LABORATORIO DE DATOS

Trabajo práctico 2

CPQ

Luna Praino
Camila Molteni Ceccón
Agustin Quintriqueo

INTRODUCCIÓN

En el presente informe trabajamos con un conjunto de datos de imágenes llamado “Emnist”. El mismo es un conjunto de dígitos de caracteres escritos a mano derivados de la base de datos “NIST Special Database 19”. Cada fila del dataset representa una imagen de 28x28 de una letra. Cada elemento de la fila representa un píxel de la misma, éstos toman valores que van desde el 0 hasta el 255. Donde el 0 corresponde al fondo (llamémosle “negros”), el 255 a los “blancos” y los valores intermedios a los “grises”.

Cada letra toma 2400 formas distintas pues estamos trabajando con un subconjunto del dataset original. Éstas varían en forma, grosor y demás, es decir, varía la cantidad de 0, 255 e intermedios que toma cada píxel. Por ejemplo, la ‘H’ de la fila 14 no es igual a la ‘H’ de la fila 8.

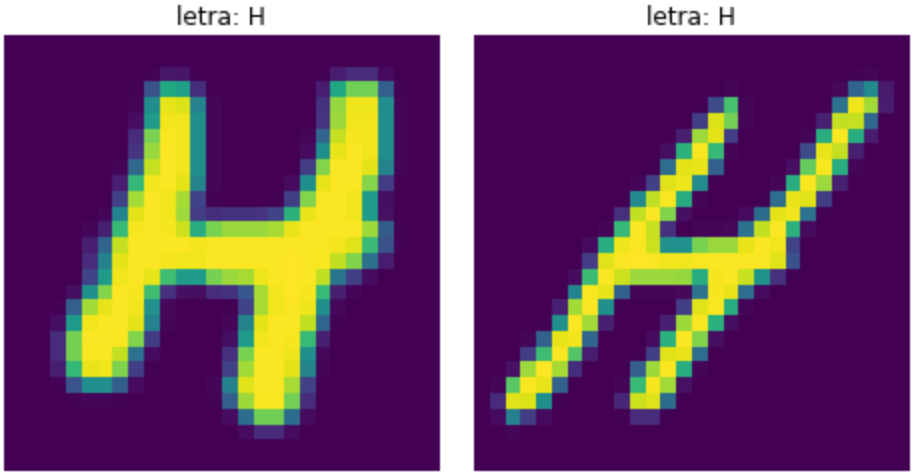

Figura 1: a la izquierda, la H representada en la fila 8, a la derecha en la fila 14 .

Comenzando con el análisis exploratorio, en primer lugar, podemos afirmar que a partir de los atributos que describen la foto (valores numéricos) no podemos predecir directamente qué letra va a representar. Pero sí podemos deducir qué forma va a tomar, es decir, si va a ser más o menos ancha, alta, o aproximar dónde se van a encontrar posicionadas. Para esto, hicimos algunos gráficos para ver qué valores van tomando los píxeles en diferentes filas de cada imagen.

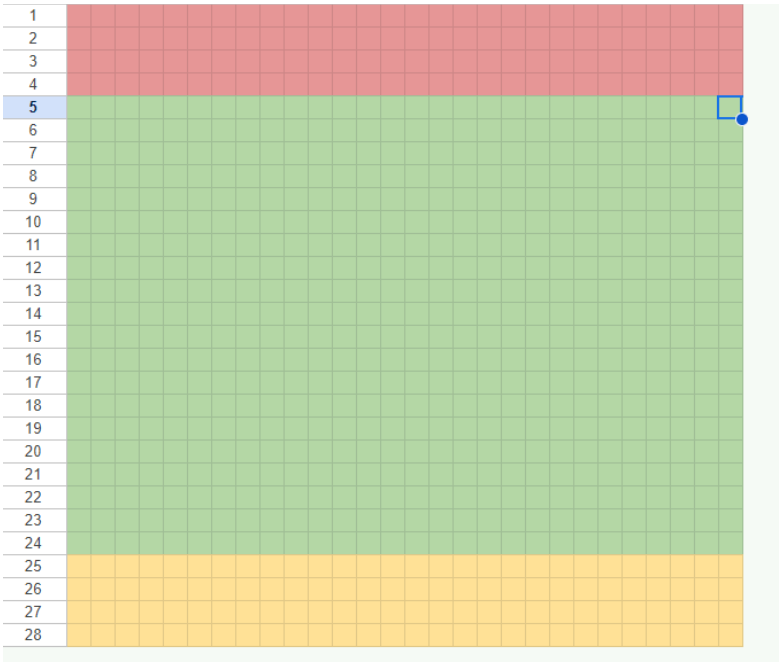

Figura 2: representación de la imagen de 28x28. En rojo, un rango de filas desde la 1 hasta la 4. En verde, desde la 5 hasta la 24. En amarillo, desde la 25 hasta la 28.

En la figura 2 podemos apreciar a qué nos referimos a continuación cuando hablamos de filas de la imagen.

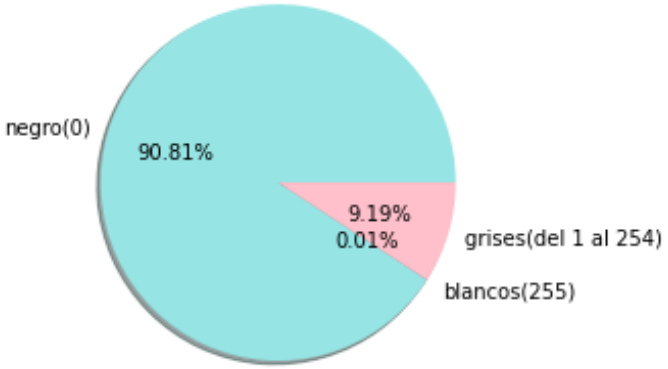

Figura 3: gráfico de torta donde las porciones representan el porcentaje de blancos, negros y grises que toman los píxeles desde la fila 1 hasta la 4 de la imagen de 28x28.

En el caso de la figura 3, se puede ver claramente que no hay mucha información relevante sobre las letras, porque el mayor porcentaje corresponde al “negro”, o sea, al fondo.

Que haya un pequeño porcentaje de “grises” y “blancos” nos muestra que el comienzo de algunas letras se encuentra posicionado dentro de este rango de filas.

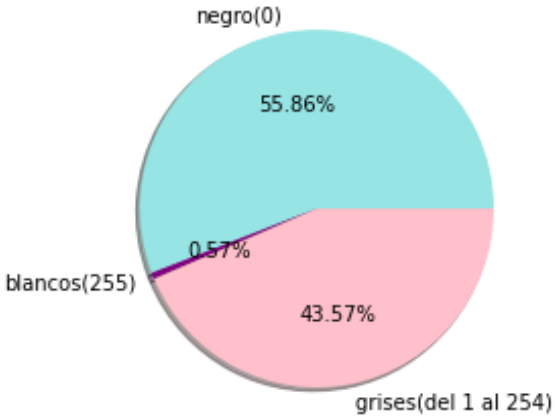

Figura 4: gráfico de torta donde las porciones representan el porcentaje de blancos, negros y grises que toman los píxeles desde la fila 4 hasta la 24 de la imagen de 28x28.

Bajo las condiciones representadas en la figura 4 podemos observar que la mayor distribución de blancos y grises se encuentra dentro de este rango de filas, son los valores más relevantes porque son los que describen a las letras en sí pues tienen información de la estructura y forma de las mismas. Estos píxeles son importantes para distinguir las diferentes letras y formas que toman.

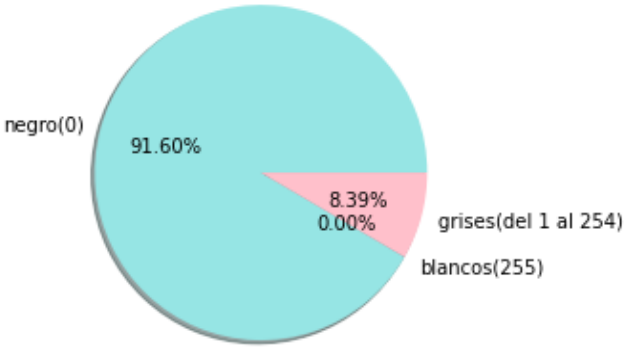

Figura 5: gráfico de torta donde las porciones representan el porcentaje de blancos, negros y grises que toman los píxeles desde la fila 24 hasta la 28 de la imagen de 28x28.

En la instancia que propone la figura 5 sucede algo parecido a lo visto en la figura 3, es decir, no hay mucha información significativa sobre las letras porque el mayor porcentaje del gráfico corresponde al fondo.

En conclusión, la descripción de la cantidad de grises por fila proporciona una visión general de cómo se distribuyen las características de las letras en la imagen, esto puede servirnos para identificar patrones en común entre letras y determinar los atributos relevantes que este caso serían los atributos desde el 112 (fila 4) hasta el 672 (fila 24) .

Siguiendo con el análisis, queremos ver si hay letras que son parecidas entre sí. Para esto tomamos las letras ‘E’, ‘L’, y ‘M’. En principio, calculamos para todas el promedio de cada píxel, e hicimos un gráfico de la forma promedio de cada una.

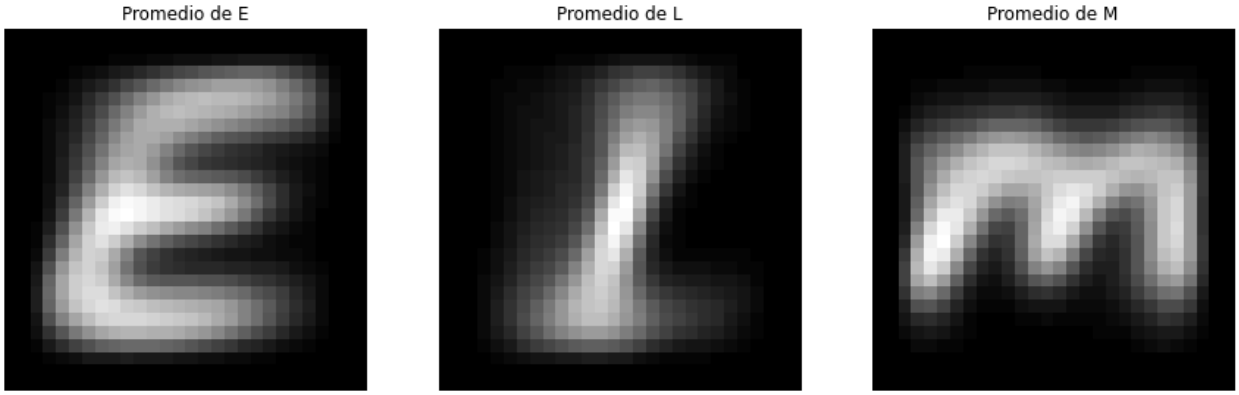

Figura 6: “E” promedio, “L” promedio y “M” promedio, respectivamente.

En la figura 6 podemos observar que, si bien las imágenes promedio de estas letras son bastante diferentes entre sí, encontramos algunas similitudes.

A simple vista vemos que la E y la L tienen una distribución de píxeles similares en la zona superior e inferior de las imágenes.

En la zona central de la L y la M, parece que comparten blancos.

La E y la M no tienen grandes similitudes en su estructura.

Para ver estas comparaciones más claramente, calculamos el módulo de las diferencias entre estos promedios, y las graficamos.

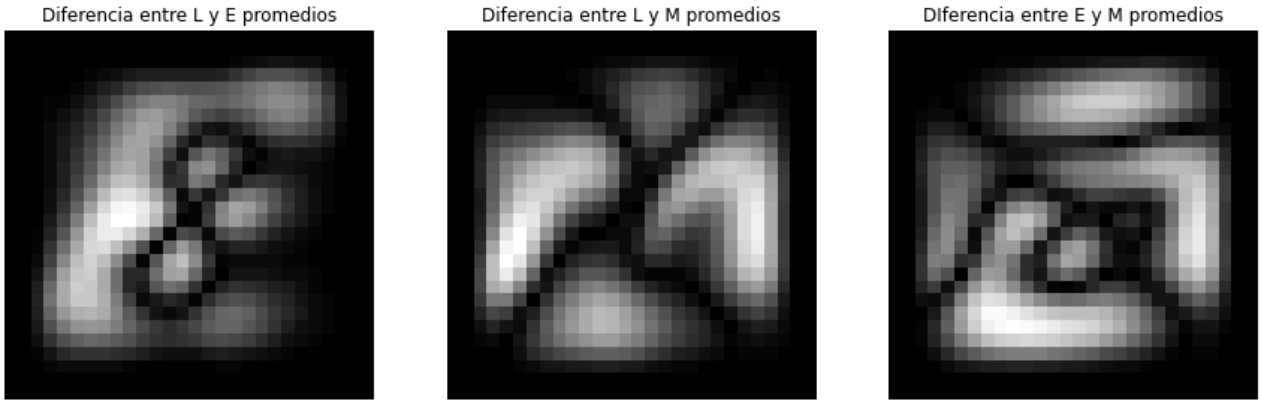

Figura 7: módulo de la diferencia entre L y E promedios, entre L y M promedios, y entre E y M promedios, respectivamente.

A partir de la figura 7 podemos derivar que las zonas que se encuentran en negro son aquellas que las letras promedio tienen en común, es decir, aquellas tales que el resultado de la métrica utilizada para calcularlas es igual a 0

Luego, tomamos la clase ‘C’ con el fin de determinar si todas las imágenes son muy similares entre sí. Para esto, visualizamos la C promedio.

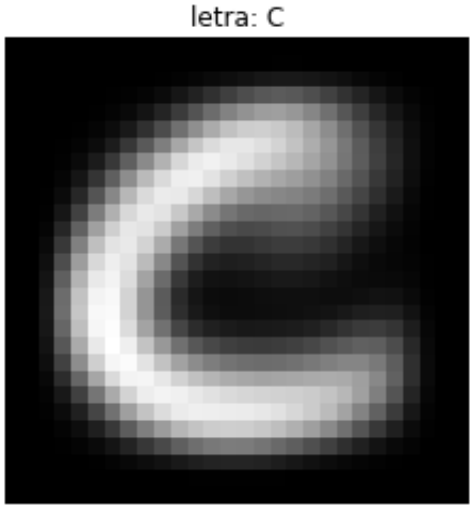

Figura 8: imagen correspondiente a la ‘C’ promedio.

En el caso de la figura 8, como puede distinguirse claramente una ‘C’ en la imagen. Este detalle nos permite deducir que, aunque existen variaciones en la parte superior de la foto, todos los tipos de letra promediadas muestran una gran similitud entre sí.

Como este dataset estaba compuesto por imágenes, planteaba una diferencia respecto a los datos vistos en clase, pues los mismos tenían información más concreta. En el caso particular del dataset de Titanic, por ejemplo, una columna determinaba si el pasajero sobrevivió o no y esto nos permitía trabajar a partir de estos datos.

En el caso de “emnist_letters”, para realizar un análisis concreto, fue necesario ver algunas imágenes, cómo se demostraban las mismas según los valores que tomaba cada píxel, etcétera, porque necesitamos la visualización concreta de las imágenes.

DESARROLLO

Pasando a la sección de experimentos realizados, en el primero de ellos se trató de ajustar diversos modelos de KNN de manera tal , que dada una imagen, se pueda determinar si la misma corresponde a una letra ‘L’ o ‘A’.

Para ajustar a cada modelo con los respectivos datos de train y luego realizar comparaciones de los resultados utilizando los conjuntos de test generados, seleccionamos diversos atributos que representan píxeles de las imágenes correspondientes a las letras mencionadas.

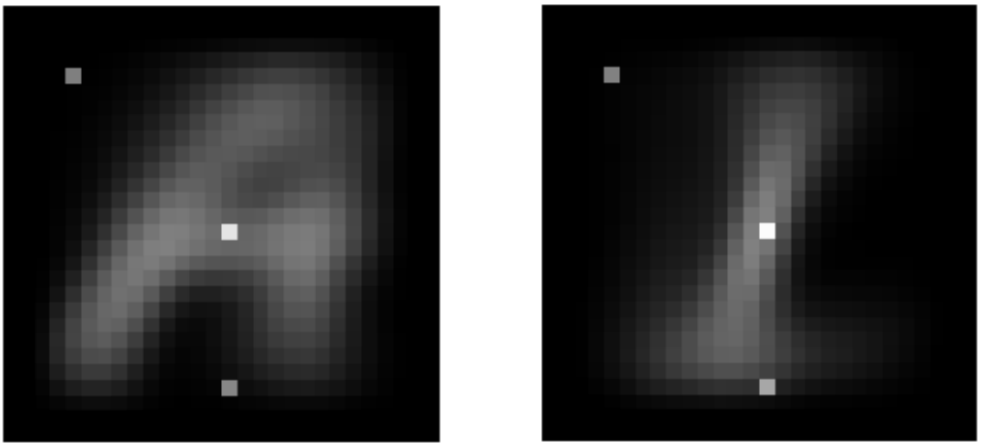

Figura 9: imágenes de las letras promedio ‘A’ y ‘L’ con los píxeles seleccionados (atributos) para entrenar los respectivos modelos de knn.

En segundo lugar entrenamos árboles de decisión utilizando los criterios de “entropy” y “gini”. En cada caso, utilizamos árboles con profundidades 1, 2, 3 y 5 para ambos criterios. Un ejemplo de árbol sería la figura 10, cuyo criterio es entropy y su profundidad es 2.

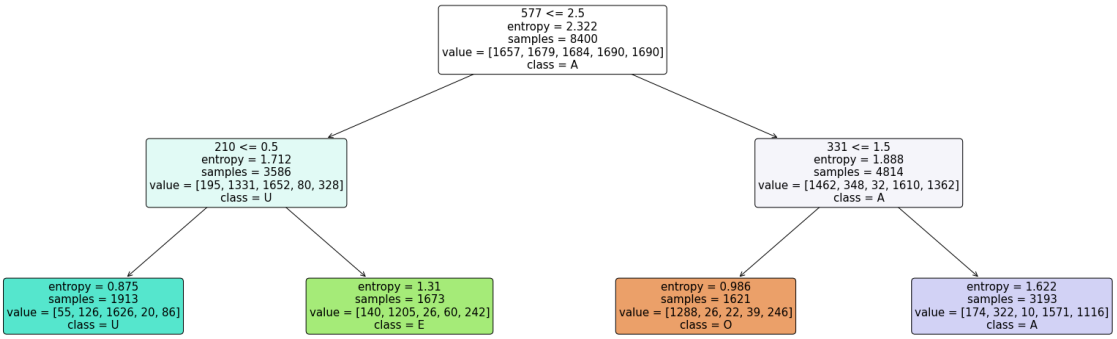

Figura 10: la imagen corresponde a un árbol de decisión cuyos hiperparametros son profundidad 1 y criterio entropy.

Luego, evaluamos los respectivos árboles utilizando KFold Cross-Validation. Además, calculamos con funciones proporcionadas por la biblioteca sklearn, el accuracy para ambos criterios.

Finalizando con la actividad, comparamos ambos criterios y pudimos notar, a partir de un gráfico(exhibido más adelante), cómo varían los scores promedio en cada instancia.

CONCLUSIÓN

Para ir concluyendo el informe, en esta oportunidad trabajamos con un dataset compuesto por imágenes, lo cual presentó una diferencia respecto a los vistos en clase. Para realizar un análisis concreto de “emnist_letters”, tuvimos que visualizar y entender cómo se comportaban las imágenes según los valores que tomaban sus píxeles.

Como enumeramos anteriormente, observamos que no es tan sencillo predecir qué letra íbamos a ver a partir de los valores de sus píxeles, pero sí podíamos imaginar qué forma iba a tomar.

Retomando la sección de experimentos realizados, más concretamente, el referido a los modelos de KNN, en base a la comparación hecha entre los distintos casos pudimos notar que se tiene una mayor exactitud a medida que aumenta la cantidad de atributos seleccionados.

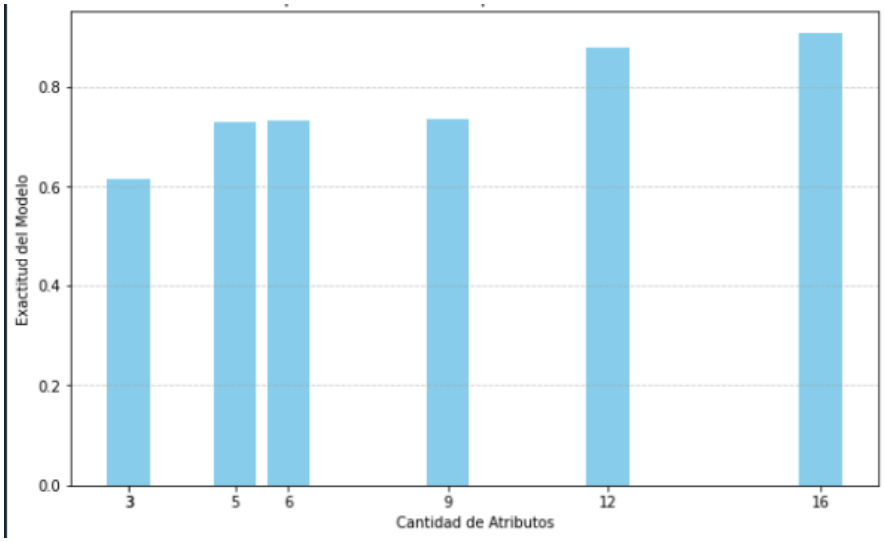

Figura 11: imagen correspondiente a la comparativa de la exactitud de modelos en base a la cantidad de atributos

De la figura 11 podemos notar como el modelo que consta de 16 atributos, tuvo una mayor exactitud notoria por sobre el resto de los modelos.

Respecto a los árboles entrenados bajo los criterios de “entropy” y “gini”, realizamos un gráfico que compara los scores promedio por profundidad en cada caso.

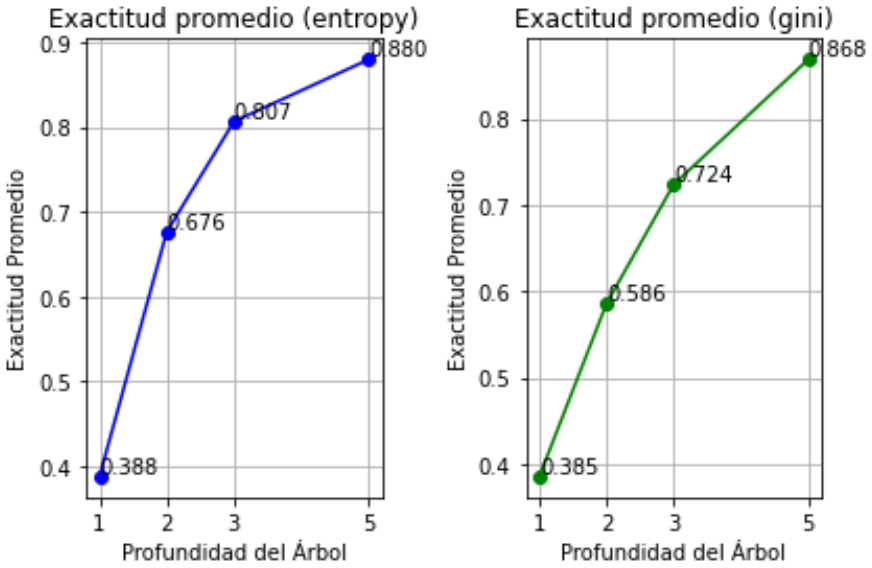

Figura 12: gráfico comparativo entre las exactitudes promedio de los árboles de “entropy” y “gini”.

Luego de lo visto en la figura 12, concluimos que, si bien en ambos casos la precisión aumenta con la profundidad, el mejor modelo obtenido se dio con el árbol de “entropy” con profundidad 5, aunque las exactitudes no difieren en grandes medidas.

Después de determinar cuál fue el mejor modelo, calculamos para el mismo, el score para el conjunto de desarrollo y evaluación, los resultados obtenidos fueron 0.889 y 0.88, respectivamente. Los resultados obtenidos son muy similares, esto sugiere que el modelo no está sobreajustado.

Para finalizar, luego de llevar a cabo una serie de procesos de análisis sobre el dataset estudiado, como la realización de gráficos, árboles bajo dos criterios distintos, experimentos en base al modelo de KNN, podemos concluir que el análisis de datos puede variar dependiendo del tipo de datos y su estructura.