Entrega 1 – Propuesta Proyecto Final del Curso Grupo 9

Motivación

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.

Aplicación de ML

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 Maquina 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 (de lo cual hablaremos más adelante). Para no profundizar aún en este tema, 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 clasificación.

La segunda aplicación propuesta es la implementación de un modelo de ***K-Medoides*** para intentar 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 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.

Datos para el desarrollo

Lo inicial es comentar acerca de la base que se va a utilizar para llevar a cabo la implementación de los modelos presentados anteriormente. Dicha base es presentada por la International Skin Imaging Collaboration, y está compuesta por un total de 2.357 imágenes clasificadas en los diferentes tipos de posibles cánceres de piel, sin embargo, como se comentó anteriormente esta clasificación no se va a tener en cuenta dados los supuestos propuestos de los requerimientos para el etiquetado manual de las fotos.

Tabla 1 - Tipos de cáncer de piel en la base

Tipo de cáncer
Actinic keratosis
Basal cell carcinoma
Dermatofibroma
Melanoma
Nevus
Pigmented benign keratosis
Seborrheic keratosis
Squamous cell carcinoma
Vascular lesión

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. Es posible entonces afirmar que, el dataset resultante tiene un tamaño de (2.357, 150.528).

A continuación, se presentan algunos ejemplos para las fotos que se pueden encontrar dentro del dataset;

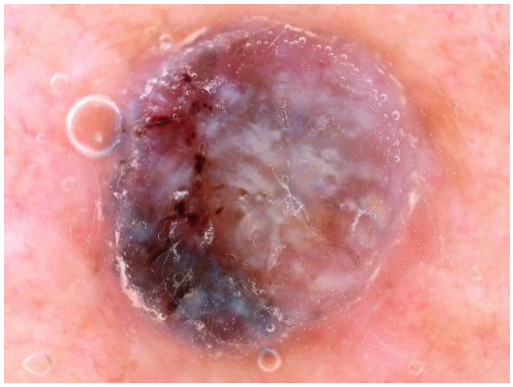

Ilustración 1 - Ejemplo de Melanoma

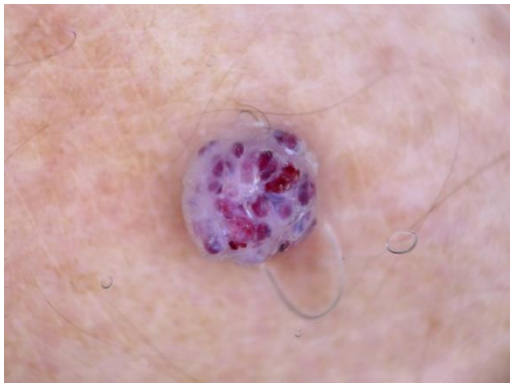

Ilustración 2 - Ejemplo de una lesión vascular

El dataset es posible de obtener completo mediante la aplicación del siguiente comando; *kaggle kernels output vsr123456/skin-cancer-detection-vikas-sharma -p /path/to/dest*

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 2 - 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 3 - 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	Pendiente
8	Descripción detallada de los datos	Alejandro	Pendiente
9	Elaboración del documento final de propuesta	Johan	Pendiente
10	Configuración del Repositorio de GitHub	Johan	Pendiente

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