Implementación de un modelo de Machine Learning para la detección y clasificación de cáncer de piel.

Universidad Autónoma de Chihuahua.
Facultad de Ingeniería.
2025.

Asignatura: Data Science.
Profesor: Manuel Alberto Chávez Salcido.

Alumno: Rodrigo García Núñez.

UNIVERSIDAD AUTÓNOMA DE CHIHUAHUA

Título: Implementación de un modelo de Machine Learning para la detección y clasificación de cáncer de piel.

Dominio del Problema:

Objetivos

- Desarrollar un sistema de clasificación automática de lesiones de cáncer de piel utilizando técnicas de Machine Learning, capaz de distinguir entre lesiones benignas y malignas.
- Recopilar y preprocesar un conjunto de datos conformados por imágenes de lesiones de piel de alta calidad y datos clínicos relevantes.
- Implementar y entrenar modelos de aprendizaje profundo, como redes neuronales convolucionales (CNN), para la clasificación de lesiones de piel.
- Evaluar el rendimiento de los modelos utilizando métricas como precisión, sensibilidad y especificidad.

Descripción del problema

El cáncer de piel es una de las formas de cáncer más comunes a nivel mundial, y su detección temprana es crucial para aumentar las tasas de supervivencia. El diagnóstico tradicional de cáncer de piel se basa en la evaluación visual de lesiones cutáneas por parte de dermatólogos, seguida de biopsias en casos sospechosos. Sin embargo, este proceso presenta varios desafíos:

- Subjetividad y variabilidad:
    - La evaluación visual puede ser subjetiva y variar entre diferentes profesionales de la salud, lo que lleva a diagnósticos inconsistentes.

- Dificultad en la diferenciación:
    - La diferenciación visual entre lesiones benignas y malignas puede ser difícil, especialmente en etapas tempranas.

- Aumento de la incidencia:
    - El aumento de la incidencia del cáncer de piel genera una carga mayor sobre los profesionales de la salud.

Estos desafíos resaltan la necesidad de desarrollar herramientas de diagnóstico más precisas, objetivas y accesibles. La aplicación de técnicas de Machine Learning, tiene el potencial de abordar estos desafíos al:

- Automatizar el análisis de imágenes y datos clínicos:
    - Permite el análisis rápido y objetivo de grandes volúmenes de datos de lesiones cutáneas, incluyendo imágenes y datos clínicos relevantes.

- Mejorar la precisión del diagnóstico:
    - Los modelos de Machine Learning pueden aprender patrones complejos y sutiles que son difíciles de detectar por el ojo humano.

- Facilitar el diagnóstico temprano:
    - Al proporcionar una herramienta de apoyo para los profesionales de la salud, se puede acelerar el proceso de diagnóstico y tratamiento.

Por lo tanto, este proyecto busca abordar el problema de la detección temprana y precisa del cáncer de piel mediante el desarrollo e implementación de un modelo de Machine Learning que pueda clasificar automáticamente las lesiones cutáneas a partir de datos conformados por imágenes y datos clínicos relevantes sobre. Cabe mencionar que este sistema no busca reemplazar el trabajo de profesionales de la salúd, sino ser un sistema auxiliar en el proceso de diagnóstico de lesiones cutáneas.

Aplicaciones Similares

Como se mencionó anteriormente, el problema de detección de cáncer de piel resulta en un buen desafío, por lo que ya hay aplicaciones que abordan este problema utilizando técnicas de aprendizaje automático, de entre las cuales se encuentran:

- **SkinVision:**
    - Esta aplicación utiliza algoritmos de inteligencia artificial para analizar fotos de lesiones cutáneas y proporcionar una evaluación de riesgo de cáncer de piel.
- **DermaCam:**
    - Desarrollada en México, esta aplicación móvil se enfoca en la detección de melanomas mediante el análisis de la coloración y otras características de la piel usando inteligencia artificial.
- **Medic Scanner:**
    - Esta aplicación ayuda a analizar los lunares en la piel para los tipos más comunes de cáncer de piel, incluido el melanoma.
- **FotoSkin:**

- Aplicación gratuita desarrollada por investigadores de la Universidad de Alcalá, que tiene por objetivo dar respuesta a una de las mayores necesidades de los médicos para mejorar la precisión en el diagnóstico precoz del cáncer de piel, conocer su evolución.

- **SkinScreener:**
  - SkinScreener utiliza IA para identificar el cáncer de piel con una precisión excepcional, brindando a personas en zonas rurales o con recursos limitados la oportunidad de detectar cambios malignos en la piel en fases tempranas.

Propuesta de Solución usando Ciencia de Datos.

1. Modelo de Aprendizaje Profundo:

- **Arquitectura:**
  - Se implementará una Red Neuronal Convolucional (CNN) debido a su eficacia comprobada en el análisis de imágenes.
  - Se explorarán modelos como ResNet, EfficientNet o DenseNet, ajustando la profundidad y complejidad según el rendimiento en el dataset específico.

2. Conjunto de Datos:

- **Fuentes:**
  - Se utilizará el dataset HAM10000 de ISIC Archive, pues es un dataset muy extenso y detallado que cuenta con imágenes e información representativa para el diagnóstico de lesiones de la piel.
- **Preprocesamiento:**
  - Las imágenes se normalizarán para uniformizar el conjunto de datos.
  - Los datos clínicos serán sometidos a un proceso de limpieza de datos para ser ingresados al modelo.

3. Métricas de Evaluación:

- **Precisión, Sensibilidad y Especificidad:**
  - Estas métricas medirán la capacidad del modelo para clasificar correctamente las lesiones benignas y malignas.
- **Matriz de Confusión:**
  - Esta matriz permitirá visualizar los falsos positivos y falsos negativos, lo que ayudará a identificar áreas de mejora.

4. Herramientas y Tecnologías:

- **Python:**
    - Se utilizará Python como lenguaje de programación principal debido a su amplia disponibilidad de bibliotecas de ciencia de datos.
- **TensorFlow o PyTorch:**
    - Se utilizarán estas bibliotecas de aprendizaje profundo para implementar y entrenar el modelo CNN.
- **Keras:**
    - Keras será utilizado como una interfaz de alto nivel para tensorflow o pytorch, para facilitar el desarrollo del modelo.
- **OpenCV:**
    - Se utilizará esta librería para el preprocesamiento de las imágenes.

6. Interfaz de usuario:

- Se creará una interfaz de usuario en donde los usuarios puedan ingresar las imágenes, y el modelo les entregue el resultado y la probabilidad de que la lesión sea maligna.

Tipos de datos necesarios.

Para desarrollar un modelo de Machine Learning efectivo para la detección y clasificación del cáncer de piel, se necesitan varios tipos de datos, tanto de imágenes como clínicos. Los datos necesarios para el proyecto son:

1. Datos de imágenes dermatoscopias:

- Imágenes de alta resolución:
    - Imágenes claras y nítidas de lesiones cutáneas, capturadas con dermatoscopios.
    - Es crucial que las imágenes tengan una resolución suficiente para permitir la identificación de detalles finos, como estructuras vasculares y patrones de pigmentación.

- Variedad de lesiones:
    - Imágenes de una amplia gama de lesiones cutáneas, incluyendo melanomas, carcinomas basocelulares, carcinomas de células escamosas y lesiones benignas como nevos y queratosis seborreicas.
    - Esta variedad es esencial para entrenar un modelo robusto que pueda diferenciar entre diferentes tipos de lesiones.

- Imágenes con anotaciones:
    - Cada imagen debe estar etiquetada con la clasificación correcta de la lesión (melanoma, benigna).

- Las anotaciones pueden incluir también la segmentación de la lesión, delimitando el área exacta de la lesión en la imagen.

2. Datos clínicos:

- Información del paciente:
  - Edad, sexo y etnia del paciente.
  - Análisis Sanguíneos.
  - Historial médico, incluyendo antecedentes de cáncer de piel.
  - Factores de riesgo, como exposición al sol y antecedentes familiares de melanoma.

- Características de la lesión:
  - Ubicación de la lesión en el cuerpo.
  - Tamaño, forma y color de la lesión.
  - Presencia de características específicas, como asimetría, bordes irregulares y cambios en el color.

- Resultados de biopsias:
  - Información de las biopsias de lesiones sospechosas, que proporcionan un diagnóstico definitivo.
  - Esta información es crucial para validar la precisión del modelo de Machine Learning.

Tipo de algoritmo a implementar

Para la detección y clasificación de cáncer de piel a partir de imágenes dermatoscopias, los algoritmos de aprendizaje profundo, y en particular las redes neuronales convolucionales (CNN), han demostrado ser los más efectivos dadas las siguientes características:

- Extracción automática de características:
  - Las CNNs son especialmente buenas para analizar imágenes porque pueden aprender automáticamente a identificar patrones y características relevantes, como bordes, texturas y formas, que son cruciales para distinguir entre lesiones benignas y malignas.

- Alto rendimiento:
  - Las CNNs convolucionales han logrado una precisión impresionante en tareas de clasificación de imágenes.

- Capacidad de aprendizaje jerárquico:

- Las CNNs pueden aprender características jerárquicas, lo que significa que pueden identificar patrones simples en las primeras capas y luego combinar esos patrones para identificar características más complejas en las capas posteriores.

CNN's Recomendables:

- ResNet (Redes Residuales):
  - ResNet es conocida por su capacidad para entrenar redes muy profundas, lo que permite aprender características muy complejas.
  - Las redes ResNet evitan el problema de la "desaparición del gradiente", lo que les permite mantener un buen rendimiento incluso con muchas capas.

- EfficientNet:
  - EfficientNet es una familia de modelos que se centra en lograr un equilibrio óptimo entre precisión y eficiencia computacional.
  - Estos modelos son muy eficientes en el uso de recursos, lo que los hace adecuados para aplicaciones móviles o dispositivos con recursos limitados.

- DenseNet (Redes Densamente Conectadas):
  - DenseNet conecta cada capa con todas las capas anteriores, lo que mejora el flujo de información y reduce el problema de la "desaparición del gradiente".
  - DenseNet tiende a tener un buen rendimiento con conjuntos de datos más pequeños.

En resumen, las redes neuronales convolucionales son los algoritmos más recomendables para el problema de detección y clasificación de cáncer de piel por su capacidad de análisis de imágenes.

Tipo de aprendizaje a Utilizar

El aprendizaje supervisado es preferible en el contexto de la detección y clasificación del cáncer de piel debido a su capacidad para aprender de datos etiquetados y generar predicciones precisas.
1. Precisión y exactitud:

- Datos etiquetados:
  - El aprendizaje supervisado se basa en conjuntos de datos etiquetados, donde cada imagen de lesión cutánea está asociada con un diagnóstico confirmado (por ejemplo, melanoma, benigno).

- Esto permite al modelo aprender la relación directa entre las características de la imagen y el diagnóstico correspondiente.

- Predicciones precisas:
  - Al entrenar con datos etiquetados, el modelo puede aprender a reconocer patrones sutiles y complejos que son indicativos de lesiones malignas.
  - Esto conduce a predicciones más precisas y confiables en comparación con el aprendizaje no supervisado, que no tiene acceso a esta información de diagnóstico.

2. Clasificación y diagnóstico:

- Tareas de clasificación:
  - La detección y clasificación del cáncer de piel son inherentemente tareas de clasificación, donde el objetivo es asignar una etiqueta de clase (por ejemplo, maligno, benigno) a una imagen de lesión cutánea.
  - El aprendizaje supervisado es especialmente adecuado para este tipo de tareas, ya que puede aprender a discriminar entre diferentes clases de lesiones.

- Apoyo al diagnóstico:
  - El aprendizaje supervisado puede proporcionar una herramienta de apoyo valiosa para los dermatólogos, ayudándoles a realizar diagnósticos más precisos y consistentes.
  - Esto es especialmente importante en la detección temprana del cáncer de piel, donde el diagnóstico preciso es crucial para mejorar las tasas de supervivencia.

3. Validación y evaluación:

- Métricas de evaluación:
  - El aprendizaje supervisado permite el uso de métricas de evaluación bien definidas, como precisión, sensibilidad, especificidad y área bajo la curva ROC (AUC), para medir el rendimiento del modelo.
  - Estas métricas proporcionan una evaluación objetiva de la precisión del modelo y su capacidad para generalizar a nuevos datos.

- Validación cruzada:
  - El aprendizaje supervisado facilita el uso de técnicas de validación cruzada, que ayudan a evitar el sobreajuste y garantizan que el modelo sea capaz de generalizar datos no vistos.

En resumen, el aprendizaje supervisado es la mejor opción para la detección y clasificación del cáncer de piel debido a su capacidad para aprender de datos etiquetados, generar predicciones precisas y proporcionar una herramienta de apoyo valiosa para los profesionales de la salud.

Comprensión de Datos

Para el desarrollo de este proyecto, se emplearán los conjuntos de datos HAM10000 de ISIC Archive, el cual representa una vasta colección de datos necesarios para la implementación del modelo de Machine Learning destinado a la clasificación de diferentes diagnósticos de cáncer de piel. Se aplicarán las siguientes estrategias para garantizar la calidad y eficacia en el uso de los datos:

1. Calidad de los Datos:

- **Verificación y Limpieza:**
  - Se realizará una revisión exhaustiva del conjunto de datos para identificar y corregir posibles errores, inconsistencias o valores atípicos.
  - Se implementarán técnicas de limpieza de datos para manejar valores faltantes y garantizar la coherencia de la información.
- **Anotaciones y Metadatos:**
  - Se validarán las anotaciones de las lesiones para asegurar su precisión.
  - Se analizarán los metadatos disponibles (edad, sexo, ubicación de la lesión, etc.) para evaluar su relevancia y utilidad en el modelo.

2. División del Conjunto de Datos:

- **Partición Estratificada:**
  - Se dividirá el conjunto de datos en tres subconjuntos: entrenamiento, validación y prueba.
  - Se utilizará una partición estratificada para asegurar que cada subconjunto mantenga la distribución de clases original.

- **Propósito de Cada Subconjunto:**
  - **Entrenamiento:** Se utilizará para entrenar el modelo de Machine Learning.
  - **Validación:** Se utilizará para ajustar los hiperparámetros del modelo y evaluar su rendimiento durante el entrenamiento.

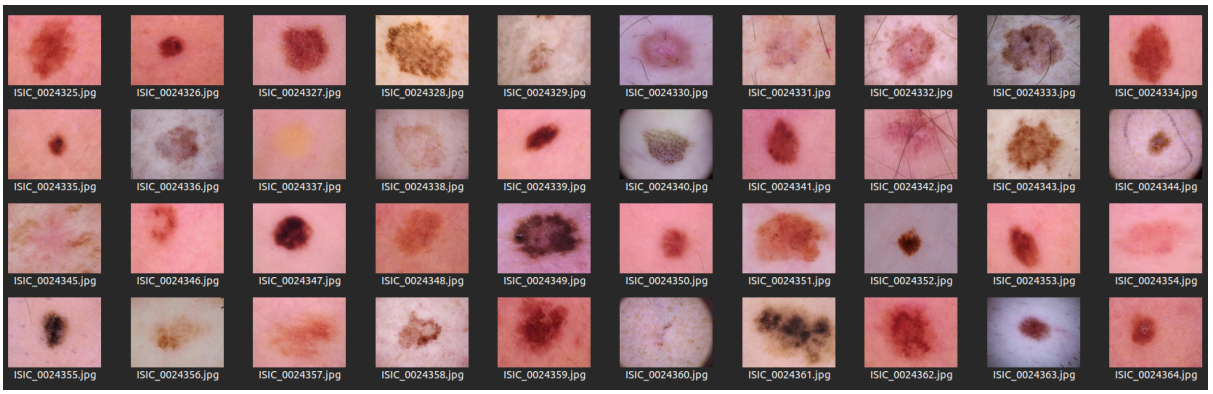
- **Prueba:** Se utilizará para evaluar el rendimiento final del modelo con datos no vistos.

3. Balance de Clases:

- **Análisis de Distribución:**
  - Se analizará la distribución de las clases que representan los diagnósticos de cáncer de piel para determinar si existe un desbalance significativo.
- **Técnicas de Balanceo:**
  - En caso de desbalance, se aplicarán técnicas como:
    - **Sobremuestreo (Oversampling):** Generación de muestras sintéticas de la clase minoritaria (por ejemplo, SMOTE).
    - **Submuestreo (Undersampling):** Reducción de muestras de la clase mayoritaria.
    - **Ponderación de Clases:** Asignación de pesos diferentes a las clases durante el entrenamiento.

Recopilación de Datos

Se tiene una tabla compuesta de 17 atributos y 11, 720 registros; cada registro asociado mediante id a una imágen. DataSet recuperado el 11 de marzo de 2025. El DataSet es proporcionado por The International Skin Imaging Collaboration (ISIC), academia destinada al uso de imágenes digitales relacionadas al Cáncer de Piel, con el fin de apoyar el desarrollo de tecnologías dedicadas a reducir la mortalidad del Cáncer de Piel. Enlace al DataSet: isic-archive.



Ejemplo de imágenes relacionadas por ID al DatatSet.

Descripción de Datos

Atributo	Tipo de Dato	Descripción	Entrda/Salida
isic_id	String	Identificación única de la lesión cutánea del registro de ISIC. Este id asocia los metadatos con las imágenes.	Entrada
age_approx	Entero	Edad aproximada. Rango de edades que va desde los 5 a los 85 años.	Entrada
anatom_site_general	String/categórico	Sitio general en el que se halló la lesión.	Entrada
anatom_site_special	String/categórico	Sitio particular en el que se halló la lesión.	Entrada
bening_malignant	String/categórico	Diagnóstico general de la lesión cutanea	salida
concomitant_biopsy	Booleano	Si se extrajo o no la lesión para realizar biopsia.	Entrada
diagnosis	String/categórico	Diagnóstico específico definitivo de la lesión.	Salida
diagnosis_1	String/categórico	Primera instancia de diagnóstico.	Entrada
diagnosis_2	String/categórico	Segunda instancia de diagnóstico.	Entrada
diagnosis_3	String/categórico	Tercera instancia de diagnóstico.	Entrada
diagnosis_confirm_type	String/categórico	Método utilizado para confirmar el	Entrada

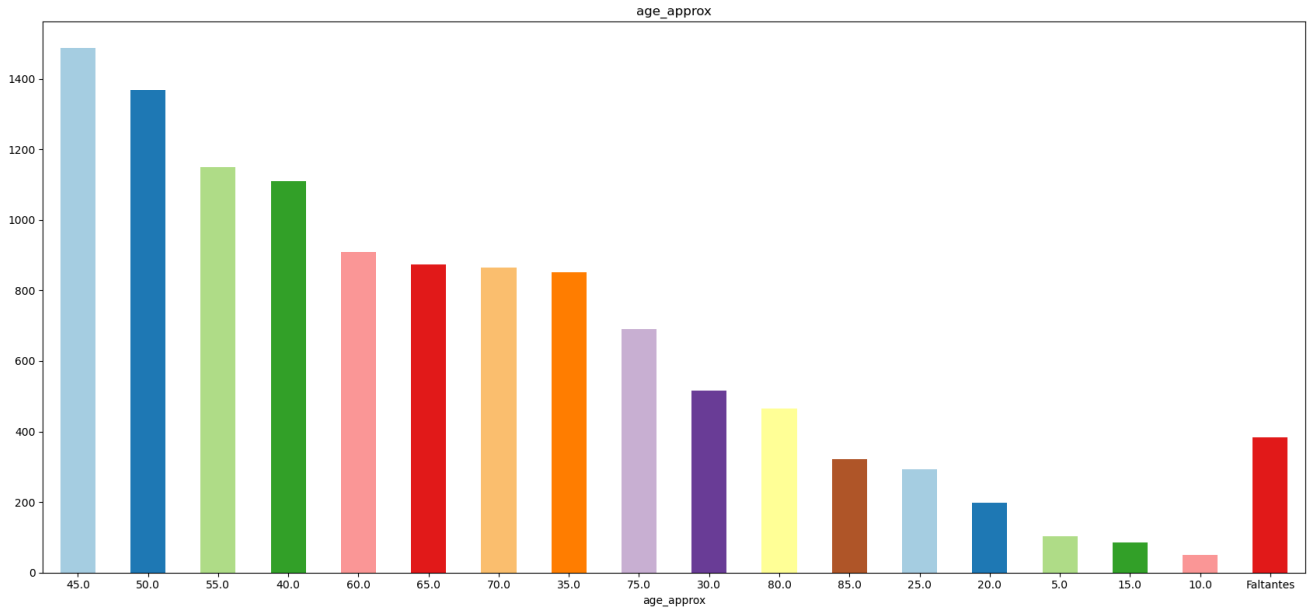
		diagnóstico.	
image_type	String	Imágen Dermatoscopica con la que se presenta la lesión cutánea.	Entrada
lesion_id	String	Identificador único dentro del DataSet Ham10000.	Entrada
melanocytic	Booleano	Indica si la lesión está compuesta o no de melanocitos	Entrada
sex	String/categórico	Género del paciente	Entrada

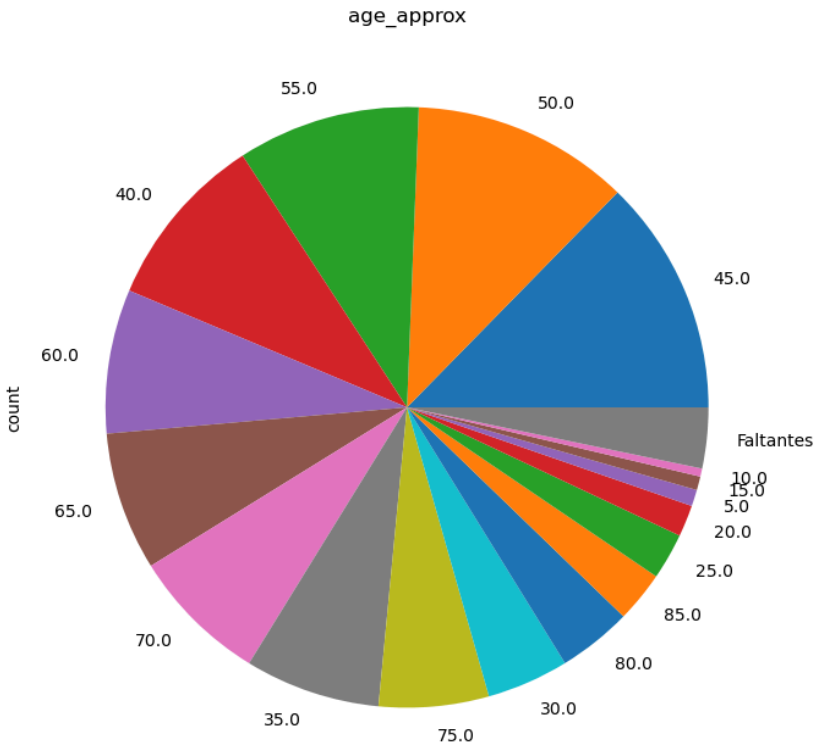
Tabla de descripción de Columnas.

Exploración de Datos

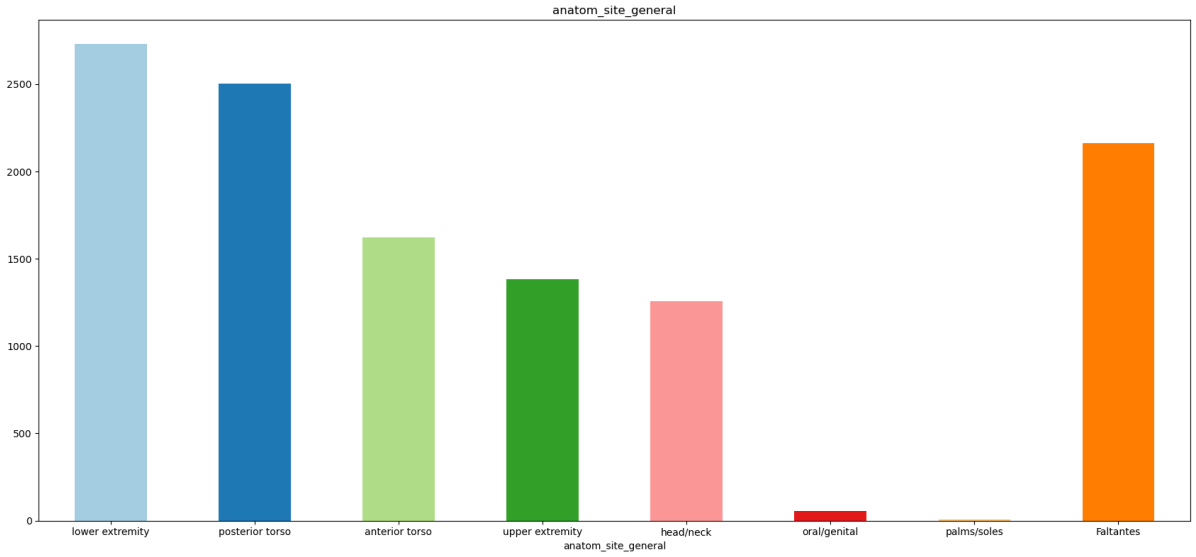
A continuación se graficaron todas las columnas del dataset mediante gráficos de barras y de pastel, con el fin de observar la distribución de los datos por columna.

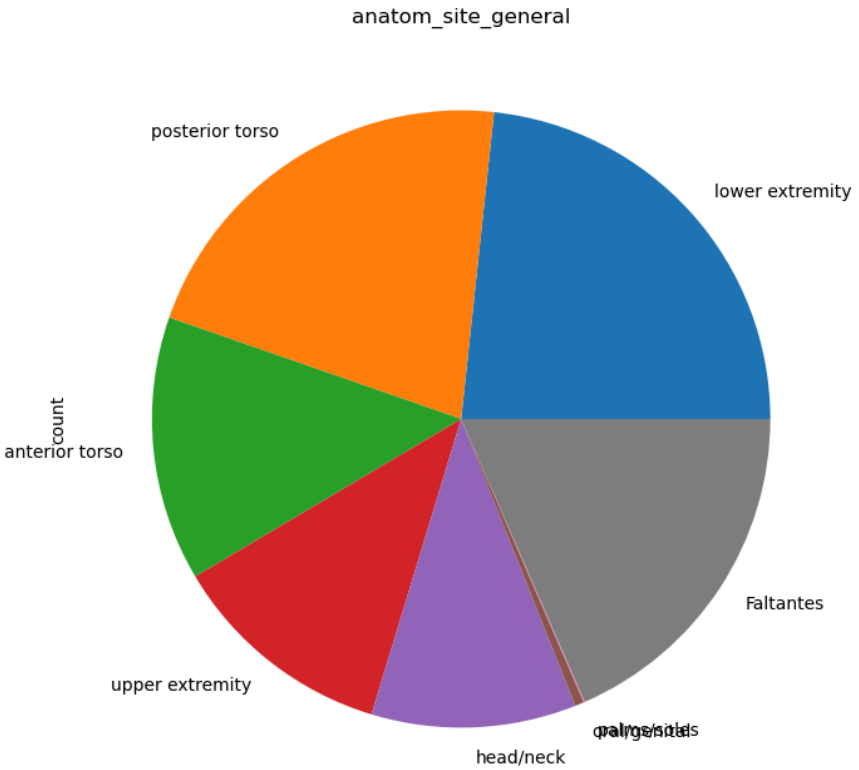
age_approx: 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80 u 85 años


age_approx
1400
1200
1000
800
600
400
200
0
45.0
50.0
55.0
40.0
60.0
65.0
70.0
35.0
75.0
30.0
80.0
85.0
25.0
20.0
5.0
15.0
10.0
Faltantes
age_approx

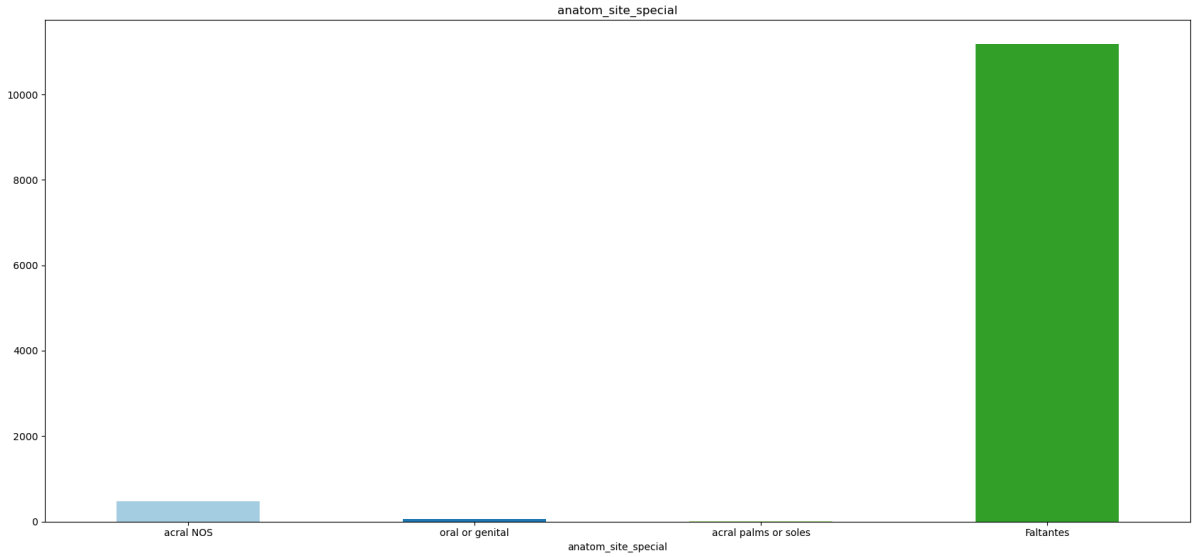

age_approx
count
55.0
50.0
40.0
45.0
60.0
Faltantes
10.0
15.0
5.0
20.0
65.0
25.0
85.0
70.0
80.0
35.0
30.0
75.0

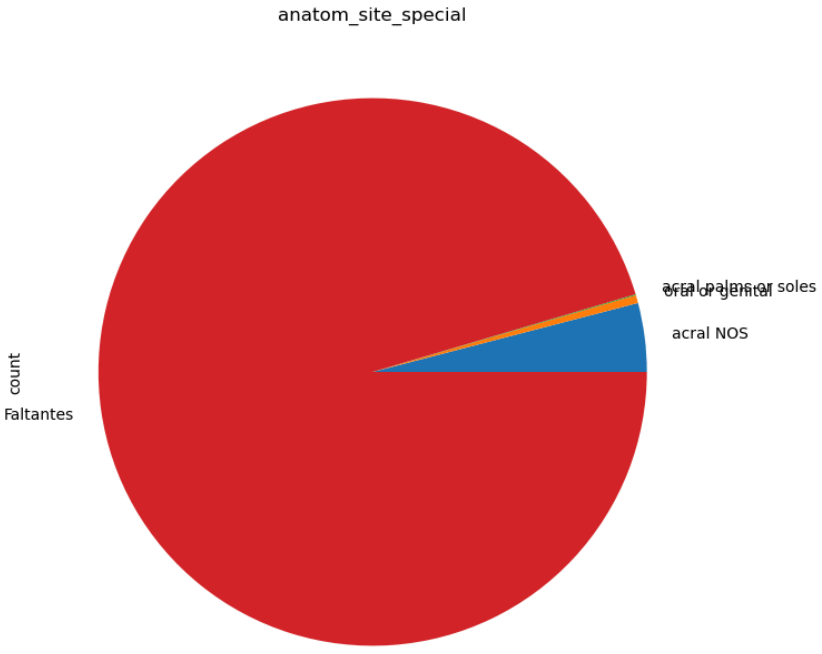
anatom_site_general: lower extremity, posterior torso, anterior torso, extremidad superior, head/neck, oral/genital, palms/soles.



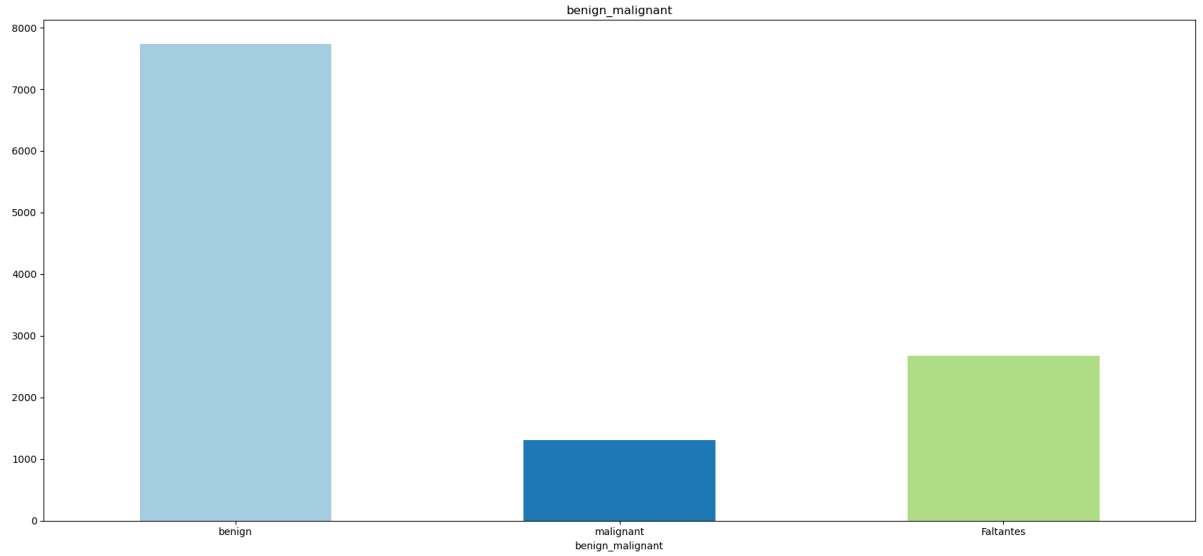


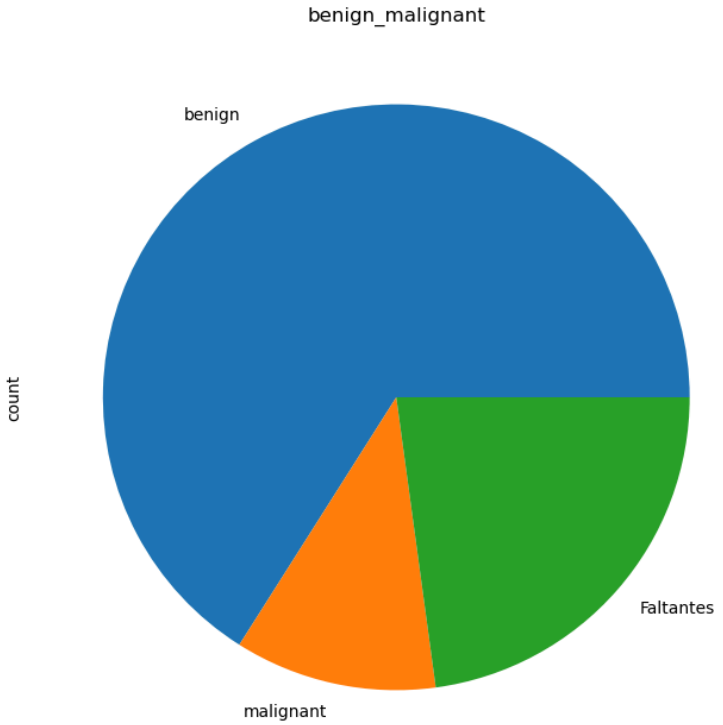
anatom_site_especial: acral_NOS (Not Otherwise Specified), oral or genital, acral palms or soles.



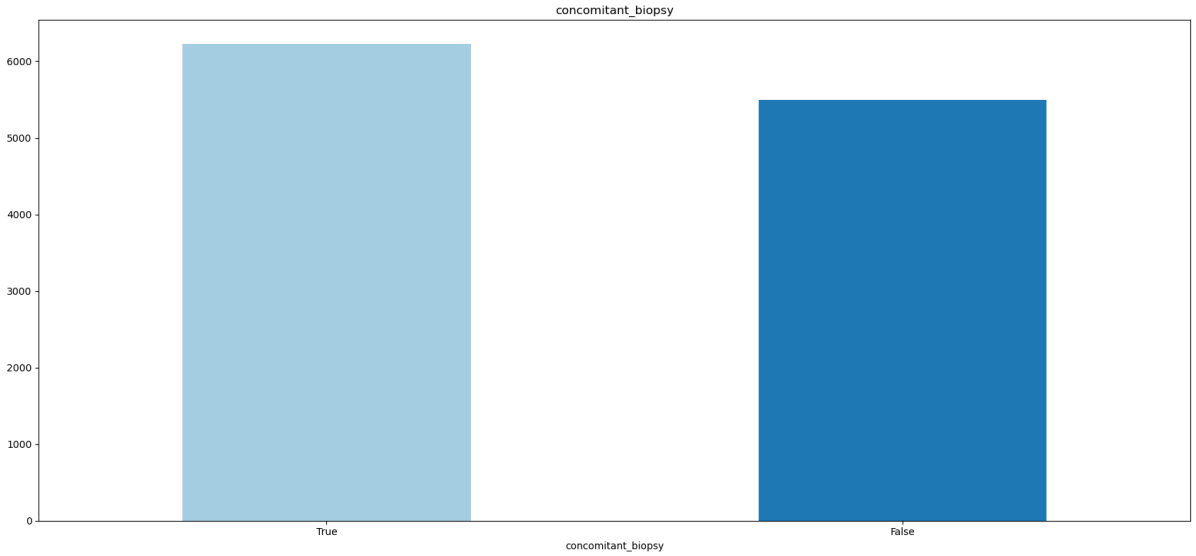


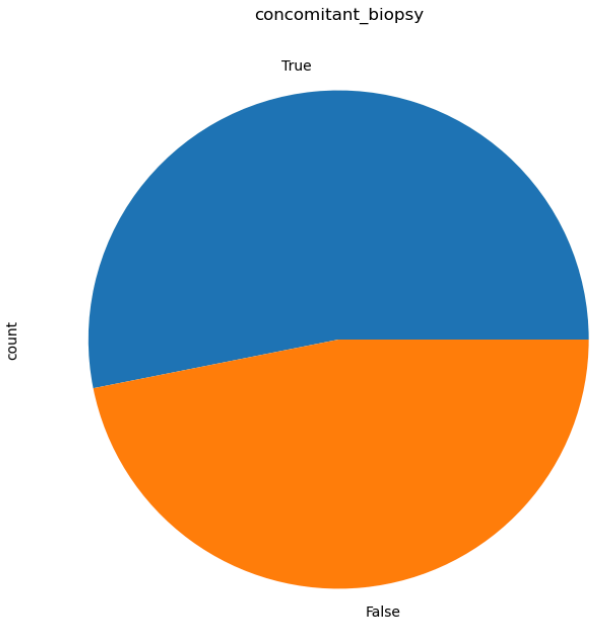
bening_malignant: bening, malignant.


benign_malignant
8000
7000
6000
5000
4000
3000
2000
1000
0
benign
malignant
Faltantes
benign_malignant

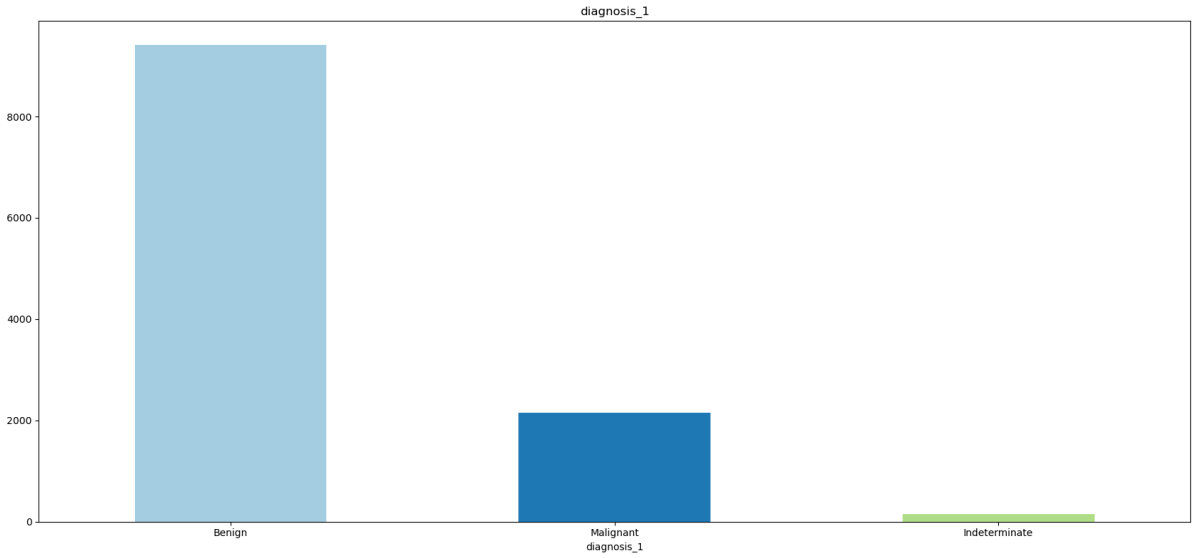

benign_malignant
benign
count
Faltantes
malignant

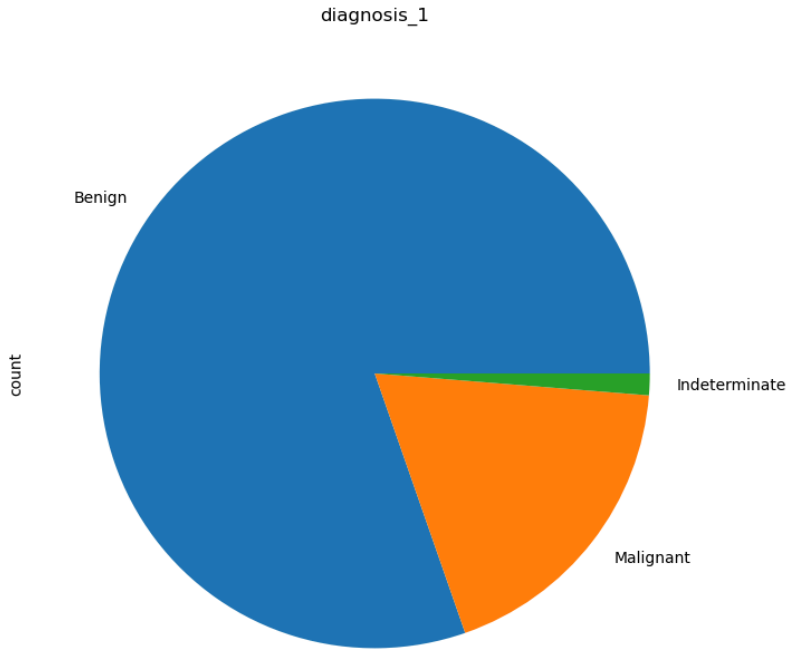
concomitant_biopsy: True, False



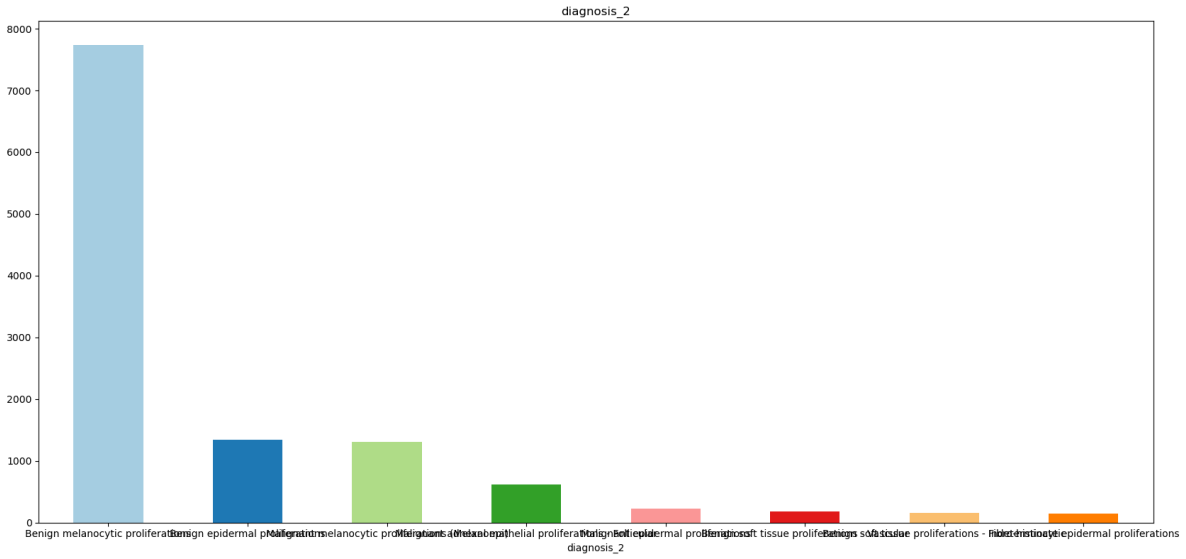


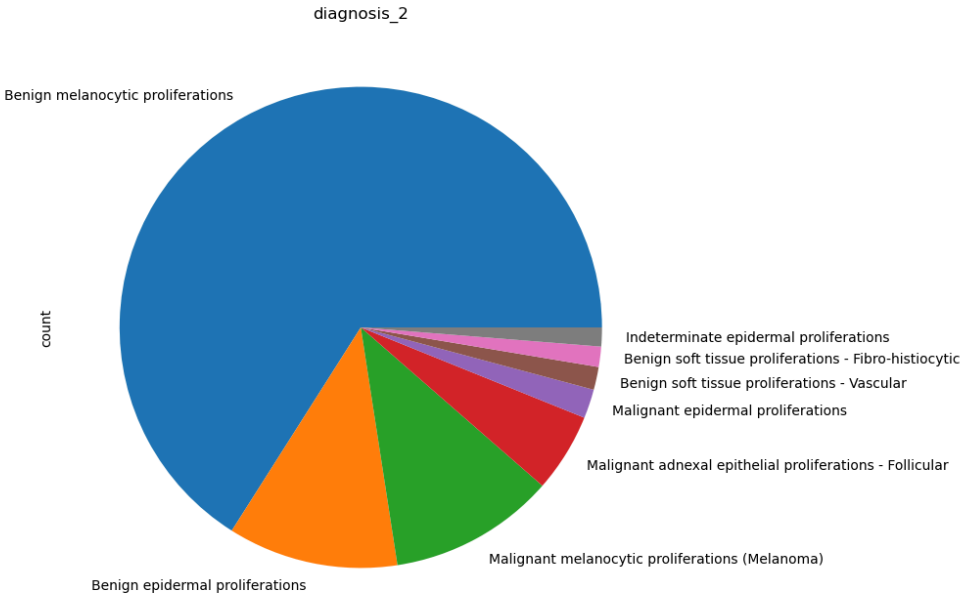
diagnosis_1: Benign, Malignant, indeterminate.



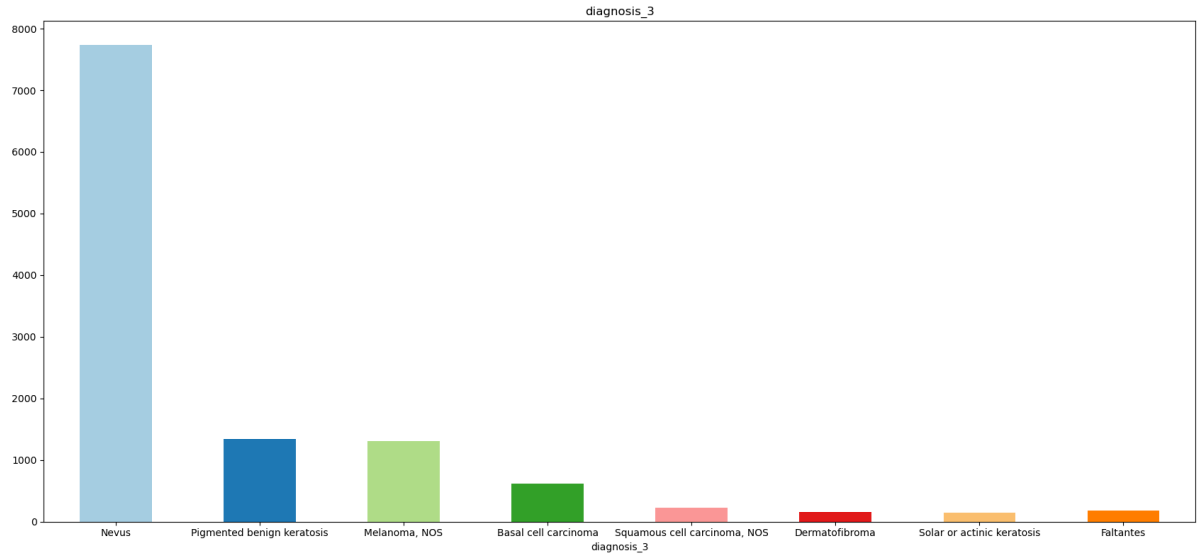


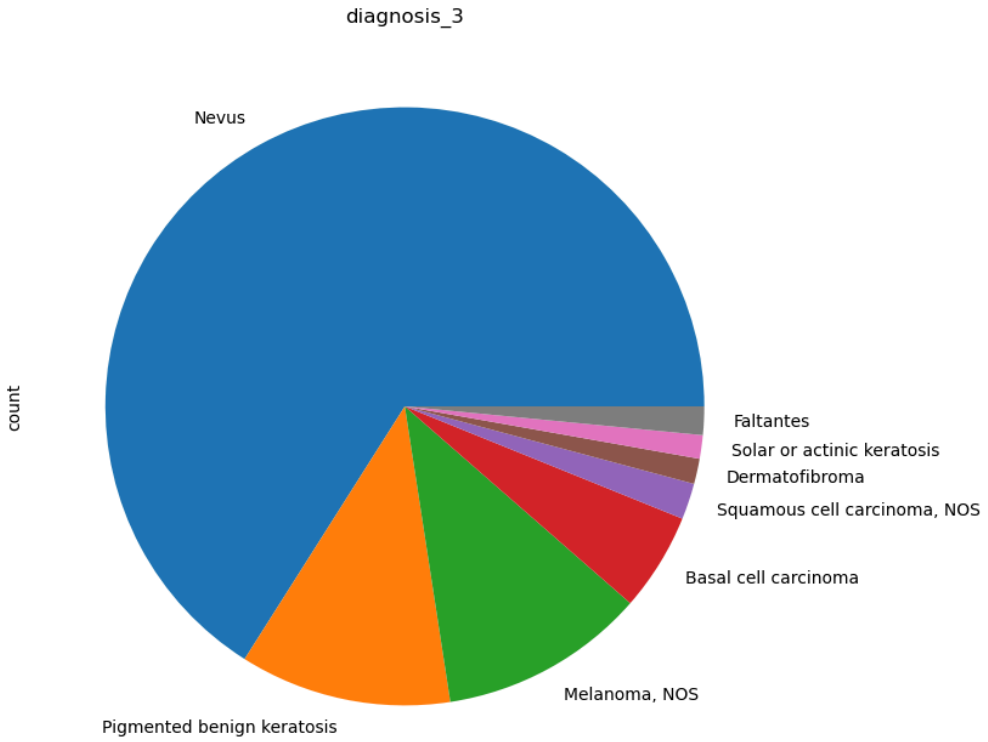
diagnosis_2: Benign melanocytic proliferations, Malignant melanocytic proliferations (Melanoma), Benign epidermal proliferations, Benign soft tissue proliferations - Fibro-histiocytic, Malignant epidermal proliferations, Malignant adnexal epithelial proliferations - Follicular, Benign soft tissue proliferations - Vascular, Indeterminate epidermal proliferations.



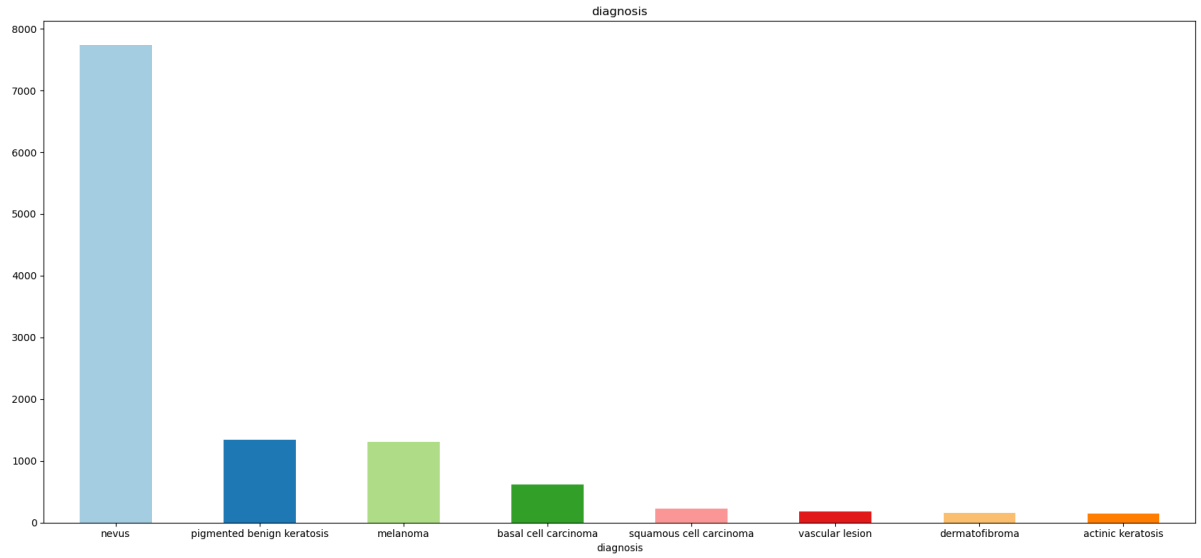


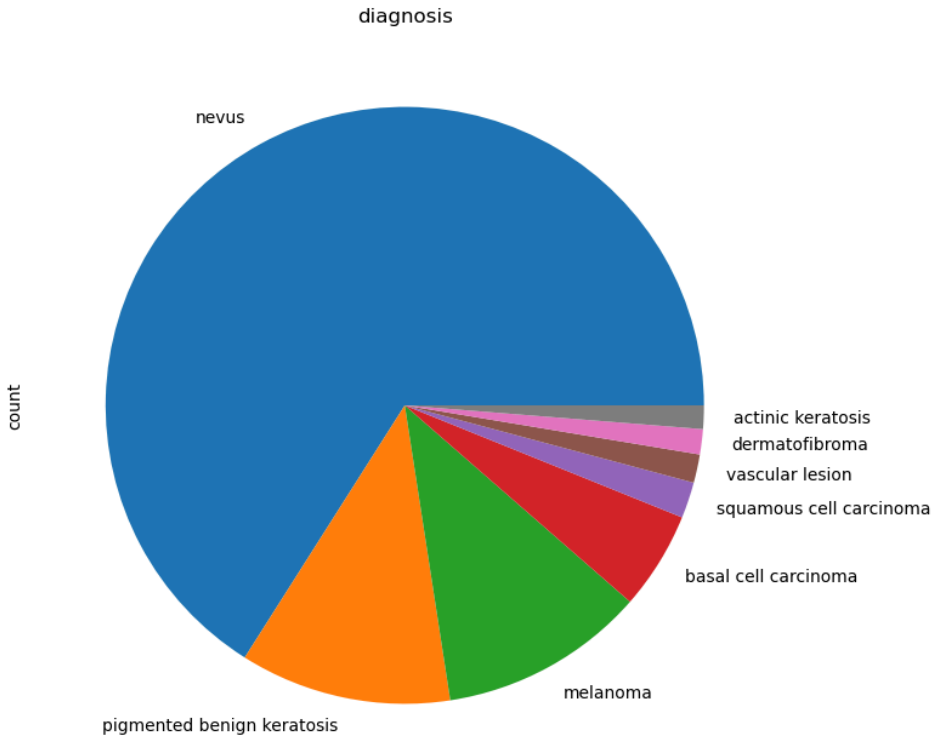
diagnosis_3: Nevus, Melanoma NOS, Pigmented benign keratosis, Dermatofibroma, Squamous cell carcinoma NOS, Basal cell carcinoma, Solar or actinic keratosis'



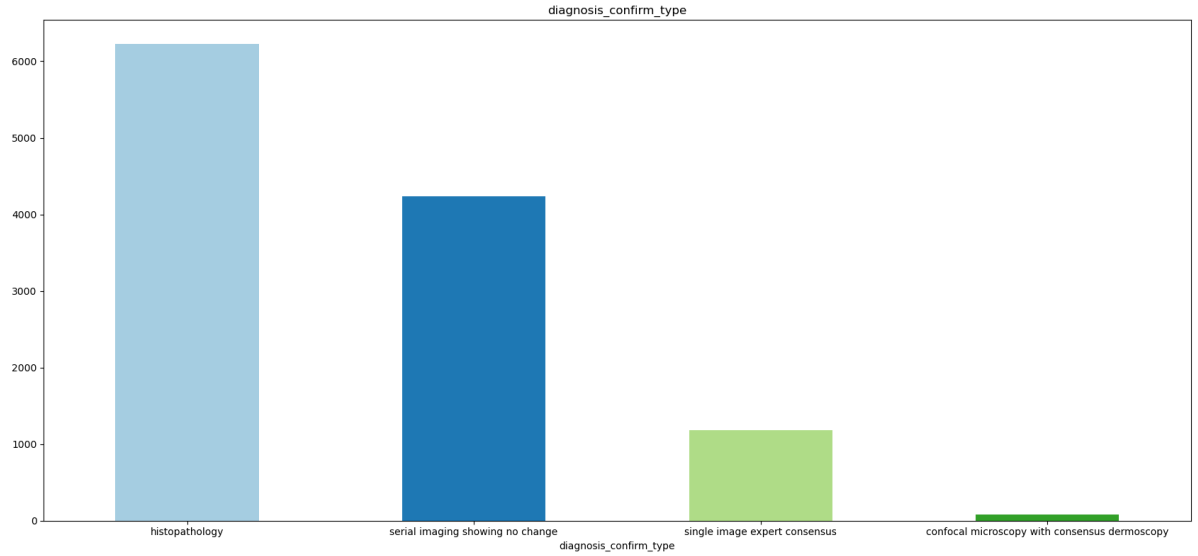


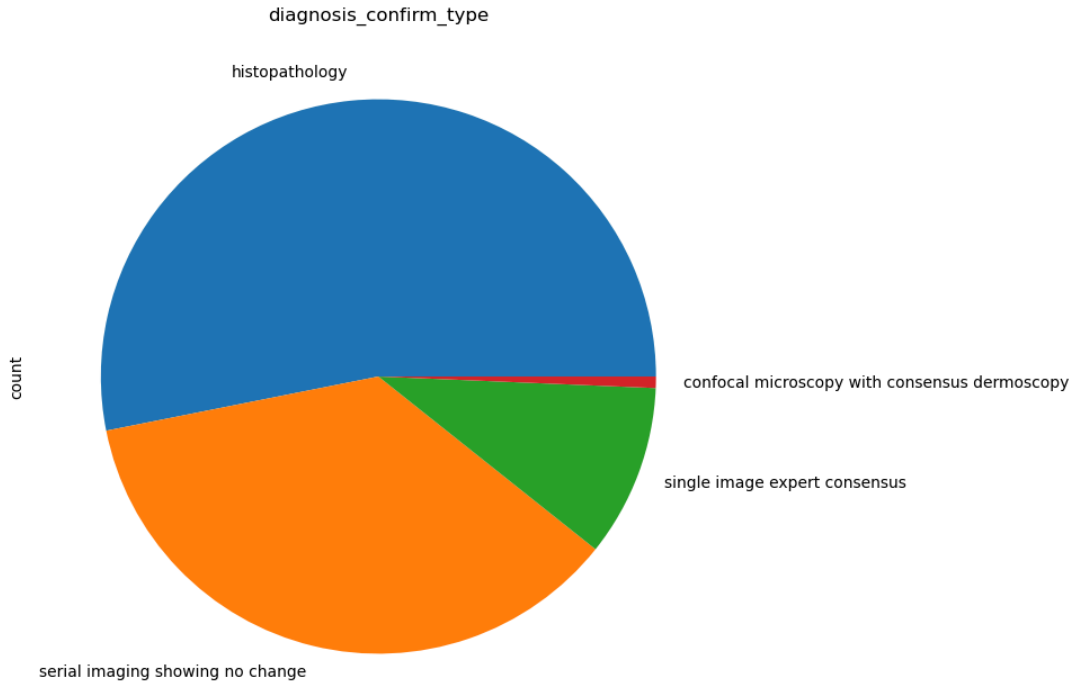
diagnosis: nevus, melanoma, pigmented benign keratosis, dermatofibroma, squamous cell carcinoma, basal cell carcinoma, vascular lesion, actinic keratosis.



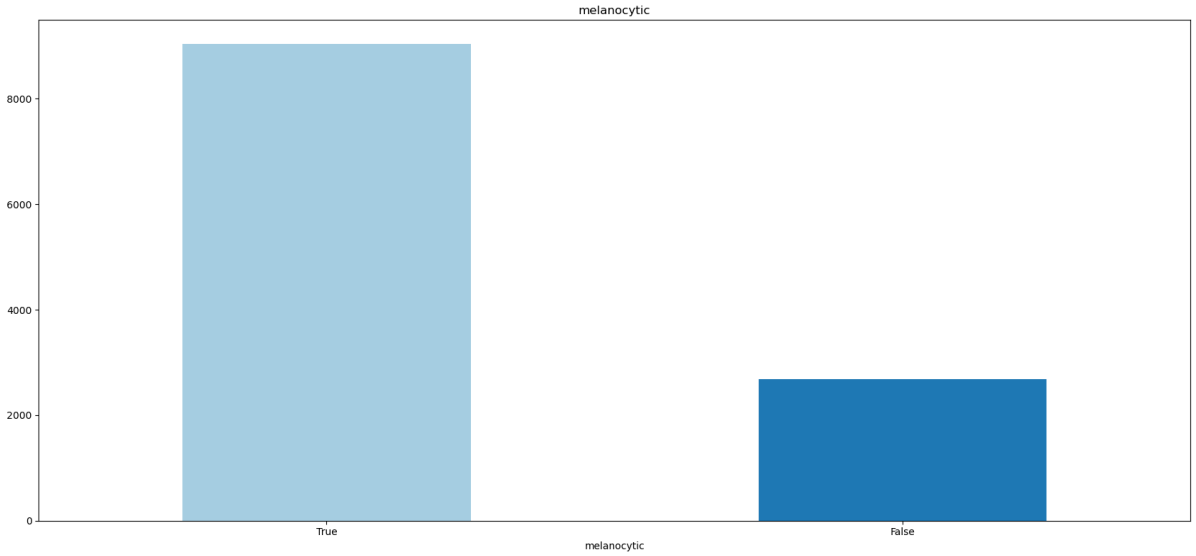


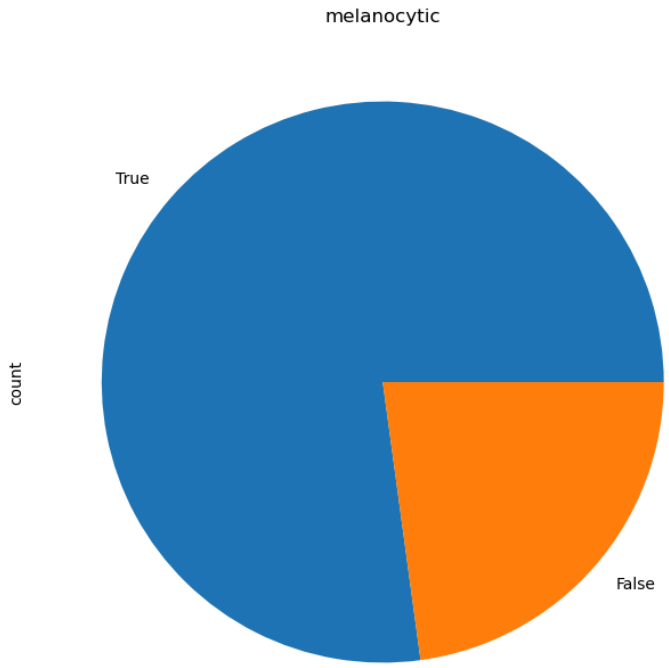
diagnosis_confirm_type: serial imaging showing no change, histopathology, single image, expert consensus, confocal microscopy with consensus dermoscopy



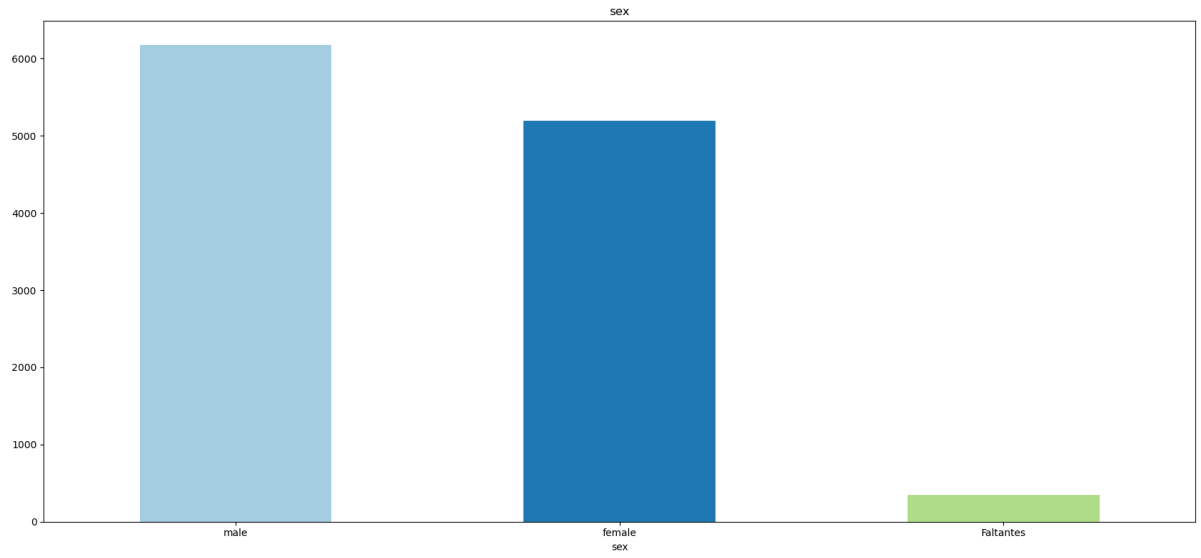


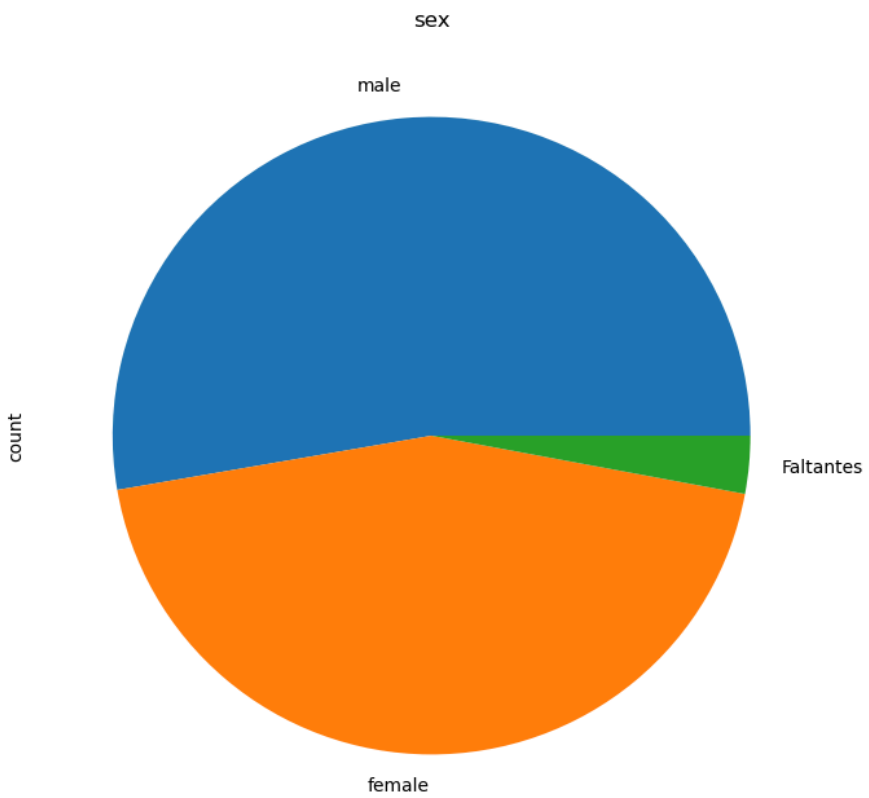
melanocitic: True, False


melanocytic
8000
6000
4000
2000
0
True
False
melanocytic


melanocytic
True
False
count

sex: male, female





Datos Faltantes

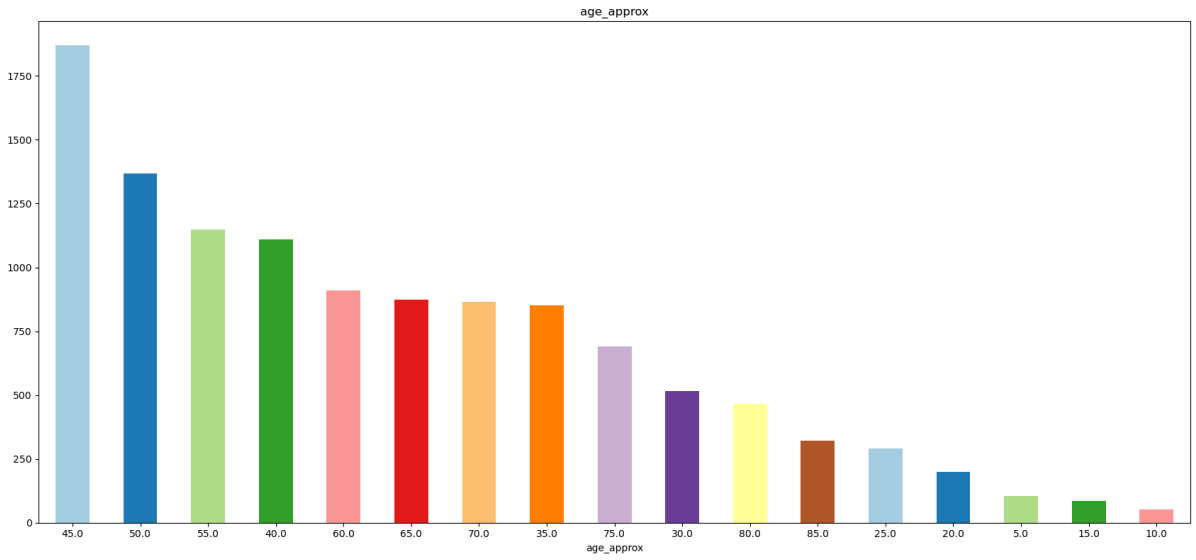
Lo primero que se notó al explorar la integridad de los datos, fue que algunas columnas faltaban de clases en algunos registros. Las columnas en las que se detectaron valores faltantes fueron:

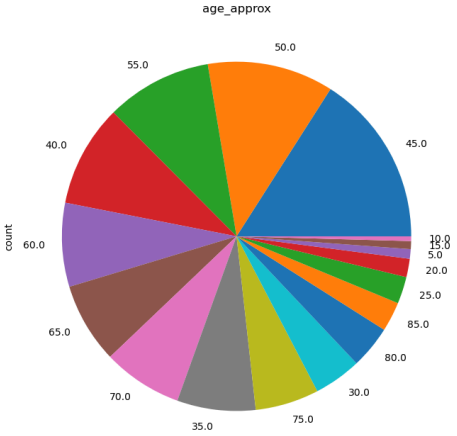
- age_approx: 383 faltantes,
- anatom_site_general: 2162 faltantes,
- anatom_site_special: 11183 faltantes,
- benign_malignant: 2678 faltantes,
- diagnosis_3: 180 faltantes,
- sex: 343 faltantes.

Para mitigar los efectos sobre el modelo de este problema, se rellenaron los datos faltantes con la moda de su columna correspondiente. Por ejemplo, para la columna “sex”, cuya moda es “male”, los 343 registros faltantes de este atributo se rellenan con “male”. Cabe mencionar que para la columnas cuya moda fue la misma ausencia del atributo, el dato faltante se rellena con la categoría “Faltante”; tal es el caso de las columna “anatom_site_special”, por lo que será eliminada del dataset que consumirá el modelo debido a que no proveerá de información de la que se pueda aprender.

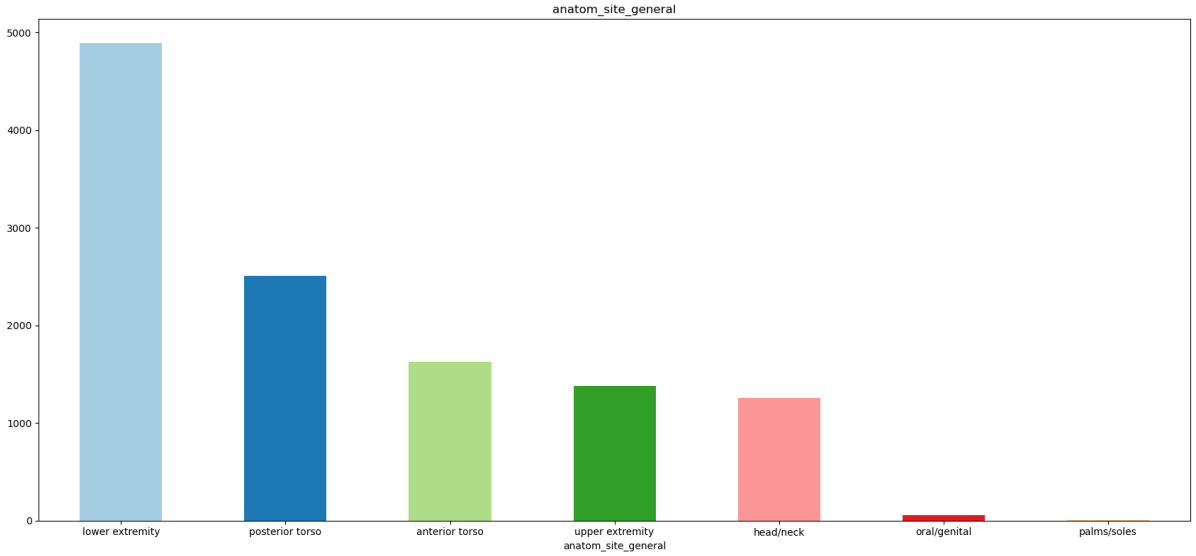
A continuación se muestran las gráficas actualizadas después del relleno de datos:

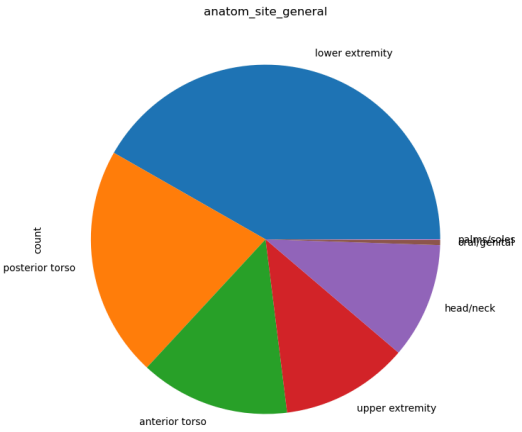
- **age_approx**



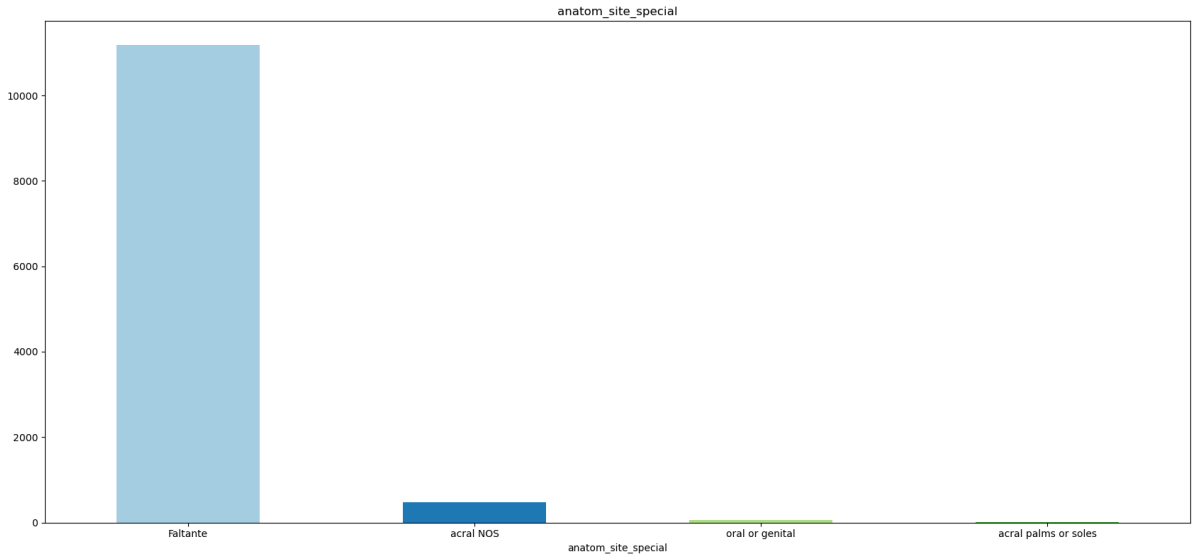


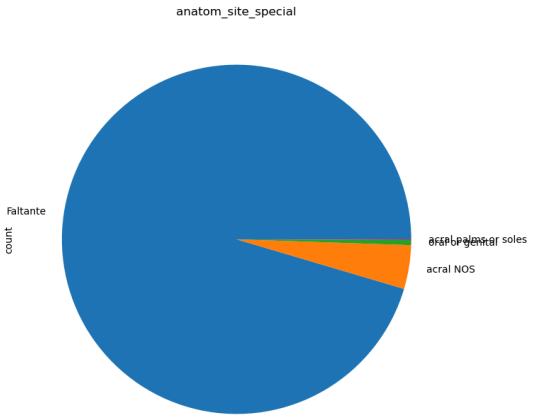
- **anatom_site_general**



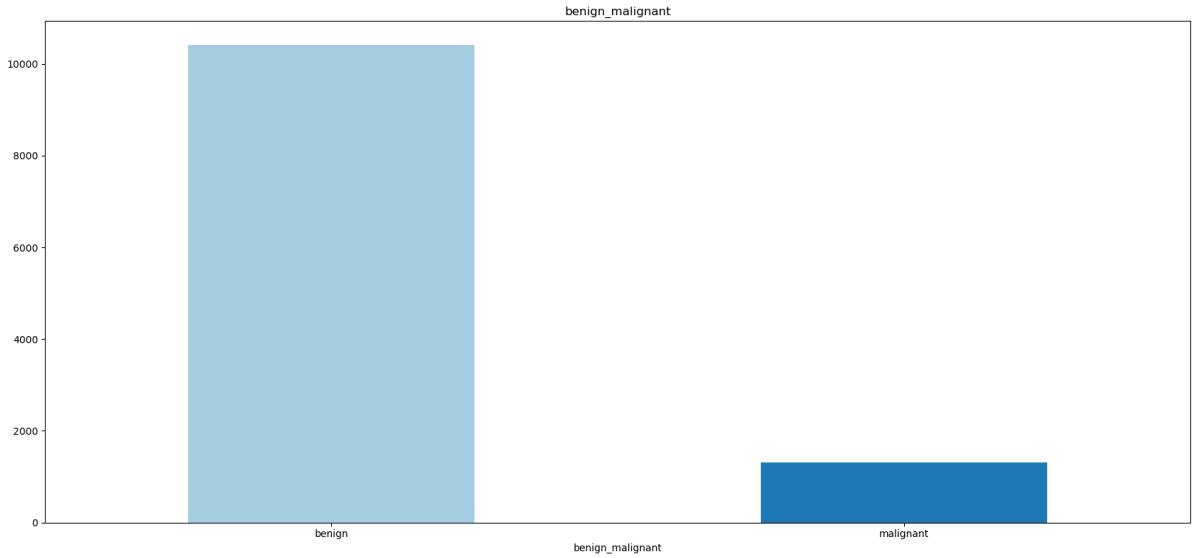


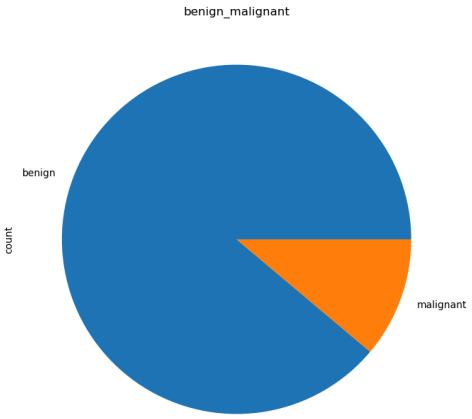
- **anatom_site_special**



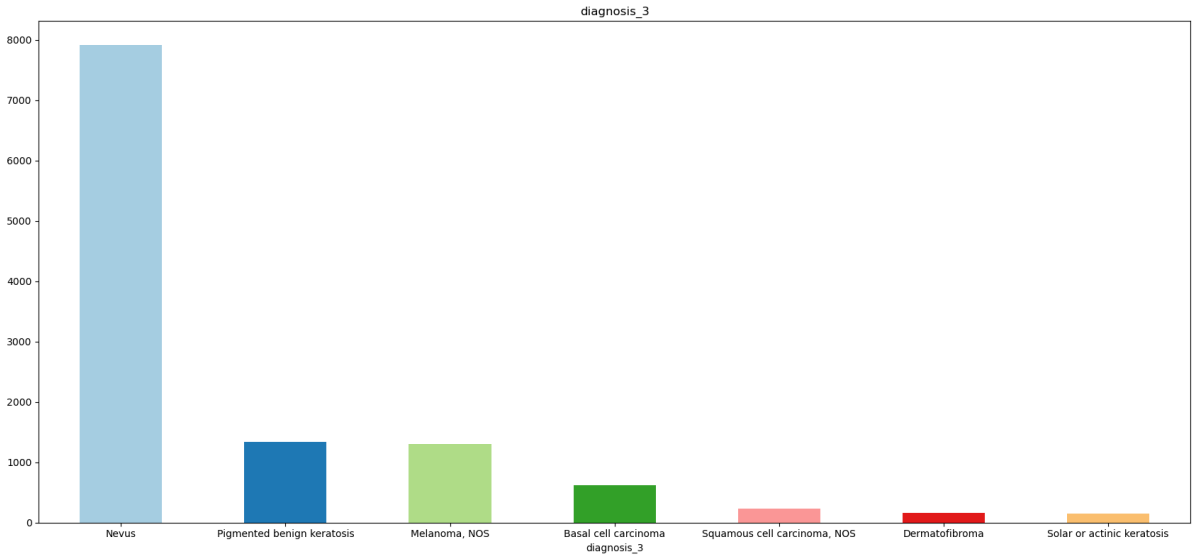


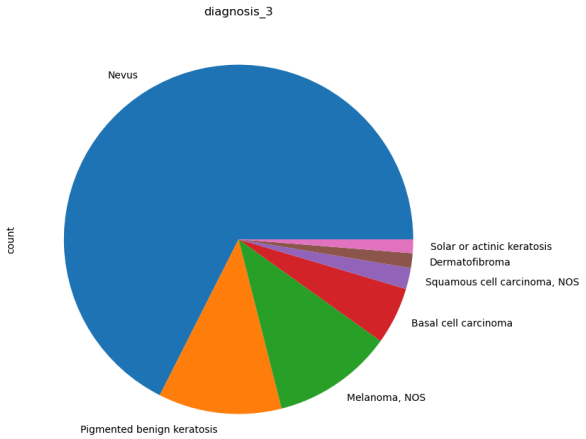
- **benign_malignant**



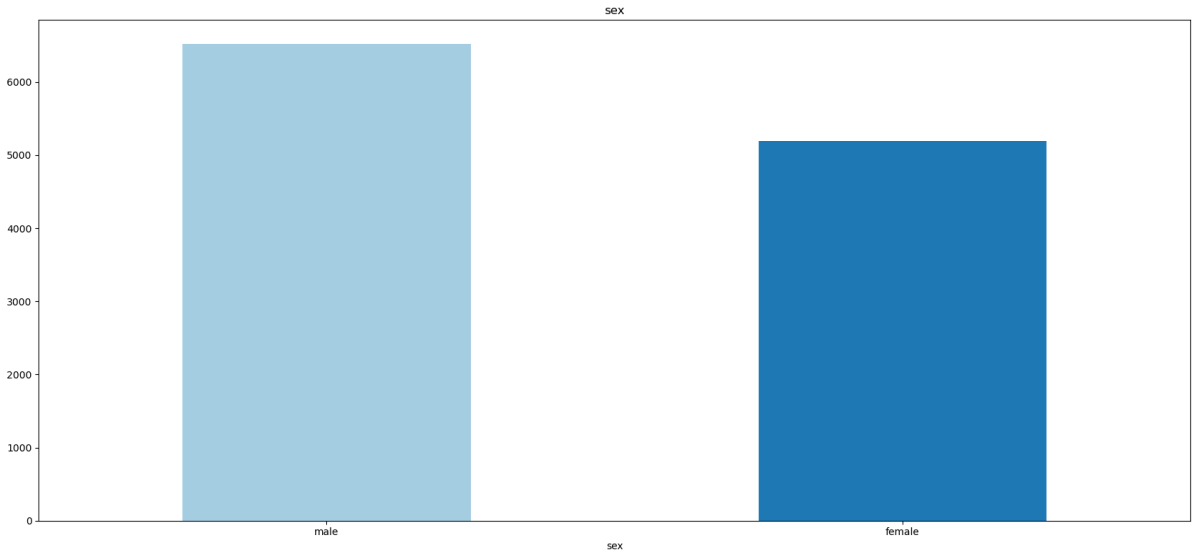

benign_malignant
benign
malignant
count

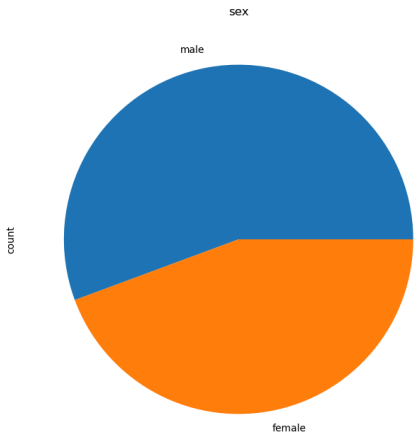
- **diagnosis_3**


diagnosis_3
8000
7000
6000
5000
4000
3000
2000
1000
0
Nevus
Pigmented benign keratosis
Melanoma, NOS
Basal cell carcinoma
Squamous cell carcinoma, NOS
Dermatofibroma
Solar or actinic keratosis
diagnosis_3


diagnosis_3
Nevus
count
Solar or actinic keratosis
Dermatofibroma
Squamous cell carcinoma, NOS
Basal cell carcinoma
Melanoma, NOS
Pigmented benign keratosis

- **sex**


sex
6000
5000
4000
3000
2000
1000
0
male
female
sex

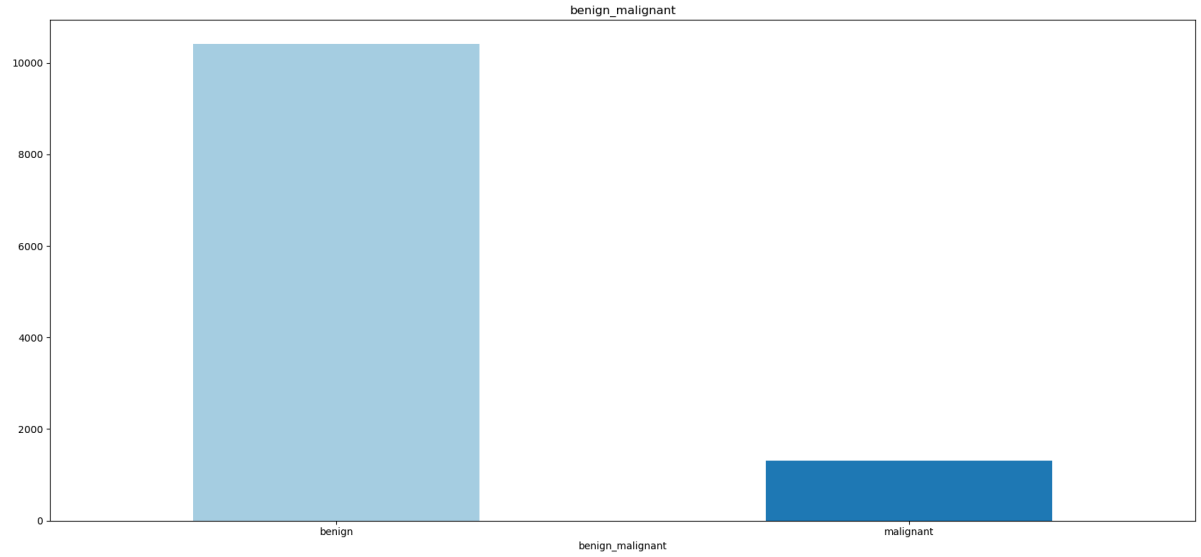


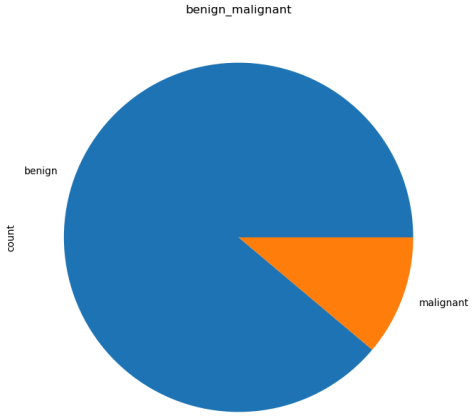
Balance de Clases

Como se mencionó en los objetivos del proyecto, se espera poder identificar si una lesión cutánea es benigna o maligna, e identificar el posible tipo de lesión que se trate. Originalmente, el DataSet Ham10000 nos proporciona las columnas “bening_malignant” y “diagnosis”, las cuales fungen como columnas objetivo del proyecto.

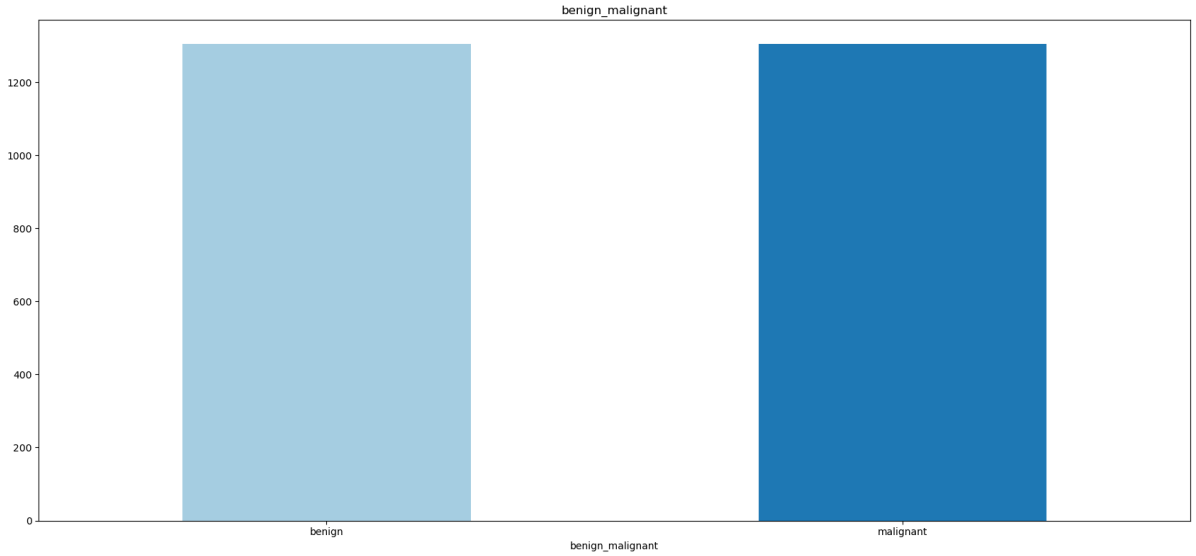
Estas columnas se encuentran desbalanceadas originalmente, por lo que se utilizó la técnica de submuestreo sobre las clases dominantes de cada columna objetivo. A continuación, se muestra la comparación entre las gráficas de las columnas objetivo, antes y después del proceso de balanceo.:

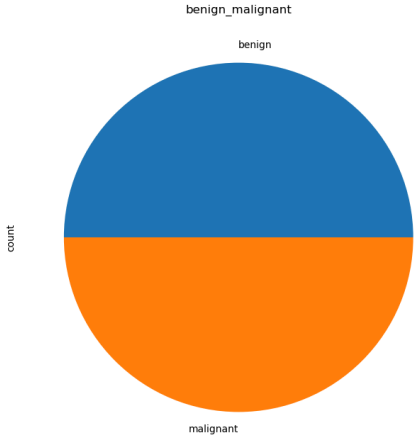
- **bening_malignant**
  - *pre-balanceo*



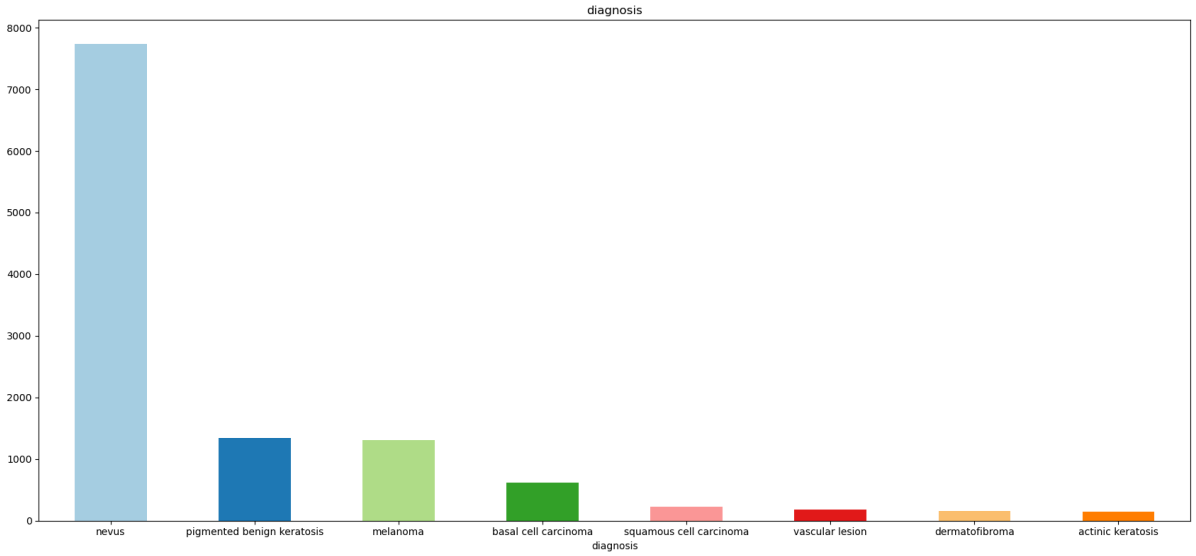


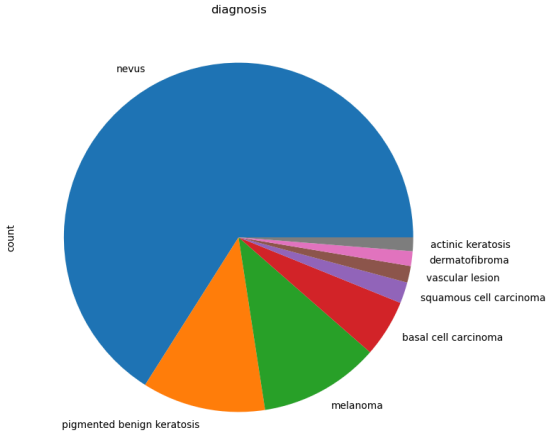
- *post-balanceo*



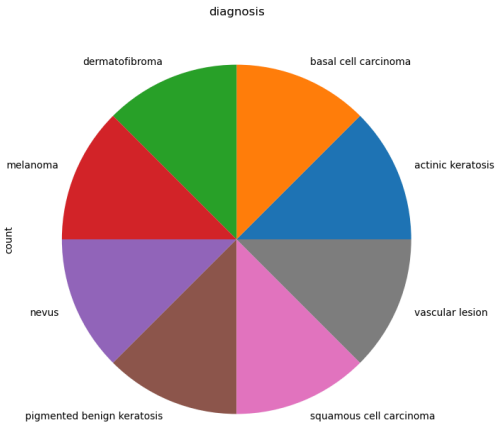


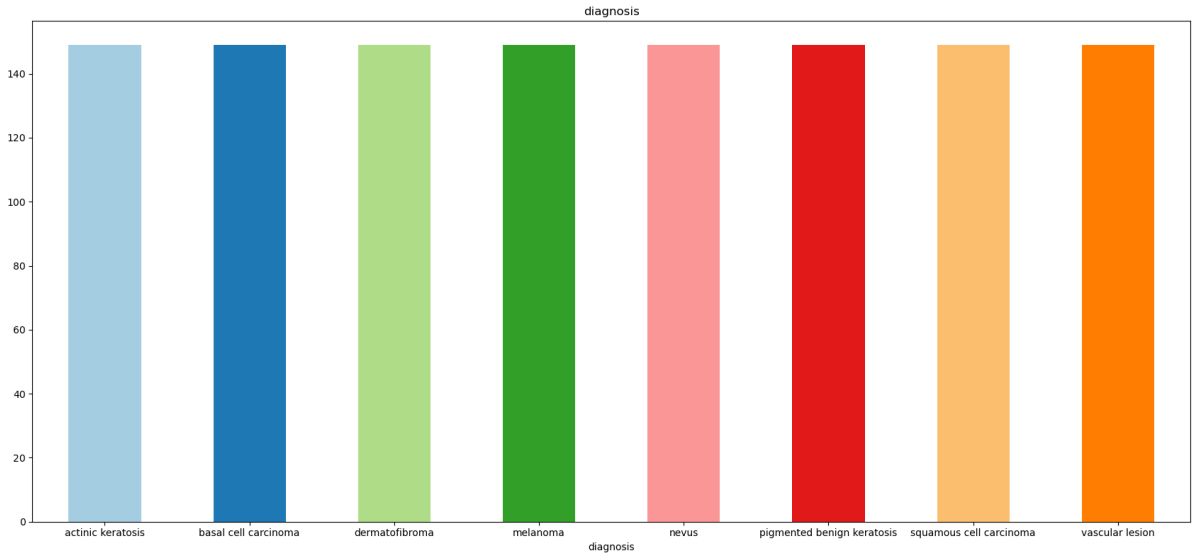
- **diagnosis**
  - *pre-balanceo*





- *post-balanceo*





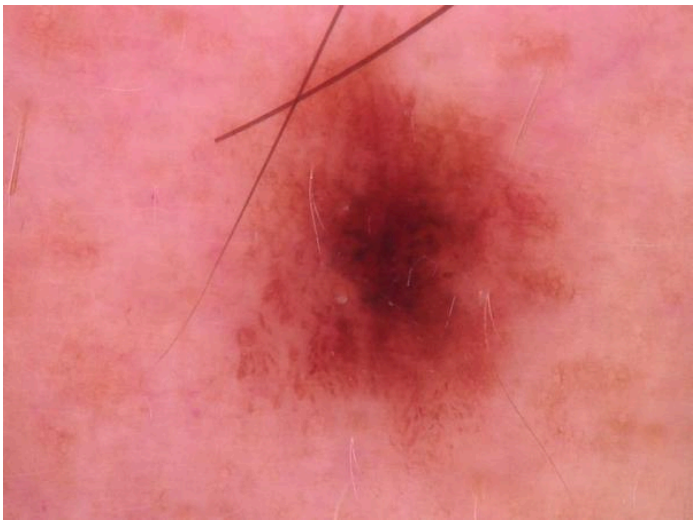
Data Augmentation

Al hacer undersampling en contra de las clases mayoritarias, tal es el caso de las clases “benign”, en el objetivo de “diagnosis”, “benign”, en el objetivo de “benign-malignant”, se obtuvieron datasets bastante reducidos en cantidad de registros, siendo 1193 registros para el objetivo de “diagnosis”, y 4611 para “bening_malignant”. De modo que, aunque se consiguieron datasets balanceados, se consideran como pobres de información como para poder extraer la información necesaria para cada tipo de diagnóstico. Para mitigar este problema, y aprovechando el hecho de que el modelo no dependerá solamente de metadatos, sino que también procesará imágenes, se usó la estrategía de Data Augmentation para generar nuevos registros, cuyas respectivas imágenes son derivadas de

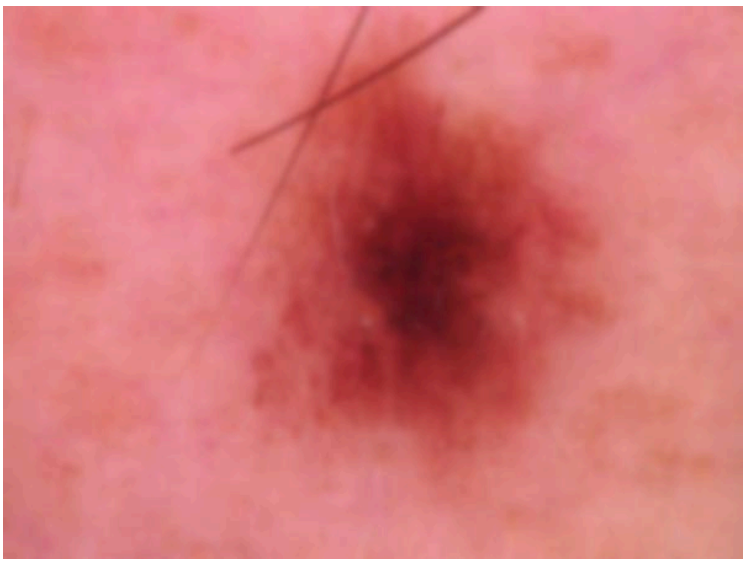
imágenes de registros originales. Tal que, aunque los nuevos registros comparten metadatos, sus imágenes son diferentes.

De cada imágen se obtuvieron 3 imágenes derivadas. Dichas imágenes se obtienen al:

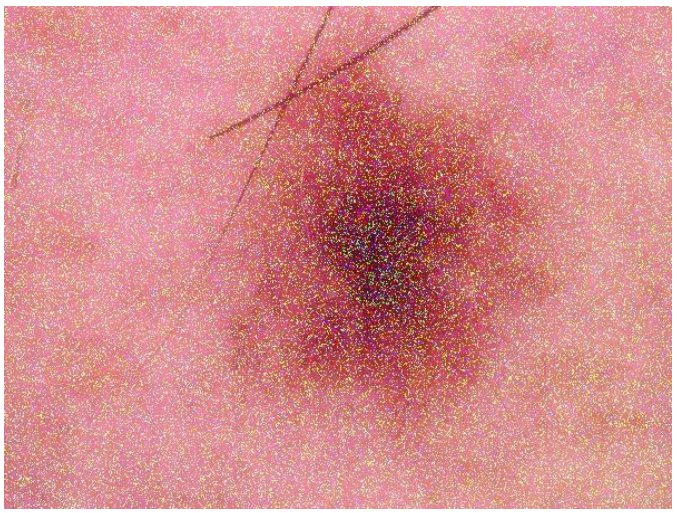
- Rotar 90 grados a la imágen original
- Aplicar Ruido Gaussiano,
- Aplicar un suavizado.



Ejemplo de Imágen Original.



Imágen suavizada.



Imágen con Ruido Gaussiano.

Imágen Rotada 90 grados.

Como resultado de aplicar esta estrategia, para el objetivo de “diagnosis” se logró obtener un dataset balanceado con 4, 768 registros.

Codificación de datos

Desde la exploración del DataSet, se nota que la mayoría de las columnas están compuestas por datos de tipo string/multicategóricos. Para evitar que el modelo

trabaje con este tipo de datos, se aplicaron estrategias de codificación de datos para dichas columnas.

Para las columnas que contienen datos de tipo string/multicategóricos, aplicamos la estrategia Hot_One encoding, la cual crea columnas binarias de tipo int para cada categoría del atributo. Las columnas afectadas son: anatom_site_genreal, diagnosis_1, diagnosis_2, diagnosis_3 y diagnosis_confirm_type.

Para las columnas de atributos binarios, simplemente se pasan las categorías de string/booleanos a int. Las columnas afectadas son: benign_malignant, concominant_biopsy, melanocytic y sex.

Además, dado que la columna “age_aprox” sigue una distribución normal, se le aplicó una estandarización para mejorar la presentación de los datos ante el modelo.

Finalmente, a las imágenes se les aplica una normalización al dividir los valores de los píxeles entre 255 para mantenerlos en un rango de [0,1].

Con estos cambios, se obtuvo un dataset con 36 columnas y 4768 registros para el objetivo “diagnosis”.

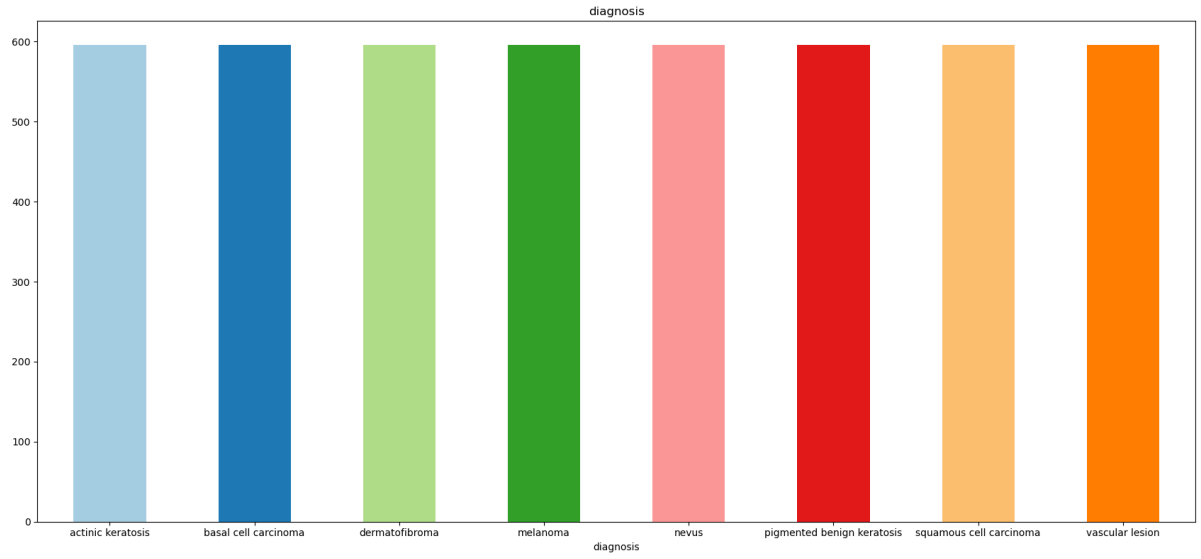


Gráfico de barras de la columna objetivo “diagnosis”, tras aumento de datos.

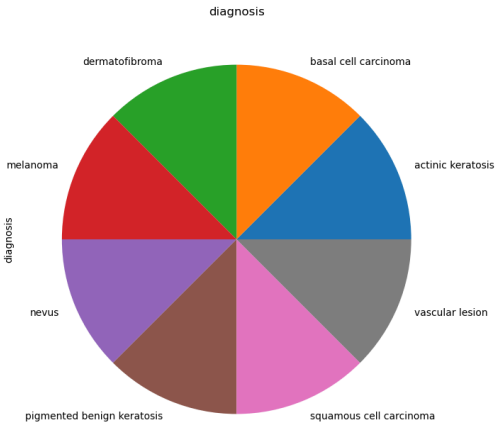


Gráfico de Pastel de la columna objetivo “diagnosis”, tras aumento de datos.

Preparación de Datos

Como se mencionó anteriormente, para el entrenamiento, validación y pruebas del modelo, se utilizarán tres dataset respectivos a cada tarea.

Tanto el conjunto de datos de entrenamiento como de validación son muestras que se dividen entre sí el conjunto de datos obtenido al final de la fase de aumento de datos. De esto modo, ambos datasets, de entrenamiento y de validación, cuentan con 2384 registros con clases balanceadas.

El dataset de prueba se obtiene de los datos restantes que no entraron al dataset de entrenamiento en la fase de balanceo. Cabe mencionar que, dado a que es un restante de un proceso de balanceo por undersampling, este dataset de prueba no se encuentra balanceado.

Finalmente, antes de pasar los metadatos al modelo, de estos se eliminan las columnas “ISIC_ID”, encargada de relacionar los metadatos con las imágenes, y “benign_malignant” el cual podría causar Data_leakage, derivando en problemas de aprendizaje.

Modelo

Implementación de Modelo.

Como se mencionó anteriormente, el modelo debe ser capaz de trabajar con imágenes y metadatos. Para cumplir con este requisito, se implementa un red neuronal convolucional que recibe a las imágenes, cuya salida será la entrada de una red neuronal densa. A esta entrada también se le integran los metadatos, de modo que el resultado final será resultado tanto de características de las imágenes, como de metadatos.

Para implementar la red neuronal convolucional que procesa metadatos e imágenes, se utilizó la api funcional de tensorflow-keras, la cual permite implementar modelos con características especiales, como es el caso de nuestro modelo.

La estructura del modelo es la siguiente:

- **Estructura de entrada:**
  - image_input: Imagen de entrada con tamaño 32×32 píxeles y 3 canales (RGB).
  - metadata_input: Vector de metadatos con 33 columnas.

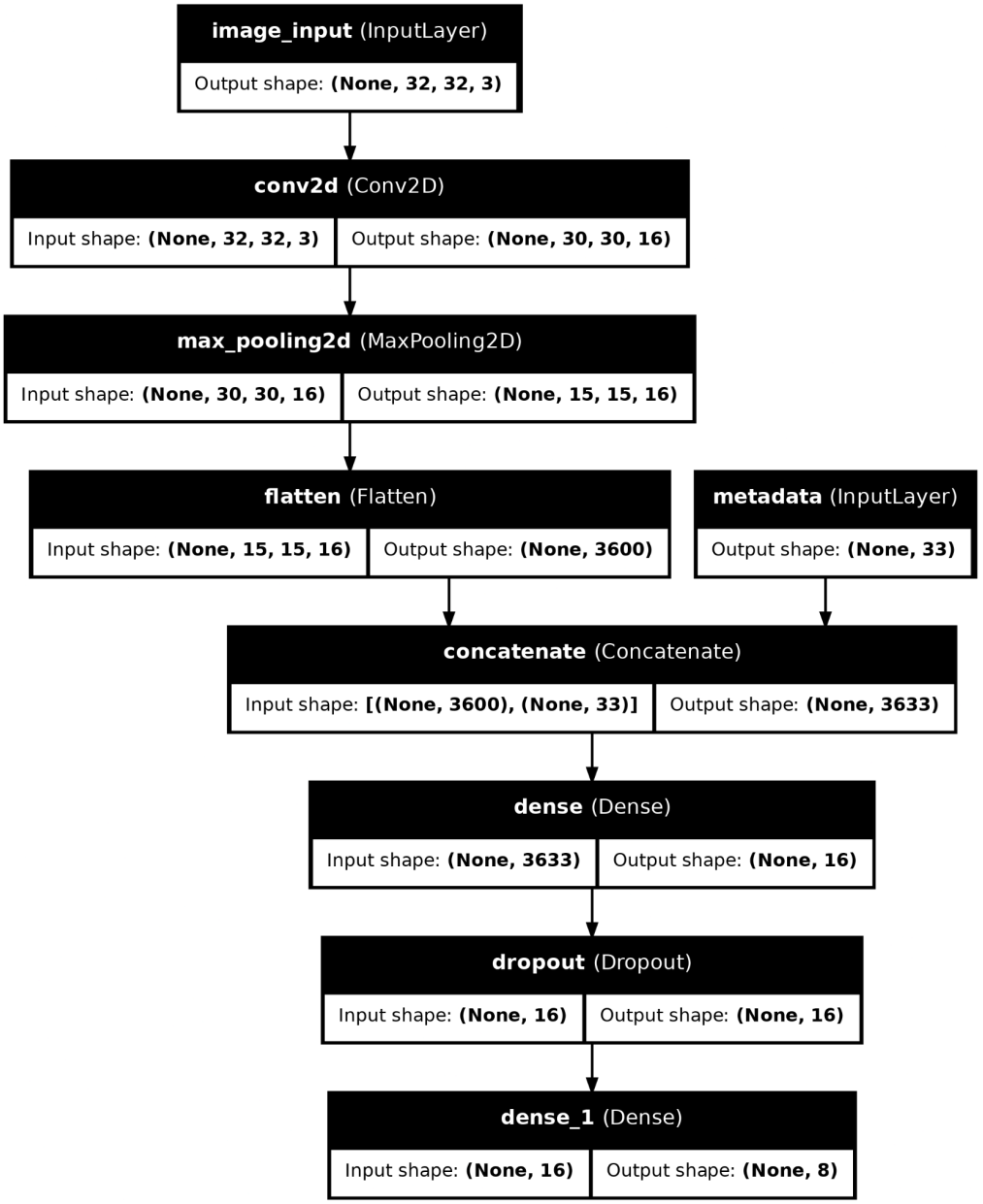
- **Segmento Convolucional:**
  - Conv2D(16, (3, 3), activation='leaky_relu'): Extrae 16 mapas de características con un filtro de 3x3 y activación Leaky ReLU.
  - MaxPooling2D((2, 2)): Reduce las dimensiones espaciales a la mitad.
  - Flatten(): Convierte los mapas de características en un vector unidimensional.

- **Integración de imágen con metadatos:**
  - Concatenate()([x, metadata_input]): Concatena el vector de características visuales con los metadatos, generando una representación combinada.

- **Segmento de red neuronal densa:**
  - Dense(16, activation='leaky_relu'): Capa completamente conectada que transforma la representación combinada.
  - Dropout(0.5): Regulariza la red apagando aleatoriamente el 50% de las neuronas durante el entrenamiento, reduciendo el sobreajuste.

- **Salida final:**
  - Dense(8, activation=output_activation): Capa de salida con 8 neuronas y una función de activación softmax para obtener la clase con mayor probabilidad.

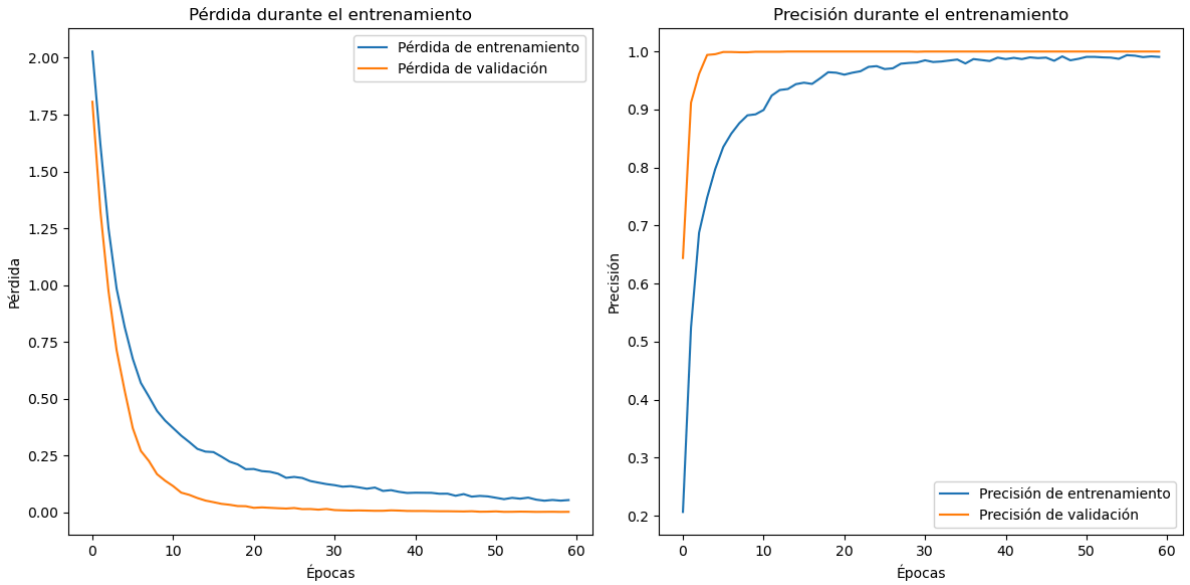
Visualización del Modelo



Disposición del Modelo HÍbrido.

Performance del Modelo

Historial del entrenamiento

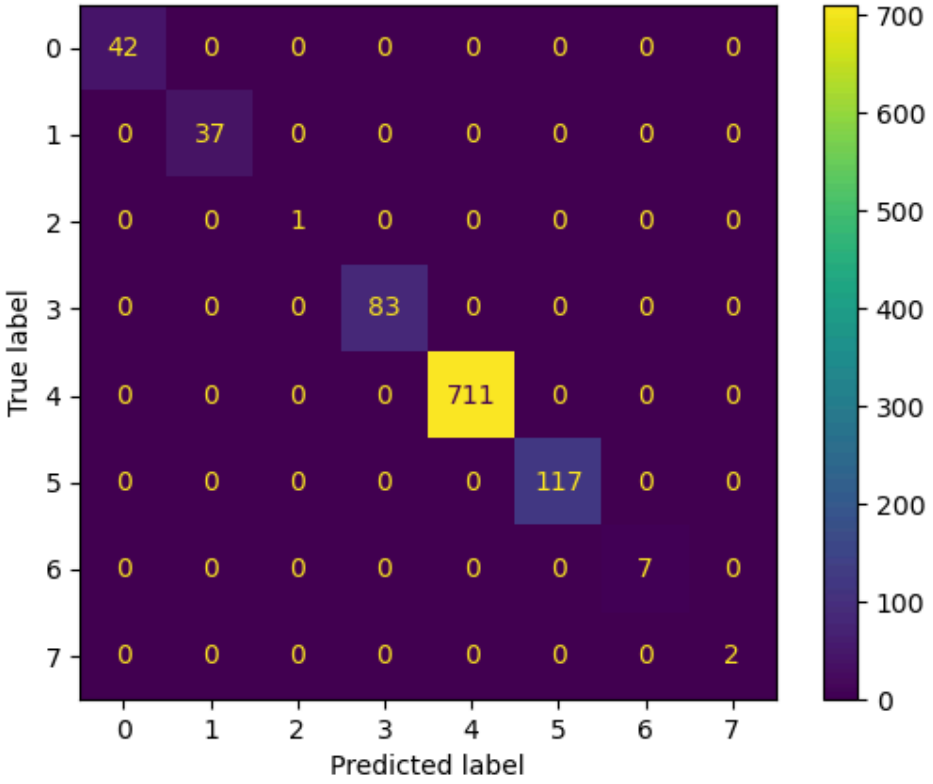


Historial de entrenamiento de 60 generaciones.

Al terminar el entrenamiento, el modelo obtuvo las siguientes métricas:

- categorical_accuracy: 0.9857
- loss: 0.0545 -
- val_categorical_accuracy: 1.0000
- val_loss: 0.0024
- Generaciones: 60

Matriz de Confusión con datos de prueba



Matriz de Confusión con datos de prueba

De la matriz de confusión, podemos notar que, aunque el dataset de validación no fue balanceado, el modelo clasificó bien todos los diagnósticos de las clases.

Métricas:

- Tiempo total de predicción de 1000 registros: 0.25922679901123047
- Tiempo por predicción única: 0.0002592267990112305
- Precisión: 1.000
- Sensibilidad: 1.000
- Especificidad por clase:
  - Especificidad clase 0: 1.0000
  - Especificidad clase 1: 1.0000
  - Especificidad clase 2: 1.0000
  - Especificidad clase 3: 1.0000
  - Especificidad clase 4: 1.0000
  - Especificidad clase 5: 1.0000

- Especificidad clase 6: 1.0000
- Especificidad clase 7: 1.0000

Para ver la codificación completa:

- https://github.com/RodrigoGarciaNunez/Skin_cancer_classifier

Implementación en Google Cloud

Este modelo fue importado a una aplicación alojada en un contenedor de google cloud run.

Dicha aplicación fue realizada utilizando el framework “flask”.

El link de la aplicación es:

- https://skincancerapp-935771581787.northamerica-south1.run.app

Fuentes:

Organizaciones y recursos adicionales:

- Skin Cancer Foundation:
  - https://www.skincancer.org/
- National Cancer Institute:
  - https://www.cancer.gov/
- American Academy of Dermatology:
  - https://www.aad.org/

Aplicaciones Similares:

- Skin Vision:
  - https://www.skinvision.com/
- DermaCam:
  - https://saluddigital.com/en/noticias/mexicanos-desarrollan-app-movil-para-la-deteccion-de-melanomas-a-traves-de-inteligencia-artificial/
- Medic Scanner
  - https://www.appbrain.com/app/medic-scanner-skin-analyze/health.medicalscanner.app
- FotoSkin:
  - https://fotoskin.softonic.com/android
- SkinScreener:
  - https://skinscreener.com/language/en/skincancerprevention-via-app/

DataSets

- Hamm10000
  - https://www.kaggle.com/datasets/kmader/skin-cancer-mnist-ham10000/data
- ISIC Archive
  - https://www.isic-archive.com

Herramientas

- Tensorflow-keras:
  - https://www.tensorflow.org/guide/keras/functional_api
- Google Cloud Run:
  - https://cloud.google.com/sdk/gcloud/reference/run