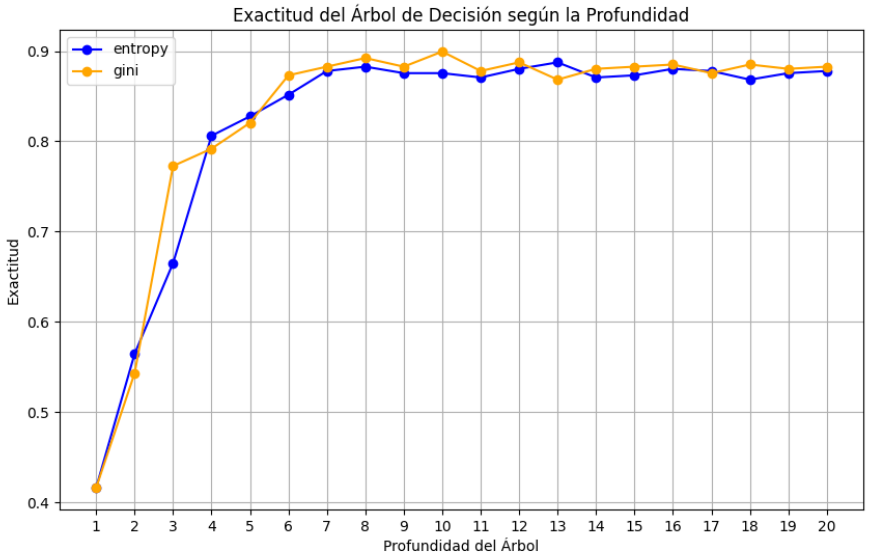

Figure 6: Exactitud De Arbol De Decision Segun La Profundidad

El propósito principal es analizar cómo varía la exactitud del modelo en función de los siguientes factores:
Profundidad del árbol: Variando desde 1 hasta 20.
Criterios de división: Comparando los criterios de Entropía y Gini.
Algunas de las principales preguntas que buscamos responder con este análisis es: ¿Qué profundidad del árbol es la más adecuada para este conjunto de datos? y ¿Con qué criterio conviene clasificar? Sabemos que, a medida que se aumenta la profundidad del árbol, el modelo tiene la capacidad de aprender patrones más complejos. Sin embargo, llega un punto en el que el modelo comienza a sobre ajustarse a los datos de entrenamiento, lo que puede resultar en una disminución de su capacidad de generalización para nuevos datos.
A partir del análisis realizado y de los gráficos generados para comparar visualmente el rendimiento de los modelos, podemos concluir lo siguiente:

1. La profundidad máxima óptima para la clasificación es 4. A partir de esta profundidad, la mejora en la exactitud se vuelve más lenta. A medida que la profundidad aumenta, el modelo comienza a ajustarse excesivamente a los datos de entrenamiento, lo que impide una mejor generalización a nuevos datos.

2. Además, observamos que utilizando el criterio de Entropía, el modelo alcanza la mejor exactitud en la profundidad óptima. Esto sugiere que, para tareas futuras de clasificación utilizando datos hold-out, sería recomendable emplear un árbol de decisión con una profundidad máxima de 4 y el criterio de Entropía.

   La profundidad máxima óptima para la clasificación es 4. A partir de esta profundidad, la mejora en la exactitud se vuelve más lenta. A medida que la profundidad aumenta, el modelo comienza a ajustarse excesivamente a los datos de entrenamiento, lo que impide una mejor generalización a nuevos datos.
   Además, observamos que utilizando el criterio de Entropía, el modelo alcanza la mejor exactitud en la profundidad óptima. Esto sugiere que, para tareas futuras de clasificación utilizando datos hold-out, sería recomendable emplear un árbol de decisión con una profundidad máxima de 4 y el criterio de Entropía.

2.3 Ejercicio III

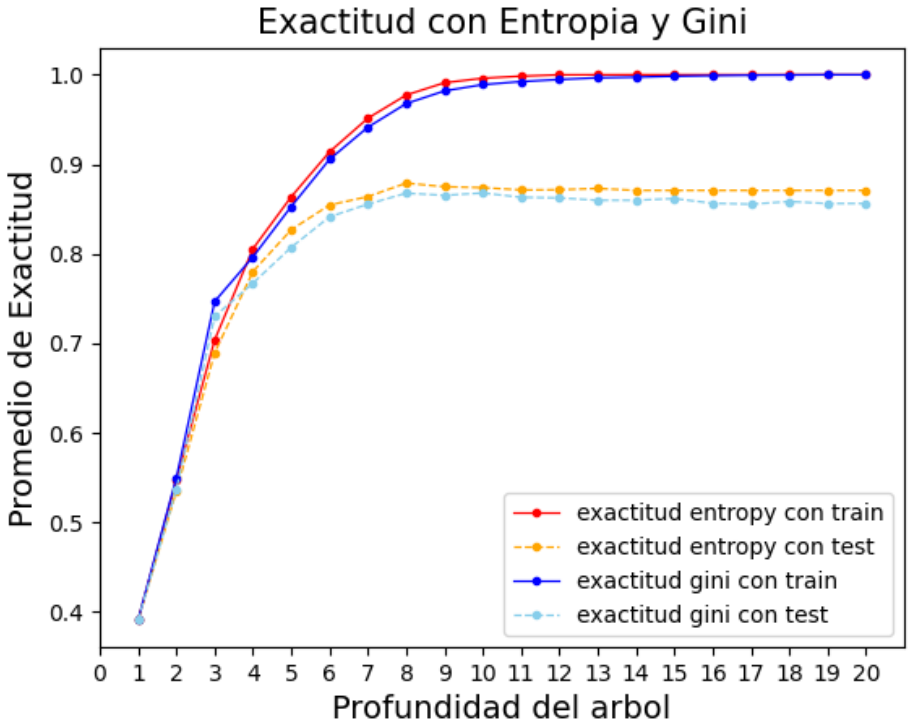

Figure 7: Exactitud Con Entropia Y Gini

Para utilizar el método de validación cruzada con K-folding, utilizamos los datos de desarrollo previamente separados y decidimos emplear 5 pliegues (folds) para el análisis.
Creamos una función que evalúa distintos modelos de árbol de decisión, considerando diferentes parámetros como las profundidades y los criterios de división (Gini y Entropía).
Calculamos la exactitud lograda por cada modelo y recopilamos esta información en un diccionario, que luego convertimos en un DataFrame.
A continuación, decidimos realizar consultas SQL para obtener el promedio de exactitud para cada profundidad, separando los resultados según los criterios de Gini y Entropía. Esto nos permitió realizar un análisis visual sobre cómo varía la exactitud según la profundidad para ambos criterios, tanto en los datos de entrenamiento como en los de validación.
Este análisis visual nos ayudó a identificar el punto donde el rendimiento del modelo comienza a estabilizarse o donde podría haber sobreajuste.

Vemos así que a partir de la profundidad 4 la exactitud que genera el árbol con los datos de entrenamiento empieza a separarse visualmente con la exactitud que toman los datos de validación, esto nos indica que hay un sobre ajuste a partir de ese punto como habíamos previsto en el ítem anterior.
Después de realizar el análisis y evaluar distintas configuraciones de hiper parámetros (profundidad del árbol y criterio de división), determinamos que la mejor configuración para la clasificación mediante un árbol de decisión es la siguiente:
profundidad maxima(max depth)=4.
Criterio de division = Entropy.
Esta configuración se eligió debido a que, aunque la exactitud continúa mejorando con profundidades mayores, a partir de la profundidad 4, el rendimiento en los datos de validación se estabiliza y la mejora en la exactitud es más lenta. A su vez, el modelo con profundidad 4 muestra un buen balance entre complejidad y capacidad de generalización, evitando el sobreajuste que podría ocurrir con mayores profundidades.
En términos de rendimiento, con esta configuración se alcanzó una exactitud de 0.81 en el conjunto de validación, lo que indica un buen nivel de precisión para este conjunto de datos.
Además, el criterio de Entropy resultó en un mejor desempeño en comparación con el criterio Gini, lo que sugiere que la métrica de la entropía, que busca un equilibrio más fino en la partición de los datos, es más adecuada para este conjunto específico.

2.4 Ejercicio IV

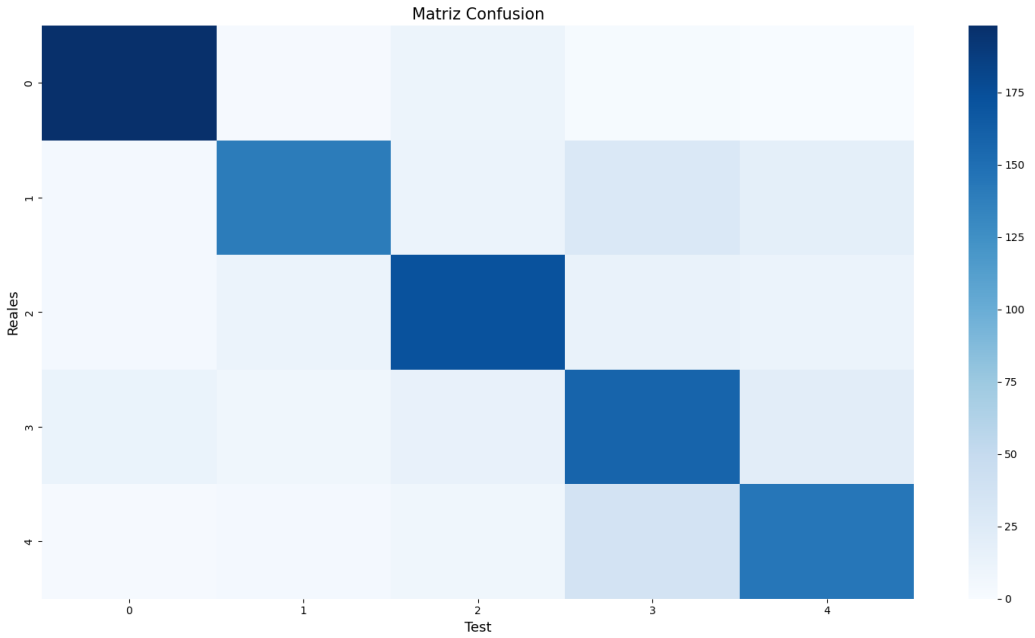

Figure 8: Enter Caption

Una vez evaluados los distintos modelos y seleccionado el óptimo, decidimos realizar el entrenamiento utilizando los datos de desarrollo y el árbol de decisión con los hiperparámetros fijados en los ítems anteriores. Posteriormente, predecimos las clases con los datos de hold-out y analizamos su rendimiento. Para evaluar el desempeño del modelo, utilizamos la matriz de confusión. Observamos que en los sectores que no son de color azul, es decir, los de color mas claro los valores no son cero, lo cual sería lo ideal. En cambio, los valores indican que el modelo está clasificando un número como otro. El modelo clasifica de manera más precisa la clase 1, con un total de 214 datos, de los cuales 198 fueron clasificados correctamente. Sin embargo, los demás tienen un margen de error considerable. Para complementar este análisis, calculamos la exactitud de nuestro modelo, que resultó ser: Exactitud = 0.7785 Este valor indica que el modelo logra predecir correctamente una cantidad significativa de datos por clase, aunque todavía tiene un margen de error considerable. Creemos que este margen de error se debe a las grandes diferencias que existen entre las imágenes de la misma clase, especialmente considerando las distintas inclinaciones y definiciones de las imágenes.