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

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, 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. Los retos principales se derivan del preprocesamiento de las imágenes (es decir, como se traduce una imagen a números que son interpretables por un computador) y como hacer este preprocesamiento de una forma óptima.

Introducción

Existen evidencias estadísticas que concluyen que el cáncer de piel es el más común entre todos los tipos de cáncer, y que a pesar esto, 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).

Según los estudios también presentados por la Sociedad Americana Contra el Cáncer de Estados Unidos, el tratamiento de esta enfermedad se puede realizar en etapas tempranas únicamente con cirugía, mientras que en etapas más avanzadas se vuelven mucho más difíciles de tratar.

Basados en los datos presentados anteriormente, 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 se realiza el planteamiento de 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 haciendo uso de sus similitudes.

Estado del Arte

Manhas 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, por ejemplo).

Investigaciones anteriores han tenido éxitos considerables reduciendo la dimensionalidad de los datos de entrada, como los obtenidos por (Thomas Martini 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.

Descripción de los datos

Lo inicial es comentar acerca de la base que se va a utilizar para llevar a cabo la implementación de los modelos propuestos anteriormente. Dicha base es presentada por la International Skin Imaging Collaboration (ISIC) la cuál es una asociación que se encarga de recopilar y facilitar el acceso a imágenes de piel para que estas puedan ser usadas por académicos y profesionales en el desarrollo para la detección temprana de melanoma en las personas (ISIC, 2022), y puede ser descargada directamente desde su página web. 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:

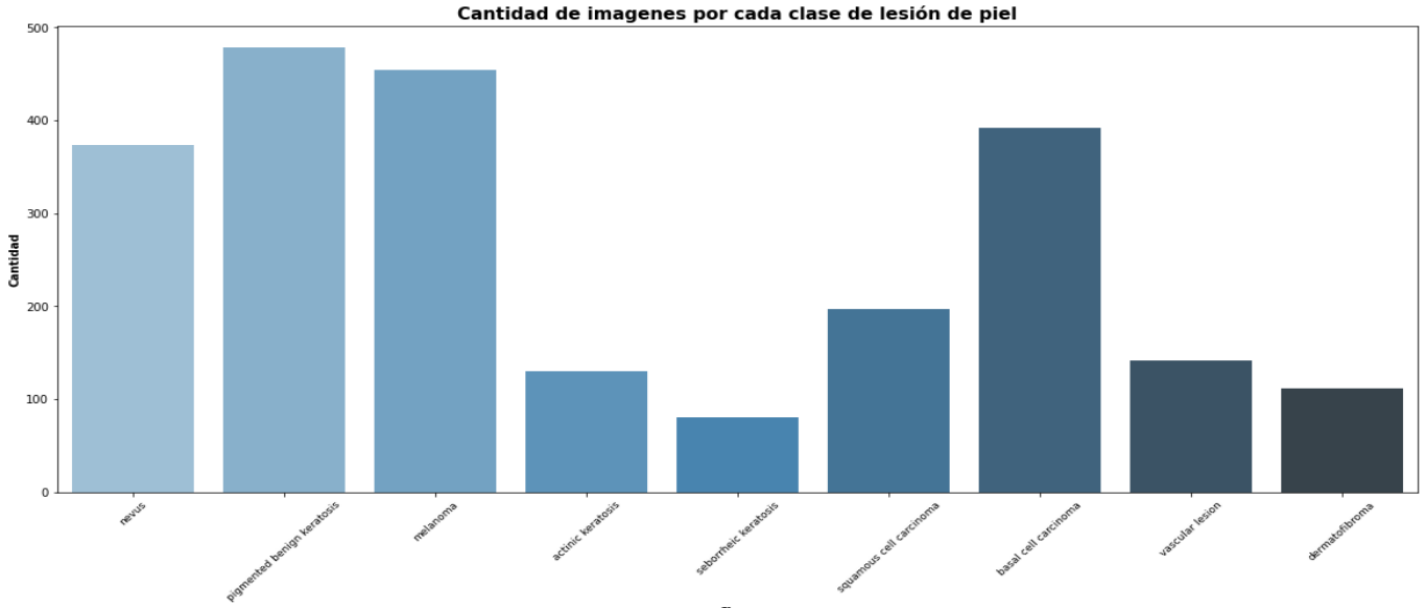

Ilustración 1- Cantidad de imágenes por cada clase de lesión de piel

Ilustración 2 - Ejemplo de Melanoma

Ilustración 3 - Ejemplo de una lesión vascular

Dichas imágenes están compuestas cada una por 3 canales representados en RGB, 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).

Dada la dimensión de datos y para realizar la lectura de las fotos en Python se realizó el apoyo de la función de *image_dataset_from_directory* de la librería *TensorFlow*, la cual permite redimensionar el tamaño de cada imagen y agruparlos en un objeto batch. Dicho lo anterior, las dimensiones con las cuales se decide realizar el cargue de cada de las imágenes es 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).

Propuesta Metodológica

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

La primera motivación 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. Se parte del concepto de que el tamaño que compone cada una de las imágenes es demasiado grande al estar compuestas por un formato RGB, y que, al tener un total de 2.357 imágenes en total, es imposible para un computador tradicional hacer el procesamiento completo del total de los datos. 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 el 5% total de la varianza de las fotos completas, y de esta forma poder correr de forma óptima los diferentes modelos de clustering.

La segunda aplicación propuesta es la implementación de tres modelos diferentes de Clustering, entre los cuales tenemos uno de ***K-Medoides*** (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 para ver qué tan bien se agrupan las imágenes con los diferentes modelos propuestos.

Plan de trabajo

Se realizó un acuerdo entre los integrantes del equipo para la división de las tareas de acuerdo a los roles asignados en la primera entrega. Dichos roles son los que se presentan en la siguiente tabla;

Tabla 1 - Roles de cada integrante.

Rol	Encargado
Ingeniero de Datos (ETL, Depuración, análisis)	Alejandro Murcia
Ingeniero de ML (Modelo, tunning)	Juan Camilo Pérez
Comunicaciones y apoyo (Documentación, video)	Johan Santacruz

Entonces, las asignaciones finales para la primera entrega del proyecto son definidas tal y como se muestra en la siguiente tabla:

Tabla 2 - Separación de actividades

Actividades Primera Entrega del Proyecto			
Actividad	**Descripción**	**Responsable**	**Estado**
1	Exploración de ideas	Equipo	Completo
2	Documentación de las propuestas y exploración inicial de dataset	Equipo	Completo
3	Definición de Roles	Equipo	Completo
4	Selección de una idea para ser abordada como proyecto de curso	Equipo	Completo
5	Detalle de la idea a ser desarrollada	Juan Camilo	Completo
6	Definición de las actividades de la primera entrega del proyecto	Johan	Completo
7	Revisión preliminar de antecedentes en la literatura	Juan Camilo	Completo
8	Descripción detallada de los datos	Alejandro	Completo
9	Elaboración del documento final de propuesta	Johan	Completo
10	Configuración del Repositorio de GitHub	Johan	Completo

El plan propuesto para la entrega final del curso es el siguiente:

Tabla 3 - Separación de actividades

Actividades Primera Entrega del Proyecto			
Actividad	**Descripción**	**Responsable**	**Estado**
1	Calibración de modelos de clustering y reducción	Equipo	Pendiente
2	Selección de los mejores modelos	Juan Camilo	Pendiente

3	Documento final del proyecto	Johan y Alejandro	Pendiente
4	Vídeo Exposición de los resultados obtenidos	Johan y Alejandro	Pendiente

Repositorio

Para llevar registro de los resultados obtenidos y del código creado para este proyecto, se ha configurado el siguiente repositorio que será alimentado en la medida que el equipo logré progresos en los objetivos propuestos:

https://github.com/jsantacruzl/AprendizajeNoSupervisado_G9

Se crea un archivo .readme que facilita la navegación y uso de este por parte del equipo de trabajo y público en general.

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