Análisis de imágenes de lesiones causadas por Cáncer de piel: Una aproximación al diagnóstico no supervisado de la gravedad de la enfermedad

Proyecto Final Grupo 9

(Pérez, Juan Camilo; Santacruz, Johan; Murcia Pinilla, Alejandro)

Resumen - Los algoritmos de aprendizaje no supervisado pueden facilitar en gran medida el trabajo clínico necesario para la identificación de cáncer en la piel, basándose únicamente en imágenes de las heridas causadas por esta enfermedad en los pacientes. Utilizando las técnicas de reducción de dimensionalidad, se intentará demostrar que es posible retener el 95% de la varianza de las imágenes del banco de datos, en un espacio menor al original (pasando de 67500 columnas o variables originales a un número mucho menor). Luego, aplicando técnicas de clústering (Kmeans, Clustering jerarquico, DBSCAN), se pueden generar agrupaciones de imágenes con características específicas, que posteriormente ayudaran a determinar qué tan probable es que una imagen que pertenece una u otra agrupación, sea considerada como un cáncer maligno o de otro tipo.

I. Introducción y estado del arte

Existen evidencias estadísticas que concluyen que el cáncer de piel es el más común entre todos los tipos de cáncer y existen tipos que pueden ser tanto malignos como benignos, lo cual cambiará completamente en las atenciones que debería recibir una persona para su correcto tratamiento. Además, se ha identificado al melanoma como el tipo que representa solamente un 1% del total de los casos, pero al cual se le atribuye la gran mayoría de las muertes por este tipo de cáncer.

Se identifica también que, el riesgo de padecer un cáncer de piel de tipo melanoma es del 2.6% para las personas de razas blancas, siendo 1 de cada 38 personas afectadas por esta condición, mientras que, para personas de color es del 0.1% y para los hispanos del 0.6%. Adicional a esto, según la Sociedad Americana Contra el Cáncer en los Estados Unidos, para el año 2022 se diagnosticaron alrededor de 100.000 casos nuevos de melanoma, y que el riesgo de morir es para una población de 7.650 personas (American Cancer Society, 2022).

Dado lo anterior, el riesgo latente en las personas de presentar este tipo de cáncer de piel es relativamente alto, y uno de las mejores formas de reducir y mitigar el riesgo de muerte por esta enfermedad es la adecuada identificación temprana.

Es por esto, que la ***motivación de este proyecto*** es: *"aplicar metodologías de Aprendizaje de Máquina para poder ayudar tanto a las personas como a los doctores, a la posible identificación temprana de anormalidades en la piel que puedan conducir a la presencia de melanoma o cualquier tipo de cáncer de piel maligno en las personas, por medio de la agrupación de las imágenes de las heridas, haciendo uso de sus similitudes"*.

Manhas (2022) propone un marco de trabajo útil para trabajar diagnósticos de cáncer utilizando técnicas de Machine y Deep Learning. Partiendo del conjunto de entrada de datos, realiza un preprocesamiento de las imágenes. Seguidamente realiza segmentación. Luego, Aplica técnicas de extracción y selección de características. Finalmente, utiliza el resultado de todo el proceso como entrada para algoritmos de clasificación de aprendizaje supervisado para obtener una predicción si una nueva imagen es o no causada por cáncer de piel (Manhas, 2022). Aunque esa idea general es similar a la que aquí se expone, no usaremos algoritmos de aprendizaje supervisado, sino que se usarán las características de cada clúster generado con el fin de estimar la probabilidad de un diagnóstico (basándonos en la cantidad de imágenes de dicho clúster que son causadas por cáncer maligno contra el total de imágenes del clúster).

Investigaciones anteriores han tenido éxitos considerables reduciendo la dimensionalidad de los datos de entrada, como los obtenidos por Jorgensen (2008) quienes utilizando combinaciones de algoritmos de aprendizaje no supervisado como PCA y algoritmos de aprendizaje supervisado como SVM lograron aciertos en los diagnósticos cercanos al 80%.

Por último (Bechelli & Delhommelle, 2022) realizaron una investigación detallada de los resultados obtenidos en diagnóstico de cáncer de piel utilizando las imágenes de las lesiones, enfocándose principalmente en la clasificación de cáncer maligno y benigno bajo algoritmos de aprendizaje supervisado como LDA y Redes Neuronales Recurrentes.

La principal diferencia contra la propuesta de este documento es que no usaremos aprendizaje supervisado y haremos mayor énfasis en tratar de reducir la dimensionalidad del problema original, para posteriormente agrupar las imágenes por sus atributos intrínsecos.

II. Materiales y Métodos

Los datos que se van a usar para el desarrollo de este proyecto es una recopilación parcial de todas las imágenes con las que cuenta ISIC, y está compuesta por un total de 2.357 imágenes clasificadas en los diferentes tipos de posibles cánceres de piel (9 clases), siguiendo la distribución que se muestra en la siguiente gráfica:

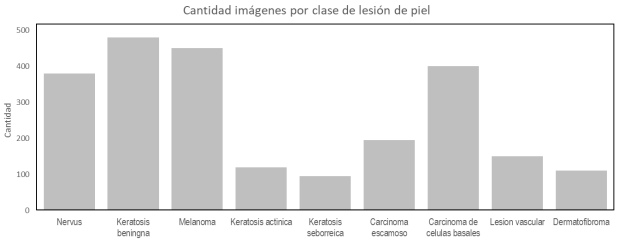


Ilustración 1- Cantidad de imágenes por cada clase de lesión de piel

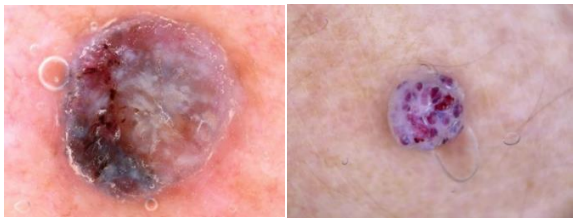


Ilustración 2 - Ejemplos de Melanoma (izquierda) - Lesión vascular (Derecha)

Por tal motivo, no es necesario realizar limpieza de los datos. Los datos de entrada son imágenes que están compuestas cada una por 3 canales representados en RGB (valores numéricos), y cada una presenta dimensiones de 224*224, generando así un total de 150.528 pixeles de información para cada una. Es posible entonces afirmar que, el dataset resultante tiene un tamaño de (2.357, 150.528). No se cuenta con variables categóricas ni tampoco es necesario imputar datos faltantes.

Dada la dimensión de datos y para realizar la lectura de las fotos en Python se realizó un proceso inicial de reducción de la dimensión de estas, con el apoyo de la función de image_dataset_from_directory de la librería TensorFlow. Esta permite redimensionar el tamaño de cada imagen y agruparlos en un objeto batch. Con esto, se modifica el dataset de entrada y se decide reducir las dimensiones de las imágenes a un tamaño de 150*150, generando así un total de 67.500 pixeles de información para cada imagen. Es posible entonces afirmar que, el dataset resultante final tiene un tamaño de (2.357, 67.500).

III. Propuesta Metodológica

Tal y como se planteó en la motivación del proyecto, la solución a la problemática en la identificación temprana de melanoma en las personas podría tener una solución basada en metodologías de Aprendizaje de Máquina No Supervisado.

La razón para hacer uso de estas metodologías o algoritmos es debido a las dimensiones que componen cada una de las imágenes, y al tamaño comprendido por todas las imágenes para realizar el entrenamiento de cualquier modelo.

Es por esta razón, que la primera aplicación será la implementación de un algoritmo de PCA para buscar reducir el tamaño de los datos a entrenar, intentando penalizar por máximo entre el 1% y el 5% del total de la varianza de las fotos completas, y de esta forma poder ejecutar de forma óptima los diferentes modelos de clustering con los limitados recursos computacionales disponibles.

La segunda aplicación propuesta es la implementación de tres modelos diferentes de Clustering, entre los cuales tenemos uno de ***Kmeans*** (para grupos convexos); otro de ***clustering jerarquico*** (para grupos no convexos); y uno final de ***DBSCAN*** (para obtener un resultado diferente con grupos no convexos). Dichos modelos se implementarán con la intención de lograr agrupar las imágenes de las lesiones de forma tal que los patrones similares entre estas nos permitan dar una alerta temprana de la presencia de un cáncer de piel maligno o benigno en una persona, o en caso similar, servir a un médico especialista para apoyarse en su toma de decisiones.

Es de aclarar que se parte de la idea de que el etiquetado manual de estas imágenes requiere de la revisión de un especialista en el tema, y que al ser tantas imágenes requiere de un costo en tiempo significativamente alto, por lo cual se intenta no acudir a algoritmos de aprendizaje de maquina supervisados, sin embargo, aprovechando que la base que vamos a utilizar para este ejercicio se encuentra etiquetada, estas etiquetas van a ser usadas como métrica de desempeño para evaluar qué tan bien se agrupan las imágenes con los diferentes modelos propuestos.

Al final, lo que se espera con la implementación de los modelos propuestos anteriormente, es obtener un algoritmo de clasificación que nos permita agrupar las imágenes en cánceres de piel de tipo maligno, diferentes a las agrupadas con tipos de cáncer benigno. De manera alternativa, en caso de no lograr que las imágenes se distribuyan de acuerdo con la expectativa inicial, esperamos que los clusters resultantes nos permitan clasificar una determinada foto con un nivel de riesgo de que sea un cáncer maligno, en otras palabras, si el cluster agrupa un 80% de fotos malignas pero lo mezcla con un 20% de fotos benignas, buscaremos darle una alerta de riesgo “alto” al usuario debido a su alta probabilidad de que sea maligno. Si se obtiene lo opuesto, sería un riesgo “bajo”. Finalmente, si el

clúster agrupa un 50%-50% de ambas clases, se llega a un resultado no concluyente.

IV. Resultados Iniciales

Reducción con PCA, buscando explicar al menos el 95% de la varianza de los datos originales:

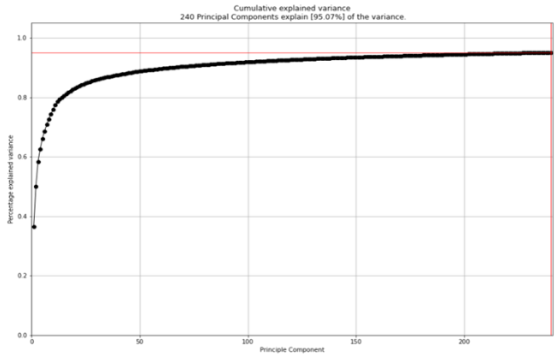


Ilustración 3 – Resultado Parcial de PCA

Con lo anterior, se confirma que, utilizando 240 componentes principales y siguiendo el criterio de la varianza explicada, es posible explicar el 95% de la varianza del conjunto total de imágenes de entrenamiento, que como se ha dicho anteriormente, tienen tamaño 150*150. También se construyó un escenario alternativo en donde se retuvo el 99% de la varianza usando los primeros 1000 componentes principales. Ambos conjuntos de entrada fueron usados para evaluar los resultados obtenidos.

Ahora bien, es posible realizar un primer análisis preliminar de clustering utilizando la librería AgglomerativeClustering. Seleccionamos 9 clusters porque son 9 clases las que componen el dataset de entrada. Posteriormente podemos validar por medio de una matriz de confusión que tan bien o mal han quedado distribuidas las imágenes en los clusters (recordando que tenemos marcadas todas las imágenes):

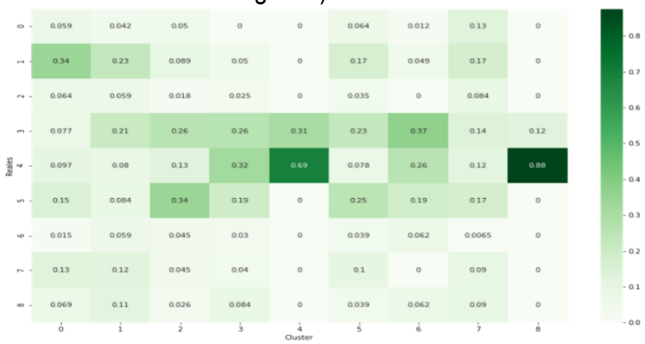


Ilustración 4 – Resultado Parcial de Matriz de confusión para 9 clusters con AgglomerativeClustering

Los resultados parciales obtenidos muestran que, por ejemplo, la clase 4 está siendo incluida principalmente en el cluster 4. Las demás clases son agrupadas en clusters diferentes a las que le corresponden.

V. Refinamiento de los resultados

Se da inicio entonces al modelamiento de los dos primeros algoritmos propuestos; ***Kmeans y Cluster Jerárquico***, buscando dar una primera aproximación a clústeres con formas convexas y con formas no convexas. Es importante decir que se dio la idea de empezar desde un objetivo macro y a medida de ir encontrando resultados, ir haciendo más específicos los objetivos; es decir, que la primera parte para este proyecto se basó en separar las imágenes de los tipos de cáncer de piel únicamente en 2 tipos diferentes que fueron Malignos o Benignos, y a medida de encontrar una agrupación que separe de forma adecuada estos dos tipos, ver cómo clasificar cada uno de los tipos de forma intra-clúster.

Para ***Kmeans*** se realizó un barrido de los diferentes parámetros, variación de los PCAs (240 PCA con 95% de varianza explicada y 1000 PCA con 99% de varianza explicada) y entrenar con estos parámetros y variar entre el número de clústeres generados, empezando en 2 y llegando hasta 8. Lo anterior nos deja resultados como los siguientes:

Beningno	**0.5**	**0.59**
Maligno	**0.5**	**0.41**
	0	1
	Cluster	

Beningno	**0.6**	**0.6**	**0.47**	**0.64**	**0.47**
Maligno	**0.4**	**0.4**	**0.53**	**0.36**	**0.53**
	0	1	2	3	4
	Cluster				

Ilustración 5 – Resultado Parcial de Matrices de confusión para 2 y 5 clusters con Kmeans con 240 PCA

En las dos matrices de confusión el eje y representa la categoría real. Por otro lado, el eje x hace referencia a los clústeres generados por el algoritmo, de tal forma que la primera imagen sería la matriz de confusión para un k de 2, mientras que el de la imagen siguiente sería el resultante para un k de 5.

Lo anterior nos permite concluir que cuando sacamos 2 clústeres, la agrupación resultante no nos deja acercarnos a nuestro objetivo, ya que las imágenes quedan distribuidas equitativamente entre malignos y benignos para los dos grupos, y esto no nos permite definir un nivel de riesgo para ninguno de los dos casos. Tal vez como parte de lo que esperábamos en este punto, una agrupación aun haciendo uso de un k igual a 5 tampoco nos permite acercarnos al objetivo definido, y esto se debe a que la intuición nos dice que un agrupamiento en formas convexas no es la mejor

solución para este tipo de imágenes, sin embargo, aquí podríamos empezar a acercarnos un poco más a lo deseado que es encontrar una diferenciación entre los clústeres y poder definir un nivel de riesgo de acuerdo a esto, por ejemplo, decir que el clúster 3 es de un riesgo menor a los restantes dada la probabilidad de que la foto que pertenezca a este clúster sea benigno y no maligno.

Para ver cómo varían los resultados de este modelo, se procedió a cambiar el número de PCA de 240 a 1.000.

Beningno	**0.5**	**0.59**
Maligno	**0.5**	**0.41**
	0	1
	Cluster	

Beningno	**0.47**	**0.6**	**0.64**	**0.48**	**0.6**
Maligno	**0.53**	**0.4**	**0.36**	**0.52**	**0.4**
	0	1	2	3	4
	Cluster				

Ilustración 6 – Resultado Parcial de Matrices de confusión para 2 y 5 clusters con Kmeans con 1.000 PCA

De acuerdo a los resultados anteriores, no se ve ningún tipo de mejoría en el modelo. Por otro lado, se realizó la implementación del ***clustering jerárquico***, para el cuál de entrada esperamos una segmentación mejor a la encontrada anteriormente debido a la capacidad de este método para encontrar formas no convexas en los datos. En este caso también se hizo una variación entre los diferentes parámetros como; La afinidad, la cual dejamos establecida como euclidiana debido a su comportamiento final frente al resto, y el tipo de enlace a usar, el cual se definió como completo. Nuevamente se realizó la aplicación de un barrido sobre 2 a 8 clústeres, buscando identificar el que mejor se adapte a nuestro objetivo, y los resultados se pueden ver en la siguiente matriz de confusión;

Beningno	**0.55**	**0**	**0**	**0**
Maligno	**0.45**	**1**	**1**	**1**
	0	1	2	3
	Cluster			

Ilustración 7 – Resultado Parcial de Matrices para 4 clusters con clustering jerárquico con 240 PCA

Encontramos que esta clasificación nos permite sacar mejores conclusiones de acuerdo con lo que buscamos, aunque claramente quisiéramos obtener resultados mejores. De esta matriz podemos concluir entonces que el clúster 1, 2 y 3 sería para nosotros un riesgo alto, ya que todo lo que pertenezca a este clúster probablemente va a ser maligno, mientras que el clúster 0 podría clasificarse como no concluyente, ya que la probabilidad de que sea maligno se ve considerablemente reducida.

Beningno	**0.48**	**0.61**	**0.58**	**0.56**
Maligno	**0.52**	**0.39**	**0.42**	**0.44**
	0	1	2	3
	Cluster			

Ilustración 8 – Resultado Parcial de Matrices para 4 clusters con clustering jerárquico con 1.000 PCA

Al cambiar el número de PCA's de 240 a 1.000 vemos que ahora todos los clusters parecieran ser no concluyentes, donde los clusters 1 a 3 tienden a tener mayor probabilidad de ser cáncer benigno, mientras que el cluster 0 esta equitativamente distribuido.

Dado que los modelos anteriores no resuelven correctamente la problemática planteada, daremos paso al último algoritmo planteado dentro del alcance de este proyecto: la implementación de un método de **DBSCAN**, el cuál suena tentador con los resultados, ya que es otro de los métodos que permite realizar una segmentación con formas no convexas. El algoritmo DBSCAN (por su nombre en inglés Density-based spatial clustering of applications with noise) agrupa los datos en función de las densidades de las observaciones, mientras maneja el ruido de manera eficiente. DBSCAN entonces se diferencia de Kmeans o Clustering Jerárquico porque puede incorporar la intuición de agrupar los diferentes tipos de cáncer a partir de los patrones visuales de las lesiones de los pacientes.

Se puede calcular la distancia desde cada punto a su vecino más cercano usando el parámetro *Neighbors*. Este método devuelve dos matrices, una que contiene la distancia a los n_neighbors puntos más cercanos y la otra que contiene el índice para cada uno de esos puntos.

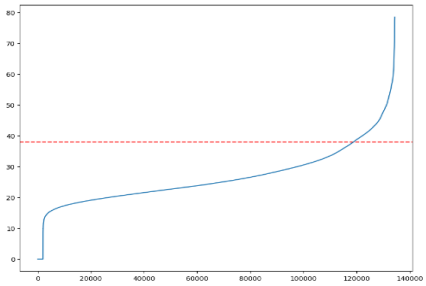


Ilustración 9 – Grafica exponencial con DBSCAN con 240 PCA

En la gráfica se puede ver que la parte exponencial de la curva comienza alrededor de una distancia de 38. A continuación, se muestra uno de los resultados obtenidos:

Beningno	**0.38**	**0.55**
Maligno	**0.62**	**0.45**
	0	1
	Cluster	

Ilustración 10 – Resultado de Matrices para 1 clusters con DBSCAN y n_samples=5 y 240 PCA.

De esta matriz podemos concluir entonces que el clúster 0 sería para nosotros un riesgo alto, ya que las imágenes que pertenezca a este clúster probablemente van a tratarse de cancer maligno, mientras que el clúster 1 podría clasificarse como no concluyente, aunque con mayor probabilidad en que en la clasificación resulte ser beningno. Se procede a aumentar el número de PCA de 240 a 1.000.

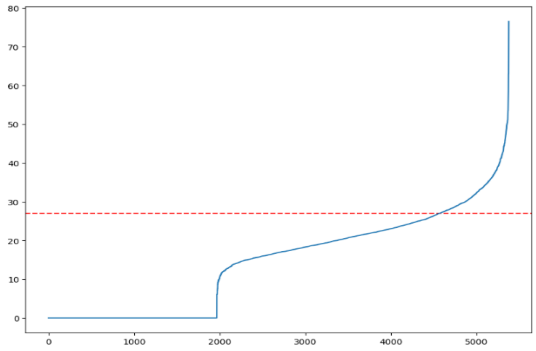


Ilustración 11 – Gráfica DBSCAN con 1.000 PCA

Beningno	**0.5**	**0.55**	**0.72**	**0.92**	**1**	**0.33**	**1**
Maligno	**0.5**	**0.45**	**0.28**	**0.077**	**0**	**0.67**	**0**
	0	1	2	3	4	5	6
	Cluster						

Ilustración 12 – Resultado Parcial de Matrices para 1 clústeres con DBSCAN y n_samples=5 y 1.000 PCA.

En esta nueva iteración encontramos que los clústeres 2,3,4 y 6 tienen una buena clusterización ya que permite identificar aquellos casos de riesgo bajo o cáncer benigno. Por su parte el clúster 5 logra identificar aquellos casos mayor probabilidad de padecer algún tipo de cáncer maligno (riesgo alto) y los clústeres 0 y 1 no son concluyentes puesto que las probabilidades de cada segmento son equiprobables.

Como trabajo futuro, se puede plantear la exploración de otro tipo de algoritmos, como Gaussian Mixture y comparar contra los resultados ya obtenidos con el fin de seleccionar la mejor alternativa. También, se deben explorar alternativas que permitan que los clúster cuenten con una mayor frecuencia de imágenes, ya que en algunos casos (específicamente para DBSCAN) algunos clusters solamente se conforman con un porcentaje de imágenes muy bajo, mientras que otros (sobre todo los equiprobables) suelen acumular la mayor cantidad de imágenes. Esto implica directamente que los resultados obtenidos hasta este punto experimentales y que requieren de más trabajo para poder ser utilizados en la industria médica.

VI. Conclusiones

- La reducción de la dimensionalidad facilita la elaboración de análisis no supervisados en entornos de recursos computacionalmente limitados de una manera eficiente, sin sacrificar excesivamente la información más importante de un conjunto de datos.
- Es posible construir clústeres que permitan sugerir la temprana detección de cáncer de piel usando únicamente las imágenes de las heridas causadas por esta enfermedad, aunque en algunos escenarios, el resultado puede no ser concluyente.
- Los mejores resultados fueron obtenidos utilizando DBSCAN, utilizando 1000 componentes principales que lograron explicar el 99% de la varianza de las imágenes, logrando la generación de 7 clusters. 2 de ellos son equiprobables y no concluyentes. 1 está asociado con el cáncer maligno y 4 están asociados al cáncer benigno.
- Al graficar los resultados de las clasificaciones en dos ejes (analizando los componentes principales por pares) notamos que existe solapamiento en casi todos ellos, por lo que es complicado para algoritmos de clustering como los usados en este ejercicio lograr una clara división para las clases deseadas. Además, si comparamos los resultados obtenidos en general con la aproximación no supervisada contra los que encontramos en el estado del arte (la mayoría de ellos usando aprendizaje supervisado) se hace evidente que este ultimo enfoque logra mejores resultados que los obtenidos en este documento.

VII. Bibliografía

- American Cancer Society. (2022). ACERCA DEL CÁNCER DE PIEL TIPO MELANOMA. Obtenido de Estadísticas importantes sobre el cáncer de piel tipo melanoma: https://www.cancer.org/es/cancer/cancer-de-piel-tipo-melanoma/acerca/estadisticas-clave.html
- Bechelli, S., & Delhommelle, J. (2022). ProQuest. Obtenido de https://www.proquest.com/docview/2642338340?accountid=34489&parentSessionId=AfN2El0ZTN7N86%2BDc7iZ7Qgm1D7gpjU6gOXDN6Z3jK4%3D&pq-origsite=primo
- ISIC. (9 de 2022). ISIC. Obtenido de https://www.isic-archive.com/#!/topWithHeader/tightContentTop/about/aboutIsicOverview
- Manhas, J. G. (4 de 9 de 2022). A Review on Automated Cancer Detection in Medical Images using Machine Learning and Deep Learning based Computational Techniques: Challenges and Opportunities. Obtenido de Spinger Link: https://doi-org.ezproxy.uniandes.edu.co/10.1007/s11831-021-09676-6
- Thomas Martini Jorgensen, A. T. (2008). EBSCO. Obtenido de https://web-p-ebscohost-com.ezproxy.uniandes.edu.co/ehost/pdfviewer/pdfviewer?vid=0&sid=f23c8e7e-9b98-4e85-91c6-1442ee6f0af6%40redis

VIII. ANEXOS

a. Ilustraciones

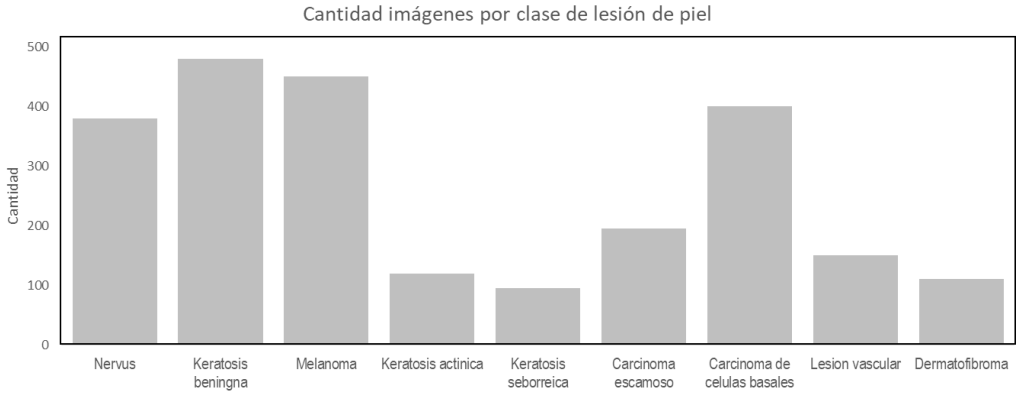


Ilustración 1- Cantidad de imágenes por cada clase de lesión de piel

```
Number of Train Images 2239
Number of Test Images 118
##################################################
Train Images Breakdown {'actinic keratosis': 114, 'basal cell carcinoma': 376, 'dermatofibroma': 95, 'melanoma': 438, 'nevus':
357, 'pigmented benign keratosis': 462, 'seborrheic keratosis': 77, 'squamous cell carcinoma': 181, 'vascular lesion': 139}
_____
Test Images Breakdown {'actinic keratosis': 16, 'basal cell carcinoma': 16, 'dermatofibroma': 16, 'melanoma': 16, 'nevus': 16,
'pigmented benign keratosis': 16, 'seborrheic keratosis': 3, 'squamous cell carcinoma': 16, 'vascular lesion': 3}
##################################################
Classes ['actinic keratosis', 'seborrheic keratosis', 'basal cell carcinoma', 'pigmented benign keratosis', 'nevus', 'squamous
cell carcinoma', 'melanoma', 'vascular lesion', 'dermatofibroma']
```

Ilustración 2- Cantidad de imágenes por cada clase de lesión de piel y por subset (Train/Test)

Ilustración 3 - Ejemplos de Melanoma (izquierda) - Lesión vascular (Derecha)

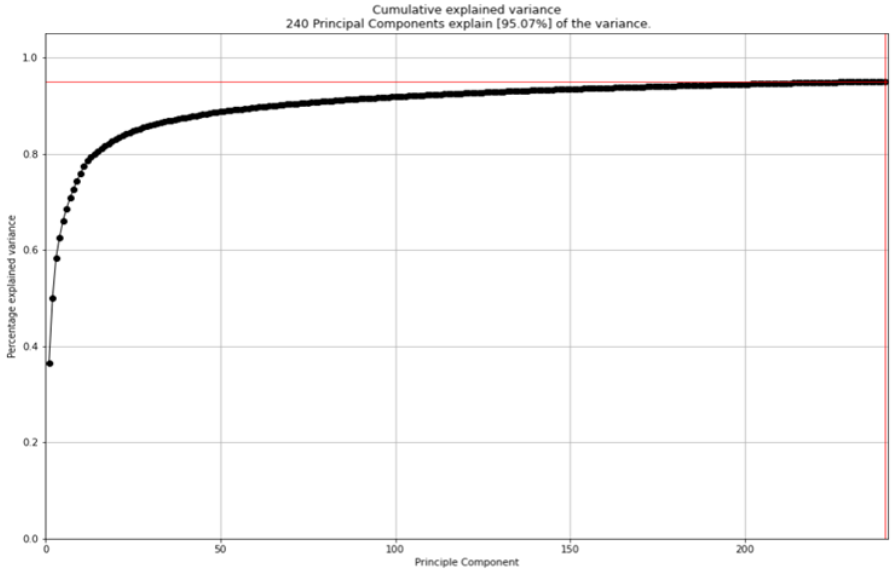


Ilustración 4 – Resultado Parcial de PCA

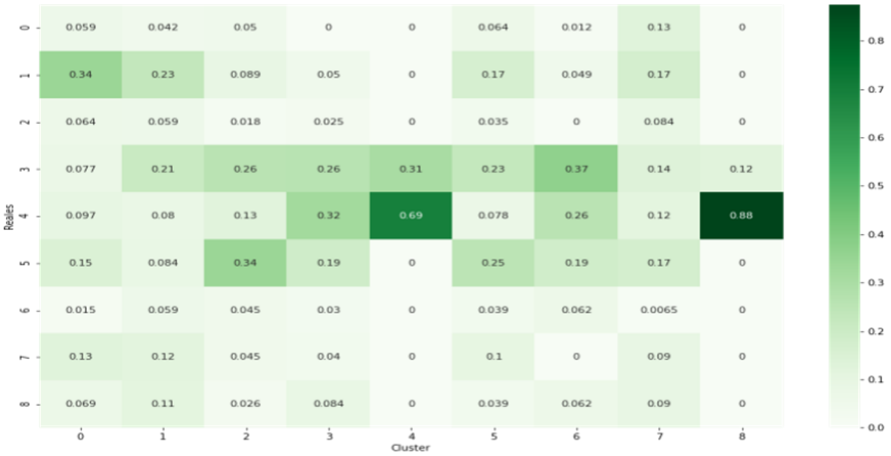
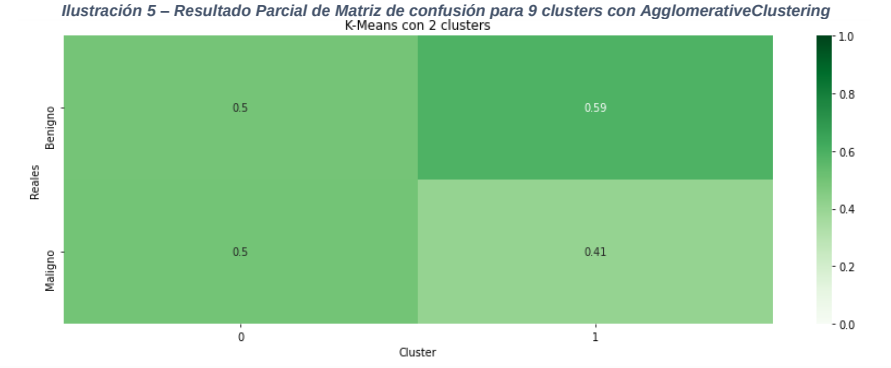


Ilustración 5 – Resultado Parcial de Matriz de confusión para 9 clusters con AgglomerativeClustering



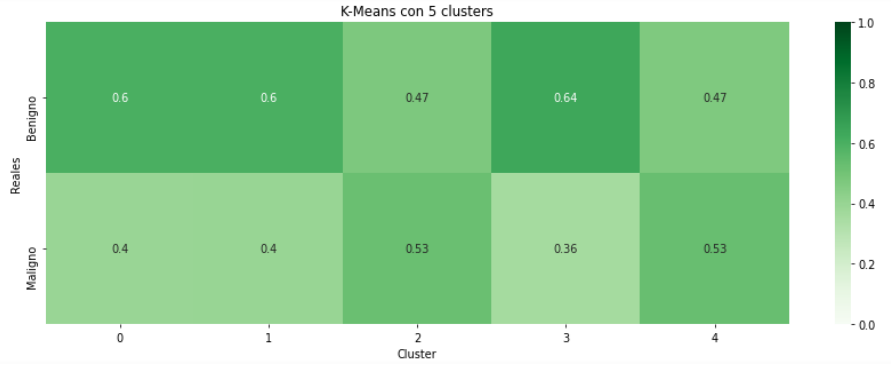


Ilustración 7 – Resultado Parcial de Matrices de confusión para 2 y 5 clusters con Kmeans con 240 PCA

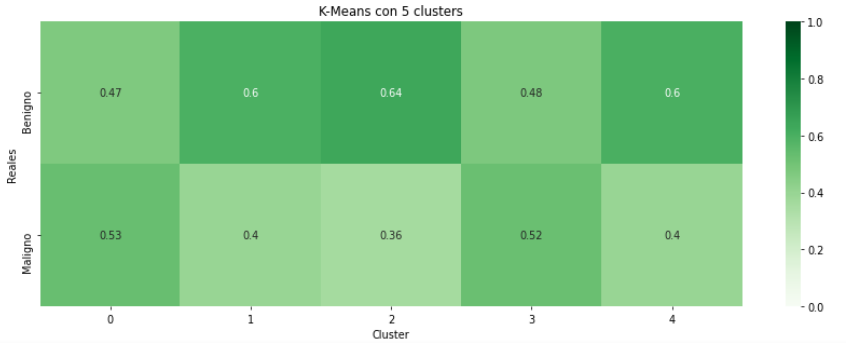


Ilustración 8 – Resultado Parcial de Matrices de confusión para 2 y 5 clusters con Kmeans con 1.000 PCA

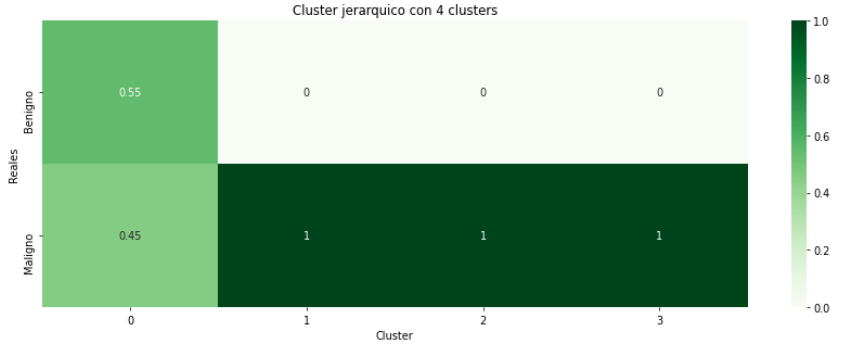


Ilustración 9 – Resultado Parcial de Matrices para 4 clusters con clustering jerárquico

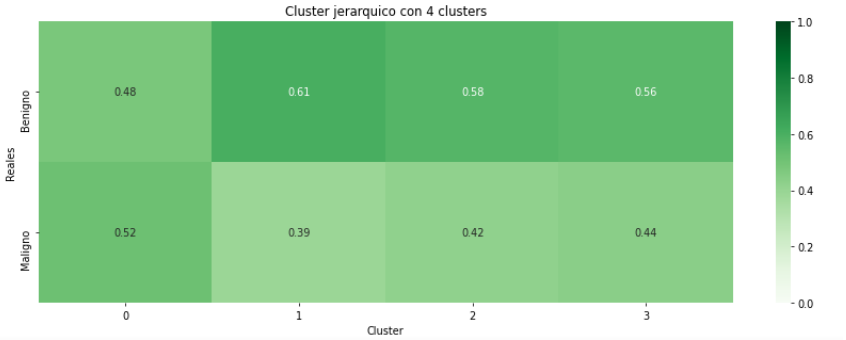


Ilustración 10 – Resultado Parcial de Matrices para 4

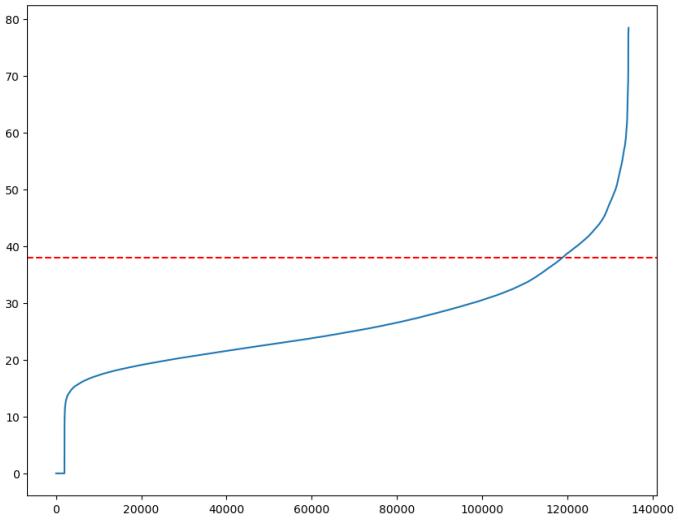


Ilustración 11 – Grafica exponencial con DBSCAN con 240 PCA

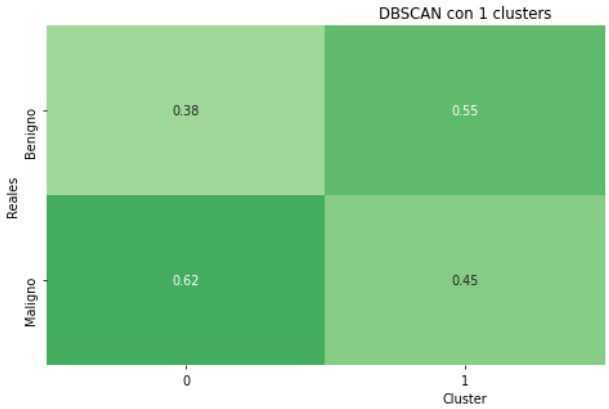


Ilustración 12 – Resultado Parcial de Matrices para 1 clusters con DBSCAN y n_samples=5 y 240 PCA.

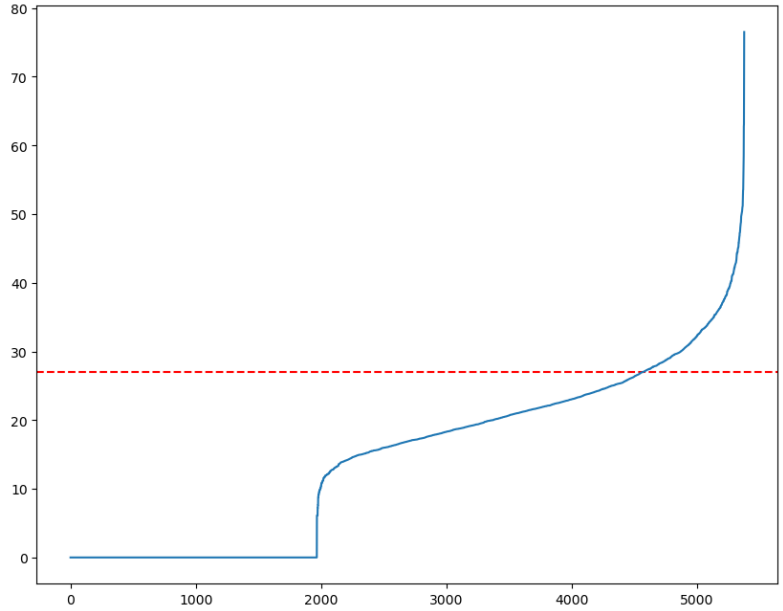


Ilustración 13 – Grafica exponencial con DBSCAN con 1.000 PCA

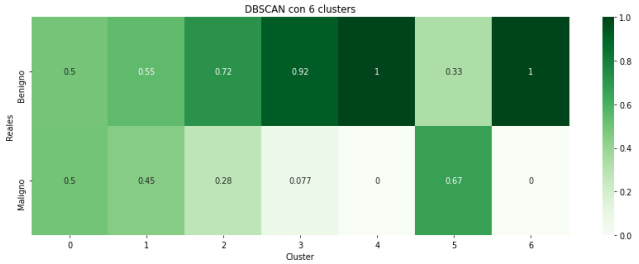


Ilustración 14 – Resultado Parcial de Matrices para 1 clusters con DBSCAN y n_samples=5 y 1.000 PCA.