DEPARTAMENTO DE COMPUTACION

Facultad de Ciencias Exactas y Naturales - UBA

Moderlo de KNN

13 de noviembre de 2023 materia

Grupo

Integrante	LU	Correo electrónico
Pavez Cayupel, Richard Pavez	1019/22	r3ch1rd2013@gmail.com
Munho Vital, Facundo Nicolás	151/21	facundomunho@gmail.com
Caceres Blanco, Juan Manuel	273/14	caceres.blancojm@hotmail.com

Facultad de Ciencias Exactas y Naturales
Universidad de Buenos Aires
Ciudad Universitaria - (Pabellón I/Planta Baja)
Intendente Güiraldes 2610 - C1428EGA
Ciudad Autónoma de Buenos Aires - Rep. Argentina
Tel/Fax: (++54 +11) 4576-3300
http://www.exactas.uba.ar

Resumen

En este trabajo exploramos el dataset Fashion MNIST y aplicamos en él los modelos de aprendizaje automático KNN y Árboles de decisión. Usamos el modelo KNN para predecir si una una nueva prenda es una remera o un pantalón y usamos Árboles de decisión para predecir de qué tipo es una nueva prenda. Para esto usamos 3 tratamientos distintos sombre los datos de entrenamiento y evaluamos su rendimiento usando distintos hiperparámetros, Modelo píxeles relevantes Modelo correlación y Modelo distancias arquetípicas.

1. Introducción

En análisis de los datos, unos de los métodos para poder entender y predecir el comportamiento de los objetos de estudio es la clasificación de los datos. Por ello vamos a explorar la efectividad de algunas técnicas computacionales, problemas y soluciones asociados a estos métodos. Para ello vamos a utilizar el dataset Fashion MNIST [1] extraído de Kagle, el cual cuenta con 60000 imágenes de 28x28 píxeles, compuestas por 10 tipos de prendas con 6000 imágenes cada una. El formato del dataset se presenta como una tabla, con 785 atributos, donde el primero corresponde al atributo "**label**"que referencia el tipo de prenda con un entero del 0 al 9, y los 784 atributos restantes son los píxeles que conforman las imágenes, en la figura 1 se observa un ejemplo de cada prenda y en la tabla 1 las primeras 5 filas de la tabla con los primeros y últimos atributos. Además en la tabla 2 se presenta la equivalencia entre los números del atributo "*labelz* el tipo de prenda que representa.

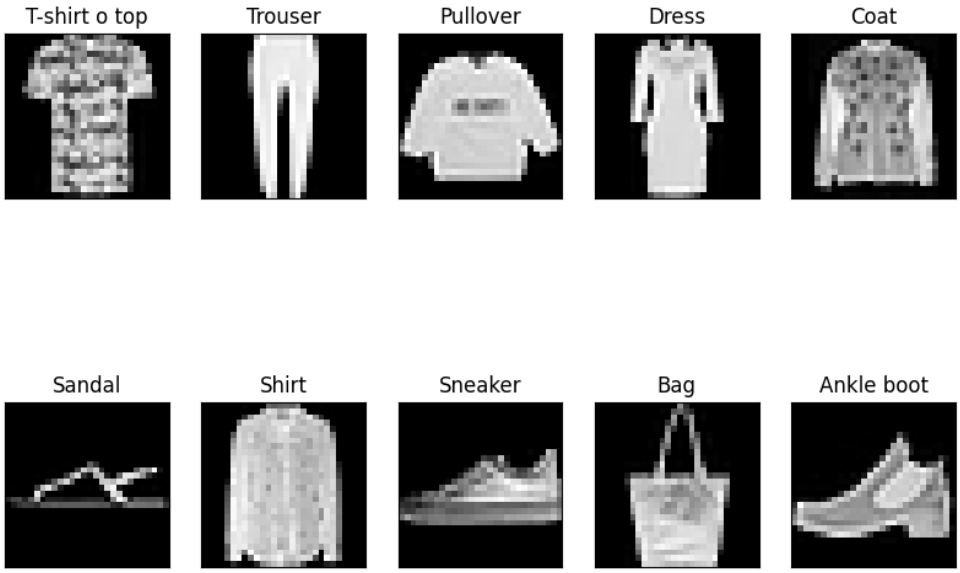


Figura 1. Imágenes de ejemplo del dataset Fashion MNIST [1]

Index	label	pixel1	pixel2	pixel3	⋯	pixel781	pixel782	pixel783	pixel784
0	2	0	0	0	⋯	0	0	0	0
1	9	0	0	0	⋯	0	0	0	0
2	6	0	0	0	⋯	0	0	0	0
3	0	0	0	0	⋯	0	0	0	0
4	3	0	0	0	⋯	0	0	0	0

Tabla 1. Primeras 4 filas del dataset Fashion MNIST [1]

Codigo	0	1	2	3	4	5	6	7	8	9
Tipo De Prenda	T-shirt o top	Trouser	Pullover	Dress	Coat	Sandal	Shirt	Sneaker	Bag	Ankle boot

Tabla 2. Tabla de equivalencias entre el codigo de las prendas del dataset Fashion MNIST [1]y el tipo de prenda que le corresponde segun Kagle

En particular, este informe evalúa como método de clasificación a las técnicas de *k vecinos más cercanos* (KNN) [2] y arboles de decisión [3], para ambos métodos utilizamos la implementación OS proveída por **Scikit-learn** [4],[5]. Es necesario para la lectura del proceso, tener a mano algunas definiciones estadísticas como lo son el promedio/media (1) de los valor de un conjunto, la mediana, que es el valor que separa a la mitad los valores de un conjunto y la desviación estándar. (2)

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{1}$$

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{2}$$

Donde n es el número de elementos en el conjunto de datos, x_i representa cada uno de los elementos en el conjunto de datos, desde $i = 1$ hasta $i = n$, $\bar{x}$ es el promedio o media de los datos y σ es la desviación estándar del conjunto.

También es de utilidad para este informe recordar algunos aspectos del álgebra lineal. Algo que podemos hacer con las imágenes presentadas, es pensarlas como vectores de 784 posiciones de números naturales entre el 0 y el 255. Este conjunto de todas las posibles imágenes generan un espacio vectorial, es más es fácil demostrar que este espacio es en realidad normado con la norma euclídea o la norma 2 (3), ya que en realidad es un subespecio de $R^7 84$. Luego las imágenes que son muy parecidas entre si estarán muy cerca entre si bajo la norma de dicho espacio vectorial, por ejemplo las imágenes de una misma clase, esto nos permitirá armar para cada clase una region que encierra a todas las imágenes que pertenecen esa clase, esto centrado en una prenda, la prenda promedio de la clase. Esta prenda sera llamada "*arquetipo de*z se pueden ver en la figura del apéndice **??**

$$|\vec{x} - \vec{y}| = \sqrt{\sum_{i=1}^{784}(x_i - y_i)^2} \tag{3}$$

2. Exploración y limpieza de datos

2.1. Exploración de datos

Para el armado de los modelos de clasificación por aprendizaje supervisado como lo son KNN o arboles de decisión, es importante el entendimiento del comportamiento de los datos con los que entrenaremos los modelos. Como ya se vio en la introducción 1 la estructura e información se corresponden a la descripta en la pagina web de Kaggle: Fashion MNIST [1]. Por lo que en esta sección nos centraremos en entender como es la forma de los datos. Para ello lo primero que se realizo, fue un estudio del comportamiento del conjunto de imágenes de cada una de las clases de prendas. En la imagen 2 podemos observar el promedio (1), la mediana, la desviación estándar (2) y el máximo valor en cada píxel de la clase remera, estas características para la clase pantalón en la figura 3, y en la figura 4 para la clase vestido, el resto de las prendas se encuentran en el apéndice en la figura 13. De observar estas imágenes podemos extraer algunas conclusiones, las mayoría de las prendas no utilizan los píxeles de los bordes, pues el promedio y la moda de la iluminación de los píxeles suele ser baja y la desviación de iluminación de estos píxeles también tiende a ser baja. Por otro lado, las clases camisa("shirt"), camperas(çoat") pullovers son clases donde sus promedios y medianas son muy parecidas entre si, y que salvo por las mangas en la mayoría de los casos, son muy parecidos a las remeras("T-shirt o top"), por otro lado el resto de las clases son muy distintas en media y moda, permitiendo diferenciarlas con mayor facilidad entre sí que con entre las clases anteriormente nombradas. Otra conclusión interesante es ver como todas las prendas, salvo por las sandalias("sandals") y los bolsos("bag"), pertenecen a clases donde el promedio y la moda están muy bien definidos, por lo tanto las prendas dentro de dichas clases son parecidas entre si.

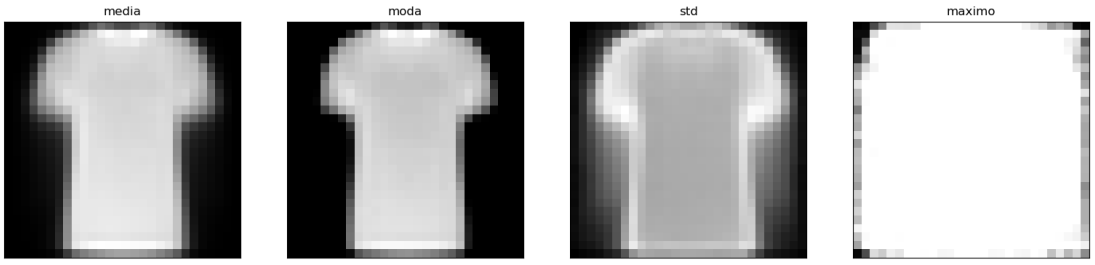


Figura 2. Medidas de resumen remera. De izquierda a derecha promedio, media, desviación estándar y máximo de cada píxel

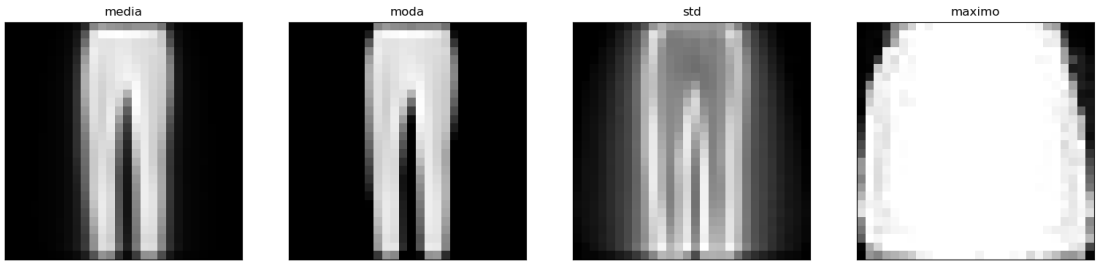


Figura 3. Medidas de resumen pantalón. De izquierda a derecha promedio, media, desviación estándar y máximo de cada píxel

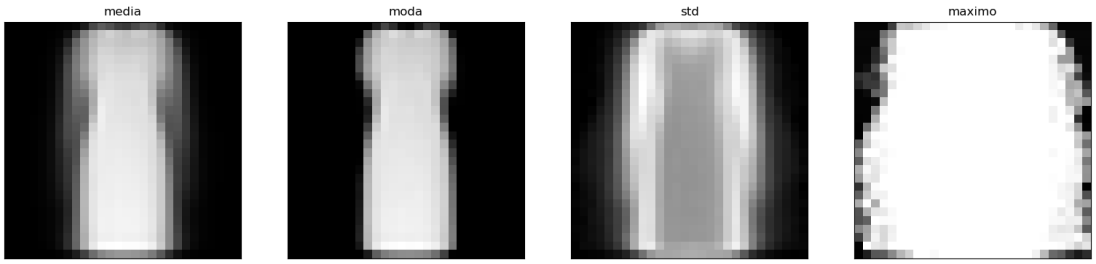
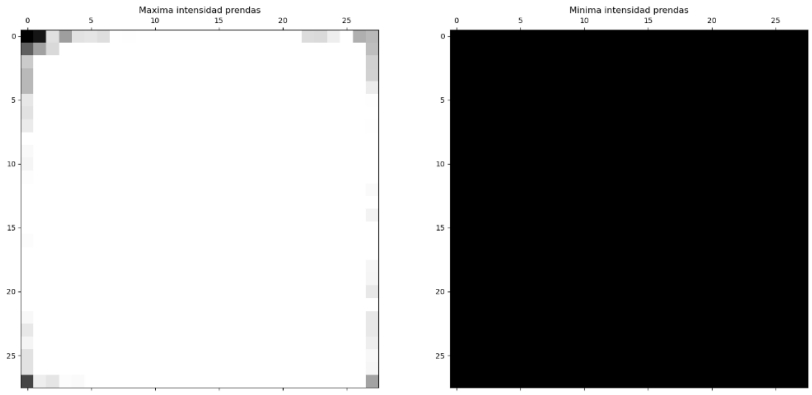
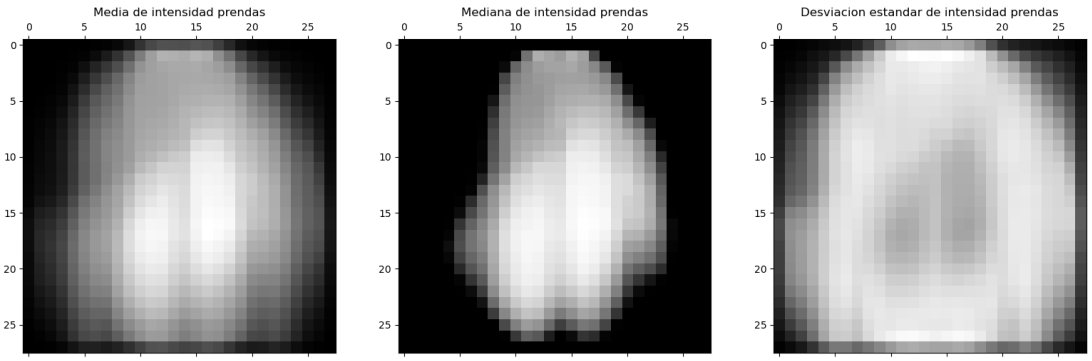


Figura 4. Medidas de resumen vestido. De izquierda a derecha promedio, media, desviación estándar y máximo de cada píxel

Otro aspecto interesante del set de datos es observar como es la distribución de la iluminación de cada píxel en general. Para ello calculamos las medidas resumen de las prendas anteriormente mencionadas, pero para todo el set de datos en conjunto, que se pueden observar en la figura 5. En 5(a) podemos observar como casi todos los píxeles alguna vez obtuvieron el valor máximo de iluminación, con la excepción de aquellos que están en los bordes, asi como en la figura de los mínimos todos en algún momento estuvieron apagados completamente. Por otro lado, en la figura 5(b), podemos observar como en promedio y en media las prendas tienen a estar distribuidas por el centro de la imagen, con una ligera forma de pera. Además, los píxeles centrales de las prendas suelen ser píxeles que no poseen mucha variabilidad, es decir tienen una desviación estándar baja, mientras que los píxeles que mas varían suelen ser los que están entorno al borde de la pera. Por otro lado, podemos ver como los bordes de las imágenes siguen siendo píxeles que tienen en promedio y moda poca iluminación y poca variabilidad de esta, y por lo tanto poca información. Sin embargo, para los métodos planteados en este informe y el dataset en cuestión no fue necesario reducir la cantidad de píxeles utilizados para que los algoritmos se ejecutaran en un tiempo razonable.



(a) A la izquierda el máximo de intensidad de cada píxel alcanzado alguna vez y a la derecha el mínimo.



(b) De izquierda a derecha promedio, media y desviación estándar de cada píxel.

Figura 5. Medidas de resumen generales

Aunque no se tomara la posibilidad de tratar los datos de la manera descripta recientemente, si se realizo un tratamiento de a las imágenes. Lo primero es la separación de los datos para el correcto entrenamiento, testeo y validación de los modelos, que realizamos de la siguiente manera: Un 10 % de los datos para la validación de los modelos, del 90 % restante, se separo un 30 %(27 % del total) para el testeo de los modelos y el 70 %(63 % del total) final para el entrenamiento de los modelos. En las siguientes subsecciones se explicaran como se seleccionaron los datos para cada modelo.

2.2. Selección de datos

2.2.1. Modelo píxeles relavantes

El método para seleccionar píxeles en este modelo varía un poco si es utilizado para diferenciar remeras de pantalones o si se utiliza para diferenciar entre las 10 clases de prendas, pero es en esencia el mismo.

Para el primero restamos la media de las remeras y la media de los pantalones y le tomamos valor absoluto (6). Podemos decir que los mayores valores son los píxeles que hacen la mayor diferencia entre una prenda y la otra, pues suelen estar iluminados en las remeras pero no en los pantalones, por lo que elegimos los 11 de mayor valor que correspondían en la remera media a regiones de axila y ombligo (píxeles que en el pantalón medio tienen un valor muy bajo). Entonces para los datos de train y test del dataframe que contiene las remeras y pantalones nos quedamos con 2 atributos que son la suma de los píxeles de axila y la suma de los píxeles de ombligo.

Para diferenciar entre los 10 tipos de prendas seleccionamos los píxeles de relevancia que obtenemos al analizar la resta entre la media de una clase en particular y el promedio de la media de las demás clases (7). Llamamos píxeles relevantes a los de mayor valor, pues muestran píxeles que aparecen más en una clase en particular y no tanto en las demás, y llamamos píxeles no relevantes a los de menor valor pues nos muestras píxeles que aparecen más en otras clases y no tanto en la clase que analizamos. Entonces con los píxeles seleccionados transformamos los datos de train y test en 10 atributos, correspondientes a cada una de las clases, que suman los píxeles relevantes para cada clase y restan los no relevantes.

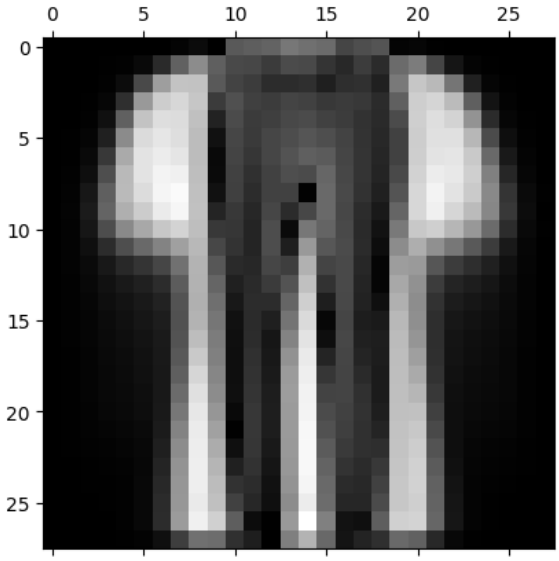
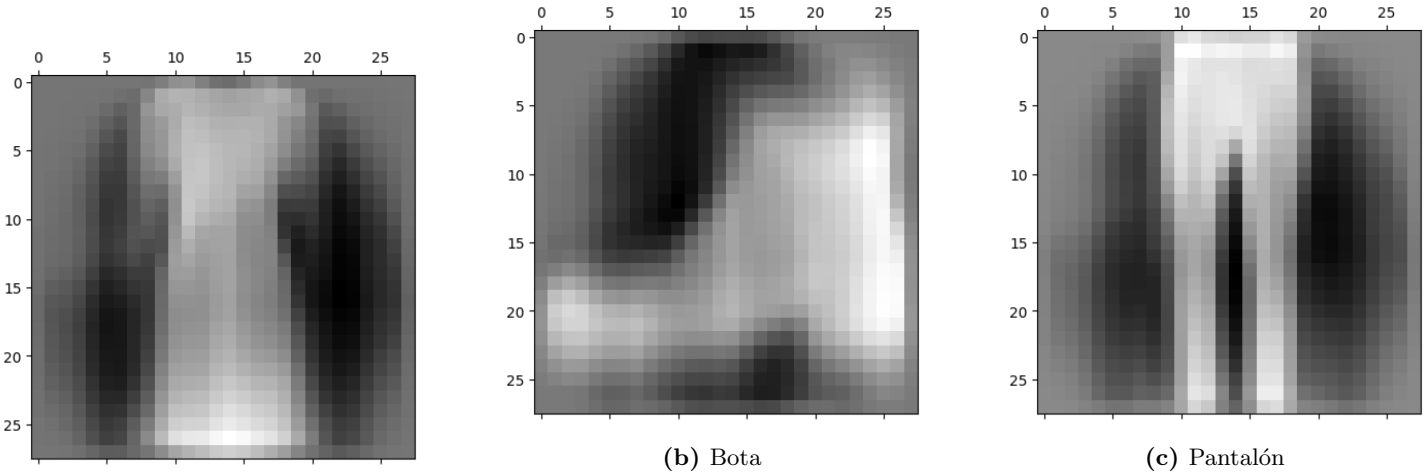


Figura 6. Diferencia absoluta entre la media de pantalones y la media de remeras



(a) Vestido

(b) Bota

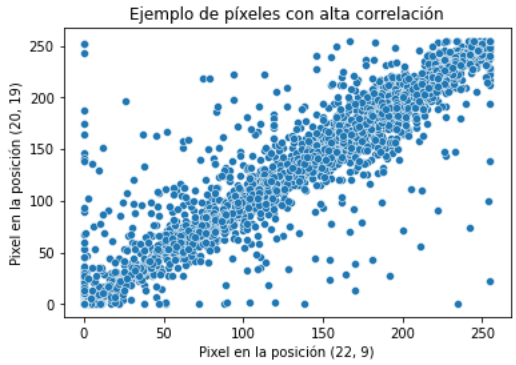
(c) Pantalón

Figura 7. Ejemplos de la media de una clase restada con el promedio de las demás

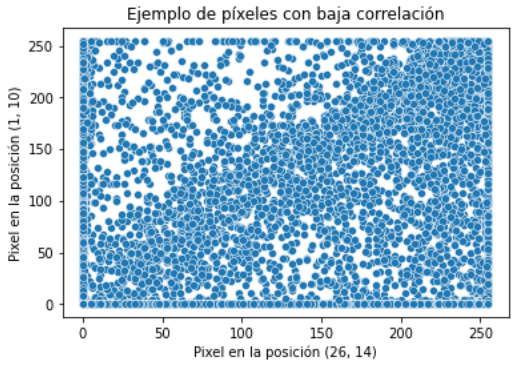
2.2.2. Modelo correlación

El objetivo es seleccionar un trío de píxeles que usaremos para entrenar un modelo de árbol y uno de KNN en la clasificación de prendas. Primero tomamos todas las imágenes del dataset y calculamos la desviación estándar de cada píxel

en dicho conjunto de imágenes, luego seleccionamos los 30 que tengan desviación estándar mas grande. La idea es que como son los que mas cambian su luminosidad de una prenda a la otra también son los que mas información tienen de cada prenda y por lo tanto de como diferenciarlas, un píxel que siempre tenga casi el mismo color no aporta información a cerca de cada prenda en si y menos de su clase. Pero los píxeles cercanos (o vecinos) tendrán una variación de luminosidad que esta correlacionada, ya que su luminosidad tendera a ser similar, estos no sirven pues aportan información repetida, entonces debido a que la iluminación de un píxel se puede inferir de la de los otros resultan ser píxeles redundantes, como seria el caso de dos píxeles en lados opuestos de la imagen, ya que al ser la mayoría de las imágenes simétricas suponen información redundante. Si se hiciera entonces un gráfico de puntos (un punto por prenda) de la luminosidad de un píxel contra la de otro estos no van a estar correlacionados si dicho gráfico esta homogéneamente cubierto de puntos, como en el caso mostrado en la figura 8(b), pero si están concentrados en un cono o una franja entonces si lo estarán, pues dada la luminosidad de un píxel se puede estimar que la del otro píxel entra dentro de dicha franja, como en el caso mostrado en la figura 8(a).



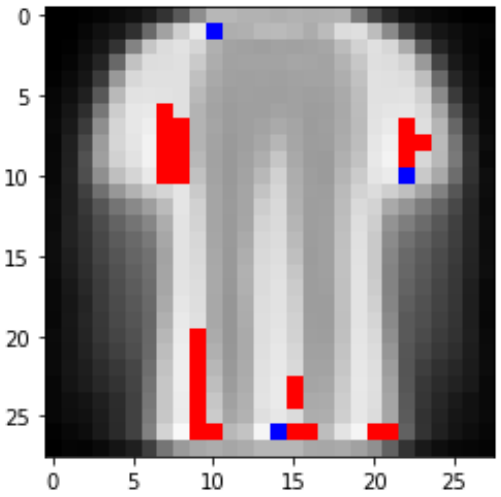
(a) Caso de los 2 píxeles con mayor correlación según nuestra métrica.



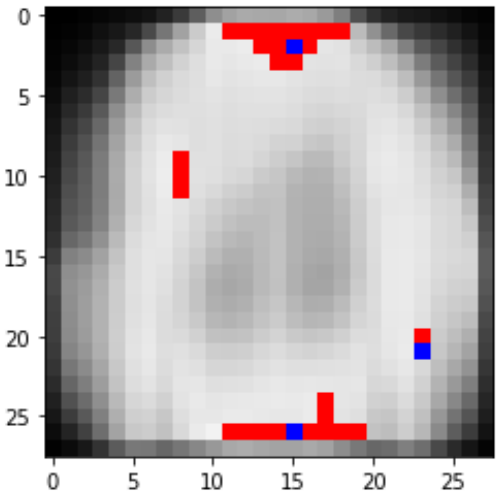
(b) Caso de los 2 píxeles con menor correlación según nuestra métrica.

Figura 8. Ejemplos de correlación entre píxeles (del dataset de pantalones y remeras), dados un par de píxeles, cada punto corresponde a la luminosidad en una prenda de un pixel graficada contra la del otro píxel.

Dicho esto, lo que se hizo es tomar la luminosidad de un píxel i y de un píxel j (de los 30 seleccionados anteriormente) en cada prenda de ropa, calcular un ajuste lineal y su error en norma 1, la idea es que el error sea muy grande porque eso quiere decir que el ajuste no es significativo, entonces no se puede calcular la iluminación del píxel i tomando al j y haciendo una cuenta, entonces a error grande corresponde una correlación baja. Pero como son 3 píxeles los que vamos a elegir, calculamos también el error del píxel i contra otro píxel k y de j contra k, sumamos los 3 errores, repetimos para toda convinacion posible i,j,k de píxeles distintos, y nos quedamos con los que tengan el error conjunto mas grande. En la figura 9 se muestra para el dataset entero (figura 9(b)) y para el caso de solo pantalones y remeras (figura 9(a)) la desviación estándar de cada píxel, los píxeles con mayor desviación y los 3 seleccionados al final.



(a) Selección de píxeles con la mayor desviación estándar entre pantalones y remeras.



(b) Selección de píxeles con la mayor desviación estándar considerando todas las clases del dataset.

Figura 9. En escala de grises la desviación entandar de cada píxel, en rojo los 27 píxeles que poseen la mayor desviación, en azul los 3 píxeles que además poseen menor correlación entre sí.

2.2.3. Modelo distancias arquetípicas

Esta sección de selección de datos se basa en la utilización del álgebra lineal para resumir la información de cada imagen, para ello lo que se le hizo a cada una de las prendas es calcular las distancias en norma 2 (3) a cada uno de los arquetipos, en la tabla 3 se pueden ver la distancia de las primeras 5 prendas a cada una de los arquetipos. De esta manera el objetivo es poder generar regiones para cada una de las clases de prendas inspiradas en la idea de espacio vectorial, donde cada prenda sería un vector.

label	remera	pantalón	pullover	vestidos	camperas	sandalias	camisetas	zapatillas	bolsos	botas
4	8.15	9.77	4.97	8.94	4.56	10.93	5.85	11.175	7.52	10.29
0	6.00	6.438	8.030	6.163	8.62	6.29	6.25	8.06	8.73	9.51
9	11.34	12.38	10.30	11.48	10.69	8.91	10.01	9.24	8.146	5.25
4	6.94	8.98	6.67	7.37	5.33	12.16	6.13	12.27	8.69	11.08
7	12.66	12.73	11.71	12.13	12.04	8.61	11.30	5.82	8.96	9.535

Tabla 3. Tabla de distancias a los arquetipos de las primeras 5 prendas del dataset.

Es relevante aclarar que para el modelo de clasificación pantalón-remera con knn solo se utilizaron las columnas de distancia remera y distancia pantalón de la tabla 3, en cambio para el modelo de clasificación multiclase, se utilizaron todas las columnas.

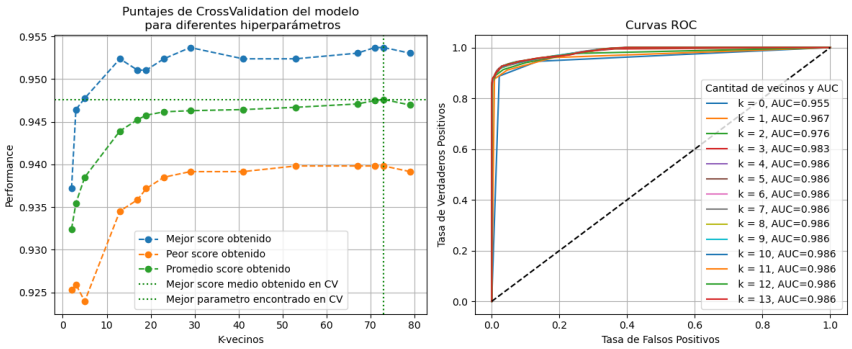
3. Modelado y discusiones

Para ambos tipos de modelos y los diferentes datos que se utilizaron para entrenarlos se siguió una misma metodología:

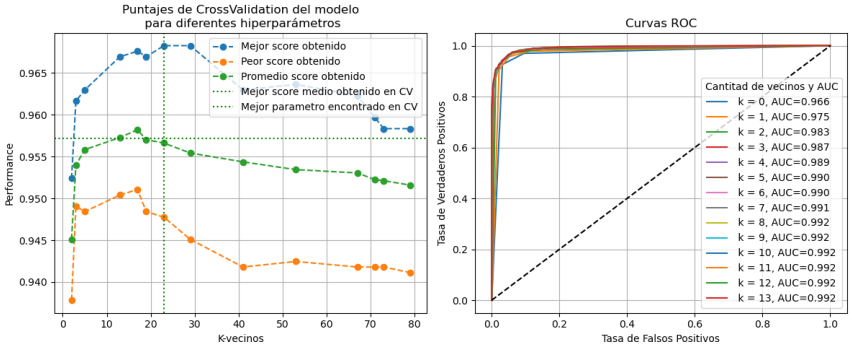
1. Se armó un modelo arbitrario para intentar observar si el modelo es posible, en todos los casos el modelo arbitrario lograba clasificar correctamente al menos el 50 % de los datos.
2. Se entrenó el modelo variando los hipermarámetros y utilizando la técnica crossvalidation para encontrar la mejor combinación posible de hiperparámetros, se graficó el score de cada combinacion de hyperparametros utilizados y las curvas ROC, en el caso del modelado pantalón remera y en el caso del modelo multiclase se realizó un reporte por clases, es decir se evaluó la perfomance de las combinaciones de hiperparametros.
3. Se eligió la mejor combinación de hiperparametros de cada selección de datos y se compararon entre si utilizando los datos de test, se elijió el modelo con la mejor performance.
4. Se valida el modelo final y se reporta la perfomance según corresponda al modelo.

3.1. Modelado Remera-Pantalon con KNN

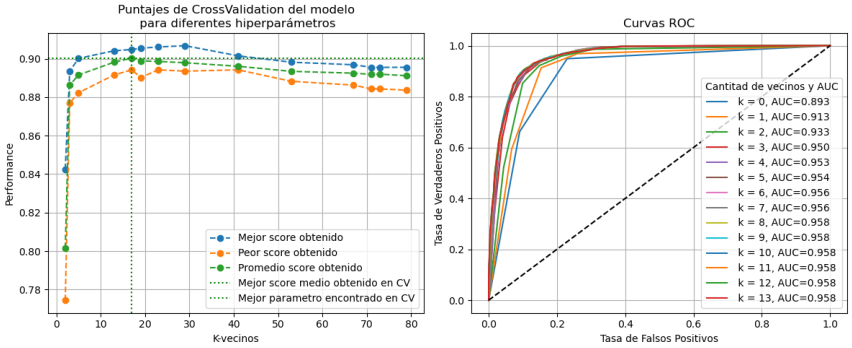
En la figura 14 podemos observar el desempeño de los modelos knn para cada tipo de datos.



(a) Modelo KNN de clasificación con los datos de Modelo píxeles arquetipos



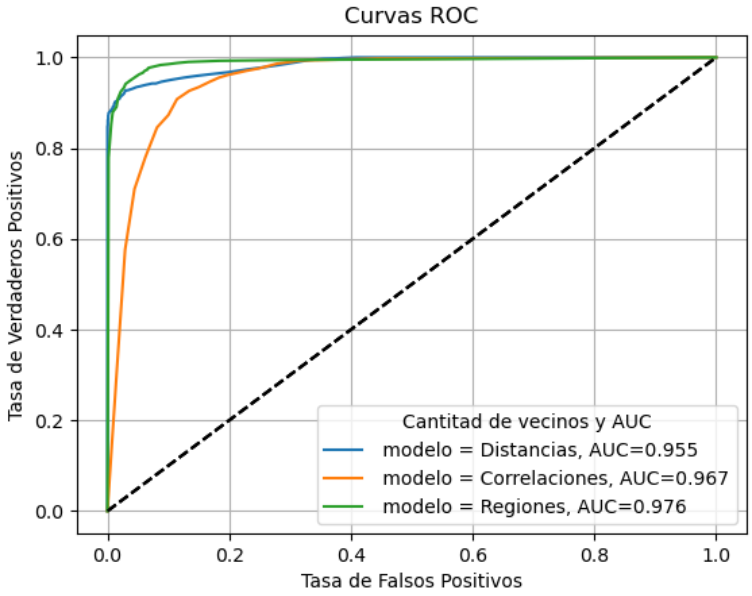
(b) Modelo KNN de clasificación con los datos de Modelo píxeles relavantes



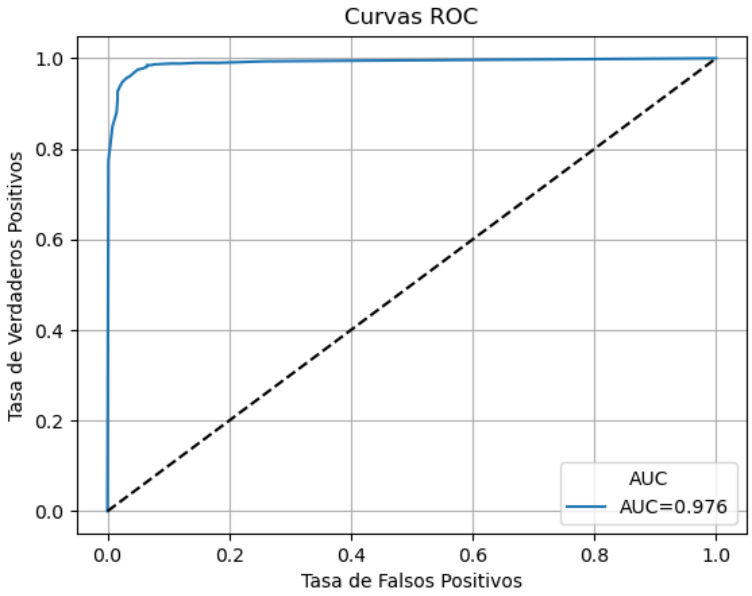
(c) Modelo KNN de clasificación con los datos de Modelo correlación

Figura 10. A la izquierda la evolución del score para diferentes hyperpametro, a la derecha las curvas ROC para cada hyperparametro

En la figura 14(a), modelo arquetipos, podemos ver como el mejor hyperparametro es k = 73, en 14(c), modelo píxeles relevalentes, podemos ver como el mejor hyperparmetro es k = 17 y en la figura 14(b), modelo píxeles correlación, podemos ver que el mejor hyperametro es k = 23. En la figura **??**, podemos observar las curvas de perfomance para cada uno de los mejores modelos. Donde podemos ver que el mejor modelo es el que esta entrenado con píxeles relevantes, llamado Regiones".



(a) Curvas Roc para los mejore modelos, el modelo "Distancia.[es] el modelo con los datos arquetipos, el modelo Çorrelaciones.[es] el modelo con los datos correlación, el modelo Regiones.[es] el modelo con los datos píxeles relevantes



(b) Curva Roc de validación del mejor modelo encontrado, entrenado con los datos pixeles relevantes y con un hyperparametro k = 17

Figura 11. A la izquierda las curvas Roc de cada uno de los mejores modelos, a la derecha la curva Roc del mejor modelo encontrado con los datos de validación

3.2. Modelado multiclase con arboles de decision

En la figura 14 podemos observar el desempeño de los modelos knn para cada tipo de datos en varianza de los score.

Podemos ver que en todos los tipos de elección el mejor hyperparametro criterion fue ."[en]tropy", luego tanto en el modelado con los datos de arquetipos y score, el mejor parámetro de altura fue de 11, en el caso de píxeles relavantes fue tambien de 11. En las tablas 4, 5 y 6, podemos ver los reportes de performance de cada modelo de clasificación para cada clase

Donde podemos ver que el mejor modelo es el de arquetipos, en la tabla 7 se pueden ver los resultados finales de la perfomance del modelo.

Es relevante recalcar que los modelos de arboles de decision son muy buenos para separar entre prendas muy diferentes como zapatillas de bolsos y zapatillas de pantalones, pero no entre prendas parecidas como camperas, remeras, camisas y pullovers.

	Class	Precision	Recall	F1-Score	Support
0	remera	0.669388	0.708642	0.688456	1620
1	pantalon	0.928709	0.892593	0.910293	1620
2	pullover	0.561179	0.540741	0.550770	1620
3	vestidos	0.690111	0.805556	0.743378	1620
4	camperas	0.558840	0.630247	0.592399	1620
5	sandalias	0.836460	0.880864	0.858088	1620
6	camisetas	0.421233	0.303704	0.352941	1620
7	zapatillas	0.823457	0.823457	0.823457	1620
8	bolsos	0.908112	0.884568	0.896185	1620
9	botas	0.885225	0.861728	0.873319	1620

Tabla 4. Reportes de performance para Modelado arquetipo.

	Class	Precision	Recall	F1-Score	Support
0	remera	0.687429	0.742593	0.713947	1620
1	pantalon	0.966948	0.920988	0.943408	1620
2	pullover	0.546119	0.599383	0.571513	1620
3	vestidos	0.686404	0.772840	0.727062	1620
4	camperas	0.537012	0.492593	0.513844	1620
5	sandalias	0.815691	0.808642	0.812151	1620
6	camisetas	0.473844	0.385802	0.425315	1620
7	zapatillas	0.846494	0.827160	0.836716	1620
8	bolsos	0.834772	0.835802	0.835287	1620
9	botas	0.853168	0.889506	0.870958	1620

Tabla 5. Reportes de performance para Modelado relavantes.

	Class	Precision	Recall	F1-Score	Support
0	remera	0.369108	0.541358	0.438939	1620
1	pantalon	0.703509	0.495062	0.581159	1620
2	pullover	0.624282	0.536420	0.577025	1620
3	vestidos	0.418235	0.617284	0.498629	1620
4	camperas	0.462103	0.116667	0.186299	1620
5	sandalias	0.389506	0.238272	0.295672	1620
6	camisetas	0.371397	0.421605	0.394912	1620
7	zapatillas	0.546006	0.666667	0.600334	1620
8	bolsos	0.423895	0.159877	0.232183	1620
9	botas	0.394077	0.747531	0.516088	1620

Tabla 6. Reportes de performance para Modelado correlacion.

	Class	Precision	Recall	F1-Score	Support
0	remera	0.684953	0.728333	0.705977	600
1	pantalon	0.936644	0.911667	0.923986	600
2	pullover	0.560068	0.551667	0.555835	600
3	vestidos	0.700719	0.811667	0.752124	600
4	camperas	0.554427	0.636667	0.592708	600
5	sandalias	0.839024	0.860000	0.849383	600
6	camisetas	0.420918	0.275000	0.332661	600
7	zapatillas	0.826599	0.818333	0.822446	600
8	bolsos	0.880524	0.896667	0.888522	600
9	botas	0.884941	0.871667	0.878254	600

Tabla 7. Tablar reporte de validacion del mejore metodo

4. Conclusiones

En conclusion, podemos observar como el modelo de KNN que mejor resultado dio es el modelo de píxeles relevantes. Además podemos ver la gran capacidad que tiene el modelo de clasificación KNN tiene para pocas dimensiones Por otro lado, podemos ver que en el caso de arboles de decisión, el modelo decae mucho en la capacidad de decidir entre prendas muy parecidas, siendo que tal vez lo mejor en este caso tal vez una unión entre el modelo de arboles de decisión y de KNN. En este ultimo modelo el mejor modelo fue el de distancias a los arquetipos.

Referencias

[1] Fashion MNIST, Kaggle, ZALANDO RESEARCH

[2] ¿Qué es KNN?, IBM

[3] Árboles de decisión, IBM

[4] KNeighborsClassifier, Sciki-learn

[5] DecisionTreeClassifier, Sciki-learn

5. Apéndice

La figura 12 consiste en los arquetipos de cada prenda generados con el promedio de estas

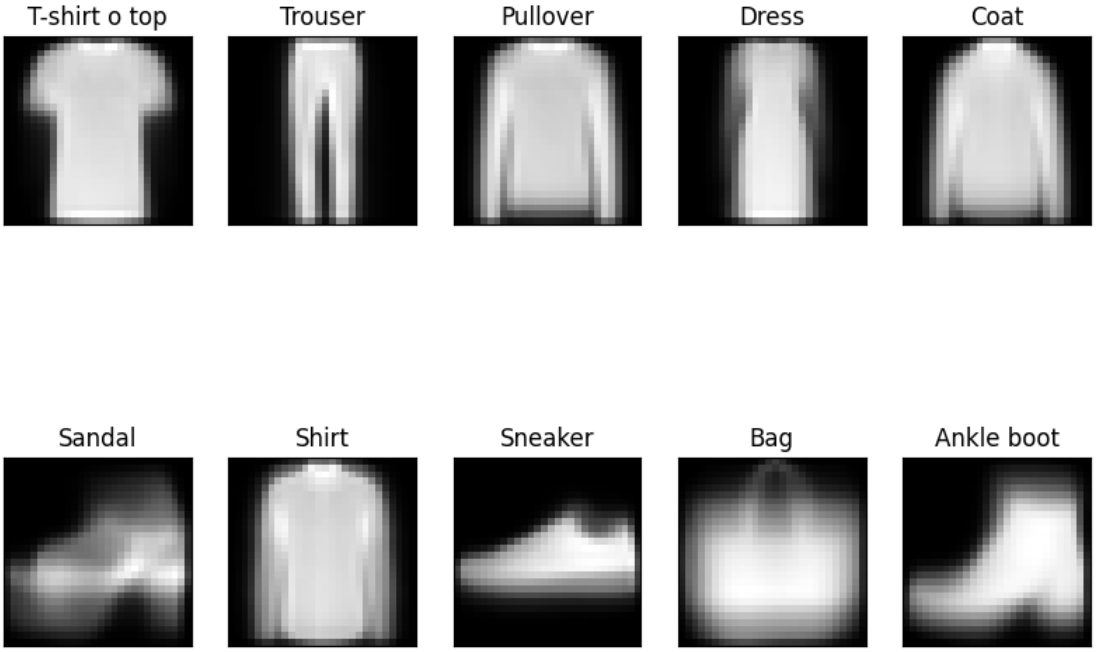
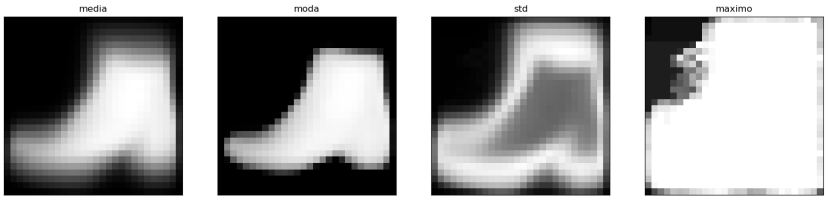
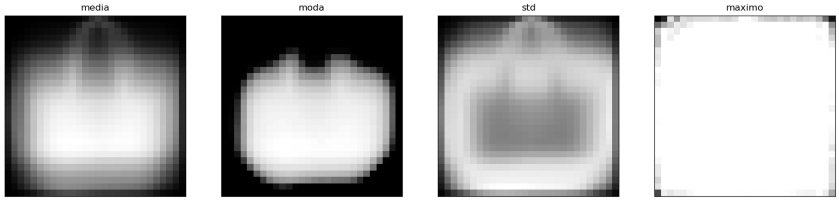


Figura 12. Arquetipos generados con el promedio de cada prenda

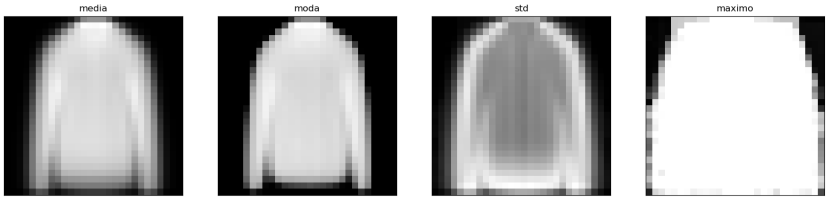
La figura 13 contiene las medidas resúmenes de cada tipo de prenda



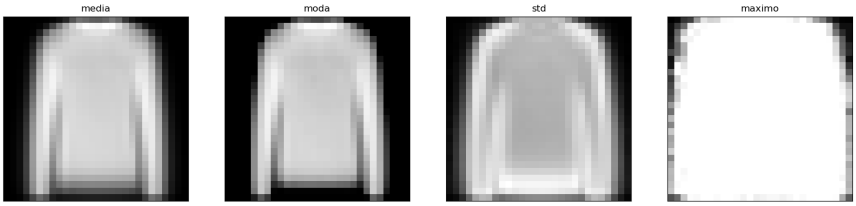
(a) Medidas de resumen botas.De izquierda a derecha promedio, media, desviación estándar y máximo de cada píxel



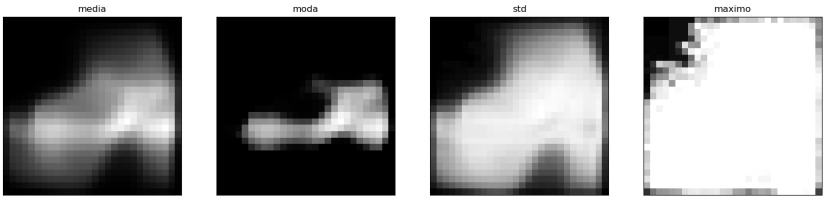
(b) Medidas de resumen bolsos.De izquierda a derecha promedio, media, desviación estándar y máximo de cada píxel



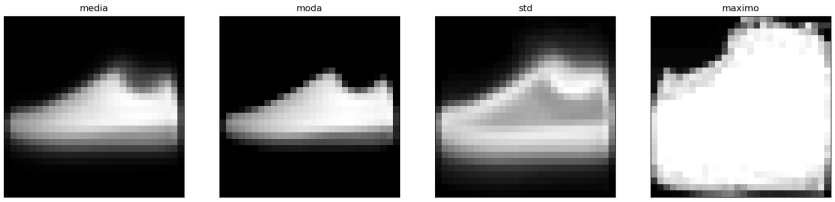
(c) Medidas de resumen camperas.De izquierda a derecha promedio, media, desviación estándar y máximo de cada píxel



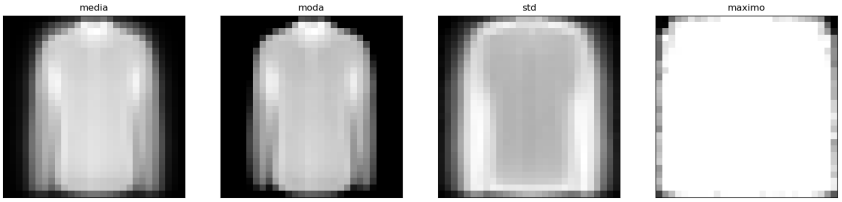
(d) Medidas de resumen pullover.De izquierda a derecha promedio, media, desviación estándar y máximo de cada píxel



(e) Medidas de resumen sandalias.De izquierda a derecha promedio, media, desviación estándar y máximo de cada píxel

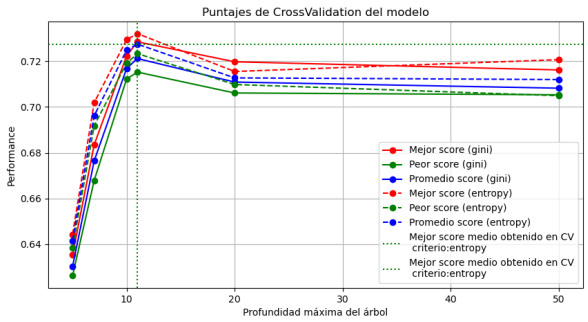


(f) Medidas de resumen zapatillas.De izquierda a derecha promedio, media, desviación estándar y máximo de cada píxel

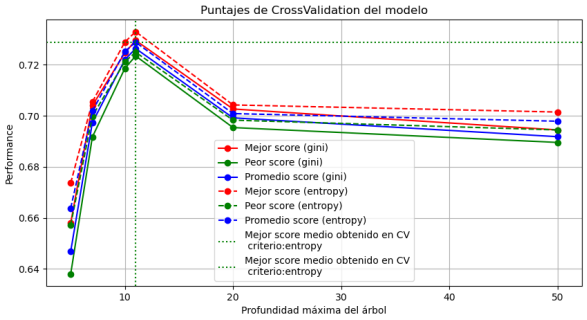


(g) Medidas de resumen camisetas.De izquierda a derecha promedio, media, desviación estándar y máximo de cada píxel

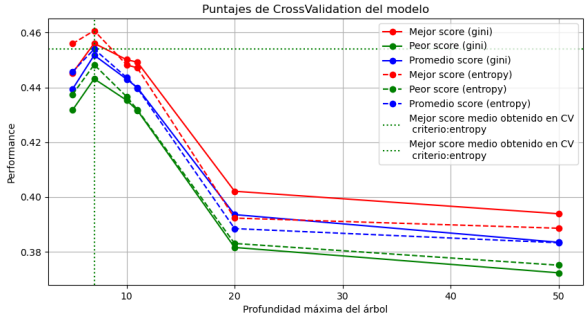
Figura 13. Medidas de resumen de todas las prendas



(a) Modelo Arboles de decisión para clasificación con los datos de Modelo píxeles arquetipos.



(b) Modelo Arboles de decisión para clasificación con los datos de Modelo píxeles relavantes.



(c) Modelo Arboles de decisión para clasificación con los datos de Modelo correlación.

Figura 14. Evolución del score variando los hyperparametros.