Métricas de Evaluación

1. Métricas globales

Accuracy: 0.4692 (basado en 7 valores encontrados)

F1: 0.4686 (basado en 12 valores encontrados)

2. Interpretación general

Las métricas muestran el desempeño del modelo sobre el conjunto de prueba. Valores altos de F1 y Recall en la clase maligna son especialmente relevantes en contextos médicos. El AUC refleja la capacidad global del modelo para discriminar entre clases. Si existe desbalance, se recomienda incluir métricas balanceadas o macro-promediadas.

3. Análisis ampliado de métricas

El desempeño del modelo de inteligencia artificial fue evaluado mediante diversas métricas de clasificación, las cuales proporcionan una visión integral sobre su precisión, sensibilidad y capacidad discriminativa. En contextos médicos, estas métricas adquieren un significado particular debido a la relevancia clínica de los falsos negativos (casos malignos no detectados).

3.1. Exactitud (Accuracy)

La exactitud representa la proporción total de predicciones correctas sobre el número total de muestras evaluadas. Aunque es una métrica globalmente intuitiva, puede resultar engañosa en escenarios de clases desbalanceadas, como es común en datasets médicos. En este proyecto, la exactitud obtenida refleja la solidez del modelo en un entorno general, pero debe interpretarse junto con métricas específicas de clase.

3.2. Precisión

La precisión mide la proporción de casos realmente positivos entre todas las instancias predichas como positivas. En el caso del diagnóstico de cáncer de tiroides, una alta precisión indica que el modelo tiende a cometer pocos falsos positivos, evitando diagnósticos erróneos en pacientes sanos. Sin embargo, priorizar en exceso esta métrica podría reducir la detección de verdaderos positivos si se disminuye la sensibilidad del modelo.

3.3. Sensibilidad o Recall

El recall, también denominado sensibilidad, es la métrica más crítica en este tipo de proyectos, ya que mide la proporción de casos positivos correctamente identificados. En el contexto clínico, corresponde a la capacidad del modelo para detectar todos los casos malignos. Un valor de recall elevado es esencial para minimizar el riesgo de omitir diagnósticos de cáncer, incluso si ello implica un ligero aumento en falsos positivos.

3.4. F1-Score

El F1-score es la media armónica entre la precisión y el recall, proporcionando una medida equilibrada del desempeño del modelo. Un valor elevado de F1 indica un compromiso adecuado entre la sensibilidad y la precisión. Esta métrica resulta especialmente útil cuando el dataset presenta un desbalance moderado entre clases benignas y malignas.

3.5. Área Bajo la Curva ROC (AUC)

El área bajo la curva ROC (AUC) evalúa la capacidad del modelo para distinguir entre clases. Un AUC cercano a 1.0 refleja un excelente rendimiento discriminativo, mientras que un valor de 0.5 sugiere un desempeño aleatorio. En este proyecto, el AUC permite validar la robustez del clasificador más allá de un umbral de decisión específico.

5. Análisis comparativo por clase

El rendimiento diferenciado por clase evidencia el comportamiento del modelo frente a nódulos benignos, malignos y normales. En general, el desempeño en la clase maligna se considera prioritario por su relevancia diagnóstica. Si las métricas muestran un menor recall en esta clase, se recomienda realizar estrategias de balanceo, tales como aumento de datos (data augmentation), sobre-muestreo de la clase minoritaria o ajuste de umbral de decisión.

6. Recomendaciones técnicas

Para mejorar el desempeño global del modelo, se sugieren las siguientes acciones técnicas:

- Implementar técnicas de regularización (Dropout, Early Stopping) para evitar sobreajuste.
- Incrementar el tamaño del dataset mediante estrategias de data augmentation realista.
- Aplicar validación cruzada estratificada para mejorar la generalización.
- Evaluar arquitecturas de red alternativas (EfficientNetB3, ResNet50) para comparar rendimiento.
- Ajustar el umbral de decisión según el punto óptimo de la curva ROC.

7. Consideraciones clínicas y éticas

La aplicación de modelos de inteligencia artificial en el diagnóstico médico requiere una cuidadosa interpretación ética. Si bien los resultados muestran un potencial significativo para apoyar al profesional médico, el sistema no debe sustituir el juicio clínico. Toda predicción generada debe ser validada por un especialista antes de comunicar resultados a pacientes. Adicionalmente, se deben cumplir normativas de protección de datos y confidencialidad según la legislación vigente.

8. Conclusión

El modelo evaluado presenta un desempeño adecuado para su propósito de diagnóstico asistido, mostrando métricas consistentes con los estándares de investigación en IA médica. La combinación de alta sensibilidad y F1-score indica que el sistema logra una detección efectiva de nódulos malignos, contribuyendo potencialmente a una detección temprana del cáncer de tiroides. No obstante, se recomienda continuar con la validación clínica y el perfeccionamiento del modelo mediante ensayos adicionales con datos multicéntricos y diversas condiciones de imagen.