INFORME DEL ANÁLISIS

Clasificación de imágenes con aprendizaje supervisado (Fashion-MNIST)

Introduccion

El presente informe aborda un análisis exploratorio del conjunto de datos Fashion-MNIST, compuesto por imágenes en escala de grises de 28×28 píxeles que representan 10 clases distintas de prendas de vestir (Figura 0 del Anexo). En cuanto al equilibrio nos encontramos con un dataset balanceado con 7000 ejemplares de cada una. A diferencia de los datasets tabulares clásicos, aquí cada atributo corresponde a un píxel, lo que exige nuevas herramientas visuales para identificar patrones útiles para tareas de clasificación.

En primer lugar, se analizó la relevancia de los atributos individuales a través de la selección de los 100 píxeles más informativos según diferentes criterios. En primer lugar evaluamos la varianza(Figura 1 del Anexo), la distribución espacial de estos píxeles muestra que las regiones externas de la imagen tienen mayor variación con respecto al promedio, mientras que los píxeles centrales tienden a ser menos discriminativos. Luego usamos el “mutual information”(Figura 2 del Anexo) que mide cuánta información comparte un píxel con la variable objetivo, que en este caso sería la clase. A diferencia del mapa por varianza (que suele marcar bordes y extremos), este criterio selecciona los píxeles que más ayudan a diferenciar entre clases, incluso si no son los que más varían.

Luego, se estudió la similitud entre clases mediante un mapa de calor de correlaciones entre los promedios de varianza por clase(Figura 3 del Anexo). Este análisis evidenció una alta similitud entre clases como la 2 (pulóver) y la 4 (camisilla), mientras que clases como la 0 (remera) y la 9 (botas) resultaron claramente diferenciables.

Por último, se examinaron las varianzas de los píxeles dentro de clases específicas (como la clase 0 y la clase 5)(Figuras 4, Figura 5 del Anexo), observando que algunas clases presentan estructuras más homogéneas que otras. Por ejemplo, la clase 0(remeras) muestra una silueta bastante definida y consistente, mientras que la clase 5 (sandalias) revela mayor dispersión entre sus instancias, lo que anticipa posibles desafíos para su clasificación automática.

Este análisis preliminar no sólo permitió reducir la dimensionalidad del problema, sino que también brindó pistas fundamentales para la elección de atributos en los modelos posteriores.

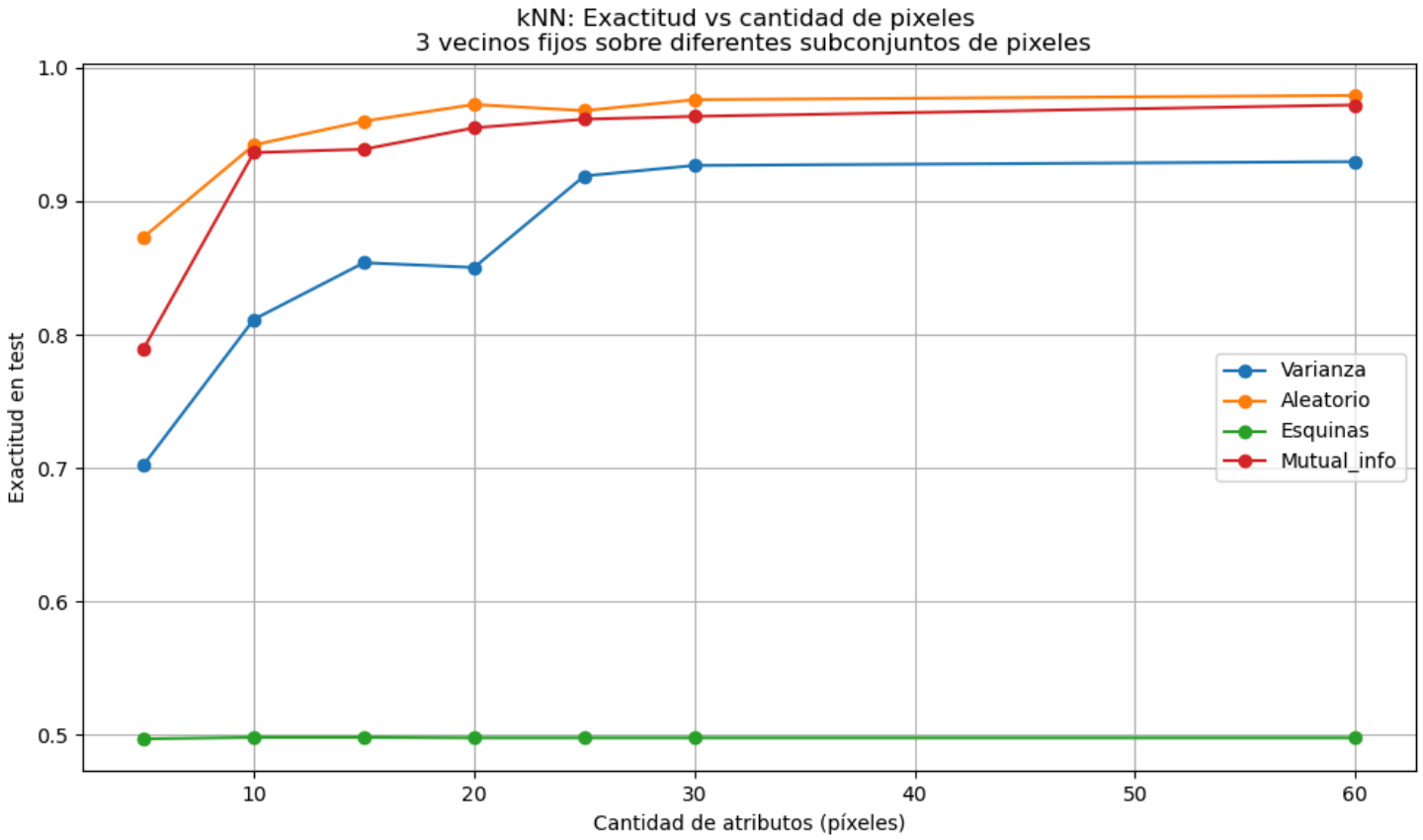
Analisis

Clasificación Binaria

A continuación, abordamos el problema de clasificación binaria entre las clases 0 y 8 del dataset Fashion-MNIST.

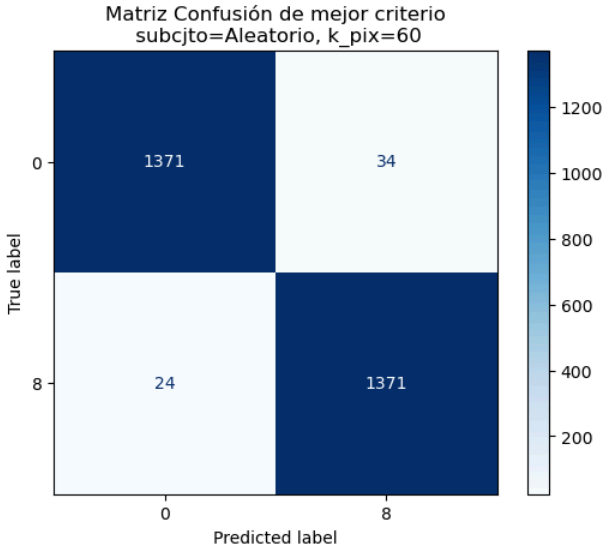
Comparación de cantidad de atributos y criterios de selección de píxeles en kNN

En este experimento empezamos comparando la performance de modelos kNN (con k=3) utilizando subconjuntos de píxeles seleccionados según cuatro criterios distintos: varianza, información mutua, selección aleatoria y píxeles ubicados en las esquinas de la imagen. El objetivo era evaluar qué tipo de atributos (píxeles) permitían predecir con mayor precisión si una imagen correspondía a la clase 0 o a la clase 8.



El gráfico de exactitud muestra que el criterio aleatorio fue el que obtuvo mejor desempeño general, alcanzando más del 97% de exactitud a partir de los 20 píxeles seleccionados. Le siguió el criterio basado en información mutua, que también mostró un crecimiento sostenido y buenos resultados, aunque siempre por debajo del aleatorio. Por su parte, la varianza ofreció mejoras a medida que aumentaba la cantidad de atributos, pero quedó algunos puntos por debajo del resto, con un máximo cercano al 93%. Y a partir de 30 píxeles los modelos no presentan una mejoría significativa.

El criterio de esquinas se mantuvo constante en una exactitud de 0.5 sin importar la cantidad de píxeles, lo que indica que los píxeles ubicados en los extremos de las imágenes no aportan información útil para discriminar entre las clases 0 y 8.



La matriz de confusión correspondiente al mejor modelo (60 píxeles seleccionados aleatoriamente) refuerza estos resultados: el modelo logró predecir correctamente 1371 imágenes de cada clase, cometiendo apenas 58 errores en total. Esto indica que el conjunto está balanceado y que el modelo no presenta sesgos hacia ninguna de las clases.

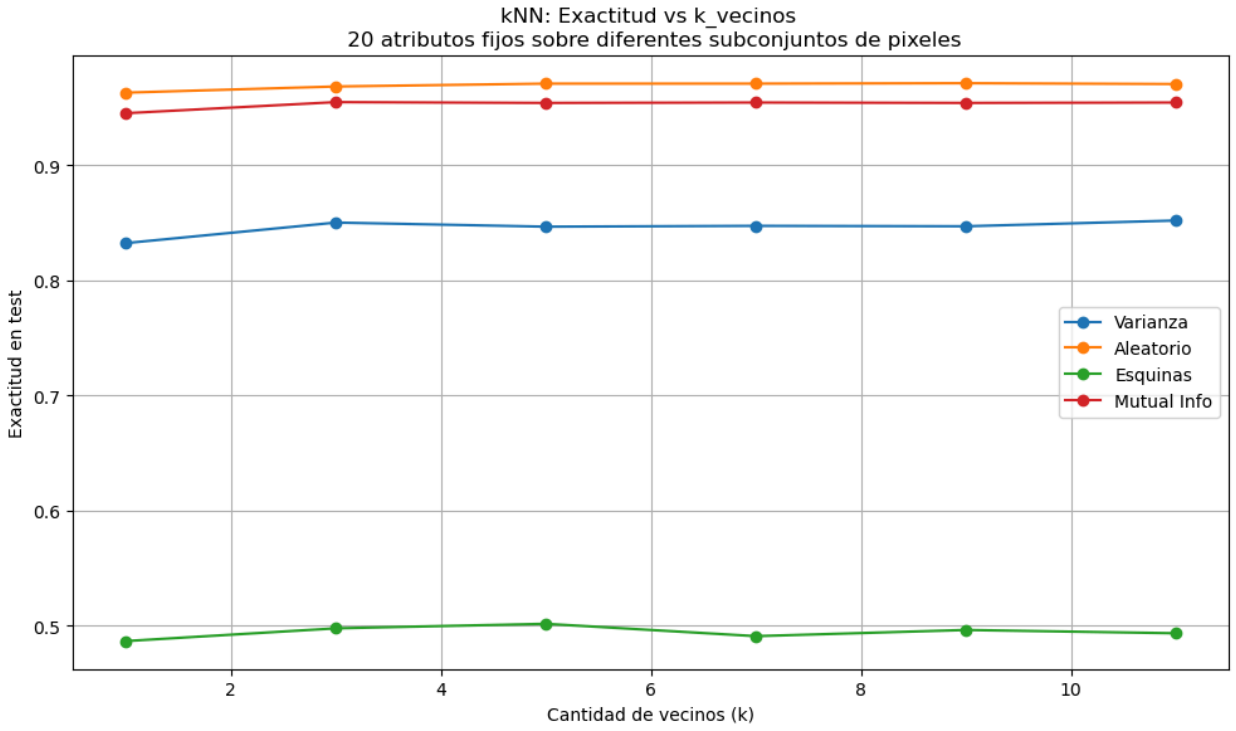
A pesar de que los métodos de selección por varianza e información mutua se aplicaron únicamente sobre las clases 0 y 8, es posible que hayan priorizado píxeles con alta actividad general dentro de esas clases, pero no necesariamente los más efectivos para diferenciarlas entre sí. En cambio, la selección aleatoria, al distribuir píxeles de manera más diversa sobre la imagen, podría haber capturado regiones clave para la discriminación entre ambas clases, incluso sin un criterio explícito de relevancia.

Por lo tanto, este experimento demuestra que:

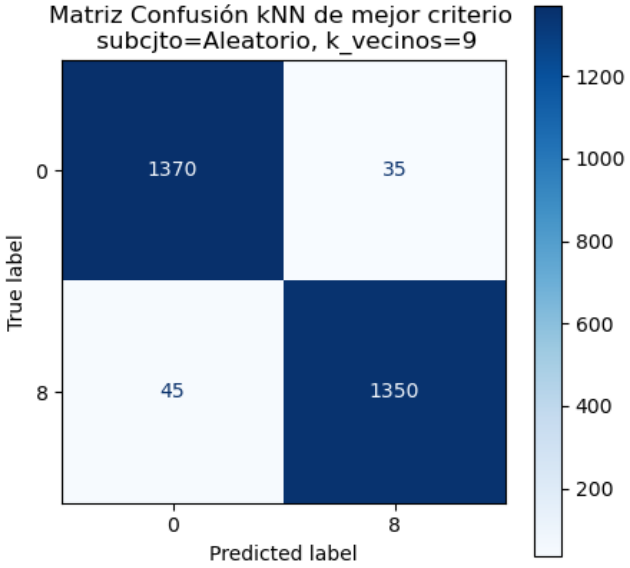
- Es posible lograr altos niveles de exactitud con solo 30 píxeles, sin usar toda la imagen.
- La selección aleatoria de píxeles puede ser sorprendentemente eficaz en este contexto.
- Este enfoque reduce la dimensionalidad y el tiempo de cómputo, sin sacrificar rendimiento.

Comparación de modelos y cantidad de vecinos en kNN

Se entrenó un modelo kNN variando la cantidad de vecinos (de 1 a 11) con una selección fija de 20 atributos, elegidos según cuatro criterios distintos: varianza, aleatorio, esquinas y mutual information.



El gráfico de exactitud muestra que, para kNN, tanto la selección aleatoria como la basada en información mutua mantienen resultados consistentemente altos (por encima del 95%) para una amplia gama de valores de k. En particular, el mejor rendimiento se obtuvo con selección aleatoria y k=9.



La matriz de confusión corresponde al mejor modelo del experimento: kNN con k=9 vecinos, usando 20 píxeles aleatorios.

Se observa una leve caída en el rendimiento respecto al mejor caso del inciso anterior (60 atributos), pero sigue siendo muy alto.

El modelo se mantiene balanceado y preciso, aunque se incrementa ligeramente la confusión entre clases al usar una menor cantidad de información (20 píxeles en lugar de 60).

Por lo tanto, este análisis muestra que:

- El modelo KNN sigue funcionando muy bien incluso con solo 20 píxeles, especialmente cuando se seleccionan aleatoriamente.
- Variar la cantidad de vecinos no tuvo un impacto crítico en el rendimiento: la exactitud fue estable y alta para k ≥ 3.
- Se observa que los píxeles seleccionados aleatoriamente superan a criterios más estructurados, como varianza, mutual info o esquinas.

Comparación preliminar con modelo de regresión logística y decision tree (extra en el anexo)

Se hizo una comparación con otros métodos de clasificación en el anexo. Las matrices de confusión de estos modelos alternativos se incluyen en el Anexo(Figuras 6, 7 y 8) para referencia y comparación.

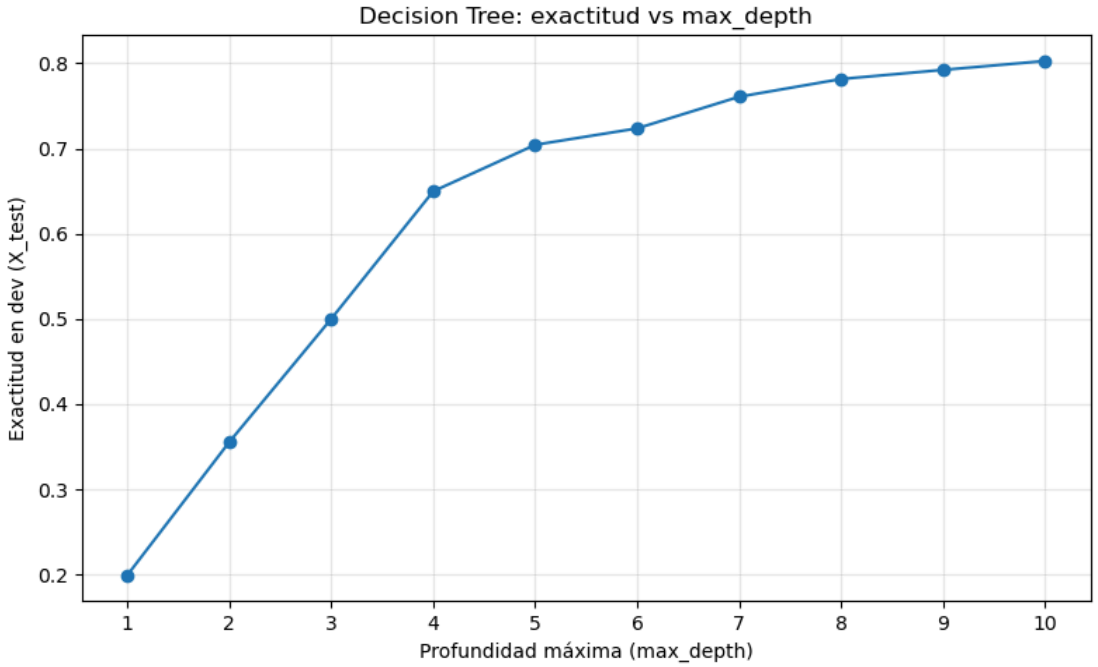
Clasificación Multiclase

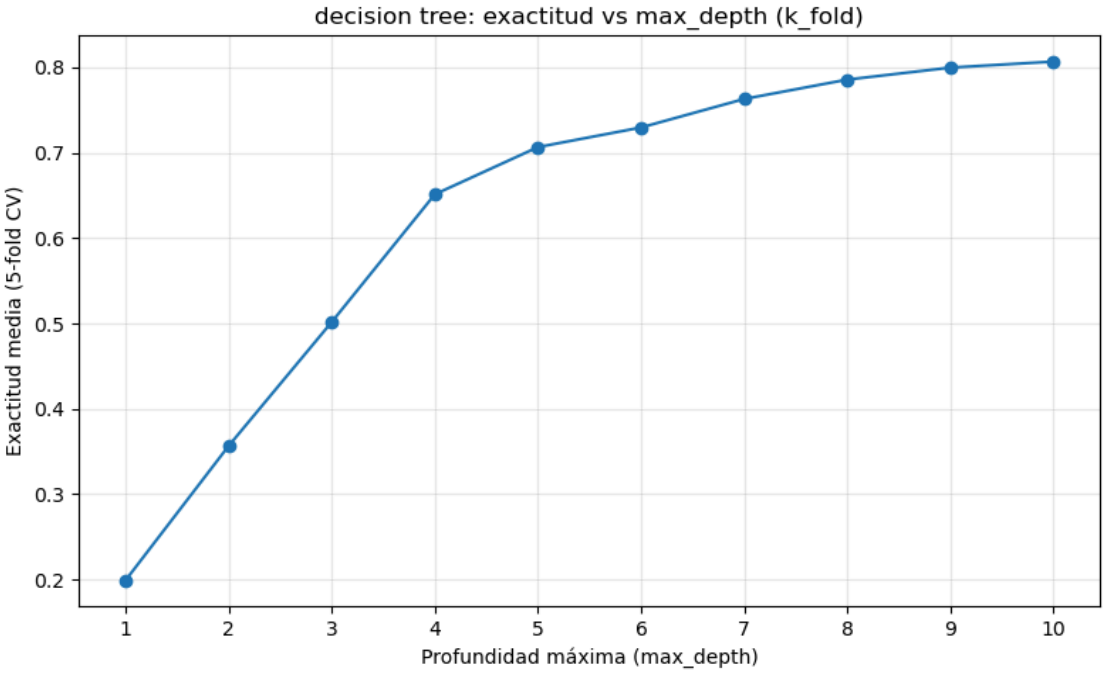
En esta sección abordamos la clasificación multiclase sobre las 10 clases del dataset Fashion-MNIST utilizando árboles de decisión. Primero dividimos el conjunto de datos en un conjunto de desarrollo y otro de validación (held-out), trabajando únicamente con el conjunto de desarrollo en los incisos a, b y c.

Evaluación de profundidad máxima en árbol de decisión

Con el objetivo de ajustar un modelo de árbol de decisión para predecir entre las 10 clases del conjunto Fashion-MNIST, se exploró el impacto de variar el hiperparámetro max_depth entre 1 y 10. Esta evaluación se realizó de dos formas complementarias:

- Conjunto de test fijo dentro del conjunto de desarrollo (X_test)
- Validación cruzada (5-fold CV) para una estimación más robusta del desempeño.





Resultados obtenidos

Ambos gráficos muestran una tendencia clara de mejora en la exactitud a medida que se incrementa la profundidad del árbol:

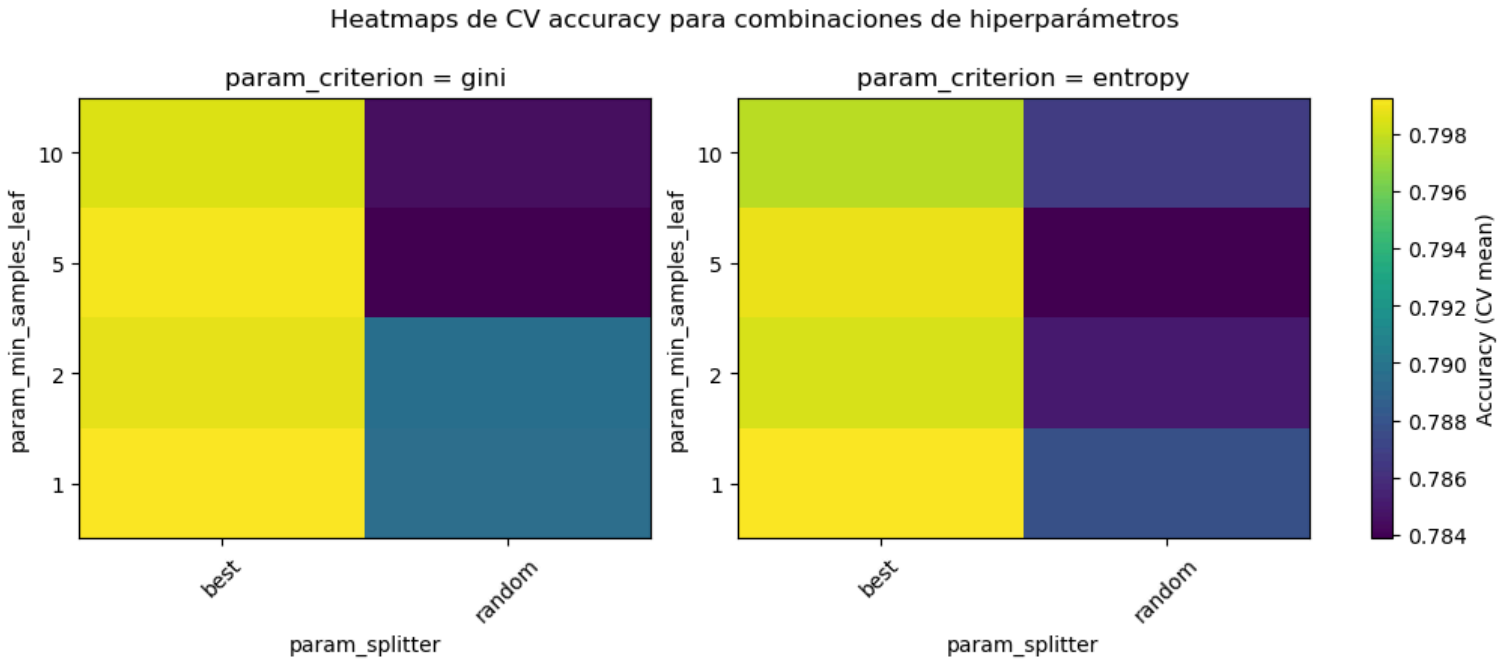
- Entre depth = 1 a 5 se observa un crecimiento abrupto de la performance, indicando que los árboles poco profundos son insuficientes para capturar la complejidad del problema.
- A partir de depth = 6, la mejora se vuelve más gradual, y para depth ≥ 9 la curva empieza a estabilizarse en ambos enfoques.
- El valor max_depth = 9 parece ser el punto de saturación, donde ya no se obtienen beneficios significativos al seguir profundizando el árbol.

Ambos métodos de evaluación (test fijo y CV) arrojaron curvas muy similares, lo que indica consistencia en el comportamiento del modelo, y permite confiar en que la profundidad 9 representa un buen balance entre capacidad predictiva y riesgo de sobreajuste.

Selección de hiper parámetros con validación cruzada

Se realizó una búsqueda exhaustiva (Grid Search) sobre los hiperparámetros del modelo de árbol de decisión: criterion, splitter y min_samples_leaf, utilizando validación cruzada con 5 folds. La profundidad máxima se fijó en 9, según lo observado en el análisis previo.

Los resultados se visualizan en la Figura, que muestra heatmaps de la exactitud promedio en CV para todas las combinaciones evaluadas. Se observó que las mejores configuraciones se obtienen utilizando splitter="best" y valores bajos de min_samples_leaf (1 o 2).



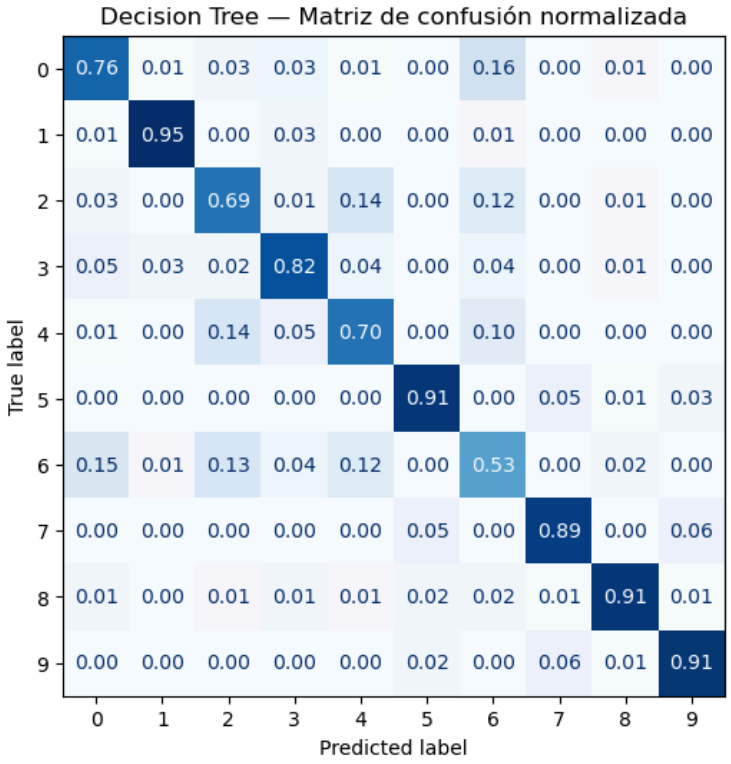
En cuanto al criterio de impureza, no se encontraron diferencias sustanciales entre gini y entropy, aunque esta última mostró un rendimiento levemente superior en su mejor configuración.

En todos los casos, el uso de splitter="random" deterioró el rendimiento, lo cual sugiere que seleccionar las divisiones de forma aleatoria no favorece la estructura del árbol en este contexto.

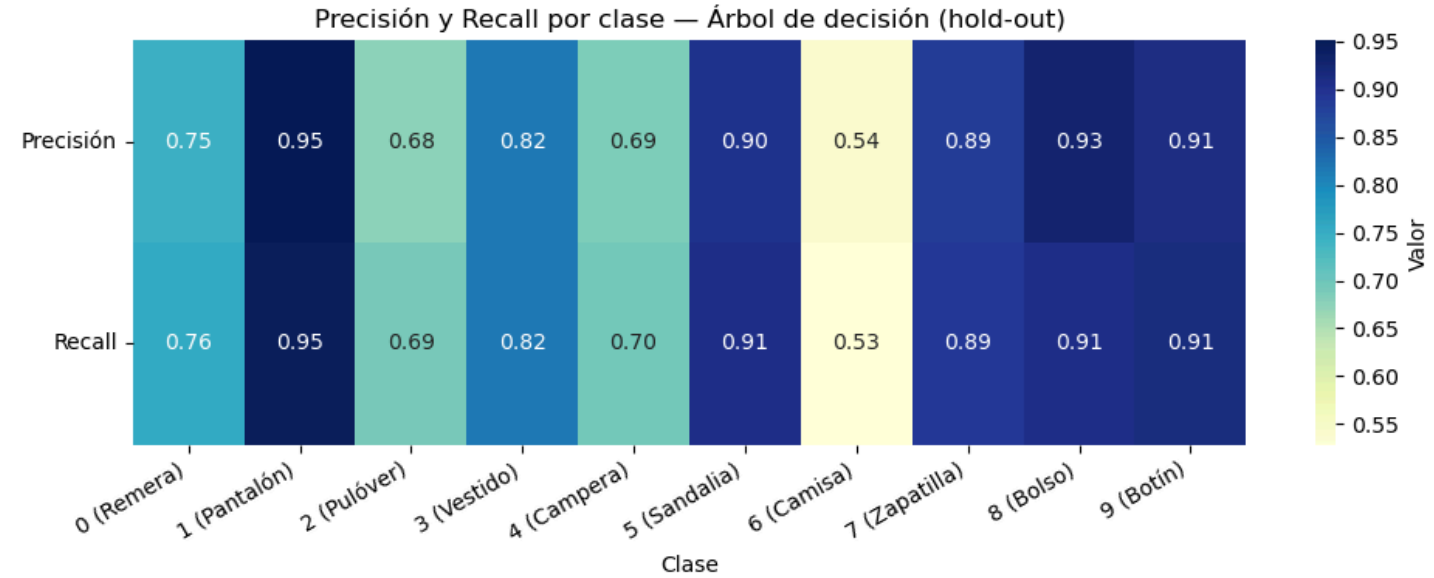
La mejor combinación de hiperparámetros fue: criterion="entropy", splitter="best", min_samples_leaf=1

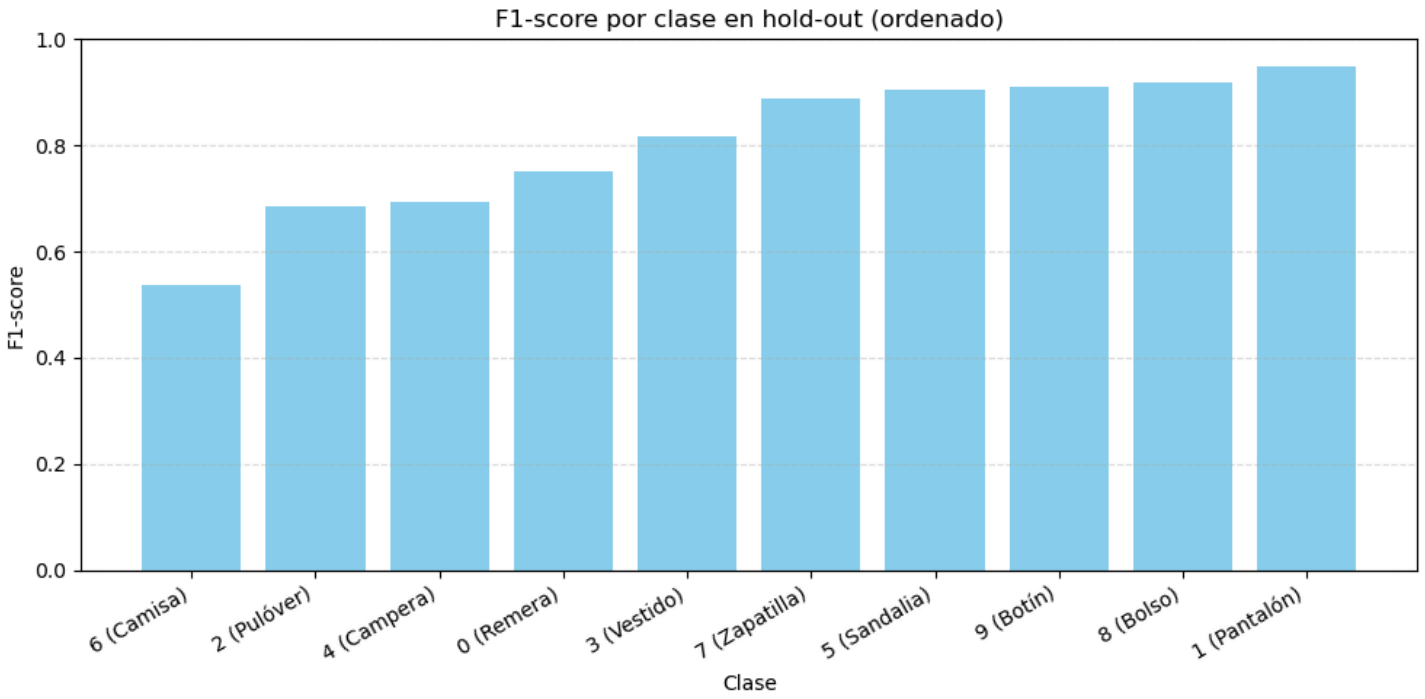
Evaluación final del modelo en el conjunto hold-out (multiclase)

El modelo de árbol de decisión con los mejores hiperparámetros obtenidos mediante validación cruzada (profundidad máxima = 9, criterion="entropy", splitter="best", min_samples_leaf=1) fue entrenado sobre el conjunto de desarrollo completo y luego evaluado sobre el conjunto hold-out. Logrando una exactitud del 82%.



La matriz de confusión normalizada muestra un rendimiento general aceptable, aunque con diferencias marcadas entre clases. Se destacan los buenos resultados para la clase 1 (pantalón), con un 95% de aciertos, así como para las clases 5 (sandalia), 8 (bolso) y 9 (botín), todas con una tasa de clasificación correcta superior al 90%. En cambio, la clase 6 (camisa) fue la más problemática, con un recall de solo 53%, siendo confundida principalmente con las clases 0 (remera), 2 (pulóver) y 4 (campera), que presentan características visuales similares.





El gráfico de F1-scores por clase confirma esta tendencia: las clases con formas bien definidas, como pantalones, bolsos o calzado, obtuvieron puntajes F1 superiores a 0.9, mientras que la clase 6 tuvo el desempeño más bajo, con un F1 cercano a 0.54. En términos generales, el modelo mostró una buena capacidad de predicción sobre clases fácilmente distinguibles, pero limitaciones al abordar clases con siluetas similares. Esto refleja una debilidad típica de los árboles de decisión en problemas de imágenes, donde la información espacial no está representada explícitamente. Los resultados obtenidos permiten identificar con claridad las clases más desafiantes, y sugieren que para mejorar su clasificación podrían ser necesarios modelos más expresivos o una selección de atributos que incorpore información espacial.

En conclusion,

- El modelo logró una exactitud del ~82% en el conjunto hold-out, con alto desempeño en clases como pantalón, bolso, botín y sandalia (F1 > 0.9).
- La clase 6 (camisa) fue la más difícil de clasificar, con un F1 ≈ 0.54, confundida con prendas de silueta similar.
- El árbol de decisión mostró limitaciones para capturar información espacial, lo que afectó la clasificación de clases visualmente parecidas.

Extra

Nuevamente dejamos una comparación preliminar con otros modelos para la clasificación multiclase en el anexo (Figura 9, Figura 10)

Conclusion

A lo largo de este trabajo se exploraron estrategias de clasificación tanto en el escenario binario (clases 0 vs 8) como en el multiclase (10 categorías) sobre el dataset Fashion-MNIST, combinando selección de atributos, ajuste de hiperparámetros y evaluación rigurosa.

Selección de atributos y reducción de dimensionalidad:

- Se compararon cuatro criterios (varianza, mutual information, esquinas y aleatorio), demostrando que con tan solo 30–60 píxeles es posible llegar a ≈ 97% de exactitud en la clasificación binaria, prácticamente igualando a modelos que usan la imagen completa.
- La selección aleatoria resultó sorprendentemente eficaz (Anexo, Fig. 9 y 10), probablemente gracias a la redundancia espacial del dataset y la diversidad de regiones cubiertas.

Clasificación binaria (clases 0 “remera” vs 8 “bolso”)

- El kNN (k=3) sobresalió con 96.86% de exactitud, F1 ≈ 0.968 y ROC AUC = 0.9835, siendo apenas superado en sensibilidad por el árbol de decisión.
- El árbol (d=5) ofreció el mejor recall (97.35%) y ROC AUC = 0.9868, con un balance muy parejo entre falsos positivos y negativos.
- La regresión logística, aun usando todos los píxeles, quedó un escalón por debajo (exactitud 95.39%, ROC AUC = 0.9843), evidenciando las ventajas de métodos no lineales o basados en proximidad en este contexto.

Clasificación multiclase (10 clases)

- Se ajustó un árbol de decisión variando max_depth (1–10), tanto en test fijo como en validación cruzada (5-fold). La exactitud creció rápidamente hasta depth = 6 y se estabilizó en ≈ 80% para depth ≥ 9.
- Mediante GridSearchCV se afinó criterion, splitter y min_samples_leaf, eligiéndose finalmente {criterion="entropy", splitter="best", min_samples_leaf=1}.
- Evaluado sobre el hold-out, este modelo alcanzó 80.49% de exactitud, con F1 > 0.90 en clases bien definidas (pantalón, sandalia, bolso, botín) y F1 ≈ 0.54 en la clase más desafiante (“camisa”), reflejando la dificultad para discriminar siluetas similares.

Análisis de errores y perspectivas

- Los heatmaps de precisión y recall por clase y las matrices de confusión (Fig. hld_out) revelan clases especialmente conflictivas (p. ej. 2 vs 4, 6 vs 0), sugiriendo que incorporar información espacial.
- Los resultados del Anexo (Fig. 9 y 10) reafirman que la elección de atributos puede tener un impacto comparable o superior al de la elección del modelo.

Comentario final:

Este trabajo demuestra que, en problemas de imágenes con alta redundancia, la reducción drástica de dimensionalidad mediante selección simple de píxeles mantiene un rendimiento casi óptimo, y que modelos sencillos (kNN, árboles) logran desempeños muy competitivos.

ANEXO

Figura 0 imágenes de cada clase

(volver)

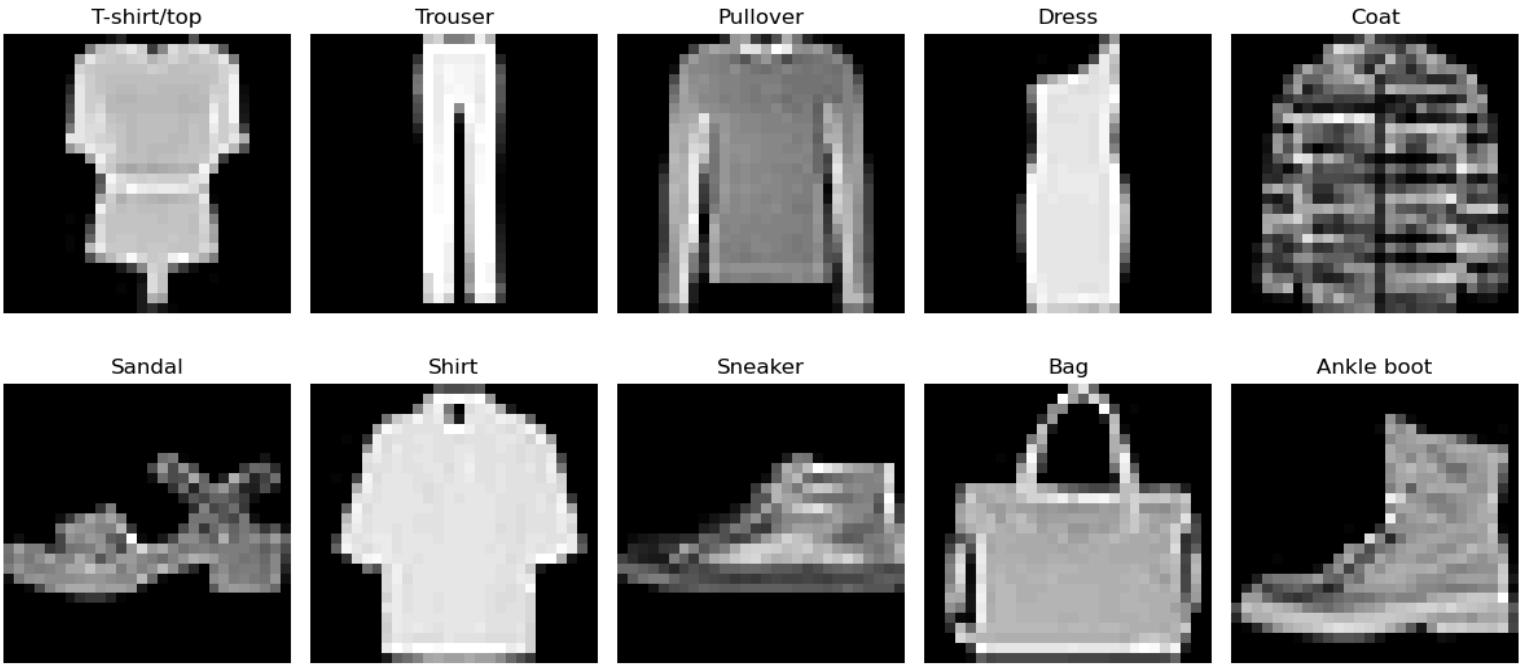


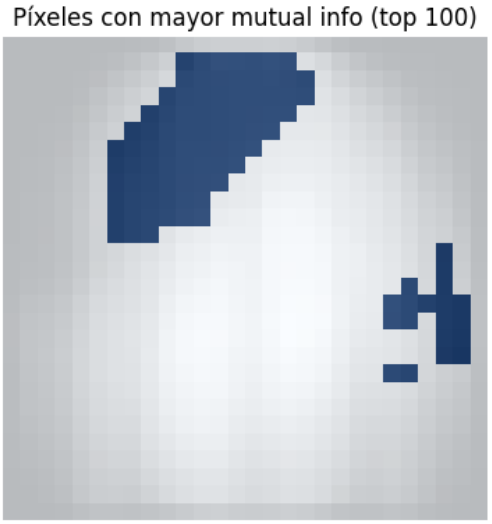
Figura 1 Pixeles más informativos según varianza

Descripción: Visualización de los 100 píxeles con mayor **varianza** a lo largo del dataset. Los píxeles seleccionados (en azul) se superponen sobre la imagen promedio, representando las regiones con mayor variabilidad en sus valores.

Utilidad: Este gráfico permite identificar las zonas de la imagen donde hay mayor variabilidad entre instancias, lo cual puede indicar regiones relevantes para diferenciar clases. Se utiliza como criterio de selección de atributos para reducir la dimensionalidad de los datos.

(volver)

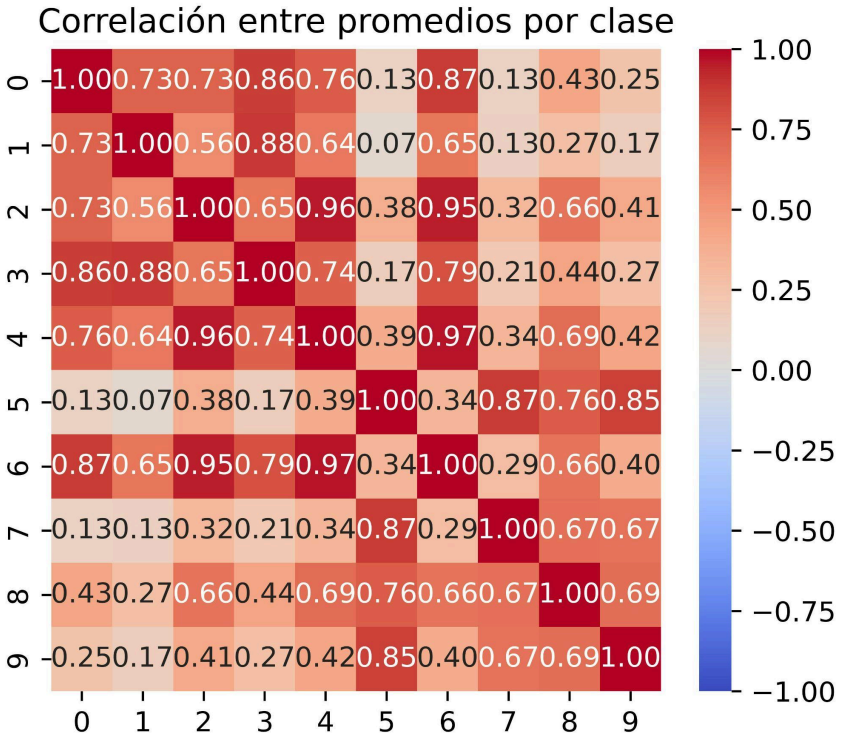
Figura 2 Pixeles más informativos según información mutua



Descripción: Muestra los 100 píxeles con mayor **información mutua** respecto a la clase, superpuestos en azul sobre la imagen promedio del dataset. La información mutua mide cuánta dependencia existe entre el valor del píxel y la clase objetivo.
Utilidad: Este criterio permite seleccionar los píxeles más útiles para la tarea de clasificación, incluso si no son los que más varían. El gráfico ayuda a visualizar qué regiones aportan más información para predecir correctamente la etiqueta de la imagen.

(volver)

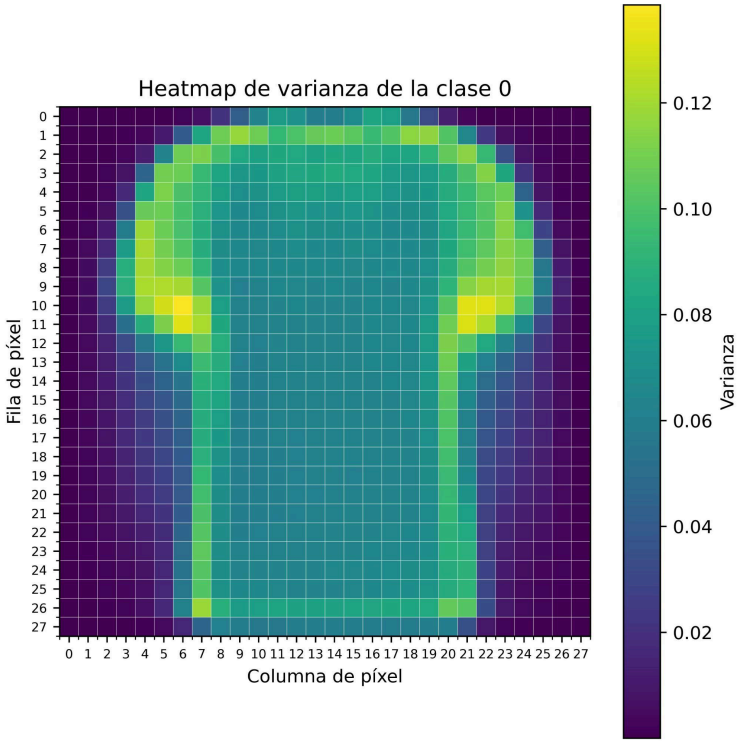
Figura 3 Correlación entre promedios por clase



Descripción: Muestra un mapa de calor con la **correlación de Pearson** entre los promedios de píxeles de cada clase. Cada celda representa el grado de similitud entre dos clases según la media de sus imágenes.
Utilidad: Este gráfico permite identificar qué clases tienen representaciones visuales similares o distintas. Por ejemplo, clases con alta correlación como 2 (pulóver) y 4 (campera) tienden a confundirse más fácilmente, mientras que pares con baja correlación (como 0 y 5) son más distinguibles para los modelos de clasificación.

(volver)

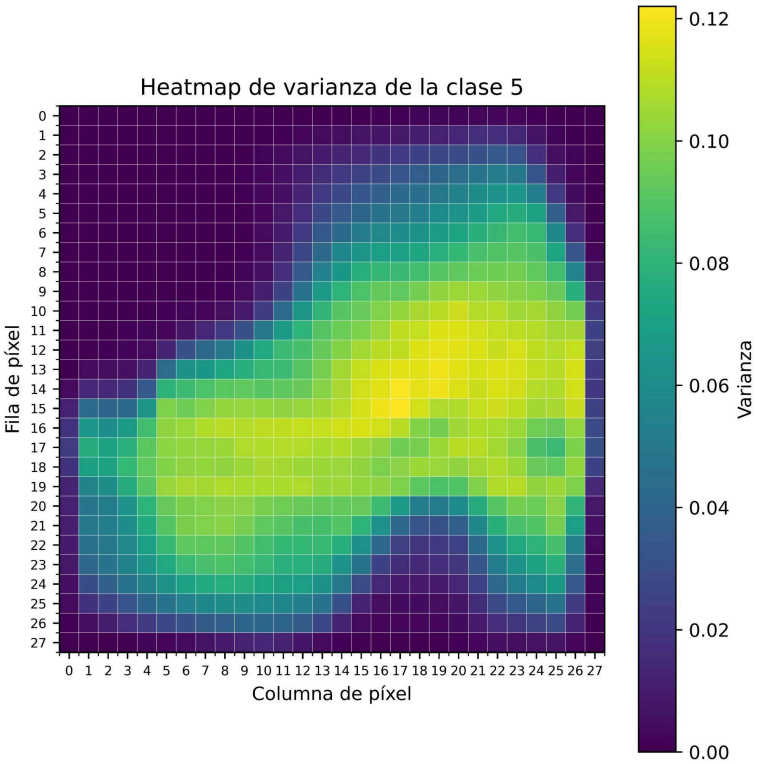
Figura 4 Heatmap de varianza de la clase 0



Descripción: Representa la varianza de cada píxel dentro del subconjunto de imágenes correspondientes a la **clase 0 (remera)**. Las regiones en amarillo indican mayor variabilidad entre instancias de esta clase, mientras que el azul oscuro representa baja variabilidad.

Utilidad: Este gráfico permite observar qué partes del objeto cambian más entre imágenes de la misma clase. En este caso, revela que la silueta de la remera es bastante estable, especialmente en el centro, lo que puede ayudar a explicar el buen rendimiento de los modelos sobre esta clase.
(volver)

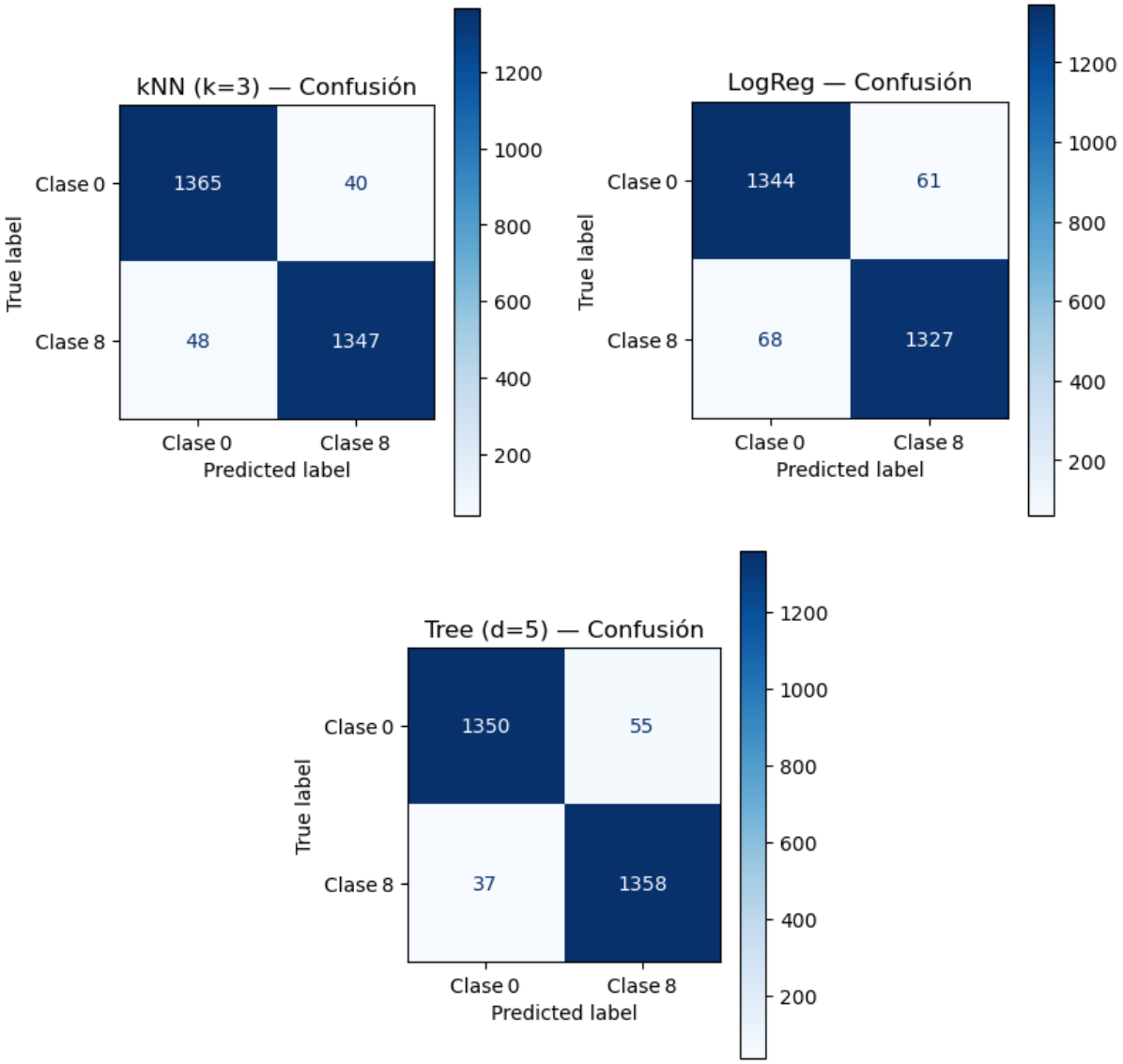
Figura 5 Heatmap de varianza de la clase 5



Descripción: Muestra la varianza de cada píxel en las imágenes de la **clase 5 (sandalia)**. Las zonas con colores más cálidos (amarillo) indican mayor dispersión entre ejemplos, mientras que las zonas oscuras son más consistentes.

Utilidad: Este gráfico evidencia que las sandalias presentan mayor variabilidad interna, con formas menos estables entre instancias. Esta dispersión puede explicar por qué modelos simples pueden tener más dificultades para clasificar correctamente esta clase en comparación con otras más homogéneas.
(volver)

Figura 6, 7 y 8 Comparación de modelos: kNN, Árbol de Decisión y Regresión Logística



Se evaluaron tres modelos sobre la tarea de clasificación binaria (clases 0 y 8) utilizando un subconjunto reducido de píxeles como atributos. Para cada modelo se calcularon métricas clave: **exactitud, precisión, recall, F1-score y ROC AUC**, además de analizar sus matrices de confusión.

Resultados comparativos

Modelo	Accuracy	Precision	Recall	F1-score	ROC AUC
kNN (k=3)	**0.9686**	0.9712	0.9656	**0.9684**	0.9835
Árbol (d=5)	0.9671	0.9611	**0.9735**	0.9672	**0.9868**
LogReg	0.9539	0.9561	0.9513	0.9536	0.9843

- El modelo **kNN** tuvo la **mejor exactitud general** (96.86%) y el mejor **F1-score**, lo que indica un **balance ideal entre precisión y recall**.

- El **árbol de decisión** fue el modelo con **mayor recall (97.35%)**: excelente para minimizar falsos negativos, a costa de una ligera caída en precisión.

- La **regresión logística** fue la más débil entre los tres, aunque sus métricas siguen siendo muy altas. Es probable que su naturaleza lineal haya limitado su capacidad de capturar relaciones más complejas en la distribución espacial de píxeles.

- En términos de **ROC AUC**, los tres modelos superaron el 0.98, indicando una muy buena capacidad de discriminación global.

Detalle por clase (según classification report)

- Tanto kNN como árbol predicen con **mayor precisión la clase 8** (la clase positiva), lo cual es deseable si el objetivo es detectar mejor esa categoría.

- LogReg tuvo un rendimiento parejo en ambas clases, pero inferior al de los otros dos modelos en cada métrica.

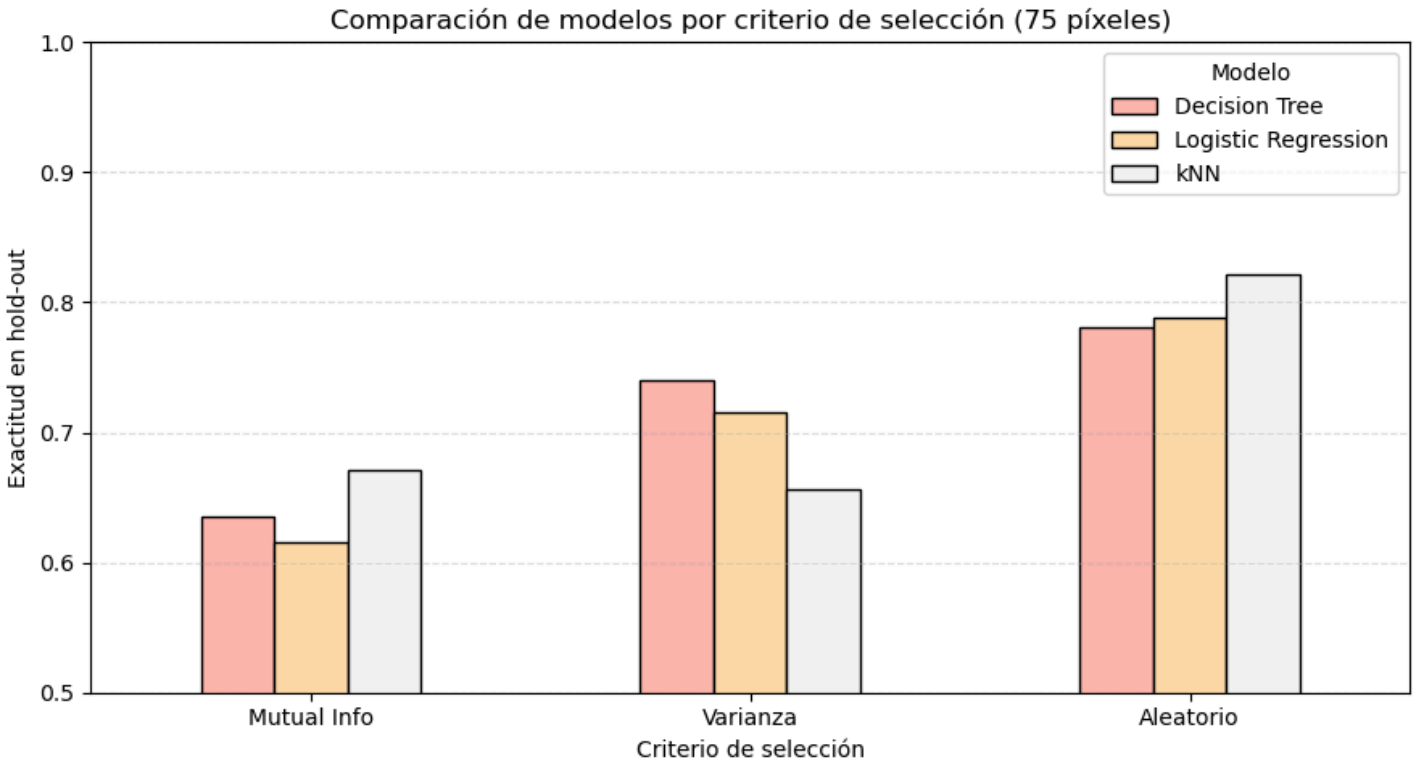
Conclusión general

Tanto **kNN como el árbol de decisión** demostraron ser modelos altamente eficaces en esta tarea binaria, logrando **muy buen rendimiento incluso con un subconjunto reducido de atributos**. El kNN se destaca por su simplicidad y balance entre métricas, mientras que el árbol resalta en sensibilidad (recall). La regresión logística, si bien sólida, fue superada por los otros enfoques.

Este análisis muestra que, incluso con pocos píxeles seleccionados, es posible construir modelos altamente precisos y balanceados para distinguir entre dos clases de imágenes, lo que **reduce el costo computacional** sin sacrificar performance.

(volver)

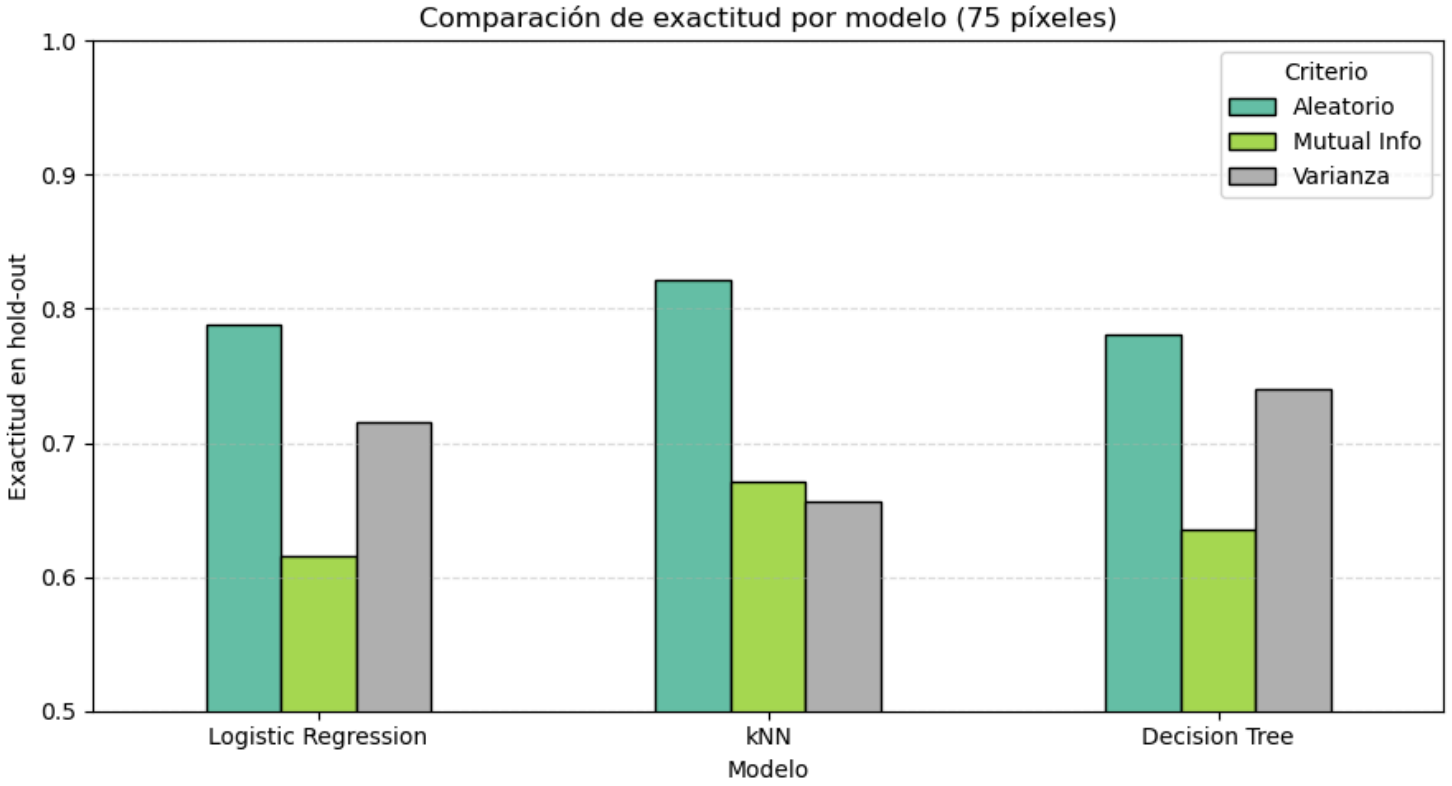
Figura 9 Comparación de exactitud por modelo (75 píxeles)



Descripción: Gráfico de barras que compara la exactitud obtenida por tres modelos de clasificación (Regresión Logística, kNN y Árbol de Decisión) utilizando 75 píxeles seleccionados bajo tres criterios distintos: aleatorio, mutual information y varianza.
Utilidad: Muestra que el criterio aleatorio fue el más efectivo para los tres modelos, especialmente en el caso de kNN, que alcanzó la mayor exactitud. Resalta que la selección estratégica de atributos puede impactar más que la elección del modelo.

(volver)

Figura 10 Comparación de modelos por criterio de selección (75 píxeles)



Descripción: Gráfico que invierte el enfoque: compara la exactitud obtenida con cada criterio de selección (mutual info, varianza, aleatorio), mostrando el desempeño de cada modelo dentro de cada grupo.
Utilidad: Refuerza que el criterio aleatorio fue el más robusto en general, con el kNN destacándose como el modelo más preciso dentro de cada criterio. Sirve para validar que, más allá del modelo, la elección de atributos tiene un peso clave en la performance.

(volver)