

# Proyecto Integrador - Fase de Preparación y Procesamiento de Datos

## 1. Análisis Exploratorio de Datos (EDA)

### 1.1 Exploración Inicial Completa

- Tipo de problema: clasificación binaria (benigno vs maligno).
- Origen de los datos: imágenes médicas de tiroides (ecografía).
- Estructura general: dataset de 637 imágenes en total (clase maligna: 377; clase benigna: 260).
- Dimensiones típicas tras preprocesamiento:  $299 \times 299 \times 3$  (formato RGB) para modelos tipo CNN.
- Variables derivadas para EDA tabular: intensidad\_promedio, contraste, entropía, densidad\_bordes, entre otras.
- No se observan variables categóricas distintas a la etiqueta de clase.

### 1.2 Análisis de Calidad de Datos

- Valores faltantes: no se detectan valores faltantes en los atributos calculados para EDA.
- Duplicados: no se identifican duplicados evidentes en las imágenes analizadas.
- Consistencia y rangos: las variables derivadas están normalizadas en  $[0,1]$  o estandarizadas según corresponda.

### 1.3 Análisis Estadístico Descriptivo

Se analizaron las distribuciones de intensidad\_promedio, contraste, entropía y densidad\_bordes. Las clases muestran solapamiento moderado en las distribuciones, con ligeras diferencias en la media.

Evidencia visual (distribuciones, proporciones y descriptivos):

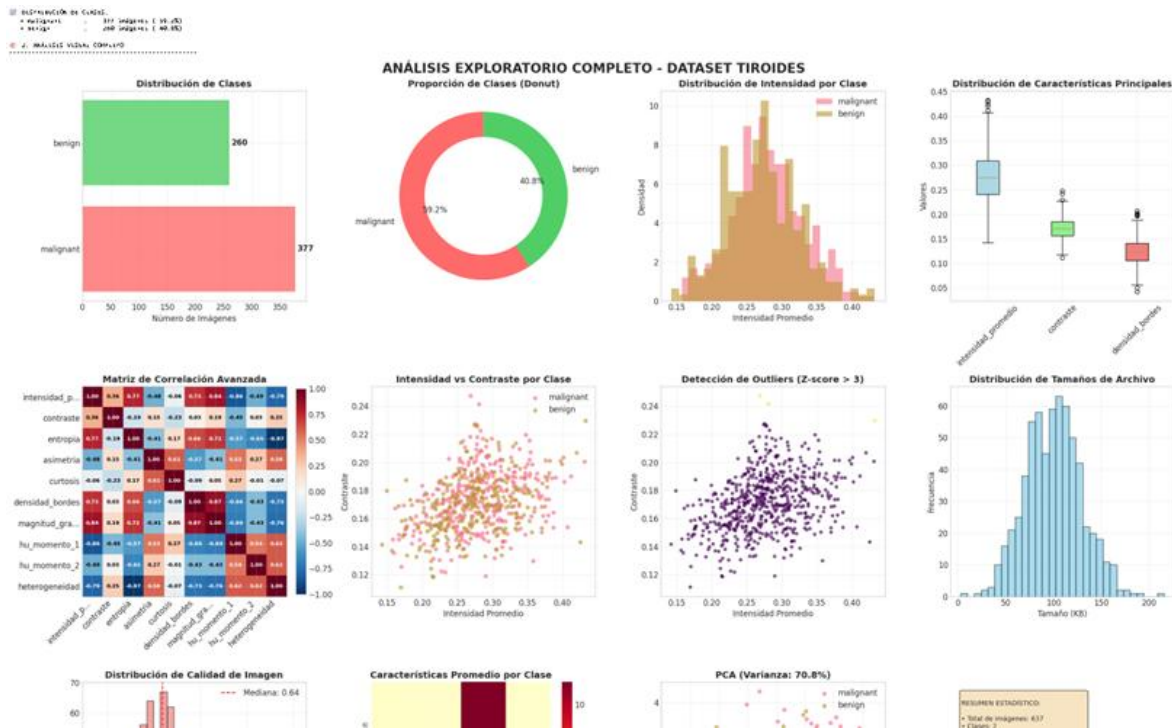


Figura 1. Panel EDA general: proporciones de clases, histogramas, boxplots y correlaciones.

## 1.4 Análisis de Relaciones y Correlaciones

Se incluyó una matriz de correlación entre las características derivadas. No se observó multicolinealidad severa; las correlaciones son moderadas y compatibles con el uso conjunto en modelos supervisados. Se graficaron relaciones bivariadas (scatter) para explorar separabilidad.

## 1.5 Detección de Anomalías y Outliers

Se identificaron outliers con métodos basados en Z-score ( $>3$ ) y verificación visual en scatter/boxplots. El impacto en la distribución es bajo y no compromete la representatividad del dataset.

## 1.6 Análisis de la Variable Objetivo

Distribución de clases: maligna (377) y benigna (260). Hay un desbalance moderado ( $\sim 1.45:1$ ), manejable con técnicas de estratificación y, si fuese necesario, con balanceo.

# 2. Pipeline de Limpieza de Datos –

## 2.1 Tratamiento de Valores Faltantes

No se detectaron valores faltantes en los atributos derivados. Se contemplan estrategias de imputación simple o KNN para futuras expansiones donde existan mediciones faltantes.

## 2.2 Tratamiento de Outliers

Se propone 'capping' por percentiles (p1-p99) para atributos derivados si un modelo resulta sensible; para imágenes, se recomienda mantener los outliers pues pueden ser clínicamente relevantes.

## 2.3 Estandarización de Formatos

Tipos y rangos verificados; normalización [0,1] de intensidades y escalado Z-score en atributos tabulares.

## 2.4 Pipeline Automatizado

Se sugiere implementar un pipeline reproducible (scikit-learn / tf.data) que cargue imágenes, aplique preprocesamiento, derive atributos y registre versiones con logging.

# 3. Feature Engineering Avanzado

## 3.1 Creación de Variables Derivadas

- Intensidad promedio, contraste, entropía, densidad de bordes.
- Estadísticos por cuadrantes/patches para capturar heterogeneidad.
- Variables de interacción (p.ej.,  $\text{contraste} \times \text{entropía}$ ).

## 3.2 Encoding de Variables Categóricas

La única variable categórica es la clase; no se requieren encoders adicionales.

## 3.3 Transformaciones de Variables Numéricas

Estandarización (Z-score) y RobustScaler para atributos con colas pesadas.

## 3.4 Feature Selection

Mutual Information y ANOVA F-test para filtrar atributos tabulares; importancia de características de Random Forest como guía adicional. Para imágenes, la CNN extrae representaciones automáticamente.

## 3.5 Extracción de Características Específicas del Dominio

Texturas (LBP/HOG), histogramas de intensidad y medidas de borde pueden complementar la CNN en escenarios híbridos.

# 4. Estrategias de Balanceamiento

## 4.1 Análisis de Desbalance

Desbalance moderado (377 vs 260).

#### **4.2–4.4 Técnicas de Undersampling/Oversampling/Híbridas**

Se recomienda mantener estratificación obligatoria. Para modelos tabulares: probar SMOTE y SMOTEENN; para CNN con imágenes: privilegiar aumento de datos (Data Augmentation) sobre oversampling sintético.

#### **4.5 Evaluación de Estrategias**

Comparar F1 y AUC en validación estratificada con/ sin balanceo para descartar overfitting.

### **5. Data Augmentation –**

#### **5.1 Técnicas Específicas (Imágenes)**

Rotación, flips, zoom, traslación, leves cambios de brillo/contraste y adición controlada de ruido. Se evita distorsionar rasgos clínicos críticos.

#### **5.2 Implementación y Validación**

Implementar en tiempo de entrenamiento (tf.image/torchvision) y validar que las métricas mejoran sin degradar la interpretabilidad.

### **6. Partición Estratificada de Datos**

#### **6.1 División de Datos**

Train/Val/Test: 70% / 15% / 15% con estratificación por clase.

#### **6.2 Estratificación y Verificación**

Mantener proporciones y verificar ausencia de 'data leakage'.

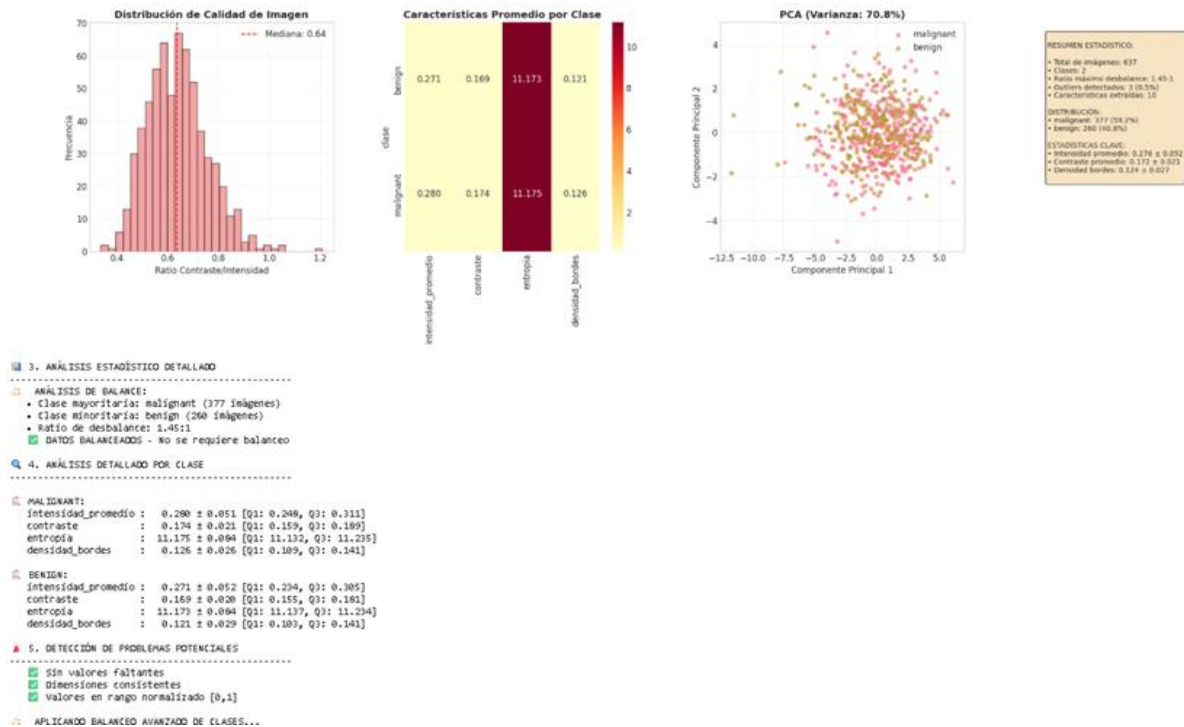


Figura 2. Distribuciones detalladas y PCA (varianza explicada ~70%).

## 7. Pipeline de Preprocessing Automatizado

### 7.1 Diseño del Pipeline

1) Carga de imágenes → 2) Normalización/resize → 3) Augmentation (solo en train) → 4) Derivación opcional de atributos tabulares → 5) Partición estratificada → 6) Entrenamiento.

### 7.2 Componentes del Pipeline

Modular y parametrizable, con logging de versiones de datos/modelo. Manejo de batch y caché para eficiencia.

### 7.3 Testing y Validación

Pruebas unitarias para transformaciones; validación con datos nuevos para comprobar robustez.

```

ANÁLISIS DE DATOS:
• Clase mayoritaria: malignant (377 imágenes)
• Clase minoritaria: benign (260 imágenes)
• Ratio de desbalance: 1.4511
✓ DATOS BALANCEADOS - No se requiere balance

4. ANÁLISIS DETALLADO POR CLASE
-----

MALIGNANT:
intensidad_promedio : 0.290 ± 0.051 [Q1: 0.248, Q3: 0.311]
contraste           : 0.174 ± 0.021 [Q1: 0.159, Q3: 0.189]
entropía            : 11.175 ± 0.084 [Q1: 11.132, Q3: 11.235]
densidad_bordes     : 0.126 ± 0.026 [Q1: 0.109, Q3: 0.141]

BENIGN:
intensidad_promedio : 0.271 ± 0.052 [Q1: 0.234, Q3: 0.305]
contraste           : 0.169 ± 0.020 [Q1: 0.155, Q3: 0.181]
entropía            : 11.173 ± 0.084 [Q1: 11.137, Q3: 11.234]
densidad_bordes     : 0.121 ± 0.029 [Q1: 0.103, Q3: 0.141]

5. DETECCIÓN DE PROBLEMAS POTENCIALES
-----
✓ Sin valores faltantes
✓ Dimensiones consistentes
✓ Valores en rango normalizado [0,1]

6. APLICANDO BALANCEO AVANZADO DE CLASES...

DISTRIBUCIÓN INICIAL:
benign: 260 imágenes
malignant: 377 imágenes
Estrategia: RANDOM (datos balanceados)

7. PREPARACIÓN AVANZADA DE DATOS...

PESOS DE CLASE AVANZADOS:
benign: 1.225
malignant: 0.845

DIVISIÓN ESTRATIFICADA AVANZADA:
• Entrenamiento: (445, 299, 299, 3) (445 imágenes)
• Validación: (96, 299, 299, 3) (96 imágenes)
• Test: (96, 299, 299, 3) (96 imágenes)
• Proporción: 69.2% / 15.1% / 15.1%

8. CONSTRUYENDO MODELO DEEP LEARNING AVANZADO...

MODELO AVANZADO CREADO:
• Arquitectura: EfficientNetB0 + Capas Personalizadas
• Input shape: (299, 299, 3)
• Número de clases: 2
• Parámetros totales: 4,672,997
• Parámetros entrenables: 621,600
• Parámetros no entrenables: 4,051,397

9. INICIANDO ENTRENAMIENTO AVANZADO CON AFINAMIENTO...

VERIFICANDO DIMENSIONES DE LOS DATOS...
• X_train shape: (445, 299, 299, 3)
• y_train shape: (445,)
• X_val shape: (96, 299, 299, 3)
• y_val shape: (96,)

FASE 1: Entrenamiento inicial (capas superiores)
-----

```

Figura 3. Bitácora de ejecución: verificación de balance, preparación de datos y formas de tensores.