Laboratorio de Datos

Trabajo Práctico 2: Fashion-MNIST

Fecha: 16/6/2025

Nombre del grupo: Datos de Labo

Nombre de los miembros:

- **Denisse Britez**
- **Julieta Samosiuk**
- **Lautaro Alvarez Bertoya**

INTRODUCCIÓN

El presente trabajo práctico se centra en el análisis y modelado del conjunto de datos Fashion-MNIST, una base de imágenes en escala de grises de 28×28 píxeles que representan prendas de vestir. Cada imagen está asociada a una etiqueta numérica del 0 al 9, correspondiente a una de las diez clases de ropa predefinidas. Este conjunto de datos, ampliamente utilizado en tareas de clasificación de imágenes, resulta ideal para introducir técnicas de aprendizaje automático y análisis visual.

El dataset Fashion-MNIST contiene 70.000 imágenes, distribuidas en diez clases distintas de prendas, codificadas con valores numéricos del 0 al 9, según:

- Clase 0: Camiseta/Top
- Clase 1: Pantalón
- Clase 2: Pullover
- Clase 3: Vestido
- Clase 4: Abrigo
- Clase 5: Sandalia
- Clase 6: Camisa
- Clase 7: Zapatilla deportiva
- Clase 8: Bolsa
- Clase 9: Botín

El objetivo general del trabajo es explorar las características del dataset y aplicar modelos de clasificación supervisada para predecir a qué clase pertenece una imagen desconocida. Para ello, se divide el desarrollo en tres partes principales.

En la primera parte se realiza un análisis exploratorio de los datos, centrado en identificar qué atributos (píxeles) resultan informativos, qué clases son más parecidas entre sí y cuán homogéneas son las imágenes dentro de una misma clase.

La segunda parte aborda un problema de clasificación binaria, en el que se busca determinar si una imagen pertenece a la clase 0 (Camiseta/Top) o a la clase 8 (Bolsa). Se evalúan distintos modelos utilizando el algoritmo k-Nearest Neighbors (kNN), probando diferentes subconjuntos de atributos y valores de *k*, y comparando su desempeño mediante métricas adecuadas.

Finalmente, la tercera parte del trabajo consiste en un problema de clasificación multiclase, cuyo objetivo es predecir la clase correspondiente entre las diez posibles. Para ello, se entrenan modelos de árbol de decisión sobre un conjunto de desarrollo, y se selecciona la mejor configuración de hiperparámetros mediante validación cruzada. Finalmente, se evalúa el modelo elegido sobre un conjunto de validación no utilizado previamente.

Este informe presenta el desarrollo completo de las tres partes, incluyendo los análisis realizados, los gráficos más relevantes, los modelos aplicados y las conclusiones obtenidas.

Análisis Exploratorio

Antes de aplicar cualquier modelo de clasificación, es fundamental comprender la estructura interna del conjunto de datos. El dataset Fashion-MNIST contiene 70.000 imágenes en escala de grises de 28×28 píxeles, representando diez clases distintas de prendas codificadas con valores numéricos del 0 al 9. Esta sección tiene como objetivo identificar atributos relevantes, estudiar similitudes entre clases y analizar la variabilidad de las imágenes dentro de una misma categoría. Para ello, se utilizan herramientas visuales que

permiten explorar tanto el comportamiento general de los datos como las particularidades de cada clase.

Relevancia de los píxeles como atributos

Se analiza la activación promedio de cada píxel en el conjunto completo de imágenes, junto con su variabilidad. Esto permite identificar qué regiones del espacio visual son potencialmente útiles para la tarea de clasificación.

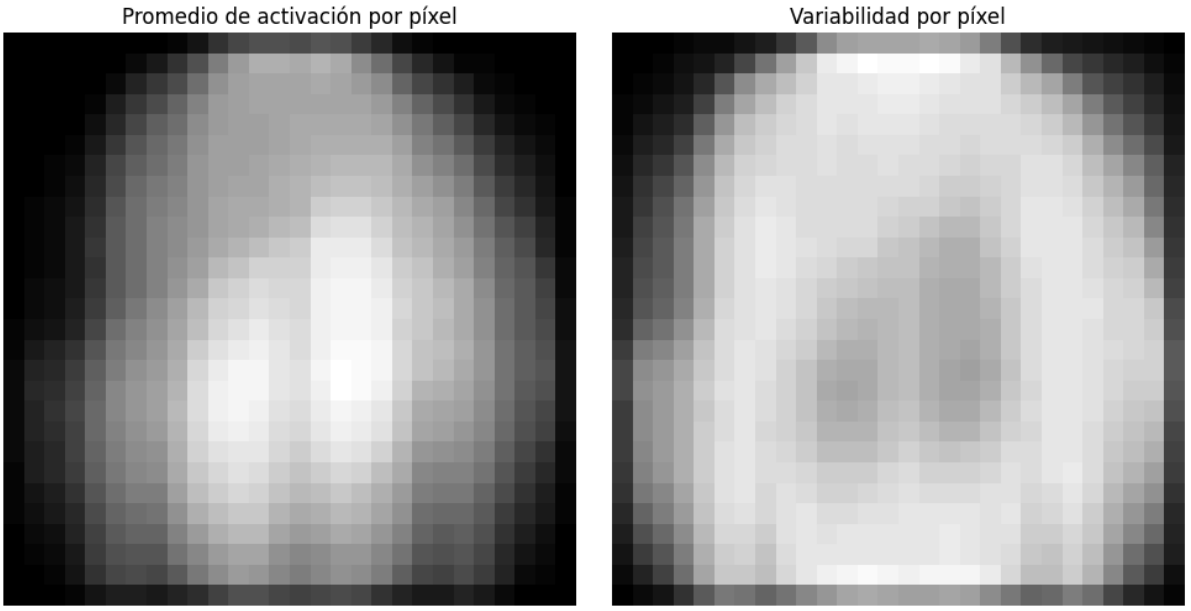

Figura 1. Promedio de activación y variabilidad por píxel en todo el dataset.

En el gráfico de activación promedio, se observa que los píxeles ubicados en la región central de las imágenes tienden a presentar valores más altos, lo cual coincide con la ubicación típica de las prendas dentro del recuadro de 28×28 píxeles. Los bordes, en cambio, permanecen en general oscuros, con valores bajos que reflejan la ausencia de contenido informativo en esas zonas.

El gráfico de variabilidad revela que los píxeles centrales no solo son los más activados, sino también los que presentan mayor variación entre imágenes. Esto indica que contienen diferencias relevantes entre clases, lo que los vuelve especialmente útiles para discriminar entre tipos de prendas. En contraste, los bordes muestran baja variabilidad, reforzando la idea de que aportan poca información para la clasificación.

Similitud entre clases

Para explorar qué clases del dataset presentan mayor o menor grado de similitud visual, se realiza una comparación entre ejemplos reales de tres categorías: Pullover (clase 2), Pantalón (clase 1) y Camisa (clase 6).

Figura 2. Comparación visual entre Pullover (clase 2), Pantalón (clase 1) y Camisa (clase 6).

Las imágenes de Pullovers y Camisas presentan una silueta ancha en la parte superior y una estructura más difusa en los bordes, con variaciones sutiles en el contorno, particularmente en la forma de las mangas o el cuello. Esta similitud puede dificultar la diferenciación automática entre ambas clases. Por otro lado, los Pantalones muestran una forma alargada y claramente simétrica, con separación entre piernas, lo cual los distingue con facilidad de las otras dos prendas. La comparación visual sugiere que, desde el punto de vista morfológico, las clases 2 y 6 son más propensas a la confusión entre sí que con la clase 1.

Variabilidad dentro de una clase

Además de comparar entre clases, resulta importante estudiar la coherencia visual dentro de una misma categoría. Como caso de estudio, se analiza la clase Bolsa (clase 8), observando distintos ejemplos para evaluar el grado de homogeneidad interna.

Figura 3. Variabilidad visual dentro de la clase 8 (Bolsa).

Las imágenes correspondientes a la clase Bolsa muestran una estructura general coherente: predominan las formas rectangulares u ovaladas centradas verticalmente en el marco de 28×28 píxeles. Algunas ilustran asas visibles, otras no; también hay diferencias en el tamaño y en el nivel de contraste. A pesar de estas variaciones, las siluetas resultan en general reconocibles como pertenecientes a la misma categoría. Esto sugiere una identidad visual suficientemente definida, aunque con cierta flexibilidad en los detalles.

Clasificación Binaria

El objetivo de esta sección es entrenar un sistema capaz de diferenciar entre dos clases específicas de prendas de vestir:

- Clase 0: Camiseta/Top
- Clase 8: Bolsa

En primera instancia, se seleccionan únicamente los casos pertenecientes a estas dos clases, con el fin de reducir la complejidad del conjunto de datos y facilitar el análisis posterior.

Análisis de atributos

A continuación se realiza un análisis de los atributos (píxeles) de las imágenes, con el objetivo de evaluar su relevancia para distinguir las clases 0 y 8. El propósito es identificar los píxeles que presentan diferencias significativas entre ambas clases y descartar aquellos que no aportan valor, optimizando así el rendimiento del modelo y reduciendo la dimensión del problema.

Este análisis, si bien similar al desarrollado en la sección de análisis exploratorio, se centra exclusivamente en las clases 0 y 8. Para ello, se calcula la intensidad promedio de cada píxel en ambas clases, y luego se realiza la diferencia entre dichos valores promedio.

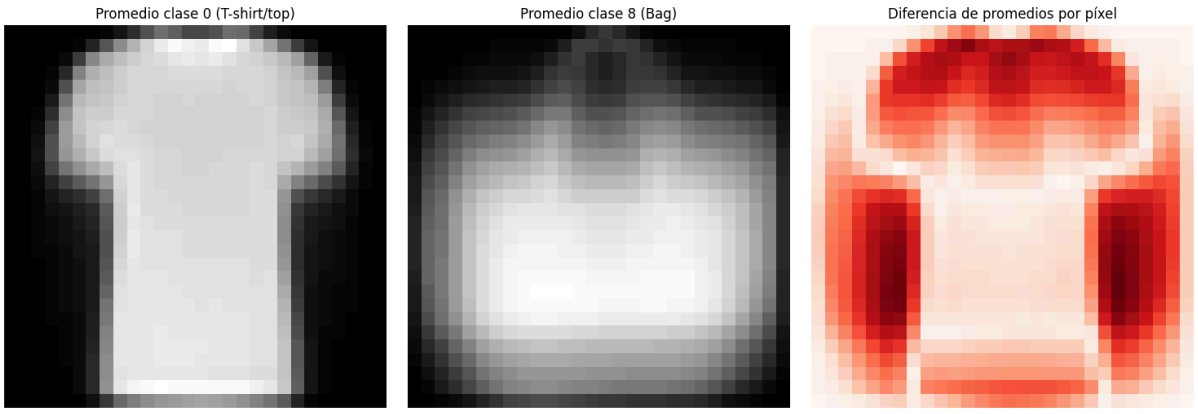

Figura 4. Diferencias promedio por píxel entre las clases 0 y 8.

El resultado se visualiza mediante un mapa de calor que resalta las regiones con mayor diferencia entre clases. Los píxeles con tonalidades más intensas de rojo indican las zonas más relevantes para la clasificación.

Selección de atributos relevantes

A partir del análisis anterior, se identifican los píxeles más relevantes mediante un criterio automático, sin intervención manual, y se seleccionan los primeros n, siendo n la cantidad de atributos con los que se desea trabajar. Esta selección tiene como objetivo evaluar si es posible entrenar un modelo eficaz utilizando únicamente un subconjunto reducido de atributos, tal como se plantea en el inciso C.

Esta estrategia apunta a reducir la complejidad computacional del modelo y a evitar el uso de información irrelevante o redundante durante el entrenamiento.

Evaluación de desempeño según grupos de píxeles

Se evaluaron distintos valores de atributos (específicamente 2, 4 y 6 píxeles seleccionados), y para cada caso se consideraron los tres primeros grupos de píxeles más relevantes, ordenados según la diferencia de promedios entre clases. En total, se analizaron 9 combinaciones distintas de atributos.

Cada conjunto fue utilizado para entrenar un modelo kNN con un valor fijo de k=10, y se calculó su desempeño mediante la métrica de exactitud (accuracy) sobre el conjunto de prueba.

Los resultados se presentan en un gráfico de barras, que permite comparar de forma visual y directa el rendimiento obtenido por cada grupo de atributos seleccionados.

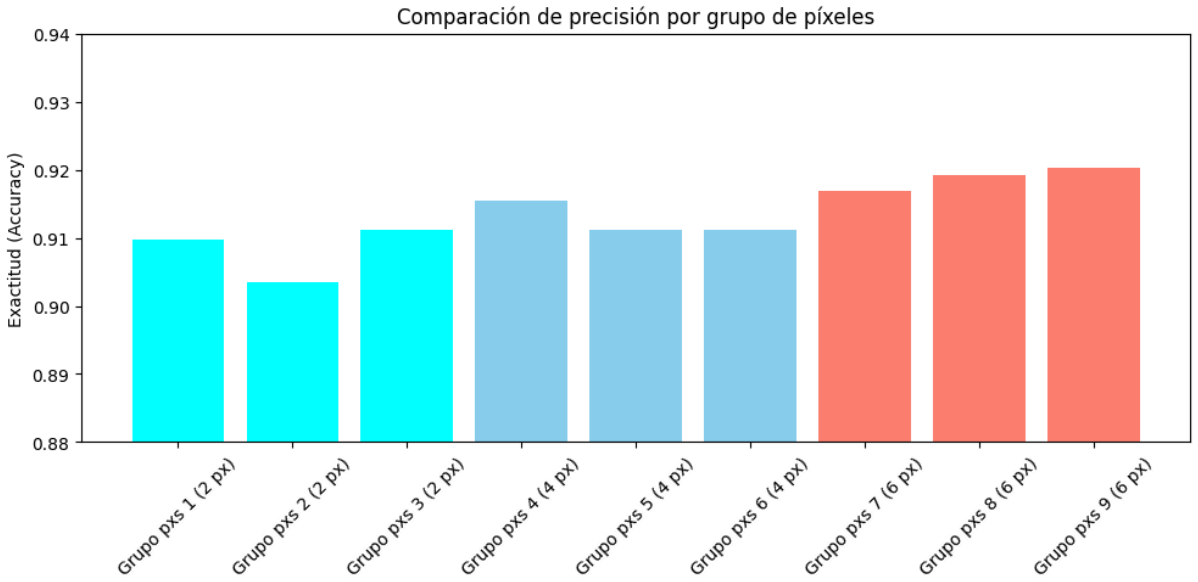

Figura 5. Precisión del modelo kNN (con k = 10) según diferentes grupos de píxeles.

A continuación, se calculó el promedio de exactitud para cada tamaño de subconjunto (2, 4 y 6 píxeles), con el objetivo de identificar cuál ofrece un mejor compromiso entre simplicidad del modelo y precisión. De acuerdo con estos promedios, se selecciona el grupo de 6 píxeles como la configuración más adecuada para el análisis posterior.

Comparación de precisión según el valor de k

Una vez determinada la mejor cantidad de atributos, se evaluó el desempeño del modelo kNN variando el valor de k entre 2 y 15. Para cada combinación de los atributos, se midió la

exactitud en los conjuntos de entrenamiento y prueba, con el fin de detectar posibles casos de sobreajuste (overfitting) o subajuste (underfitting).

Los resultados se presentan mediante gráficos de líneas que muestran por separado las curvas de exactitud para los datos de entrenamiento y test. Esto permite observar cómo varía el rendimiento del modelo según k, y facilita la identificación del valor que brinda la mejor capacidad de generalización.

Resultados Finales

Tras evaluar diversas combinaciones de atributos y valores del parámetro k, se identificó el grupo de píxeles que brinda el mejor desempeño promedio en términos de exactitud sobre los conjuntos de entrenamiento y prueba. El subconjunto óptimo resultó estar compuesto por los siguientes seis píxeles:

['pixel527', 'pixel555', 'pixel442', 'pixel454', 'pixel499', 'pixel481'].

Utilizando este conjunto, se construyó el gráfico final que muestra la evolución de la exactitud del modelo kNN al variar el número de vecinos considerados, desde k=2 hasta k=15.

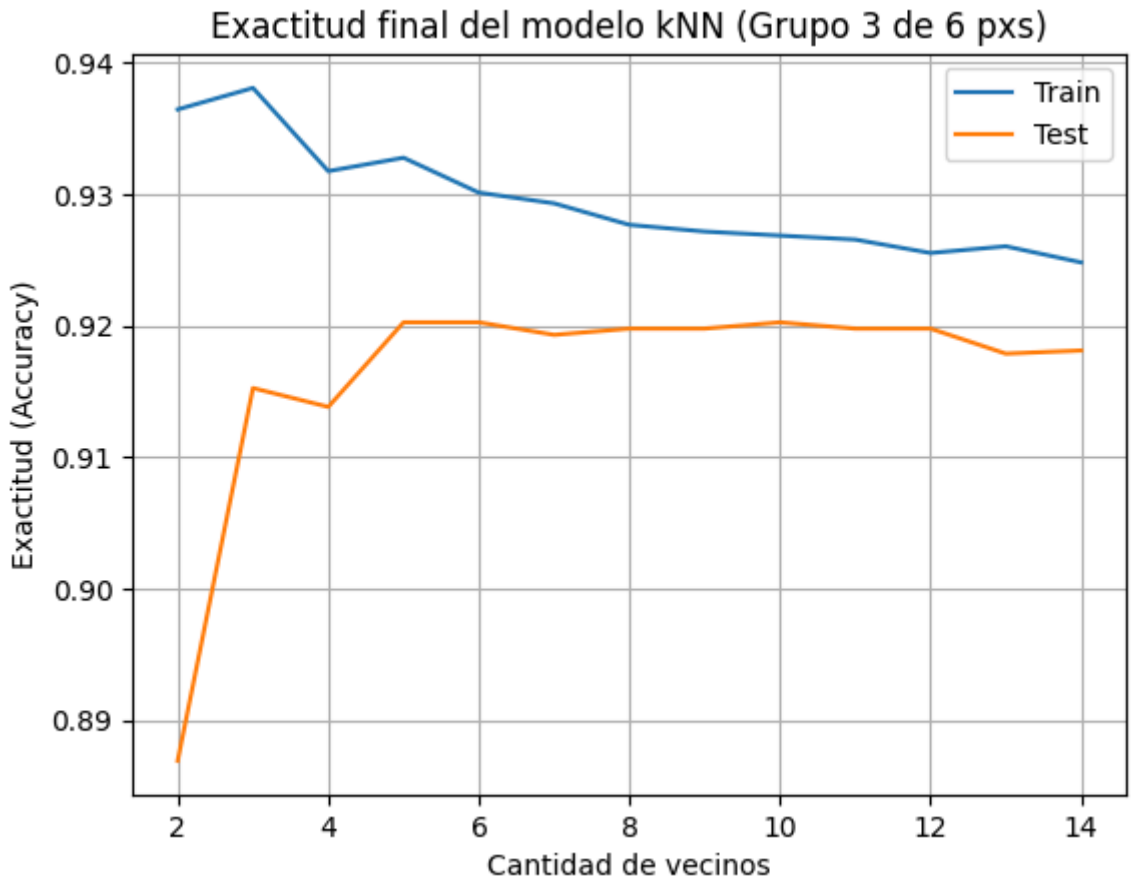

Figura 6. Precisión de kNN con 6 píxeles relevantes según valor de k.

Esta visualización nos permite analizar la relación entre la complejidad del modelo (controlada por k) y su capacidad de generalización. Los resultados muestran que, incluso trabajando con un número muy reducido de atributos (en este caso, solo seis píxeles seleccionados), es posible lograr un rendimiento competitivo en la clasificación de

imágenes, siempre que se elijan cuidadosamente los atributos y se ajuste adecuadamente el parámetro del modelo.

Aclaración: Estos resultados corresponden al conjunto de datos obtenido tras una partición aleatoria en entrenamiento y testeo. Como el muestreo tiene un componente aleatorio, es esperable que, con otra partición distinta, cambian tanto los píxeles seleccionados como los valores de accuracy obtenidos. No obstante, el análisis metodológico permanece válido.

Clasificación Multiclase

Ajuste del modelo con distintas profundidades

Se entrenaron distintos modelos de árboles de decisión variando la profundidad máxima entre 1 y 10, con el objetivo de analizar cómo este hiperparámetro afecta el rendimiento. Podemos notar que, a medida que se incrementa la profundidad, el árbol gana capacidad para aprender patrones del conjunto de entrenamiento. Esto se refleja en una mejora progresiva en la exactitud (accuracy) sobre el conjunto de desarrollo (dev).

Con profundidad 1, el modelo es extremadamente simple y su rendimiento es muy bajo (exactitud del 19.93%). Esto indica que es un caso de subajuste. A partir de profundidades intermedias (entre 5 y 7), el modelo comienza a capturar patrones más relevantes y su desempeño mejora considerablemente. Finalmente, al alcanzar profundidad 10, el modelo logra una exactitud del 84.81%.

Es importante notar que un aumento tan marcado en la profundidad puede favorecer el sobreajuste, ya que el árbol podría comenzar a memorizar patrones específicos del conjunto de desarrollo, en lugar de generalizar. Por esto, es necesario complementar este análisis con un procedimiento de validación cruzada, como se detalla en el siguiente inciso.

Selección de modelo mediante validación cruzada

Para evaluar la capacidad de generalización del modelo, se aplicó validación cruzada con 5 particiones sobre el conjunto de desarrollo. Se probaron las profundidades entre 1 y 10 y se registró la exactitud promedio. Obtuvimos como mejor resultado una profundidad 10, con una exactitud promedio del 80.57%. Aunque este valor es ligeramente menor a la exactitud de entrenamiento, indica una buena generalización, con bajo riesgo de sobreajuste. Finalmente, este método permitió una selección más confiable que basarse únicamente en los datos de entrenamiento.

Evaluación final en conjunto held-out

Una vez definida la mejor profundidad del árbol, se entrenó el modelo final utilizando todo el conjunto de desarrollo. Luego, se evaluó su desempeño sobre el conjunto held-out, formado por datos que el modelo no vio previamente. En esta evaluación, se obtuvo una exactitud del 80.50%, lo cual confirma que la configuración elegida generaliza correctamente, sin evidencias claras de sobreajuste.

Para comprender mejor el comportamiento del modelo, analizamos la matriz de confusión.

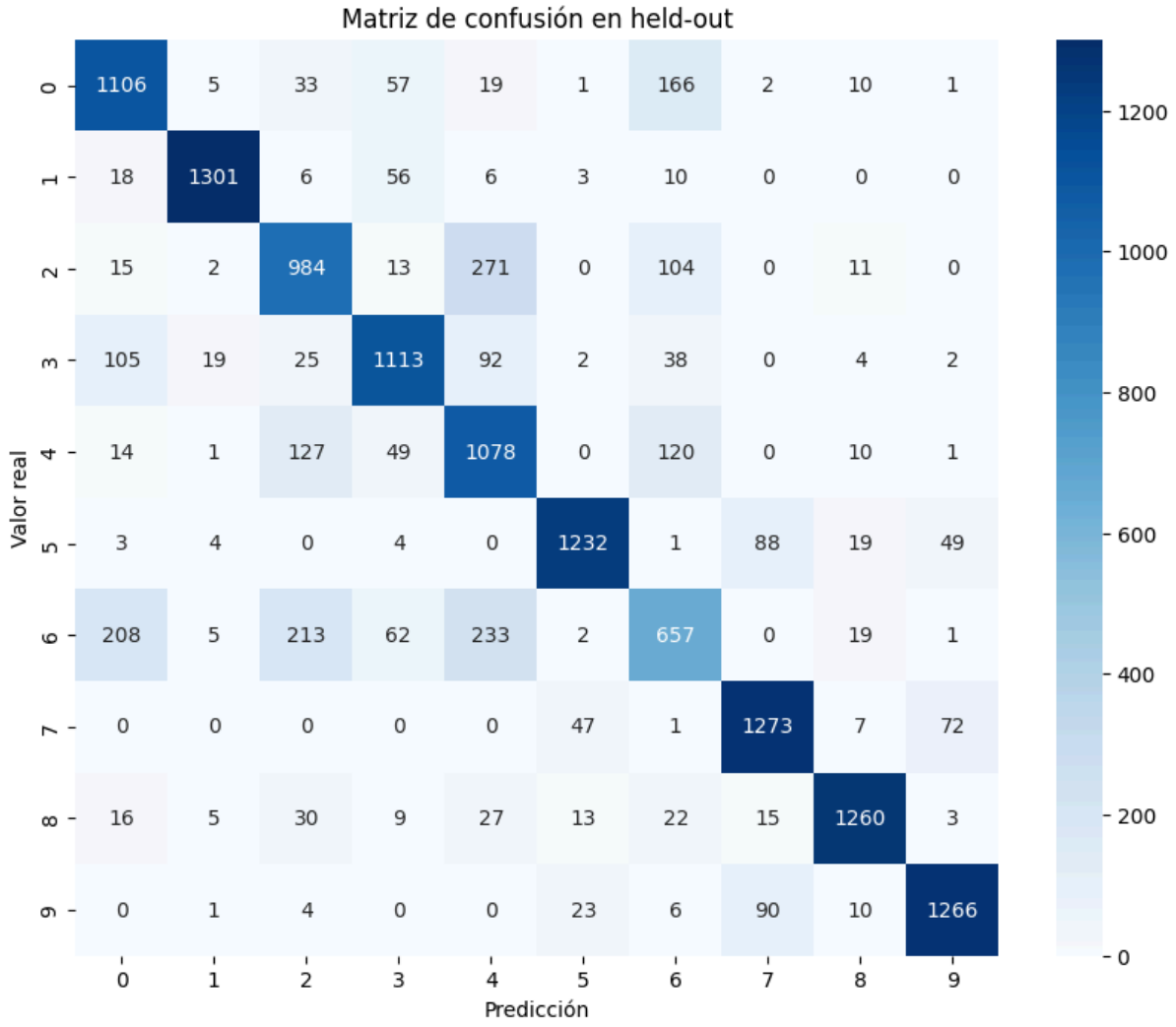

Figura 7. Matriz de confusión del modelo final evaluado sobre el conjunto held-out. Las clases más oscuras en la diagonal indican predicciones correctas; las celdas fuera de la diagonal reflejan confusiones frecuentes.

Podemos observar que, para las siguientes clases, el modelo logró identificar correctamente la mayoría de las instancias, con poco confusión respecto a otras clases:

- Clase 1: Pantalón
- Clase 5: Sandalia
- Clase 7: Zapatilla deportiva
- Clase 8: Bolsa
- Clase 9: Botín

Por otro lado, se observaron dificultades particulares en las siguientes clases:

- Clase 2 (Pullover) se confunde especialmente con la clase 4 (Abrigo), y en menor medida, con la clase 6 (Camisa).
- Clase 6 (Camisa) presenta confusiones notables con la clase 0 (Camiseta/Top), clase 2 (Pullover) y clase 4 (Abrigo).

- También se observa que la clase 0 (Camiseta/Top) tiene cierta confusión con la clase 6 (Camisa), y lo mismo ocurre entre la clase 4 (Abrigo) y la 6 (Camisa), aunque en menor medida.

Notamos que estas confusiones se dan principalmente entre prendas visualmente similares. Esto es comprensible si consideramos que los árboles de decisión tratan cada píxel como una característica independiente, sin captar patrones espaciales.

CONCLUSIÓN

A lo largo del trabajo se exploró el conjunto de datos Fashion-MNIST mediante un enfoque integral que combinó análisis exploratorio y modelado supervisado.

En la fase de análisis visual se identificaron patrones de activación y variabilidad de píxeles que permitieron entender cuáles regiones de la imagen son más relevantes para la clasificación. Este análisis guió una selección de atributos significativa, permitiendo reducir la dimensionalidad sin pérdida importante de rendimiento.

En la clasificación binaria entre camisetas/tops (clase 0) y bolsas (clase 8), se comprobó que incluso utilizando solo 6 píxeles seleccionados adecuadamente, el modelo k-Nearest Neighbors logró una precisión destacable considerando la baja dimensionalidad. Además, se observó cómo el valor del parámetro k influye en el comportamiento del modelo, destacando la importancia del ajuste fino de hiperparámetros.

Para el caso multiclase, se aplicaron árboles de decisión con distintas profundidades. Se realizó validación cruzada para prevenir el sobreajuste, y se determinó que una profundidad de 10 ofrecía el mejor balance entre complejidad y generalización. El modelo final alcanzó una precisión del 80.50% sobre un conjunto de test no visto, lo cual valida la eficacia de la metodología.

Las confusiones principales se dieron entre clases visualmente similares (por ejemplo, Pullover, Camisa y Abrigo), lo que indica que una representación puramente basada en píxeles puede limitar la capacidad del modelo para distinguir clases similares. Esto es comprensible si consideramos que los árboles de decisión tratan cada píxel como una característica independiente, sin captar patrones espaciales.

En conclusión, el trabajo muestra que es posible construir modelos efectivos para clasificación de imágenes utilizando tanto técnicas simples como herramientas de validación rigurosas. Además, se destaca la importancia del análisis exploratorio para guiar decisiones de modelado y reducir la complejidad computacional sin comprometer la calidad del resultado.