Imagen que contiene Logotipo

Descripción generada automáticamente

## Universidad de Buenos Aires

Facultad de Ciencias Exactas y Naturales

Departamento de Computación

**Trabajo Práctico 02:**

**Clasificación y validación cruzada.**

**El caso EMNIST**

**Grupo ABC**

**Alumnos:**

Abbate, Lucas Ignacio (134/21)

Becker, Guillermo (616/90)

Cevasco Rieck, Jorge Augusto (230/23)

Profesor: Turjanski, Pablo

Materia: Laboratorio de Datos

Cuatrimestre: 1C2024

## Resumen

## Introducción

En el presente trabajo, se abordará la tarea de clasificación de imágenes utilizando el conjunto de datos EMNIST, específicamente el subconjunto que contiene las letras mayúsculas manuscritas. El objetivo principal de este TP es poner en práctica los conceptos de clasificación y selección de modelos mediante la utilización de técnicas de validación cruzada.

En 1995, el NIST (National Institute of Standards and Technology) estadounidense, desarrolla la “NIST Special Database 19” (SD19), una de las primeras bases publicadas de texto manuscrito. Para su desarrollo, 3.699 empleados del Instituto, estudiantes de colegio secundario y médicos, completaron a mano un formulario (Fig. 1) que contenía números y letras que tenían que replicar debajo, además de el inicio de la Constitución Nacional Americana.

Estos formularios fueron escaneados y sus caracteres fueron separados, etiquetados y procesados (corrección de intensidad, centrado, etc.) llegando a 810.000 imágenes de 28x28 píxeles. Años más tarde, la SD19 se convirtió en la “EMNIST database” (Extended Modified NIST database), una actualización de la ya popularizada MNIST, que, además de contar con los dígitos del 0 al 9, contaba con las letras del abecedario en mayúscula y en minúscula.

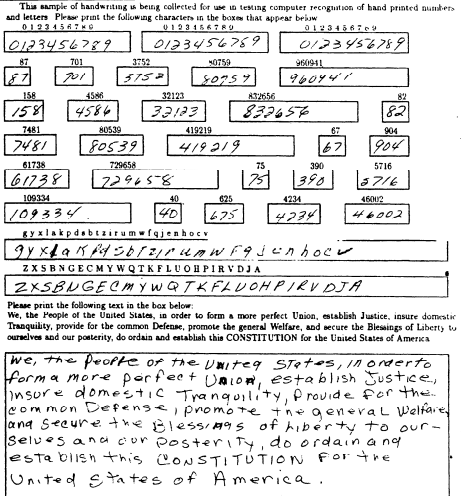


Fig. 1: Formulario para el Desarrollo de la base original del NIST. Extraído de https://www.nist.gov/system/files/documents/srd/nistsd19.pdf

Tanto MNIST como EMNIST se convirtieron en bases de datos estándar para el entrenamiento y desarrollo de modelos de aprendizaje automático: consisten en grandes cantidades de datos etiquetados, normalizados y aislados, y su procesamiento permite el desarrollo de herramientas de OCR (“optical character recognition”, reconocimiento óptico de caracteres), una aplicación concreta y de probada utilidad de estos modelos.

Desde hace décadas que distintas industrias utilizan modelos de este tipo para automatizar la categorización de textos. Uno de los ejemplos más comunes es el Servicio Postal de los Estados Unidos (USPS), que desde 2005 utiliza distintos algoritmos automáticos para “leer” las direcciones y códigos postales escritos a mano en los sobres de los correos. Se genera luego un puntaje de confiabilidad, que si supera cierto umbral es reenviado a la oficina correspondiente de dicho código postal. Caso contrario, es enviado a una oficina en la que personas leen y clasifican a mano el código postal y la dirección, generando además nuevos datos etiquetados para continuar el entrenamiento de los modelos.

En este trabajo se utilizaron distintos algoritmos para la generación de un clasificador de los caracteres. En primera instancia, se entrenó un clasificador binario, utilizando un modelo de kNN (“k -Nearest Neighbors”, o k-Vecinos más cercanos) para diferenciar las letras “L” de las letras “A”. Luego, se ajustó un clasificador multiclase, cuyo objetivo es distinguir todas las vocales. Para este último modelo se preparó un árbol de decisión, y se escogieron los hiperparámetros que mejor ajustan a nuestro set de datos utilizando validación cruzada (k-folding).

### Análisis exploratorio

La base de EMNIST utilizada en este trabajo está compuesta por 2.400 imágenes de cada letra del alfabeto inglés (62.400 en total). Cada una de esas imágenes consta de 784 valores, que representan la intensidad de cada píxel en una grilla de 28x28, en una escala que, para todas las observaciones, va de 0 a 255 (como se puede apreciar en la Fig. 2).

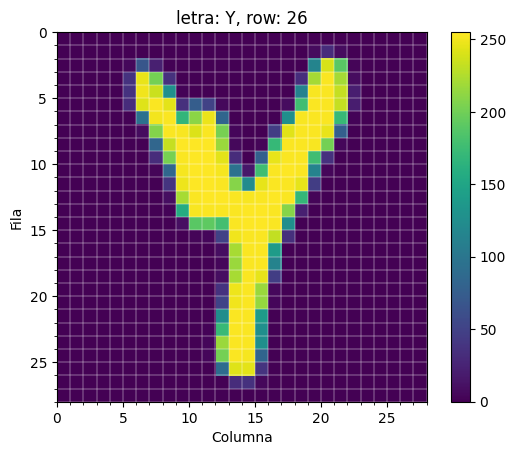


Fig. 2: representación en grilla de la 26ta observación. El color representa el valor de intensidad de dicho píxel.

Cada fila en nuestra base de datos representa una letra, y cada columna representa la intensidad de dicho píxel. Como la grilla (28x28) está representada en una serie de una dimensión, y la base utilizada no aclara si los datos vienen ordenados por filas o por columnas (es decir, si los primeros 28 valores son la primera fila o columna), se utilizó la observación graficada en la Fig. 2 para deducirlo, ya que se sabe que una “Y” debería contar, por ejemplo, en la cuarta fila, con dos picos de intensidad separados por un vacío. En efecto, se ve en la Fig. 3 que no coinciden las intensidades de los valores correspondientes a la fila 4 de la figura anterior, y a los valores 112 a 139 (es decir, los que corresponderían a la cuarta fila, si la base estuviese ordenada de fila en fila). No obstante, este último gráfico sí coincide con las intensidades de la cuarta columna de la Fig. 2.

Esto representa un desafío adicional en relación al análisis exploratorio de una relación en la que las variables pueden interpretarse de manera independiente, ya que en este caso, es necesario explorar las variables en su contexto: un atributo no es solo una variable, sino también una posición, siendo necesario además analizarla en relación a sus vecinas.

Habiendo concluido que los datos de la base estaban ordenados de “columna en columna”, se renombraron las variables con su número de fila y columna para su posterior análisis.

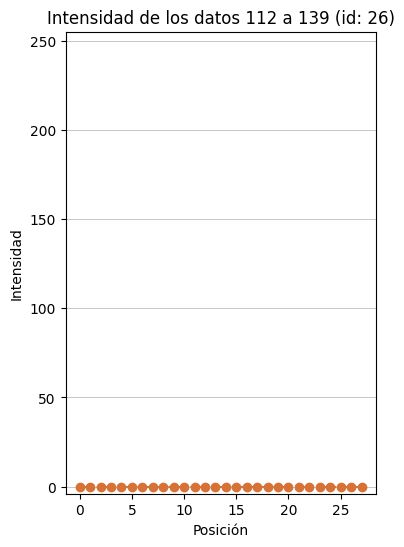
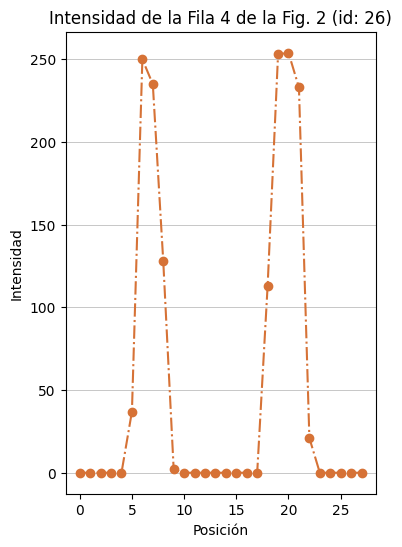


Fig. 3: gráficos correspondientes a la cuarta fila de la Fig. 2 y a los datos que deberían representar a la cuarta fila si estuviesen ordenados de fila en fila

En primera instancia, se muestran en la Fig. 4 la media, mediana y máximo de cada píxel entre todas las entradas. En el gráfico de los máximos, se ve que, aun teniendo una muestra de 86.400 imágenes, las dos primeras y últimas filas/columnas presentan pocos píxeles con intensidades significativas. Esta noción se observa aún más al analizar las medias y las medianas: las tres primeras y últimas filas/columnas dan valores prácticamente nulos, mientras que en las medianas esto se expande a mayores regiones.

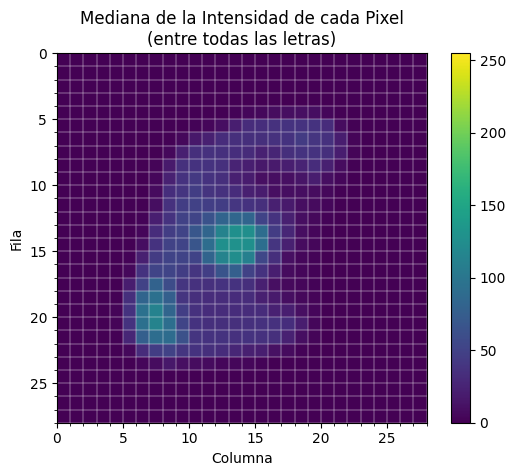
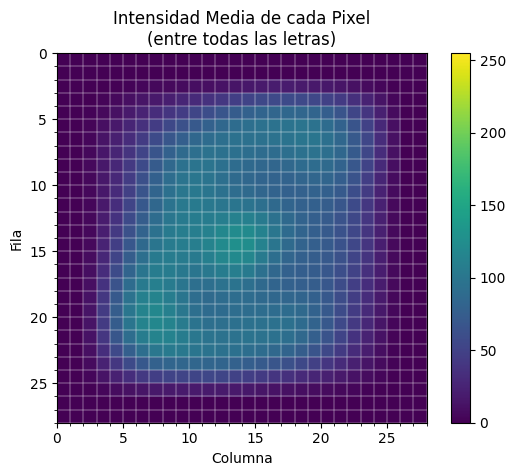
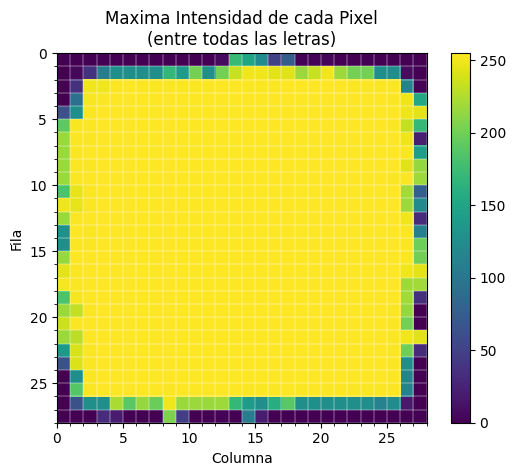


Fig. 4: Intensidad mediana, máxima y media por píxel entre todas las letras de la base.

Esto seguramente se deba a que, en la mayoría de las letras, la mayoría de los píxeles están vacíos (su intensidad es nula), pero aquellos que tienen valor tienen intensidad alta (ya que todas las imágenes están normalizadas a 255), generando entonces que la media se vea muy afectada por los valores extremos. A modo de ejemplo, se muestran en la Fig. 5 histogramas de algunos píxeles para todas las letras. Se observa que, como explicamos antes, la media suele ser significativamente mayor a la mediana, exceptuando en el caso de la columna 14 y la fila 15, donde la mediana es mucho mayor a la de los otros casos, casi empatando a la media.

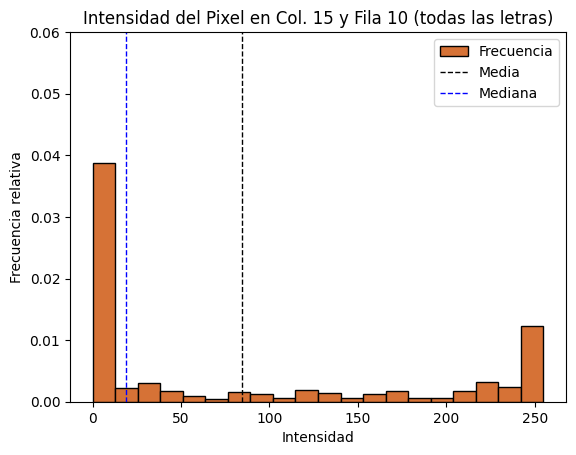
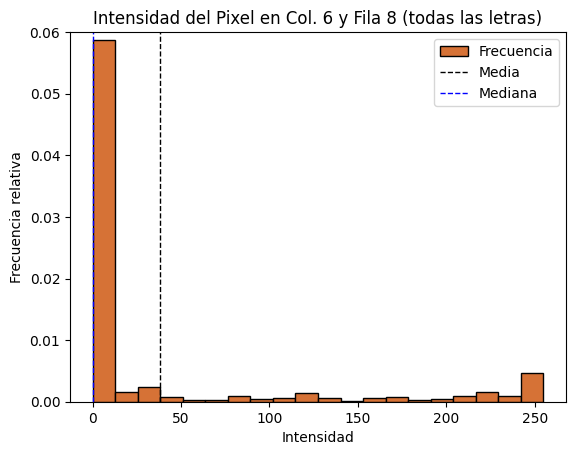
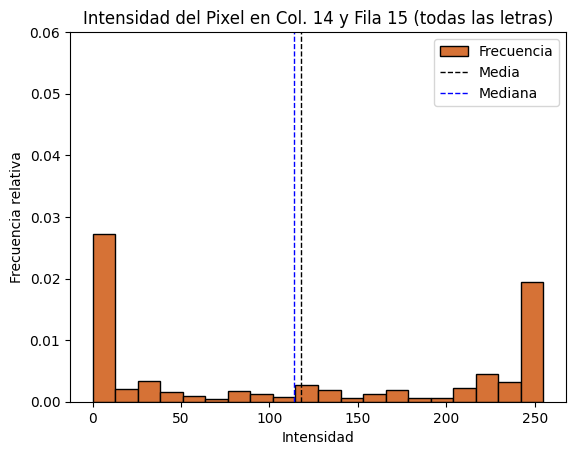
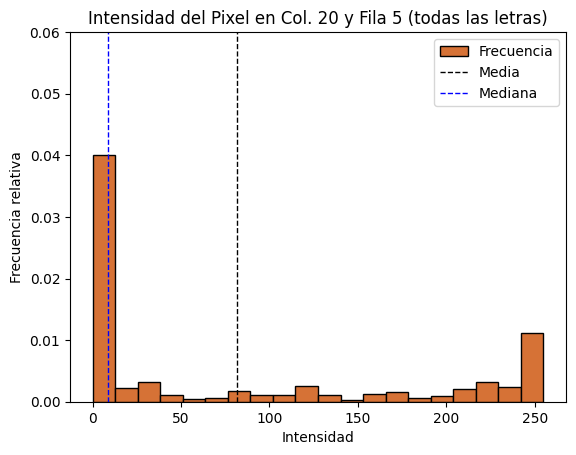


Fig. 5: Histogramas de la intensidad entre todas las letras para distintos pixeles, junto a su respectiva media y mediana

Esta diferencia entre medias y medianas se observa no sólo en las primeras y últimas filas/columnas, sino también en la intensidad de las zonas centrales: en el gráfico de las medianas se observa una figura similar a una ε, mientras que en el de las medias esta figura está mucho más difuminada por toda la zona central. Esto puede interpretarse como que, al analizar las medias, los píxeles son más similares entre sí, mientras que, analizando las medianas, se observa con mayor intensidad una separación entre clases.

|  |  |  |  |
| --- | --- | --- | --- |
| Tabla 1: Desvío de la diferencia entre medias y medianas de distintos pares de letras  Par de letras | Desvío de la diferencia entre medianas | Desvío de la diferencia entre medias | Diferencia porcentual |
| E y M | 85.26 | 61.10 | 39.6% |
| I y L | 16.14 | 25.24 | 56.4% |
| E y L | 73.97 | 44.46 | 66.4% |
| E y I | 78.97 | 53.83 | 46.7% |

En efecto, explorando la diferenciabilidad entre clases, se realizaron gráficos de la resta entre medias y medianas para distintos pares de letras. En la Fig. 6, se puede ver que las diferencias de intensidades medias, tanto para la comparación de “E” con “M”, como para “L” con “I”, son de menor magnitud que las mismas comparaciones entre medianas. Se muestran en la Error: Reference source not found los desvíos de las diferencias entre medias, medianas y su diferencia en porcentaje para estos y otros pares de letras.

Se ve que, en todos los casos, el desvío entre las diferencias de medianas es mayor al mismo para las medias.

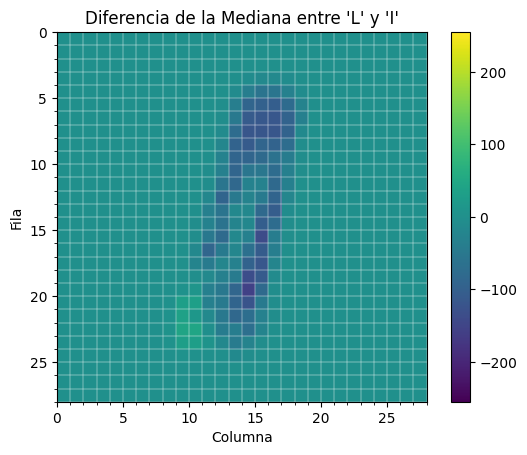
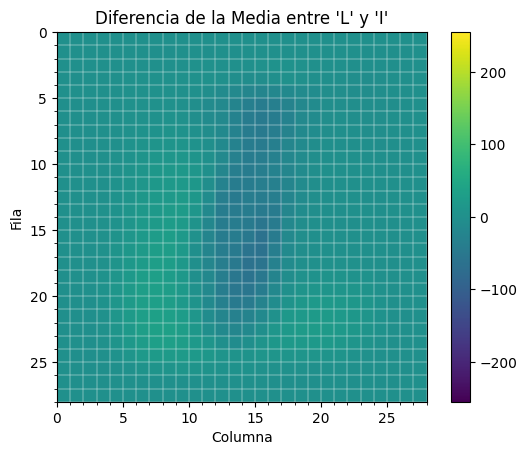
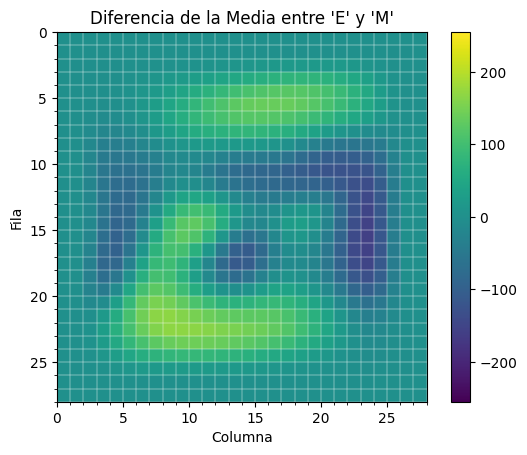
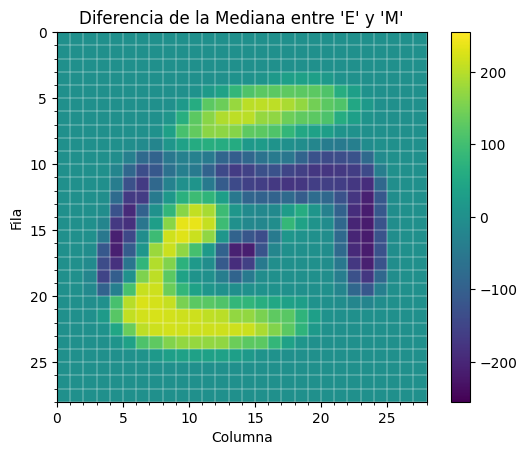


Fig. 6: Diferencias entre medias y medianas de “E” con “M” y “L” con “I”. Se puede apreciar que las diferencias entre medianas son más importantes

Cada clase presenta su propia variabilidad, y esta depende de la letra a la que refiera. A modo de ejemplo, se observan en la Fig. 7 los desvíos estándar de los píxeles para la letra C y para la letra A. Se puede ver que la letra C tiene su mayor variabilidad en una región acotada y es posible identificarla forma de la letra, mientras que para la letra A, esta variación se expande a un mayor dominio, y se pierde la identificabilidad.

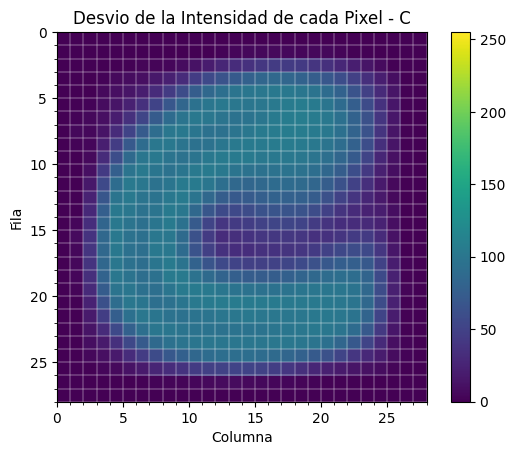
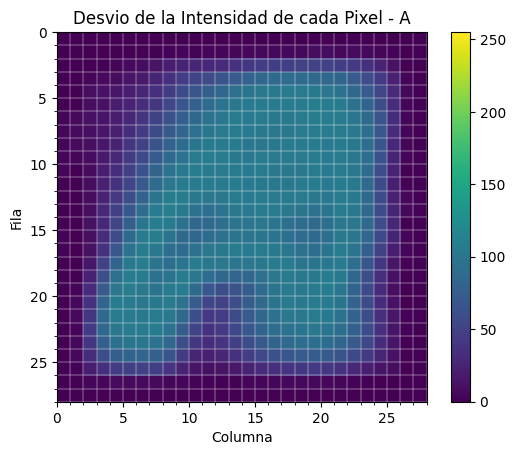


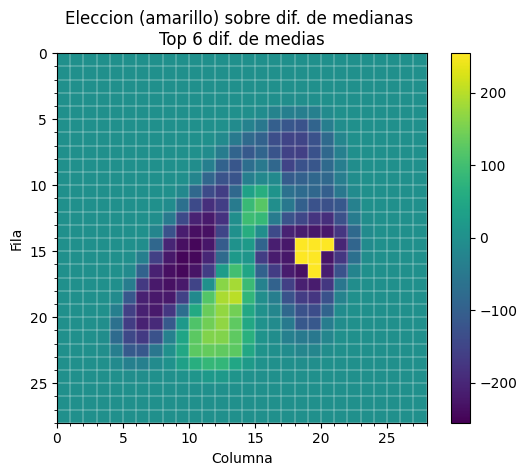
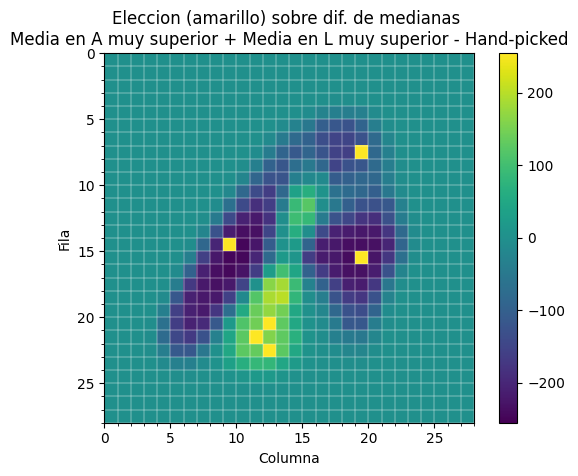
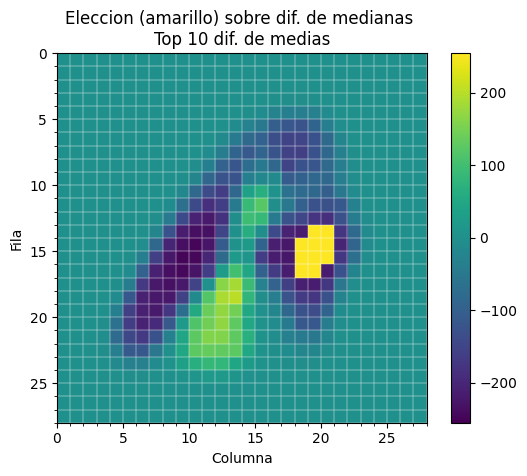
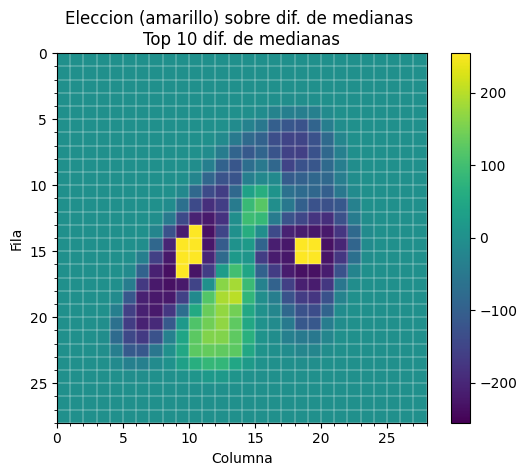
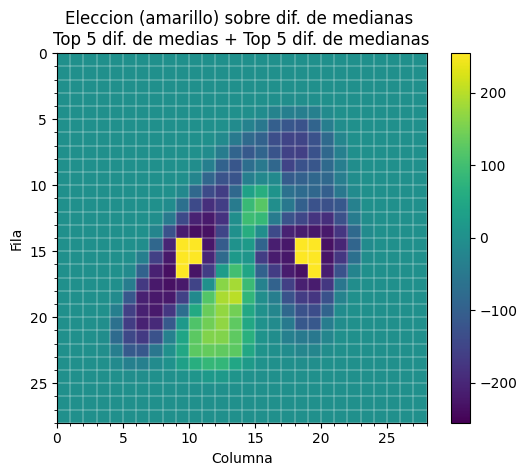
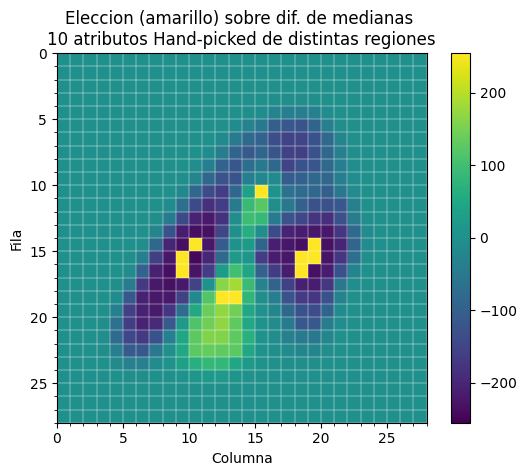
Fig. 7: Desvío de la intensidad para las letras “A” y “C”.

## Experimentos realizados

### Clasificación binaria

En primera instancia, se conservaron sólo las imágenes de las letras “L” y “A” y se dividió la nueva tabla, aislando un 30% como conjunto de test. Se observaron las medias, medianas y diferencias de estas entre ambas categorías para decidir distintas formas de escoger atributos. Se probaron diversas estrategias de selección de variables, ya sean criterios numéricos o arbitrarios. En la Fig. 8 se muestran algunos de los conjuntos de píxeles (en amarillo) utilizados para entrenar cada clasificador, sobre el gráfico de la diferencia de la mediana de cada píxel entre ambas letras.

Fig. 8: Gráfico de variables elegidas para el entrenamiento de seis de los clasificadores (en amarillo). El resto de los colores representa la diferencia de las medianas para cada píxel entre las letras “L” y “A”



Para cada uno de los conjuntos de variables seleccionados, se ajustaron modelos de clasificación de kNN utilizando 5 y 15 vecinos. Para cada clasificador, se calcularon las matrices de confusión y sus métricas derivadas: precisión, exactitud (o accuracy), exhaustividad (o recall) y F1 Score.

### Clasificación multiclase

Se conservaron sólo las imágenes de las vocales y se dividieron en un conjunto de entrenamiento y uno de test, guardando un 30% del contenido para este último. Luego, se utilizó K-folding para encontrar el mejor hiperparámetro de altura, entre 3 y 14, utilizando como métricas las medidas de impureza de Gini y la de entropía de Shannon (*entropy* en scikit-learn). Se calcularon métricas multiclase para cada altura según su criterio y se compararon los mejores de ambos, conservando así los hiperparámetros que resultaron mejores.

A modo de ejemplo, en la Fig. 9 se observan los primeros niveles de uno de los clasificadores ajustados.

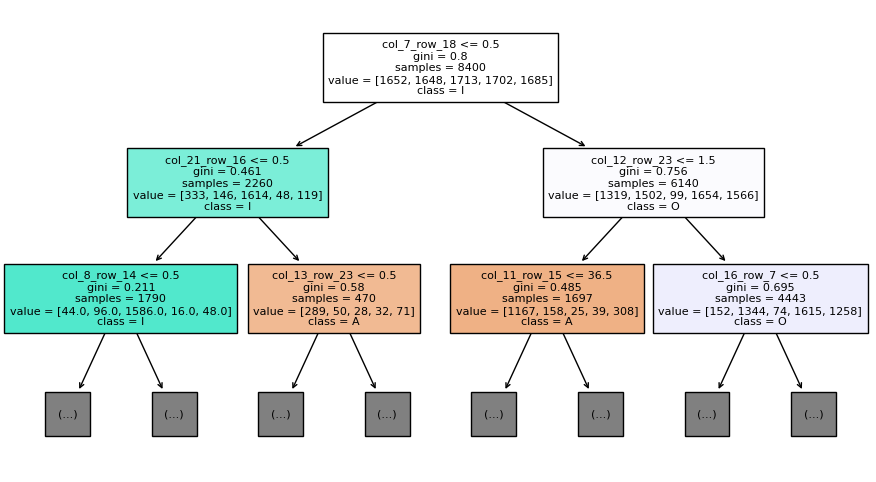


Fig. 9: Ejemplo primeras hojas de un árbol de decisión. En este caso, se utilizó Gini como métrica de impureza, y su profundidad total es 9 (no graficada)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Atributos | N\_neighbors | Accuracy | Precisión | Recall | F1 |
| Top 10 dif. de medias | 15 | 0.98056 | 0.97629 | 0.98453 | 0.98039 |
| Top 10 dif. de medias | 5 | 0.97986 | 0.97759 | 0.98172 | 0.97965 |
| Top 6 dif. de medias | 15 | 0.97639 | 0.9669 | 0.98594 | 0.97632 |
| Top 6 dif. de medias | 5 | 0.97639 | 0.97079 | 0.98172 | 0.97622 |
| Top 5 dif. de medias + Top 5 dif. de medianas | 5 | 0.97222 | 0.95771 | 0.98734 | 0.9723 |
| Top 3 dif. de medias | 15 | 0.97153 | 0.9589 | 0.98453 | 0.97155 |
| Top 10 dif. de medianas | 5 | 0.97014 | 0.95013 | 0.99156 | 0.97041 |
| Top 3 dif. de medias | 5 | 0.96875 | 0.95994 | 0.9775 | 0.96864 |
| Top 6 dif. de medianas | 5 | 0.96667 | 0.94737 | 0.98734 | 0.96694 |
| Top 5 dif. de medias + Top 5 dif. de medianas | 15 | 0.96528 | 0.94362 | 0.98875 | 0.96566 |
| 10 atributos Hand-picked de distintas regiones | 5 | 0.96319 | 0.9375 | 0.99156 | 0.96377 |

## Conclusiones

### Clasificación binaria

En la tabla 2 se observan algunos de los atributos seleccionados junto a sus métricas (completo en el archivo *clasif\_binaria\_res.csv* del Anexo). Se puede observar que la elección de píxeles no distintivos, correspondiente a las esquinas, no resulta para nada útil, pues el modelo resulta sumamente sencillo: clasifica todos los casos como la misma letra. Esto no cambia al aumentar el número de vecinos, pues se trata de píxeles que, como vimos en los máximos de la base de datos entera, no se activan en ninguna letra, por lo que no hay vecinos para comparar.

El caso donde escogimos píxeles donde la media en A es muy superior resultó eficaz bajo todas las métricas, por arriba de aquella elección sobre la media en L superior. Esto tiene sentido, pues vimos en gráficos que la diferencia de medias en los píxeles elegidos sobre A tiene mayor intensidad sobre la de píxeles elegidos sobre L, es decir, estos últimos no eran tan distintos en módulo.

Tabla 2: Algunos de los clasificadores utilizados junto a sus métricas multiclase, ordenados por F1. La escala de colores es independiente en cada métrica.

Los casos donde escogimos píxeles donde la diferencia absoluta de media era máxima resultaron los más eficaces bajo todas las métricas, mejorando leve pero consistentemente al considerar más atributos.

Al mezclar criterios el modelo parece dar prioridad a aquellos atributos donde los valores están más diferenciados, “heredando” las métricas calculadas cuando solo elegimos esos. Debido a que el modelo KNN se basa en distancias, prioriza estos atributos ya que provocan más distancia euclídea sobre el aporte del eje de los atributos más diferenciados entre las clases. Es decir, al realizar el cuadrado de la resta de estas coordenadas, habrá más peso en la clasificación.

En este caso, la diferencia al cambiar el número de vecinos de 5 a 15 no resulta significativa; en algunos casos aporta, en otros resta, pero siempre en muy baja medida.

Clasificación multiclase:

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Altura | Criterio | Acc | P\_A | P\_E | P\_I | P\_O | P\_U | R\_A | R\_E | R\_I | R\_O | R\_U |
| 3 | Gini | .69 | .78 | .56 | .92 | .48 | .79 | .83 | .03 | .92 | .92 | .72 |
| 9 | Gini | .91 | .91 | .89 | .95 | .91 | .90 | .88 | .91 | .95 | .93 | .90 |
| 14 | Gini | .91 | .90 | .89 | .94 | .91 | .89 | .88 | .91 | .95 | .92 | .89 |
| 3 | Entropy | .81 | .89 | .85 | .88 | .73 | .73 | .70 | .70 | .96 | .87 | .80 |
| 8 | Entropy | .91 | .89 | .90 | .95 | .91 | .90 | .88 | .91 | .95 | .93 | .89 |
| 11 | Entropy | .91 | .90 | .92 | .94 | .91 | .88 | .87 | .91 | .95 | .92 | .90 |
| 14 | Entropy | .91 | .90 | .91 | .94 | .91 | .89 | .87 | .91 | .95 | .92 | .90 |
| … | … | … | … | … | … | … | … | … | … | … | … | ... |

*Tabla 3: Métricas de Accuracy (acc), Precisión (P\_vocal) y Recall (R\_vocal) sobre el conjunto de train para cada par de hiperparámetros*

Como se puede observar en la Tabla 3, obtenemos que, como es de esperar, la altura más baja (para clasificar 5 clases es necesario al menos 3 decisiones) no obtiene resultados muy buenos. En esta altura el criterio de Gini brindó resultados muy bajos para algunas letras y muy altos para otras, como se puede observar con la precisión y recall, mientras que el criterio de Entropy fue bastante parejo. Es explicable, pues Gini prioriza aislar una clase antes de continuar con otra, mientras que Entropy intenta formar árboles más balanceados en general. A alturas mayores, la diferencia entre ambos criterios se mitiga considerablemente.

Eventualmente, aumentar la altura no cambia significativamente las métricas, por lo que los árboles más altos suelen ser diferenciables desde terceros decimales en adelante.

|  |  |
| --- | --- |
| Criterio | Gini |
| Altura | 9 |
| Accuracy | 0.923 |
| Promedio Precisión | 0.924 |
| Promedio Recall | 0.923 |
| Precisión A | 0.896 |
| Recall A | 0.917 |
| Precisión E | 0.904 |
| Recall E | 0.917 |
| Precisión I | 0.976 |
| Recall I | 0.947 |
| Precisión O | 0.929 |
| Recall O | 0.948 |
| Precisión U | 0.912 |
| Recall U | 0.884 |

*Tabla 4: Métricas del árbol con mejores resultados*

Como se puede ver en la Tabla 4, nos encontramos con que el árbol de altura 9 y criterio Gini resulta mejor comparando el accuracy, el promedio de precisiones y el promedio de recalls de todos los árboles. Si bien los árboles muy altos tienden a provocar overfitting, no es un fenómeno muy observable en estas alturas, pues se trata de una base de datos con muchos atributos y unas alturas proporcionalmente bajas. Es debido a la calidad de la base de datos y al bajo número de letras elegidas que nos encontramos con que estas alturas nos dan resultados tan altos.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | A | E | I | O | U |
| A | 686 | 26 | 4 | 9 | 23 |
| E | 18 | 690 | 11 | 15 | 18 |
| I | 5 | 20 | 651 | 3 | 8 |
| O | 19 | 5 | 0 | 662 | 12 |
| U | 37 | 22 | 1 | 23 | 632 |

*Tabla 5: Matriz de confusión del árbol de criterio Gini y altura 9*

En la tabla 5 se puede observar que el árbol confunde la letra A con la E y U mayormente, mientras que apenas lo hace con la I y O. Respecto a la letra E, la confunde de forma bastante pareja con todas las letras, mientras que la letra I es mayormente confundida con la E y apenas con otra letra. En el caso de la O, nos encontramos que nunca fue predecida como I, y apenas como E, con varias confusiones con la U y A. Finalmente, la letra U fue bastante confundida con la A y de forma intermedia con la E y O, y apenas con la I.