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

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 del inicio de la Constitución Nacional de EEUU.

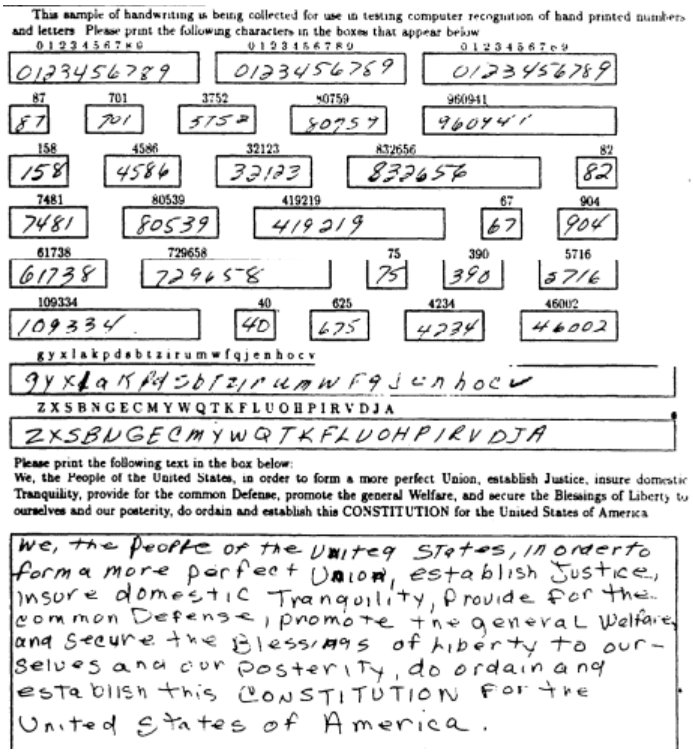

Fig. 1: Formulario para el Desarrollo de la base original del NIST. Extraído de https://www.nist.gov/system/files/documents/srd/nistsd19.pdf

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.

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 imágenes en total). Cada una de estas 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.

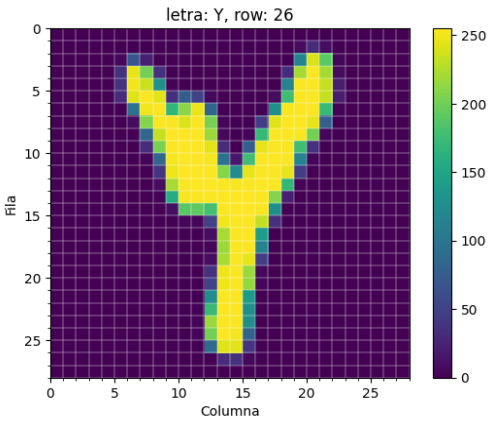

Fig. 2: representación en grilla de la 26ta observación. El color representa el valor de intensidad de dicho píxel.

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.

Esta forma de representación de los datos implica un desafío adicional en cuanto al análisis exploratorio respecto a 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: cada atributo no es solo una variable, sino que también tiene una posición, siendo necesario además analizarlo en

relación a sus vecinos. 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.

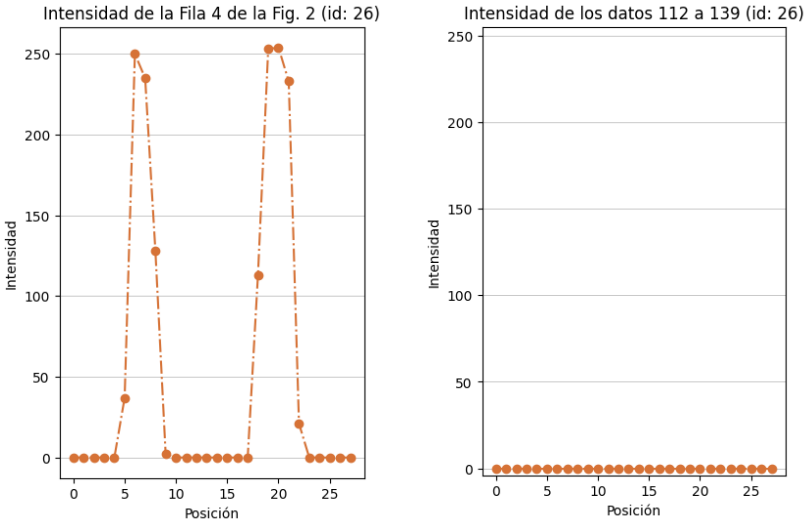

Fig. 3: gráficos correspondientes a la cuarta fila de la Fig. 2 y a los datos que deberían representar a la cuarta fila si estuviesen ordenados de fila en fila

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.

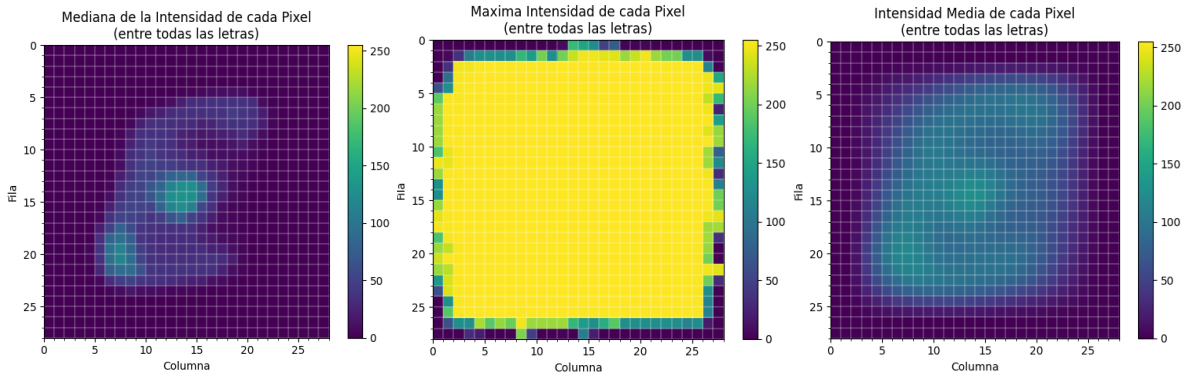

Fig. 4: Intensidad mediana, máxima y media por píxel entre todas las letras de la base.

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.

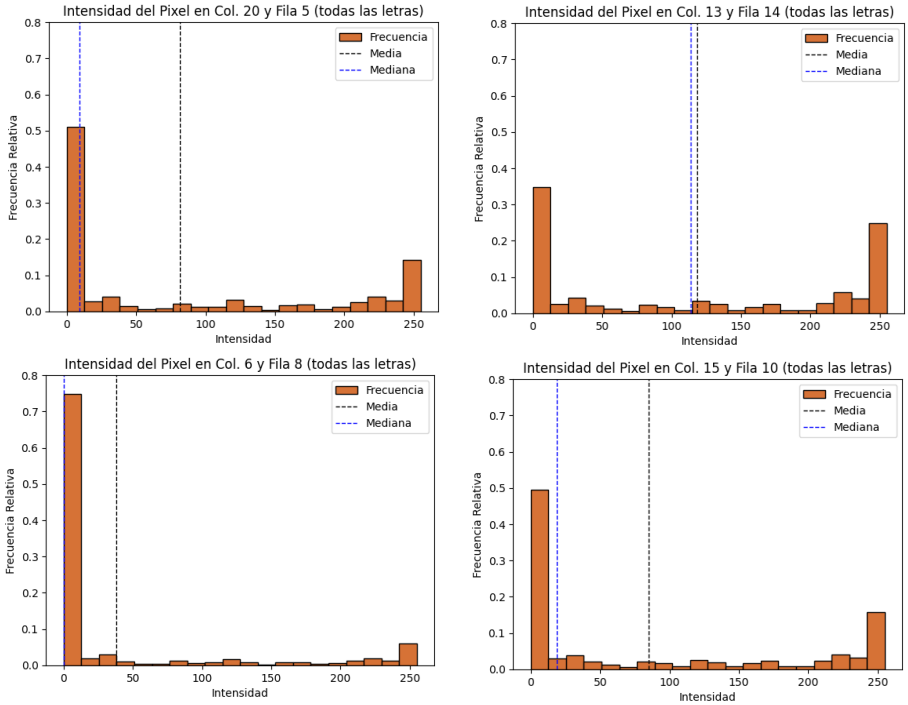

Fig. 5: Histogramas de la intensidad entre todas las letras para distintos pixeles, junto a su respectiva media y mediana

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, debido a la reducción del peso de los valores extremos.

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	25.24	16.14	56.4%
E y L	73.97	44.46	66.4%
E y I	78.97	53.83	46.7%

Tabla 1: Desvío de la diferencia entre medias, medianas y su diferencia en porcentaje para los píxeles entre distintos pares de letras

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 Tabla 1 los desvíos de las diferencias entre medias, medianas y su diferencia en porcentaje para estos y otros pares de letras. Se observa que, en todos los casos, el desvío entre las diferencias de medianas es mayor al mismo para las medias.

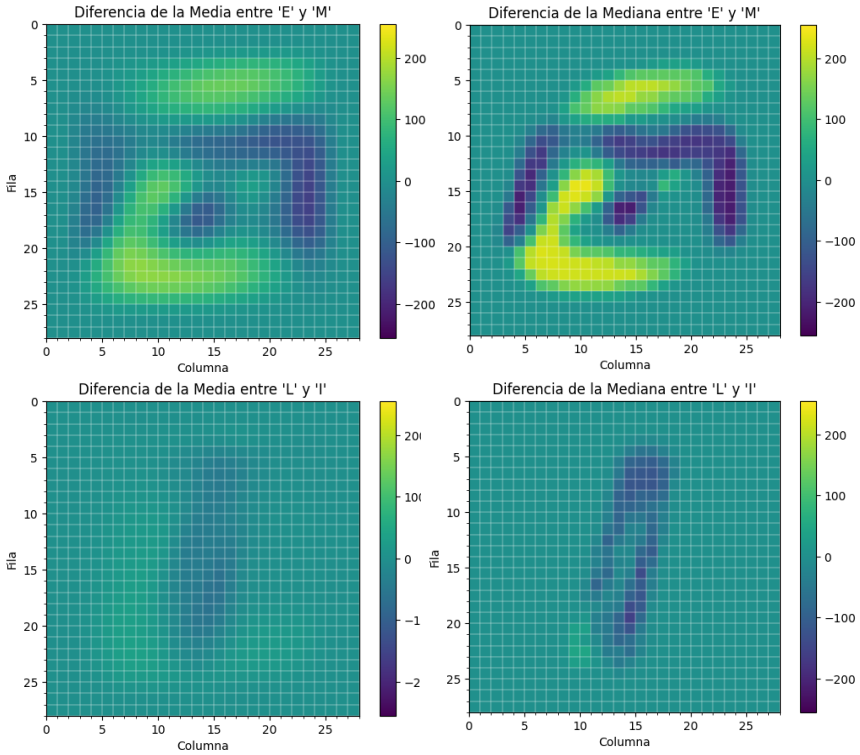

Fig. 6: Diferencias entre medias y medianas de "E" con "M" y "L" con "I". Se puede apreciar que las diferencias entre medianas son más importantes

Continuando el análisis de la Fig. 6 se observa que, tanto para la media como para la mediana, las diferencias de "E" con "M" son mucho mayores que de "L" con "I". Para compararlos numéricamente, se calcularon las medias de las distancias entre ambas medidas de centralidad. Entre "E" y "M", la distancia entre medias promedia 41.45 y 49.03 para medianas. Estas diferencias son ampliamente mayores que las de "L" con "I", las cuales promedian 9.25 y 9.29 respectivamente. Esto coincide con la noción intuitiva de que una "L" es más parecida a una "I", y por ende serán más difíciles de distinguir. También se observa que la distancia entre medianas es mayor que entre medias, lo cual sugiere que podría ser un mejor criterio de diferenciación.

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

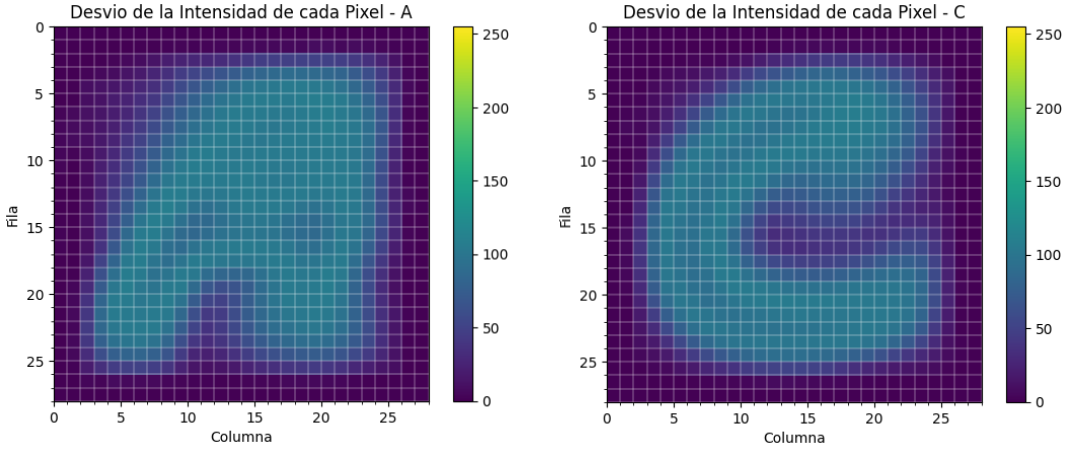

Fig. 7: Desvío de la intensidad para las letras "A" y "C".

forma de la letra, mientras que para la letra A, esta variación se expande a un mayor dominio, y se pierde la identificabilidad.

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.

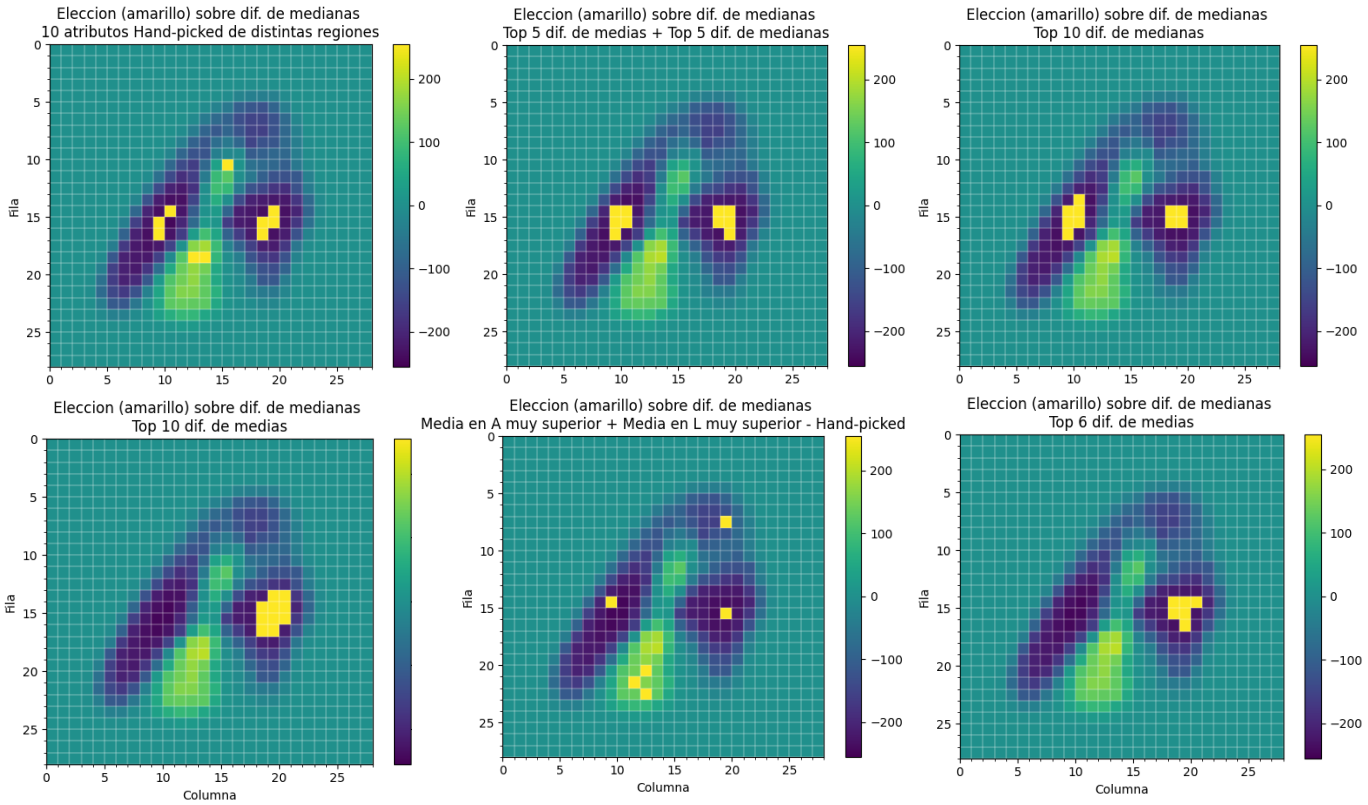

Fig. 8: Gráfico de variables elegidas para el entrenamiento de seis de los clasificadores (en amarillo). El resto de los colores representa la diferencia de las medianas para cada píxel entre las letras “L” y “A”

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 correspondientes a 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 los mejores hiperparámetros de altura, entre 3 y 14, y de medidas de impureza, entre Gini y entropía de Shannon (*entropy* en scikit-learn). Se calcularon métricas multiclase para cada altura según su criterio y se compararon los mejores de cada medición de impureza,

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.

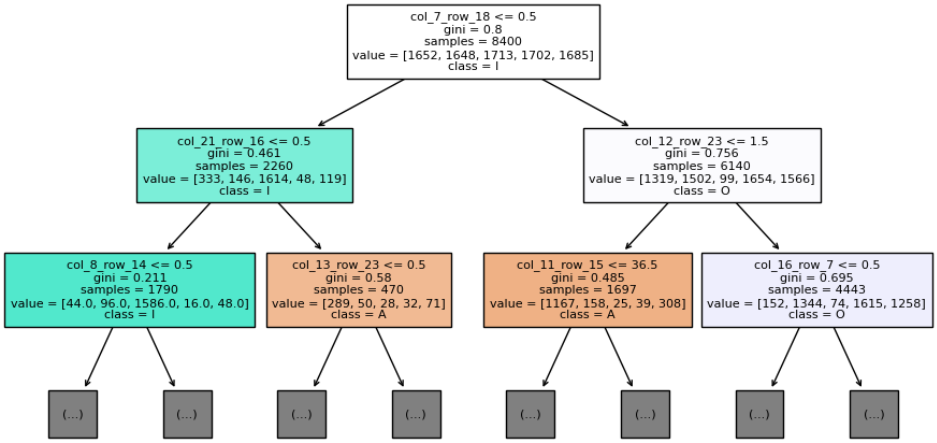

Fig. 9: Ejemplo primeras hojas de un árbol de decisión. En este caso, se utilizó Gini como métrica de impureza, y su profundidad total es 9 (no graficada)

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 aunque tenga un Recall perfecto, 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 se escogieron 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.

Atributos	N-Vecinos	Exactitud	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

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

Sin embargo, los píxeles donde la diferencia absoluta de media era máxima resultaron los más eficaces bajo todas las métricas salvo la exhaustividad, mejorando leve pero consistentemente al considerar más atributos. La exhaustividad llega a su máximo valor en los atributos con mayor diferencia de medianas, pero en estos se pierde exactitud y precisión.

Al mezclar criterios el modelo parece dar prioridad a aquellos atributos donde los valores están más diferenciados, "heredando" las métricas calculadas cuando solo se eligieron 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, y los atributos que no brindan información serán apenas tenidos en cuenta

Para este problema, la diferencia al cambiar el número de vecinos de 5 a 15 no pareciera ser importante: mejorando o empeorando en distintos casos.

A pesar de estas diferencias, se puede concluir que cualquiera de los modelos que se observan en la Tabla 2 son muy aceptables como clasificadores entre ambas letras.

Clasificación multiclase

Altura	Criterio	Accuracy	Prec. A	Prec. E	Prec. I	Prec.O	Prec. U	Prec. Media	Rec. A	Rec. E	Rec. I	Rec. O	Rec. U	Rec. Media
9	Gini	0.918	0.915	0.891	0.954	0.919	0.910	0.918	0.885	0.915	0.954	0.934	0.901	0.918
8	Entropy	0.917	0.897	0.908	0.957	0.914	0.906	0.916	0.883	0.912	0.956	0.935	0.897	0.917
11	Gini	0.915	0.909	0.900	0.946	0.914	0.906	0.915	0.885	0.907	0.959	0.928	0.897	0.915
11	Entropy	0.915	0.902	0.925	0.950	0.911	0.889	0.915	0.877	0.910	0.957	0.928	0.904	0.915
13	Entropy	0.915	0.906	0.917	0.949	0.908	0.893	0.915	0.877	0.911	0.958	0.927	0.900	0.915
10	Entropy	0.914	0.907	0.913	0.948	0.905	0.899	0.914	0.871	0.910	0.960	0.940	0.891	0.914
12	Entropy	0.914	0.905	0.920	0.944	0.911	0.892	0.914	0.875	0.911	0.957	0.927	0.902	0.914
14	Entropy	0.914	0.905	0.918	0.949	0.909	0.892	0.914	0.876	0.914	0.958	0.924	0.900	0.914
12	Gini	0.914	0.905	0.902	0.947	0.915	0.900	0.914	0.886	0.908	0.956	0.925	0.894	0.914

*Tabla 3: Precisión y Exhaustividad para cada clase y medias, y Exactitud para algunos de los hiperparámetros evaluados, ordenados por Exactitud. La escala de color es independiente en cada variable. Extraída de **clasif_multi_train_res.csv** del Anexo.*

Como se puede observar en la Tabla 3 (completa en *clasif_multi_train_res.csv* del Anexo), las alturas más bajas no obtienen buenos resultados (para clasificar 5 clases es necesario al menos 3 decisiones). En estas alturas el criterio de Gini brindó resultados muy bajos para algunas letras y muy altos para otras, como se puede observar con la precisión y la exhaustividad, mientras que el criterio de entropía fue bastante más parejo. Esto se explica porque Gini prioriza aislar una clase antes de continuar con otra, mientras que la entropía de Shannon intenta formar árboles más balanceados en general. A alturas mayores, la diferencia entre ambos criterios se mitiga considerablemente. A partir de cierta altura, aumentarla no cambia significativamente las métricas, por lo que los árboles más altos suelen ser diferenciables sólo desde terceros decimales en adelante.

Criterio	Gini
Altura	9
Accuracy	0.927
Prec. Media	0.928
Rec. Media	0.927
Prec. A	0.904
Rec. A	0.917
Prec. E	0.913

Rec. E	0.92021
Prec. I	0.96603
Rec. I	0.95197
Prec. O	0.94051
Rec. O	0.95129
Prec. U	0.91429
Rec. U	0.8951

Tabla 4: Hiperparámetros y métricas del árbol con mejores resultados.

Como se puede ver en la Tabla 3, nos encontramos con que el árbol de altura 9 y criterio Gini resulta mejor comparando la exactitud (accuracy), el promedio de precisiones y el promedio de recalls de todos los árboles. Si bien en términos generales, los árboles muy altos tienden a sobreajustar (overfitting), no es un fenómeno muy observable en estas alturas, pues se trata de una base de datos con muchos atributos y 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 dan resultados tan altos. En la Tabla 4 se muestran los hiperparámetros y las métricas de el árbol considerado como mejor clasificador

		Predichos				
		A	E	I	O	U
Reales	A	686	23	5	7	27
	E	15	692	14	12	19
	I	7	17	654	3	6
	O	18	8	0	664	8
	U	33	18	4	20	640

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

En la Tabla 5 se puede observar que el árbol confunde mayormente la letra A con la E y U, 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 (seguramente debido al tamaño que ocupa), mientras que la letra I es mayormente confundida con la E y apenas con las otras letras. En el caso de la O, se ve que nunca fue predicha 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.

En términos generales, se puede considerar que el árbol de decisión utilizado para clasificar las vocales da resultados muy buenos, teniendo tanto exactitud como promedios de exhaustividad y precisión de 0.918, es decir, por arriba del 90%.