Trabajo Práctico 2

October 2024

Materia: LABORATORIO DE DATOS

Grupo: 12
Integrante: Sabor, Solana *(449/23)* - Solanasaborschvartz@gmail.com
Integrante: Basaldua, Santos *(947/23)* - 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 siendo estas, 9 clases diferentes. Cada valor tiene asignada una imagen diferente de 28x28 pixeles, estas imagenes muestran el numero correspondiente en una escala de color Gris, tiene como atributos cada pixel y sus valores varian de 0 a 255, siendo 0 = Negro y 255 = Blanco definiendo 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 más para formar la imagen del número. Por otro lado la gran mayoria de los pixeles de los bordes toman valor 0 siendo completamente negro y no aportan informacion relevante para la clasificacion de las diferentes clases. Dicho esto podríamos descartar estos pixeles.
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 pixel teniendo en cuenta todos los dígitos del 0 al 9, luego hicimos un gráfico de HeatMap para visualizar el resultado obtenido:

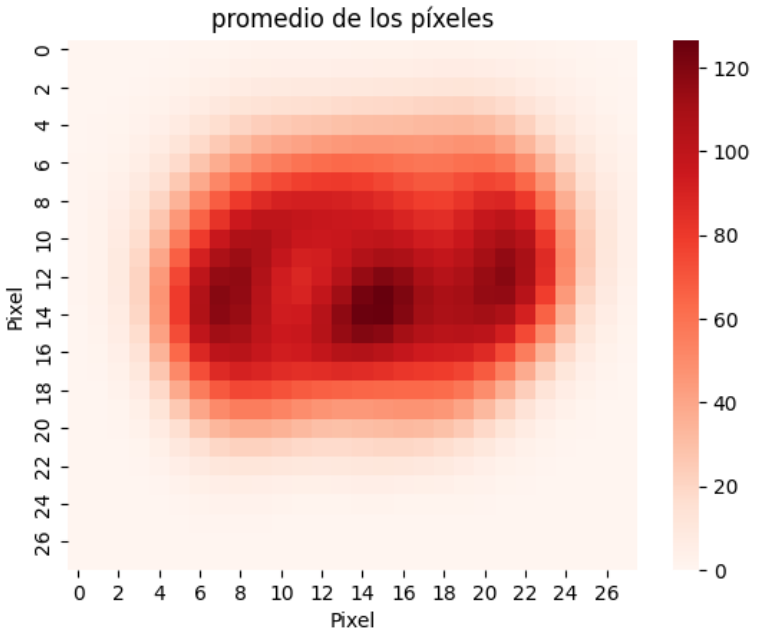

Figure 1: Promedio De Los Pixeles

Tomando como ejemplo este HeatMap vemos como varía el color de cada píxel según su valor. Estamos viendo una imagen que representa el promedio de valor para cada pixel teniendo en cuenta las 10 mil 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 clases.

1.2 Ejercicio II

Analizando las diferentes imágenes por cada dígito vemos que los dígitos que llegan a ser muy parecidos entre sí son los pares: 5 y 6, 3 y 8. Luego todos mantienen diferencias claras, los más fáciles de diferenciar son todos las que se comparen con el 7 y el 1.
Para lograr comparaciones interesantes elegimos analizar los pares 5 y 6 siendo números similares en las imágenes disponibles. Y por otro lado el 1 y 0 siendo muy distinguibles uno del otro.
Procedimos obteniendo para cada dígito el promedio de todos los valores de cada píxel de cada imagen que tienen asociada. Una vez teniendo estos

datos, las imágenes de cada dígito que tengamos van a ser una representación general de todas sus imágenes y así poder comparar de manera completa la similitud/diferencia de cada par:

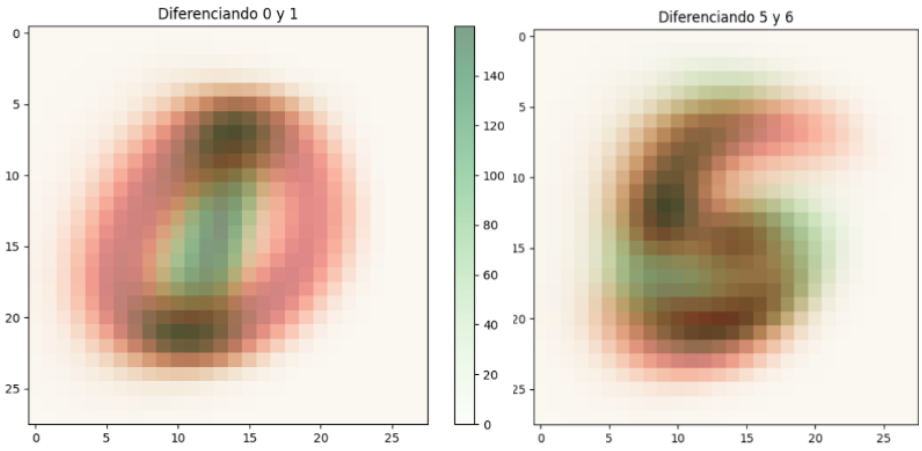

Figure 2: Diferencia Entre Digitos

Para hacer la comparación decidimos trasponer las imágenes del promedio de cada digito 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 de 1, vemos claramente que son muy diferentes uno de otro, también que son 2 grupos de pixeles muy apartados donde coinciden, y que en el resto de píxeles ni se encuentran. De esta manera encontramos la manera de visualizar la gran diferencia entre este par.
Por otro lado en el gráfico de la derecha tenemos la imagen del promedio de 6 junto con la del 5, a diferencia del par 0 y 1, notamos que la cantidad de pixeles en los que coinciden son muchos más. Visualmente el promedio de las imágenes de 5 se asemejan mucho al de las de 6 llegando a la conclusión de que, en efecto, los dígitos más similares son el par 5 y 6.

1.3 Ejercicio III

Tomando a 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 donde 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.

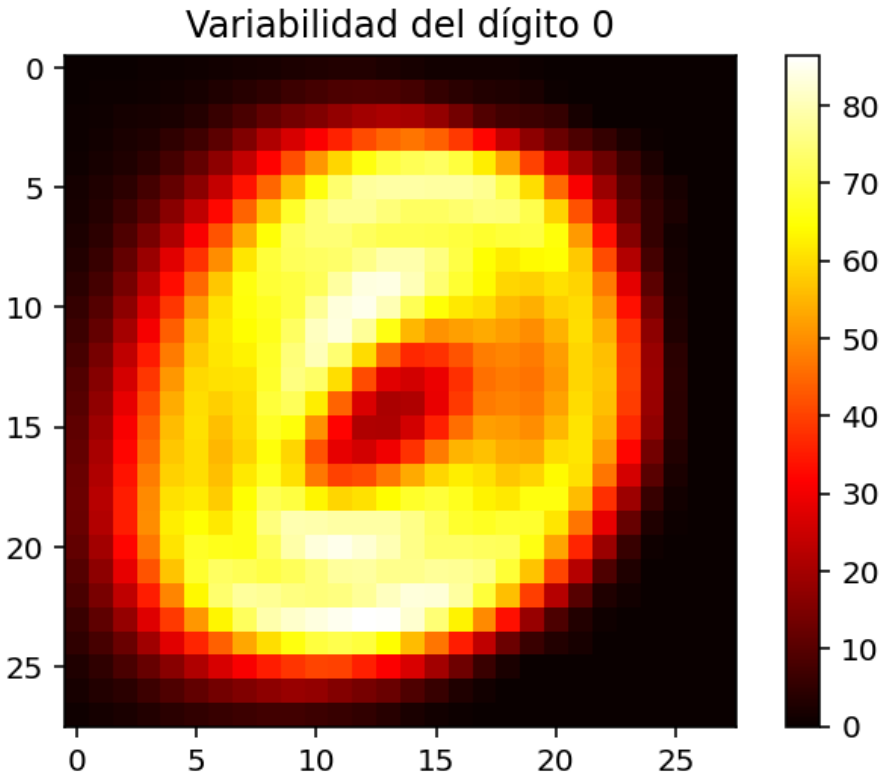

Figure 3: Variabilidad

Este es un gráfico HeatMap que representa para cada pixel de la clase 0, la desviación estándar. La desviación estándar de los valores de los píxeles en las imágenes de la clase 0 nos dicen cuán consistentes o variables son los valores de los píxeles en esas imágenes.
Lógicamente, siguiendo 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

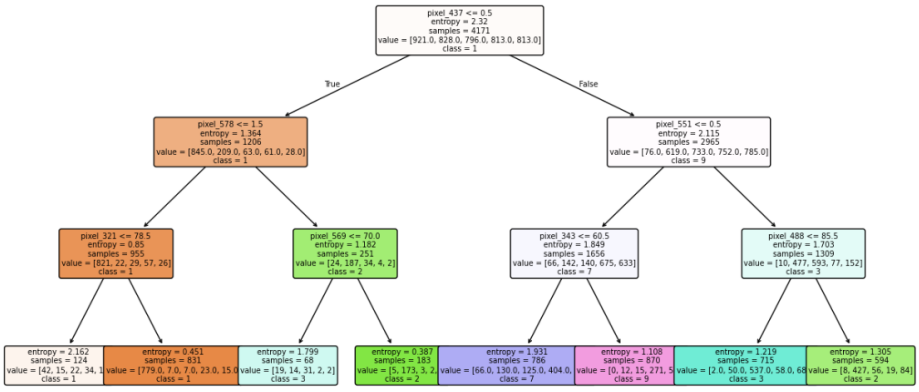

Figure 4: Entropia Longitud 3

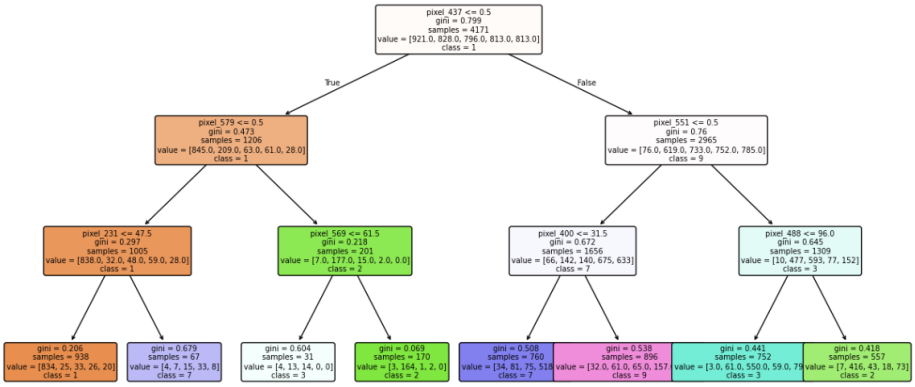

Figure 5: Gini Longitud 3

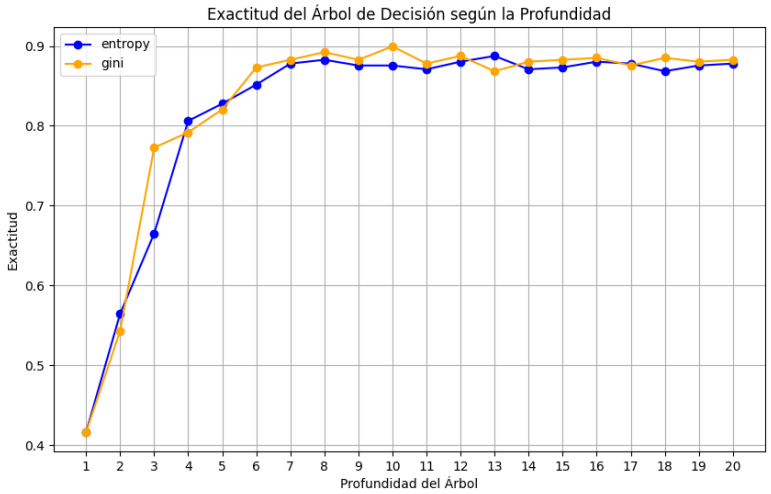

Figure 6: Exactitud De Arbol De Decision Segun La Profundidad

2.3 Ejercicio III

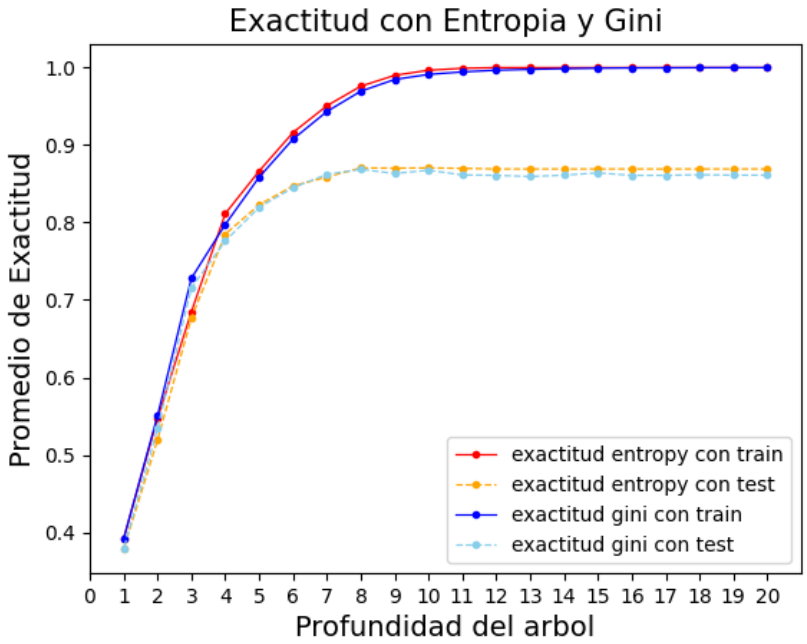

Figure 7: Exactitud Con Entropia y Gini

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.