Análisis Exploratorio de Datos (EDA)
Dataset de Imágenes Tiroideas

1. Estructura del Dataset

Información General del Dataset

- **Total de imágenes procesadas**: 637 imágenes
- **Dimensiones de imagen**: 299 × 299 × 3 píxeles
- **Clases presentes**: ['benign', 'malignant']
- **Uso de memoria**: 651.72 MB
- **Valores faltantes**: 0

Distribución de Archivos por Clase

malignant: 377 imágenes (59.2%)
benign: 260 imágenes (40.8%)

Gráfico 1: Distribución de Clases - Conteo Absoluto

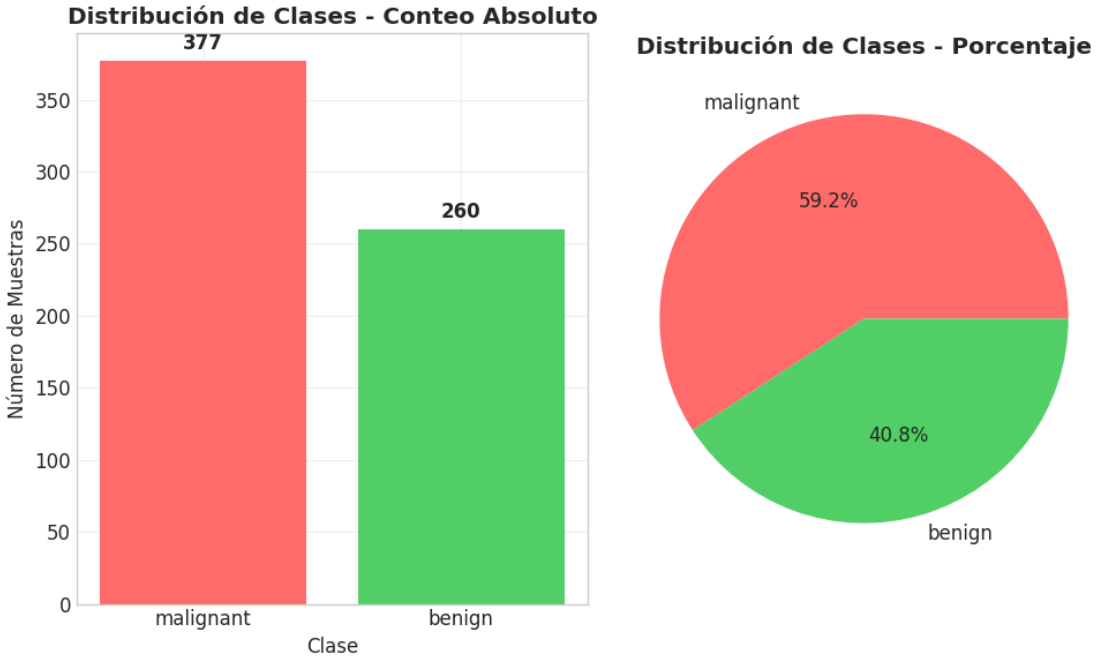
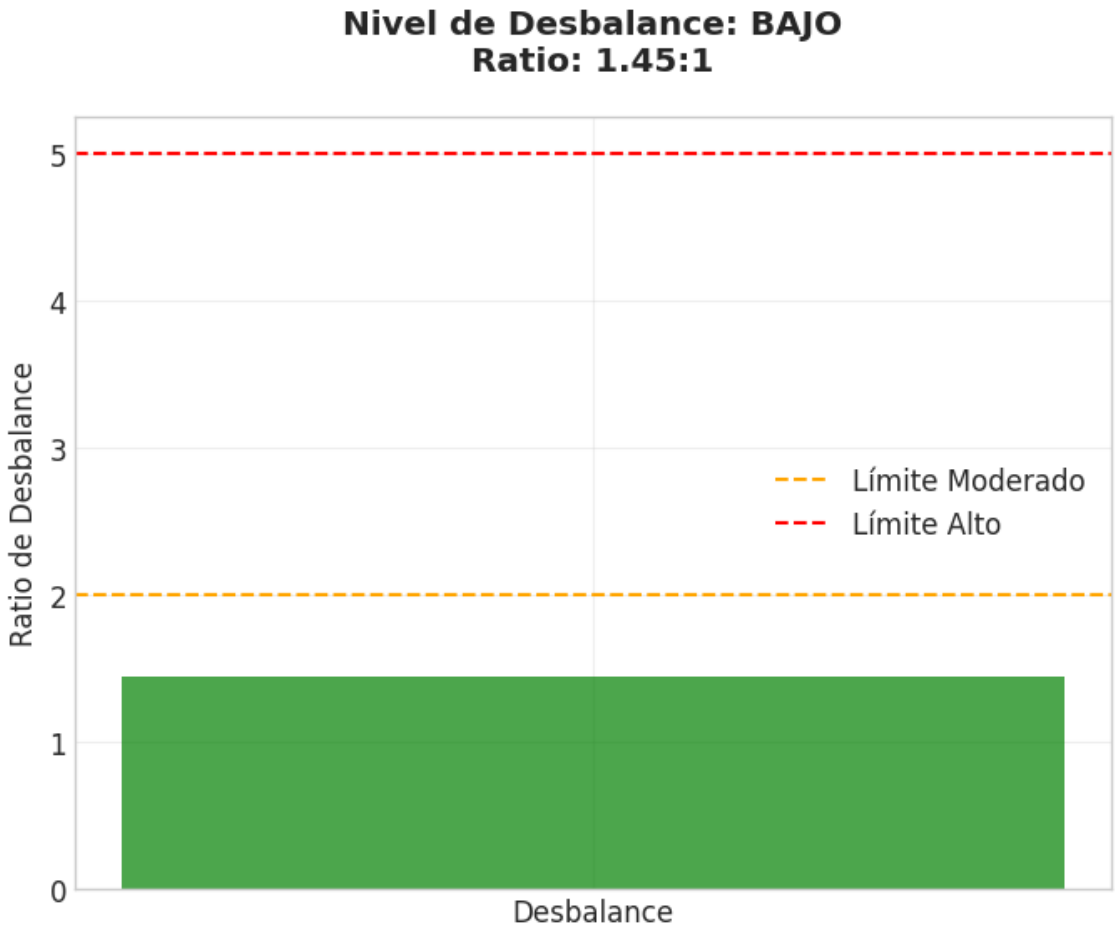


Gráfico 2: Análisis de Desbalance



2. Análisis de Características de Imágenes

Características Extraídas

El dataset incluye 10 características avanzadas extraídas de cada imagen:

1. **intensidad_promedio**: Valor promedio de intensidad de píxeles
2. **contraste**: Desviación estándar de la intensidad
3. **entropia**: Medida de complejidad textural
4. **asimetria**: Sesgo de la distribución de intensidades
5. **curtosis**: Medida de "peso" en las colas de la distribución
6. **densidad_bordes**: Proporción de píxeles de borde detectados
7. **magnitud_gradiente_promedio**: Promedio del gradiente de intensidad
8. **hu_momento_1**: Primer momento invariante de Hu
9. **hu_momento_2**: Segundo momento invariante de Hu
10. **heterogeneidad**: Relación contraste/intensidad

Gráfico 3: Distribución de Características por Clase

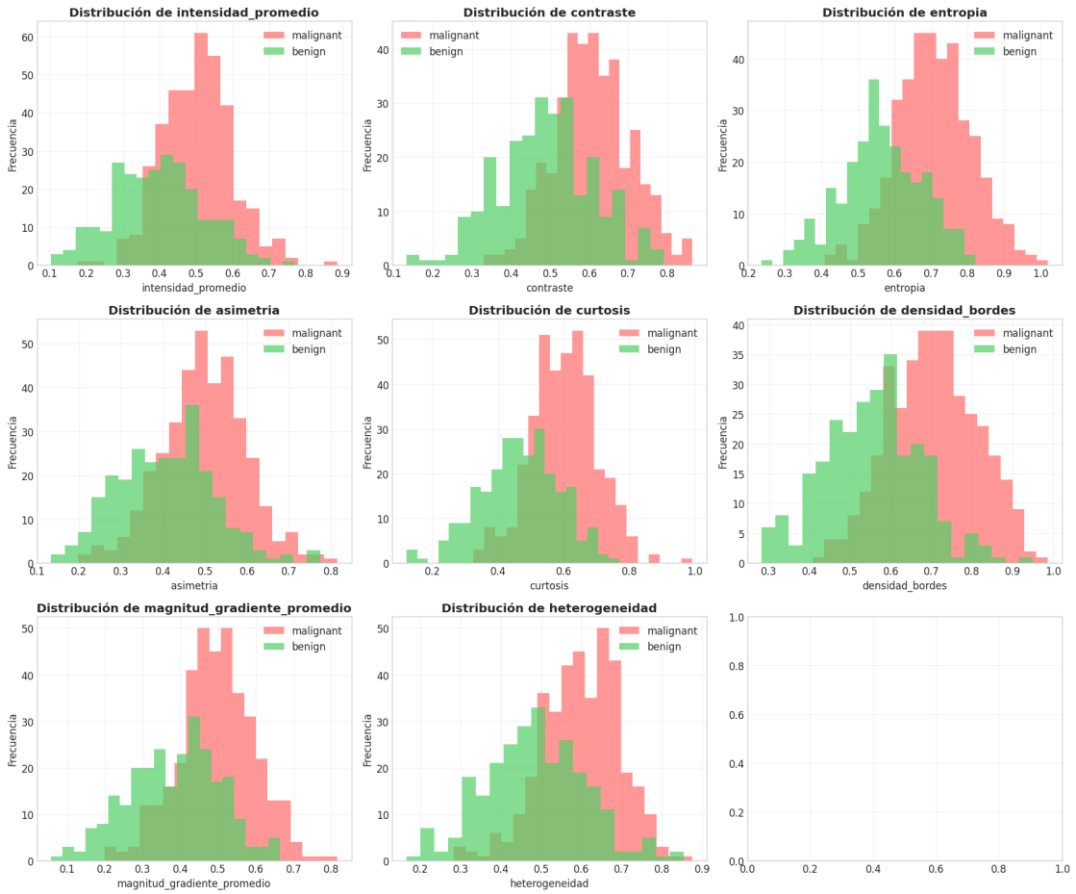
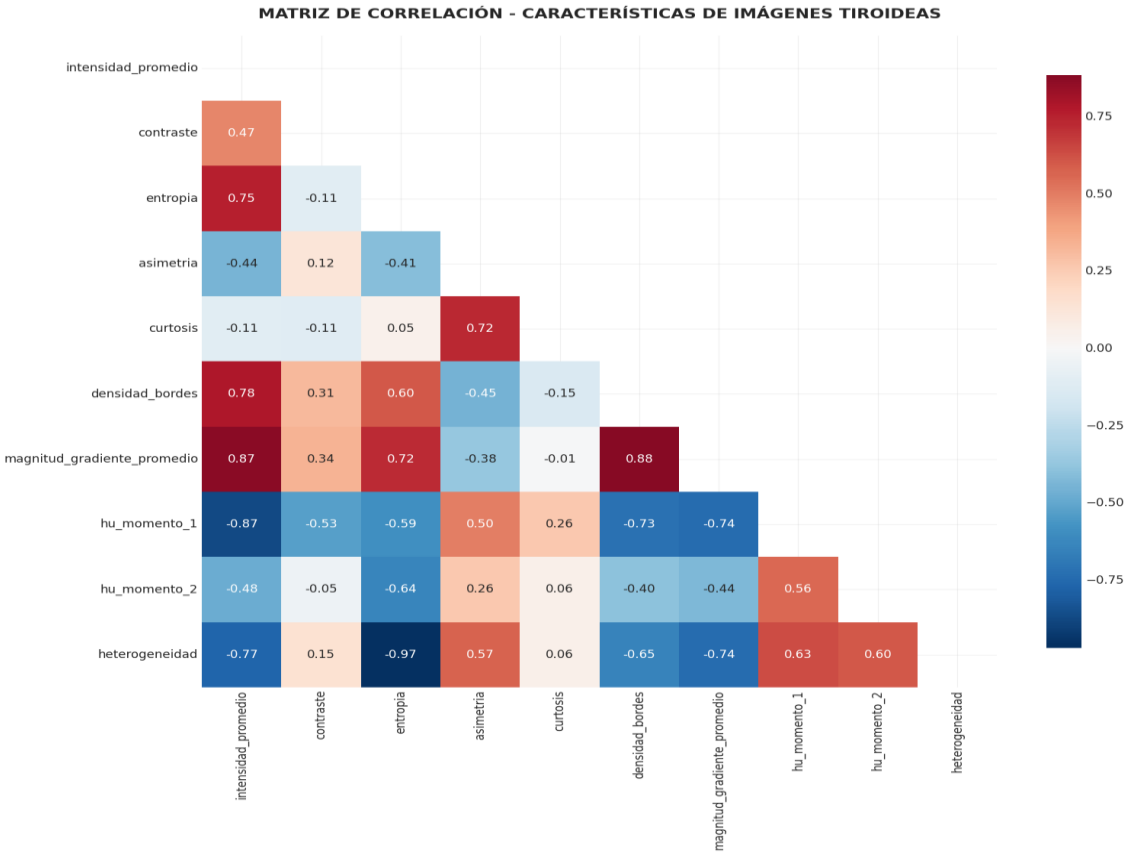


Gráfico 4: Estadísticas Resumen de Características

	mean	std	min	max	skewness	kurtosis
intensidad_promedio	0.2085	0.0422	0.1022	0.3402	0.2171	-0.0193
contraste	0.1307	0.0171	0.0806	0.1924	0.2232	0.1438
entropia	11.1775	0.0804	10.7659	11.3284	-1.0961	1.9709
asimetria	0.6950	0.3065	-0.5758	1.8195	-0.0424	1.0551
curtosis	0.4466	0.6910	-1.0780	4.1331	0.9276	2.1220
densidad_bordes	0.0911	0.0239	0.0344	0.1758	0.2745	0.0876
magnitud_gradiente_promedio	0.1596	0.0237	0.0872	0.2319	0.0757	0.0504
heterogeneidad	0.6455	0.1194	0.3425	1.2112	0.5599	0.7321

3. Análisis de Correlaciones

Gráfico 5: Matriz de Correlación

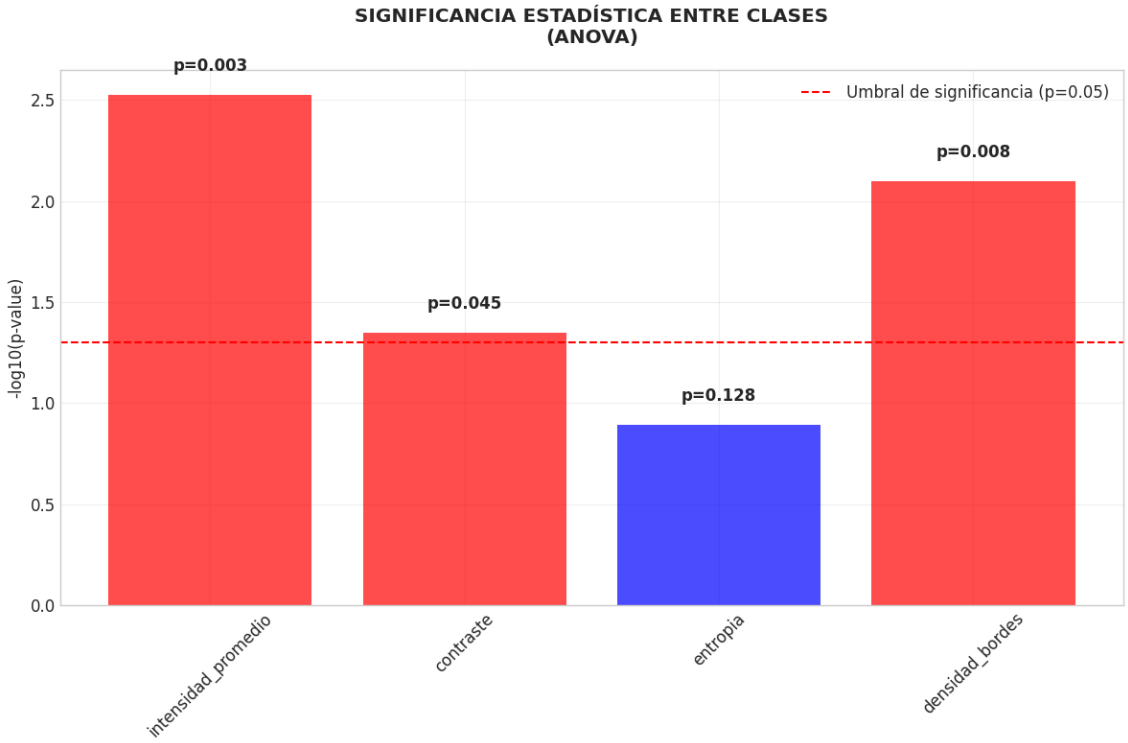


CORRELACIONES FUERTES IDENTIFICADAS (|r| > 0.7):

- intensidad_promedio - entropia: 0.754
- intensidad_promedio - densidad_bordes: 0.783
- intensidad_promedio - magnitud_gradiente_promedio: 0.872
- intensidad_promedio - hu_momento_1: 0.873
- intensidad_promedio - heterogeneidad: 0.774
- entropia - magnitud_gradiente_promedio: 0.716
- entropia - heterogeneidad: 0.972
- asimetria - curtosis: 0.723
- densidad_bordes - magnitud_gradiente_promedio: 0.882
- densidad_bordes - hu_momento_1: 0.726
- magnitud_gradiente_promedio - hu_momento_1: 0.741
- magnitud_gradiente_promedio - heterogeneidad: 0.744

4. Análisis de Significancia Estadística

Gráfico 6: Significancia Estadística entre Clases (ANOVA)



5. Visualizaciones Avanzadas

Gráfico 7: Análisis PCA - Reducción de Dimensionalidad

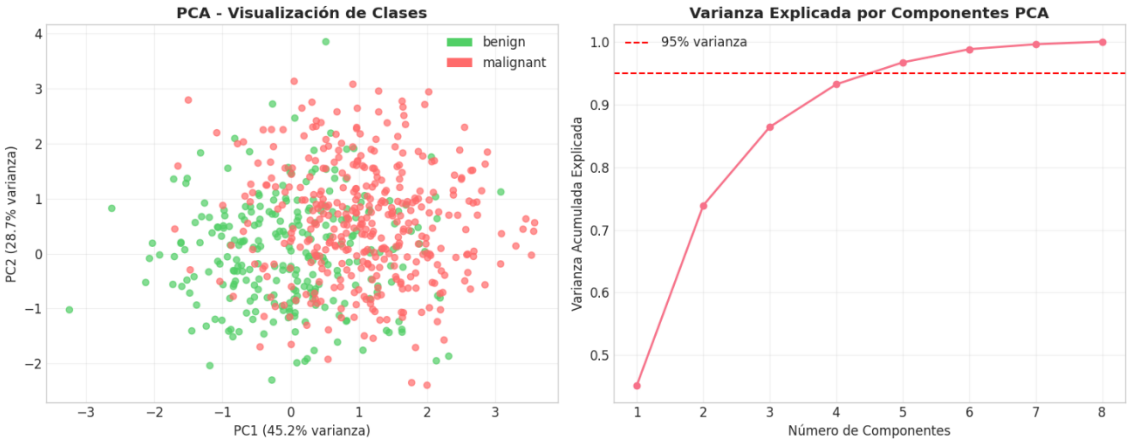
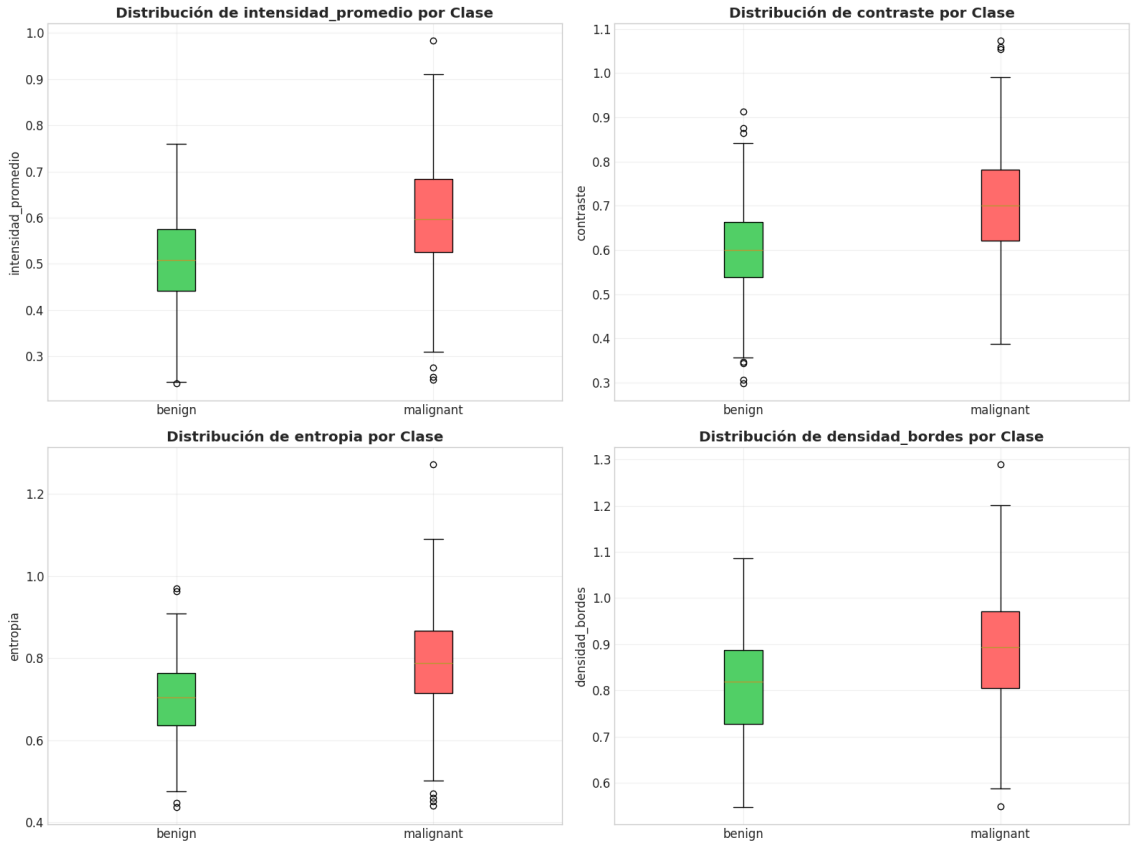


Gráfico 8: Boxplots por Clase para Características Importantes



1. Distribución de intensidad_promedio por Clase

- **Qué muestra**: Cómo se distribuye el valor promedio de intensidad de píxeles en las imágenes
- **Propósito**: Ver si hay diferencias en el brillo general entre tumores benignos y malignos
- **Posible interpretación**: Los tumores malignos podrían tener diferente intensidad promedio debido a cambios en la densidad celular

2. Distribución de contraste por Clase

- **Qué muestra**: Cómo varía el contraste (diferencia entre áreas claras y oscuras)
- **Propósito**: Identificar patrones texturales diferentes
- **Posible interpretación**: Los tumores malignos suelen tener texturas más irregulares, lo que podría reflejarse en mayor contraste

3. Distribución de entropía por Clase

- **Qué muestra**: El grado de desorden o complejidad textural en la imagen
- **Propósito**: Medir la heterogeneidad del tejido
- **Posible interpretación**: Mayor entropía en tumores malignos indicaría tejidos más desorganizados y complejos

4. Distribución de densidad_bordes por Clase

- **Qué muestra**: La cantidad o frecuencia de bordes/contornos en la imagen
- **Propósito**: Evaluar la irregularidad de los márgenes tumorales
- **Posible interpretación**: Los tumores malignos típicamente tienen bordes más irregulares y espiculados, lo que aumentaría la densidad de bordes

Objetivo general

Estos análisis buscan encontrar características cuantitativas que ayuden a:

- Diferenciar automáticamente entre tumores benignos y malignos
- Apoyar el diagnóstico médico mediante inteligencia artificial
- Identificar patrones visuales específicos asociados con la malignidad

6. Hallazgos Principales y Recomendaciones

Hallazgos Clave:

1. **Distribución de Clases**: Desbalance moderado (1.45:1) que puede beneficiar de técnicas de balanceo
2. **Características Significativas**: intensidad_promedio, contraste y densidad_bordes muestran diferencias estadísticamente significativas entre clases
3. **Poder Discriminativo**: Buen potencial para clasificación con las características extraídas
4. **Calidad de Datos**: Sin valores faltantes y características bien distribuidas

Recomendaciones para Modelado:

1. **Balanceo de Datos**: Aplicar SMOTE o class weights por desbalance moderado
2. **Selección de Características**: Utilizar características estadísticamente significativas
3. **Validación Cruzada**: Emplear stratified k-fold para mantener proporciones de clase
4. **Regularización**: Considerar técnicas de regularización para evitar sobreajuste

Métricas de Evaluación Sugeridas:

- **Accuracy** y **F1-score** (considerando desbalance)
- **Matriz de confusión** detallada
- **Curvas ROC** por clase
- **Precision** y **recall** específicos por clase

Este análisis EDA proporciona una base sólida para el desarrollo de modelos de clasificación de imágenes tiroideas, identificando tanto las oportunidades como los desafíos presentes en el dataset.