

1. Introducción

El presente reporte técnico documenta el desarrollo del **Proyecto Integrador DDTI 22**, cuyo objetivo principal es la construcción de un flujo de análisis de datos aplicados a la **clasificación de nódulos tiroideos** en imágenes ecográficas.

La glándula tiroides es una estructura esencial en el organismo, reguladora de procesos metabólicos, y los nódulos que se forman en ella pueden ser benignos o malignos. La identificación precisa de la naturaleza de un nódulo constituye un reto clínico debido a la similitud morfológica de muchas de estas lesiones. Un diagnóstico temprano y certero permite disminuir procedimientos invasivos, mejorar la calidad de vida de los pacientes y optimizar los recursos hospitalarios.

En este contexto, el proyecto se plantea como un ejercicio integrador en el cual confluyen diferentes técnicas de **ciencia de datos, procesamiento de imágenes y aprendizaje automático**. El trabajo se centra en **procesar, analizar y estructurar** un conjunto de datos de imágenes de nódulos tiroideos con el fin de preparar un pipeline robusto que sirva como base para un futuro modelo predictivo.

El alcance de este documento se limita a las etapas previas al modelado final, abarcando:

- **Exploración y análisis de los datos (EDA).**
- **Procesos de limpieza y estandarización.**
- **Creación y selección de características relevantes.**
- **Manejo del desbalance de clases.**
- **Aplicación de técnicas de Data Augmentation.**
- **Partición estratificada del dataset.**
- **Diseño de un pipeline automatizado de preprocesamiento.**

Más allá de los aspectos técnicos, este proyecto tiene también un propósito didáctico: demostrar la importancia de seguir metodologías estructuradas para abordar problemas complejos de datos en el ámbito médico.

2. Metodología General

La metodología seguida en el proyecto responde a un enfoque de **pipeline modular** que permite integrar de manera secuencial y coherente cada una de las fases del proceso. Esta decisión responde a la necesidad de mantener un sistema reproducible, escalable y fácilmente auditable.

2.1 Etapas del pipeline

1. **Análisis Exploratorio de Datos (EDA):**

Consiste en la comprensión inicial del dataset, revisión de la distribución de clases, calidad de los datos, identificación de valores faltantes o duplicados y detección de patrones preliminares.

2. **Limpieza de Datos:**

Incluye validación de formatos, normalización de intensidades de las imágenes, tratamiento de outliers y estandarización de variables derivadas.

3. **Feature Engineering:**

Se centra en la creación de nuevas variables derivadas (intensidad promedio, entropía, densidad de bordes, entre otras), así como en la selección de aquellas con mayor poder explicativo para diferenciar nódulos benignos y malignos.

4. **Balanceo de Clases:**

Dado que el dataset presenta un desbalance moderado, se aplican técnicas para mitigar este efecto, incluyendo oversampling, undersampling y métodos híbridos. En el caso particular de imágenes, se prioriza el uso de augmentation.

5. **Data Augmentation:**

Aumenta artificialmente el tamaño del dataset mediante rotaciones, flips, zooms y cambios en el brillo o contraste, mejorando la capacidad de generalización de futuros modelos.

6. **Partición Estratificada de Datos:**

División del dataset en subconjuntos de entrenamiento, validación y prueba, garantizando que la proporción de clases se mantenga en cada partición.

7. **Pipeline de Preprocesamiento Automatizado:**

Integración de todas las fases anteriores en un flujo automatizado que asegure reproducibilidad y permita aplicar el mismo proceso a nuevos datos.

2.2 Herramientas empleadas

El proyecto se implementó principalmente en **Python**, utilizando las siguientes librerías y entornos:

- **NumPy y Pandas:** manipulación de datos tabulares y creación de variables derivadas.
- **Matplotlib y Seaborn:** visualización de distribuciones, correlaciones y resultados gráficos.
- **Scikit-Learn:** algoritmos de partición, normalización, detección de outliers y técnicas de balanceo.

- **TensorFlow/Keras:** funciones de preprocesamiento y data augmentation para imágenes.
- **Jupyter Notebook:** entorno de desarrollo interactivo para la ejecución ordenada de los análisis.

2.3 Justificación metodológica

La elección de este pipeline modular tiene varias ventajas:

- **Reproducibilidad:** cualquier persona puede ejecutar el mismo flujo y obtener resultados consistentes.
- **Escalabilidad:** el pipeline puede extenderse a datasets más grandes o diferentes tipos de imágenes médicas.
- **Auditoría y control:** cada fase queda documentada y validada, lo cual es especialmente importante en contextos clínicos donde la trazabilidad es obligatoria.
- **Flexibilidad:** permite reemplazar o mejorar fases específicas (por ejemplo, cambiar la técnica de balanceo) sin alterar todo el flujo.

2.4 Desafíos previstos

Durante la planificación metodológica se identificaron los siguientes retos:

- El desbalance de clases puede afectar métricas de precisión y recall.
- La variabilidad de calidad en imágenes ecográficas puede introducir ruido.
- La elección de variables derivadas debe ser clínicamente relevante, no solo estadísticamente significativa.
- El manejo de outliers requiere equilibrio entre limpieza técnica y conservación de casos clínicos relevantes.

En síntesis, la metodología propuesta ofrece una base sólida y adaptable para abordar el problema planteado, garantizando que los resultados obtenidos sean confiables, reproducibles y útiles para una etapa posterior de modelado profundo.

3. Análisis Exploratorio de Datos (EDA)

El análisis exploratorio de datos constituye una de las fases más críticas de todo proyecto de ciencia de datos. Su objetivo es proporcionar una comprensión profunda del dataset, identificar posibles problemas de calidad y obtener información inicial sobre las relaciones entre variables. En el presente proyecto, el EDA se enfocó en caracterizar un conjunto de **637 imágenes ecográficas de nódulos tiroideos**, clasificadas en dos categorías:

- **Malignos:** 377 imágenes.
- **Benignos:** 260 imágenes.

Esta distribución implica un **desbalance moderado (1.45:1)**, lo que de entrada plantea un reto metodológico para la etapa de modelado.

3.1 Exploración inicial de los datos

La exploración comenzó con la revisión de los **metadatos del dataset**:

- **Formato:** todas las imágenes se encuentran en formato estándar compatible con librerías de procesamiento (PNG/JPEG).
- **Resolución:** se realizó un preprocesamiento para unificar las dimensiones a **299 × 299 píxeles en tres canales de color (RGB)**, lo que facilita su uso en arquitecturas de redes neuronales convolucionales (CNN).
- **Etiquetas:** la variable objetivo corresponde a un valor binario (0 = benigno, 1 = maligno).

Además de las imágenes, se derivaron **atributos tabulares** para realizar análisis estadísticos:

- Intensidad promedio.
- Contraste.
- Entropía.
- Densidad de bordes.

Estos atributos permitieron observar diferencias preliminares entre clases sin necesidad de entrenar modelos complejos.

3.2 Análisis de calidad

Uno de los primeros pasos del EDA fue revisar la **calidad de los datos**:

- **Valores faltantes:** no se detectaron valores faltantes en las etiquetas ni en los atributos derivados.
- **Duplicados:** no se identificaron imágenes duplicadas ni registros redundantes.
- **Consistencia:** todos los archivos podían abrirse y procesarse correctamente, lo que asegura la integridad del dataset.

Este hallazgo simplifica la fase de limpieza, dado que no fue necesario implementar estrategias de imputación o eliminación de registros inválidos.

3.3 Estadística descriptiva

La siguiente tabla resume las estadísticas descriptivas de los principales atributos tabulares.

| Variable | Clase | Media | Desviación Estándar | Mínimo | Máximo | Percentil 25 | Percentil 75 |
|---------------------|---------|-------|---------------------|--------|--------|--------------|--------------|
| Intensidad promedio | Benigno | 0.48 | 0.11 | 0.22 | 0.72 | 0.40 | 0.55 |
| | Maligno | 0.50 | 0.13 | 0.18 | 0.75 | 0.42 | 0.58 |
| Contraste | Benigno | 0.28 | 0.09 | 0.10 | 0.45 | 0.20 | 0.35 |
| | Maligno | 0.34 | 0.10 | 0.12 | 0.52 | 0.25 | 0.41 |
| Entropía | Benigno | 0.61 | 0.12 | 0.30 | 0.85 | 0.53 | 0.70 |
| | Maligno | 0.69 | 0.11 | 0.35 | 0.88 | 0.61 | 0.77 |
| Densidad de bordes | Benigno | 0.25 | 0.08 | 0.09 | 0.42 | 0.18 | 0.30 |
| | Maligno | 0.32 | 0.09 | 0.12 | 0.48 | 0.25 | 0.39 |

Interpretación inicial:

- Los nódulos **malignos presentan mayor contraste y entropía**, lo que refleja imágenes más heterogéneas.
- La **densidad de bordes** también es más alta en malignos, lo cual tiene sentido clínico dado que suelen tener contornos irregulares.
- La **intensidad promedio** no muestra diferencias significativas, lo que indica que no es un buen predictor por sí sola.

3.4 Análisis gráfico

Se elaboraron histogramas y diagramas de caja (boxplots) para comparar las distribuciones de cada atributo.

- **Histograma de entropía:** muestra un desplazamiento hacia valores más altos en la clase maligna.

- **Boxplot de densidad de bordes:** la mediana es claramente mayor en malignos.
- **Distribución de contraste:** ambas clases se solapan, pero los valores extremos aparecen con más frecuencia en la clase maligna.

3.5 Correlaciones y relaciones entre variables

Se calculó una matriz de correlación para los atributos derivados:

- La correlación más fuerte se observó entre **contraste y densidad de bordes ($r = 0.61$)**.
- La entropía mostró correlación moderada con contraste ($r = 0.45$).
- No se detectó multicolinealidad severa que impida el uso conjunto de estas variables.

Además, se graficaron diagramas de dispersión bivariados, donde se observó cierta **separabilidad parcial** entre clases al combinar entropía con densidad de bordes.

3.6 Outliers

Se identificaron posibles outliers en contraste y entropía mediante **Z-score > 3** y mediante la observación en boxplots. Sin embargo, al tratarse de datos médicos, se decidió **no eliminarlos**, dado que pueden corresponder a casos clínicos reales de interés (por ejemplo, nódulos con bordes extremadamente irregulares).

3.7 Variable objetivo

La variable objetivo (clase del nódulo) se distribuye de manera desigual:

- Malignos: **59%**.
- Benignos: **41%**.

Este desbalance, aunque no extremo, es suficiente para que un modelo que no lo corrija tienda a predecir con mayor frecuencia la clase maligna. Por ello, se prevé aplicar técnicas de balanceo y augmentation en fases posteriores.

4. Limpieza de Datos

La limpieza de datos es una etapa esencial en cualquier proyecto de ciencia de datos, ya que garantiza la calidad y confiabilidad del conjunto antes de proceder con el modelado. Aunque en este caso no se detectaron problemas graves de calidad durante el análisis exploratorio, fue necesario establecer un **pipeline de limpieza estructurado y reproducible** que asegure consistencia para futuras versiones del dataset.

4.1 Validación de formatos y estructura

Se verificó que todas las imágenes estuvieran en formatos estándar (PNG y JPEG), asegurando compatibilidad con librerías como *OpenCV*, *TensorFlow* y *PIL*. Además, se confirmó que los archivos pudieran abrirse sin errores, descartando la presencia de imágenes corruptas.

El paso de **redimensionamiento a 299 × 299 píxeles con 3 canales (RGB)** fue fundamental para estandarizar la entrada de datos, especialmente pensando en redes neuronales convolucionales (CNN). Esta estandarización asegura que cada imagen tenga la misma dimensión y que no se produzcan incompatibilidades durante el entrenamiento.

4.2 Tratamiento de valores faltantes

El EDA confirmó que **no existen valores faltantes** en las etiquetas ni en los atributos derivados. Sin embargo, se diseñó una estrategia de contingencia para escenarios futuros donde esto sí ocurra:

- Para atributos tabulares, se propone imputar valores con la **mediana** (robusta frente a outliers) o mediante algoritmos de imputación como **KNN Imputer**.
- Para imágenes con datos faltantes o dañados, se plantea **descartar el archivo** si el error es irreparable, dado que no tendría sentido imputar píxeles en una imagen médica sin afectar la validez clínica.

4.3 Detección y tratamiento de duplicados

No se encontraron imágenes duplicadas en este dataset. Sin embargo, se implementó un procedimiento automatizado de verificación mediante **hashing perceptual (pHash)** para asegurar que, en caso de integrar futuras colecciones de imágenes, los duplicados sean identificados y removidos automáticamente.

Esto evita que el modelo aprenda sobre instancias repetidas y reduce el riesgo de sobreajuste.

4.4 Manejo de outliers

Se identificaron outliers principalmente en las variables **contraste** y **entropía**, detectados mediante:

1. **Método estadístico:** observando valores con Z-score > 3.
2. **Método visual:** inspeccionando boxplots y diagramas de dispersión.

En lugar de eliminarlos, se decidió **conservar los outliers** debido a su posible relevancia clínica. En imágenes médicas, los casos extremos no siempre representan errores, sino que pueden ser diagnósticamente significativos (por ejemplo, un nódulo con bordes inusualmente irregulares).

No obstante, se dejó documentada una estrategia de tratamiento opcional para escenarios futuros:

- Aplicación de **capping** en percentiles (p1 y p99).
- Transformaciones robustas (por ejemplo, *RobustScaler* de *Scikit-Learn*).

4.5 Normalización y estandarización

Para garantizar comparabilidad entre variables y mejorar la estabilidad numérica de los algoritmos, se aplicaron las siguientes transformaciones:

- **Normalización de intensidades de píxeles:** cada valor se llevó al rango [0, 1]. Esto permite que el entrenamiento de modelos profundos sea más estable y evita que los gradientes se saturen.
- **Estandarización de variables tabulares:** atributos como entropía o densidad de bordes se transformaron con **Z-score** (media = 0, desviación estándar = 1). Esta técnica es especialmente útil en algoritmos sensibles a la escala de los datos.

4.6 Pipeline automatizado de limpieza

Todos los pasos de limpieza se encapsularon en un **pipeline reproducible**, diseñado en Python, que incluye:

1. Carga de imágenes.
2. Redimensionamiento a $299 \times 299 \times 3$.
3. Normalización de intensidades.
4. Cálculo de atributos derivados.
5. Validación de duplicados y valores faltantes.
6. Estándar de escalado de variables tabulares.
7. Registro automático en logs de cada ejecución.

Este pipeline asegura que cualquier nuevo lote de imágenes pase por el mismo proceso de estandarización, evitando errores humanos y asegurando la trazabilidad del proyecto.

5. Feature Engineering Avanzado

La **ingeniería de características (Feature Engineering)** es una de las fases más importantes dentro de un pipeline de ciencia de datos, pues determina qué información de los datos brutos se convierte en atributos útiles para los modelos. En este proyecto, las imágenes ecográficas de nódulos tiroideos fueron enriquecidas con variables derivadas que aportan información adicional más allá de los píxeles individuales.

El objetivo principal fue **capturar propiedades clínicas relevantes** que pudieran servir como predictores para diferenciar nódulos benignos de malignos.

5.1 Creación de variables derivadas

A partir de las imágenes, se definieron y calcularon los siguientes atributos:

1. Intensidad promedio:

- Corresponde al valor medio de los píxeles en escala de grises.
- Permite evaluar si la lesión tiende a ser más hipoecoica (oscura) o hiperecoica (clara).
- Aunque las diferencias no son radicales entre clases, puede servir como un descriptor básico.

2. Contraste:

- Calculado como la variabilidad en los niveles de intensidad dentro de una región.
- Clínicamente, los nódulos malignos suelen presentar mayor heterogeneidad interna.
- En los histogramas, la clase maligna mostró valores más altos en promedio.

3. Entropía:

- Mide el grado de desorden o aleatoriedad en la distribución de intensidades.
- Un nódulo maligno, con texturas más irregulares, tiende a tener mayor entropía.

- Se observó una diferencia clara entre benignos y malignos (medias: 0.61 vs 0.69).

4. **Densidad de bordes:**

- Se calcula aplicando un detector de bordes (por ejemplo, Sobel o Canny) y midiendo el porcentaje de píxeles detectados como borde.
- Los nódulos malignos presentan contornos más irregulares, lo cual eleva esta métrica.
- Se confirmó que esta variable presenta buen poder discriminativo.

5. **Estadísticos por regiones (quadrants/patches):**

- El análisis de subregiones de la imagen permite capturar heterogeneidad espacial.
- Se calcularon medias y desviaciones estándar en cuadrantes de la imagen, generando un conjunto de variables adicionales.

6. **Variables de interacción:**

- Se crearon combinaciones de variables, como contraste × entropía, para capturar relaciones no lineales.

Este conjunto de variables derivadas permite enriquecer el dataset más allá de los píxeles crudos, aumentando la interpretabilidad y generando un puente entre el análisis clínico y el análisis computacional.

5.2 **Selección de características**

No todas las variables derivadas aportan valor al modelo, por lo que se aplicaron técnicas de selección:

- **Pruebas estadísticas:**

- *ANOVA F-test* mostró diferencias significativas en entropía y densidad de bordes.
- *Mutual Information* indicó que entropía y contraste aportan la mayor cantidad de información para predecir la clase.

- **Modelos basados en árboles:**

- Se entrenó un *Random Forest* preliminar únicamente sobre las variables tabulares.

- El ranking de importancia de variables confirmó a entropía y densidad de bordes como los principales predictores.
- **Reducción de dimensionalidad (PCA):**
 - Se aplicó *Análisis de Componentes Principales*.
 - Los dos primeros componentes explicaron aproximadamente el **70% de la varianza total**.
 - Al graficar en un espacio bidimensional, se observó una ligera separabilidad entre clases, aunque con solapamientos inevitables.

La **Figura 2** (insertada en Word) muestra las distribuciones de las variables y el resultado del PCA.

5.3 Extracción de características específicas del dominio

Dado que las imágenes son ecográficas, se incorporaron características propias del análisis médico de texturas:

- **Local Binary Patterns (LBP):**
 - Técnica que codifica la textura local de una imagen.
 - Resulta útil para identificar patrones finos en regiones homogéneas.
 - Puede diferenciar entre nódulos con microcalcificaciones (más comunes en malignos) y aquellos con textura lisa.
- **Histogram of Oriented Gradients (HOG):**
 - Captura la orientación de bordes y contornos.
 - Clínicamente relevante, dado que los nódulos malignos tienden a presentar bordes más irregulares.
- **Medidas fractales:**
 - Se exploró la posibilidad de calcular la dimensión fractal de los bordes.
 - Nódulos malignos suelen presentar estructuras más complejas, lo que se refleja en mayor dimensión fractal.

Estas características, aunque no utilizadas en la primera iteración del pipeline, se documentaron como **potenciales mejoras futuras**.

5.4 Relevancia clínica de las variables

Cada variable derivada tiene una justificación clínica:

- **Entropía y densidad de bordes** → Irregularidad estructural y textural, asociada a malignidad.
- **Contraste** → Diferencias internas en el tejido, más marcadas en tumores malignos.
- **Intensidad promedio** → Apoyo básico para determinar ecogenicidad, aunque menos discriminante.

El objetivo no es únicamente alimentar a un modelo con datos numéricos, sino también garantizar que las características utilizadas tengan **significado clínico**, lo cual aporta interpretabilidad y aumenta la confianza en los resultados.

5.5 Conclusiones del Feature Engineering

- El proceso permitió generar un conjunto robusto de variables complementarias a los píxeles crudos.
- Se identificaron entropía y densidad de bordes como los atributos con mayor poder discriminativo.
- La reducción de dimensionalidad con PCA permitió confirmar la utilidad de las variables para diferenciar parcialmente las clases.
- Se documentaron características adicionales de texturas y bordes que pueden implementarse en fases futuras.

6. Balanceo de Clases

En todo proyecto de aprendizaje supervisado, la distribución de la variable objetivo es un factor crítico que impacta directamente en el rendimiento del modelo. En el caso de este dataset, el **59% de las imágenes corresponden a nódulos malignos (377) y el 41% a nódulos benignos (260)**. Esta proporción representa un **desbalance moderado** que, aunque no extremo, puede sesgar a los modelos hacia la clase mayoritaria.

6.1 Riesgos del desbalance

Un modelo entrenado sin corregir el desbalance podría:

- Alcanzar una **alta exactitud (accuracy)** simplemente prediciendo la clase mayoritaria, sin realmente aprender patrones discriminativos.
- Mostrar un **recall bajo en la clase minoritaria**, lo cual en un contexto médico es crítico, ya que aumentaría la tasa de falsos negativos (es decir, clasificar como benigno un nódulo que en realidad es maligno).

- Reducir la **robustez del modelo** en escenarios de validación y despliegue real.

Por lo tanto, es indispensable implementar **estrategias de balanceo** antes de pasar a la fase de entrenamiento.

6.2 Estrategias clásicas de balanceo

Existen tres enfoques principales para enfrentar el desbalance:

a) Oversampling (sobre-muestreo de la clase minoritaria)

- Consiste en aumentar artificialmente la cantidad de instancias de la clase minoritaria.
- Técnicas utilizadas:
 - **Duplicación simple:** replicar ejemplos existentes de la clase minoritaria.
 - **SMOTE (Synthetic Minority Oversampling Technique):** genera nuevas instancias interpolando entre ejemplos existentes.
 - **ADASYN:** variante de SMOTE que enfoca el oversampling en las regiones más difíciles de clasificar.

✦ Ventajas: mejora la representación de la clase minoritaria.

✦ Desventajas: puede aumentar el riesgo de sobreajuste si se generan instancias poco realistas.

b) Undersampling (sub-muestreo de la clase mayoritaria)

- Se eliminan ejemplos de la clase mayoritaria hasta equilibrar la distribución.
- Métodos comunes:
 - **Random Undersampling:** elimina ejemplos de manera aleatoria.
 - **NearMiss:** selecciona ejemplos de la clase mayoritaria más cercanos a los de la clase minoritaria.

✦ Ventajas: reduce el tamaño del dataset, acelerando el entrenamiento.

✦ Desventajas: pérdida de información valiosa, lo que puede perjudicar la capacidad predictiva del modelo.

c) Técnicas híbridas

- Combinan oversampling y undersampling.
- Ejemplo: **SMOTEENN**, que aplica SMOTE para generar nuevas instancias y luego elimina ejemplos ruidosos con Edited Nearest Neighbors.

✦ Ventaja: logra un equilibrio más natural en datasets con ruido.

✦ Desventaja: mayor complejidad y necesidad de validación adicional.

6.3 Consideraciones para datos de imágenes

Si bien las técnicas anteriores funcionan bien en datasets tabulares, en imágenes médicas presentan limitaciones:

- **SMOTE o ADASYN no son adecuados directamente** para imágenes, ya que la interpolación de píxeles puede generar artefactos poco realistas.
- El **undersampling aleatorio** en imágenes puede significar descartar ejemplos clínicamente valiosos.

Por ello, en este proyecto se decidió privilegiar el uso de **Data Augmentation** como principal estrategia de balanceo.

6.4 Estrategia aplicada en este proyecto

- Se mantuvo el dataset completo (sin eliminación de ejemplos).
- Se evitó el uso de SMOTE directo sobre imágenes.
- Se aplicó **augmentation dirigido a la clase minoritaria (benignos)** para generar más variaciones de esta categoría, equilibrando su representación.

Ejemplo de augmentations aplicados para balancear:

- Rotaciones aleatorias.
- Flips horizontales y verticales.
- Zoom y leves cambios de contraste.

Con estas transformaciones, se consiguió que la proporción entre clases fuese **prácticamente 1:1 en el conjunto de entrenamiento**, sin necesidad de eliminar ejemplos de la clase mayoritaria.

6.5 Evaluación de la estrategia

La efectividad del balanceo se evalúa no solo observando las proporciones finales de clases, sino también verificando su impacto en las métricas de validación:

- La métrica clave será el **recall en la clase minoritaria** (evitar falsos negativos).
- También se analizarán métricas balanceadas como **F1-score** y **AUC-ROC**, que reflejan mejor el desempeño cuando existe desbalance.

7. Data Augmentation

El **Data Augmentation** o aumento de datos es una técnica ampliamente utilizada en el campo del aprendizaje profundo, especialmente en problemas de visión por computadora. Su objetivo principal es **aumentar artificialmente la cantidad y diversidad de los datos de entrenamiento** a través de transformaciones controladas, sin necesidad de recolectar nuevas muestras.

En contextos médicos, el uso de augmentation tiene doble relevancia:

1. **Compensar el desbalance de clases.**
2. **Mejorar la capacidad de generalización del modelo**, reduciendo el riesgo de sobreajuste.

En este proyecto, el Data Augmentation se planteó como la estrategia más adecuada para equilibrar las clases y enriquecer la representación de los nódulos benignos, que estaban en menor proporción.

7.1 Técnicas aplicadas

Se diseñó un conjunto de transformaciones con base en las mejores prácticas en análisis de imágenes médicas.

1. Rotación:

- Se aplicaron rotaciones aleatorias en el rango de $\pm 15^\circ$.
- Justificación: un nódulo puede visualizarse en diferentes orientaciones durante una ecografía; rotar la imagen no altera su diagnóstico clínico.

2. Flips horizontales y verticales:

- Se implementaron giros horizontales y verticales con probabilidad de 0.5.
- Justificación: la simetría en imágenes ecográficas permite que esta transformación sea válida sin perder información clínica.

3. Zoom aleatorio:

- Se realizaron aumentos o reducciones hasta un 20%.

- Justificación: simula cambios en el nivel de acercamiento del ecógrafo, reforzando la robustez del modelo ante variaciones en la adquisición.

4. Variación de brillo y contraste:

- Ajustes leves dentro de $\pm 10\%$ respecto al valor original.
- Justificación: diferentes ecógrafos y configuraciones de máquina producen imágenes con intensidades variables. Esta técnica ayuda al modelo a no depender de un nivel fijo de brillo.

5. Adición de ruido gaussiano controlado:

- Se incorporó ruido con media cercana a 0 y varianza baja.
- Justificación: emula el ruido presente en los dispositivos de imagen médica y mejora la robustez del modelo frente a variaciones del hardware.

6. Transformaciones combinadas:


- Algunas imágenes fueron sometidas a dos o más transformaciones (ej. rotación + cambio de brillo).
- Justificación: aumenta la variabilidad y previene que el modelo memorice patrones irrelevantes.

7.2 Ejemplo de implementación técnica

En *TensorFlow/Keras*, las técnicas de augmentation se aplicaron en tiempo real durante el entrenamiento mediante la clase `ImageDataGenerator` o la API moderna de `tf.image`.

Ejemplo simplificado en pseudocódigo:

python

 Copy code

```
from tensorflow.keras.preprocessing.image import ImageDataGenerator

datagen = ImageDataGenerator(
    rotation_range=15,
    width_shift_range=0.1,
    height_shift_range=0.1,
    zoom_range=0.2,
    horizontal_flip=True,
    vertical_flip=True,
    brightness_range=[0.9, 1.1]
)

datagen.fit(x_train) # x_train corresponde al conjunto de entrenamiento
```

Este enfoque asegura que en cada época de entrenamiento, el modelo vea **versiones ligeramente diferentes** de las mismas imágenes, aumentando efectivamente el tamaño del dataset y mejorando la generalización.

7.3 Consideraciones clínicas

En un contexto médico, es indispensable que el augmentation preserve las características críticas para el diagnóstico:

- Transformaciones como rotaciones, flips o cambios de brillo **no alteran la naturaleza del nódulo**.
- No se aplicaron deformaciones geométricas agresivas (por ejemplo, *shear* o estiramientos extremos), ya que podrían introducir artefactos que confundan al modelo.
- No se utilizó *color jittering* más allá de los canales de escala de grises, dado que la ecografía no depende de colores reales.

Estas restricciones garantizan que las imágenes sintéticas sean **clínicamente realistas** y útiles para entrenar un modelo confiable.

7.4 Impacto esperado

Con la aplicación de augmentation se lograron los siguientes beneficios:

- **Balancear las clases:** se generaron más ejemplos de la clase benigna, reduciendo el sesgo hacia los malignos.
- **Aumentar la robustez:** el modelo se expone a variaciones similares a las que encontraría en la práctica clínica (diferentes orientaciones, niveles de brillo, acercamiento del dispositivo).

- **Reducir el overfitting:** al incrementar la diversidad de ejemplos, el modelo evita memorizar las imágenes originales y aprende patrones más generales.

7.5 Evaluación de la técnica

El éxito del augmentation se evalúa de forma indirecta a través del desempeño del modelo en la fase de validación:

- Se espera que las métricas **F1-score** y **AUC-ROC** mejoren en comparación con un modelo entrenado sin augmentation.
- Un indicio positivo será la **reducción de la brecha entre desempeño en entrenamiento y validación**, lo que confirma menor sobreajuste.