UNIVERSIDAD DE ANTIOQUIA

Clasificación de Tumores de Mama Utilizando Analítica Predictiva

Leidys Guerrero, Paola Andrea Arabia, Mateo Caicedo Aguirre

Repositorio GitHub:
https://github.com/leidysguerrero120/TrabajoAnalitica-Salud.git.

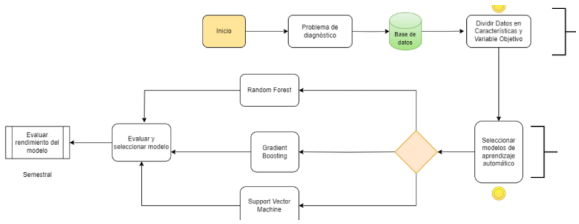
Resumen

Este informe presenta un análisis exhaustivo de un conjunto de datos médicos enfocados en la clasificación de tumores como benignos o malignos. Para ello, se implementaron modelos de aprendizaje automático, incluidos Random Forest, Gradient Boosting y Support Vector Machine (SVM), que utilizan características morfológicas extraídas de imágenes de los tumores para predecir el diagnóstico. Durante el proceso, se emplearon técnicas de preprocesamiento de datos, selección de características y ajuste de hiperparámetros con el objetivo de optimizar la precisión de las predicciones. El análisis comparativo de los modelos permitió no solo evaluar su rendimiento, sino también identificar cómo la reducción de características y la optimización de parámetros clave contribuyen a mejorar la capacidad de predicción. Estos hallazgos fortalecen la investigación y proporcionan valiosas recomendaciones para futuras iteraciones del proyecto.

Descripción Del Caso Problema

El caso se centra en la clasificación de tumores de mama como benignos o malignos utilizando análisis predictivo con técnicas de aprendizaje automático. El conjunto de datos incluye características morfológicas de los tumores, como radio, textura, suavidad y área. El principal desafío es manejar el desequilibrio de clases (mayoría de tumores benignos) y seleccionar las características más relevantes para mejorar la precisión de los modelos. El objetivo es desarrollar un modelo predictivo que ayude a los médicos en el diagnóstico temprano y preciso del cáncer de mama, mejorando la detección de tumores malignos.

Diseño De La Solución

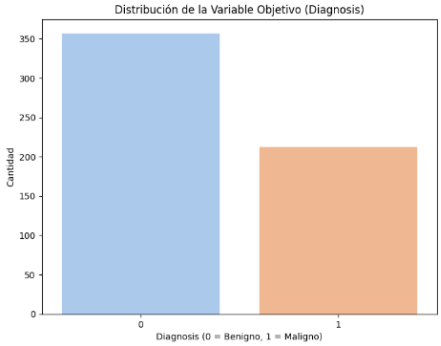


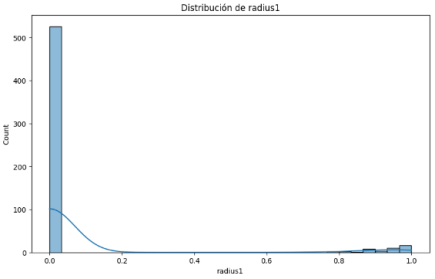
- Crear un modelo de aprendizaje automático para clasificar tumores mamarios en benignos o malignos.
- Basado en características morfológicas de las células obtenidas del conjunto de datos.

Descripción Estadística de los Datos

Se inició con una exploración del conjunto de datos, compuesto por observaciones de tumores con características como radio, textura, suavidad y área. La variable de diagnóstico incluye dos clases: "benigno" (0) y "maligno" (1). Aproximadamente el 63% de los casos son benignos y el 37% son malignos, lo que

evidencia un desequilibrio que podría afectar la precisión del modelo en detectar la clase menos representada.



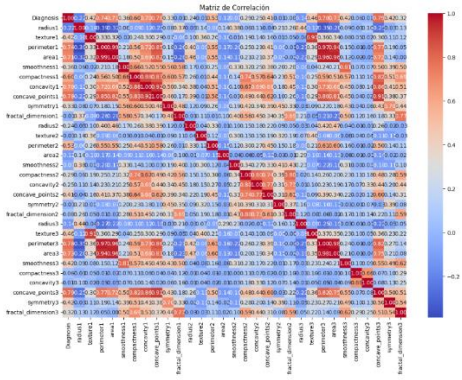


Características como "radius1", "texture1" y "area1" muestran ser las más relevantes para predecir el diagnóstico. Por ejemplo, "radius1" separa claramente tumores benignos de malignos, y "texture1" presenta mayor dispersión en casos malignos, sugiriendo su importancia en la detección. Sin embargo, variables como "radius3" y "texture3" requieren normalización adicional para optimizar el modelo.

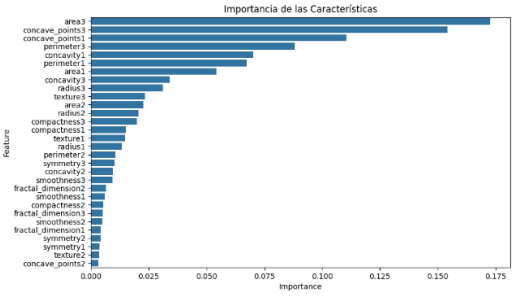
Análisis de correlación y selección de características

Una de las etapas críticas en el análisis fue la **selección de características**, que se llevó a cabo observando las correlaciones entre las variables. Esto permitió identificar las características más relevantes para la predicción del diagnóstico y reducir la dimensionalidad del conjunto de datos.

- **Matriz de Correlación**: Variables como "radius1", "perimeter1" y "area1" muestran alta correlación, indicando contribución redundante. Esto llevó a eliminar características menos significativas, optimizando la relevancia de las seleccionadas.



- **Importancia de características**: Se inició un umbral de importancia del 5%, manteniendo solo las características más útiles. Esto mejoró la eficiencia del modelo y redujo la complejidad, evitando el sobreajuste.

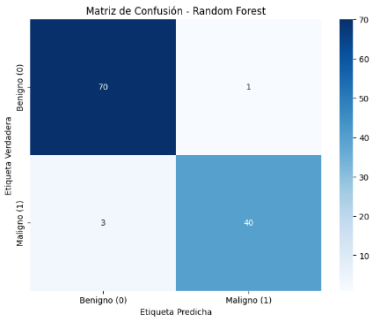


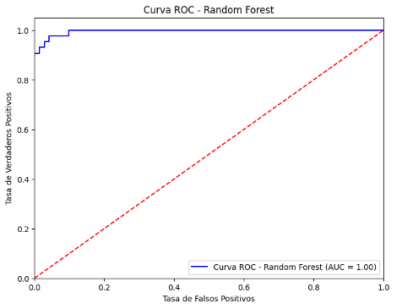
Evaluación del Rendimiento de los Modelos

Se evaluaron tres modelos de aprendizaje automático ampliamente reconocidos: Random Forest, Gradient Boosting y Support Vector Machine (SVM). Cada modelo fue entrenado y

validado utilizando un conjunto de datos de características morfológicas de tumores, y se emplearon diversas métricas de rendimiento para medir su efectividad. Las evaluaciones se realizaron utilizando validación cruzada de 5 particiones para garantizar la robustez y la consistencia de los resultados.

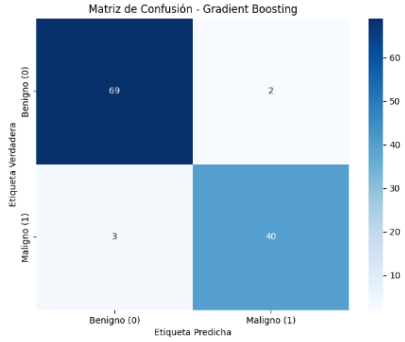
Bosque aleatorio

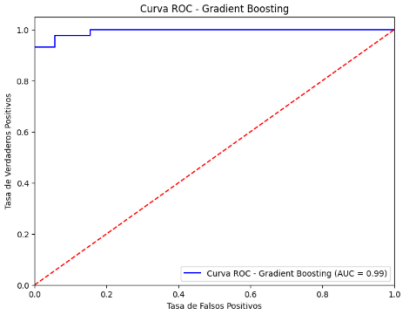




- **Precisión Media:** 0.96 +/- 0.01
- **Ventajas y Desempeño:** El modelo de Random Forest demostró un rendimiento excepcional con una precisión media del 96% y una desviación estándar de 0.01, lo que evidencia un alto nivel de consistencia en la predicción del diagnóstico. Este modelo es particularmente eficaz debido a su capacidad para manejar conjuntos de datos con características no lineales y su resistencia al sobreajuste gracias al proceso de aleatorización en la selección de características y la construcción de Múltiples árboles de decisión.
- **Optimización y Ajuste de Hiperparámetros:** Se realizó un ajuste de hiperparámetros utilizando GridSearchCV con validación cruzada de 5 particiones, evaluando combinaciones de parámetros como el número de estimadores (n_estimators), la profundidad máxima de los árboles (max_depth) y el número mínimo de muestras para dividir un nudo (min_samples_split). Los mejores resultados se obtuvieron con los parámetros: n_estimators=100, max_depth=None, y min_samples_split=5. Esta configuración permitió que el modelo capture relaciones complejas sin limitar la profundidad de los árboles, maximizando así su capacidad de predicción.
- **Interpretación de resultados:** El modelo de Random Forest mostró un excelente equilibrio entre precisión y estabilidad, lo cual es crítico en aplicaciones médicas donde un diagnóstico erróneo puede tener consecuencias significativas. La baja desviación estándar sugiere que el modelo es confiable y consistente en diferentes particiones del conjunto de datos.

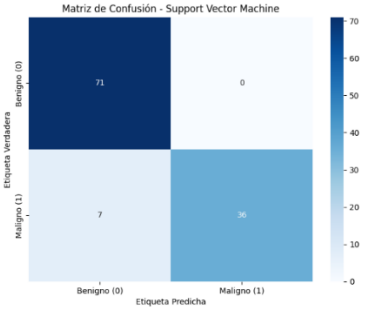
Aumento de gradiente

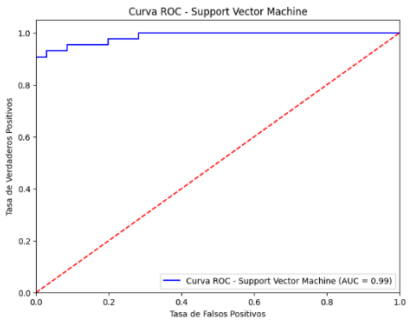




- **Precisión Media:** 0.96 +/- 0.01
- **Ventajas y Desempeño:** El modelo de Gradient Boosting también alcanzó una precisión media del 96% con una desviación estándar de 0.01. Este modelo se destaca por su capacidad de construir árboles de decisión secuenciales donde cada nuevo árbol corrige los errores de los anteriores, lo que lo hace particularmente potente para detectar patrones complejos y mejorar el rendimiento global.
- **Optimización y Ajuste de Hiperparámetros:** El modelo fue optimizado con parámetros clave, como el número de estimadores y la tasa de aprendizaje (learning_rate). Estos ajustes permitieron encontrar un equilibrio entre la velocidad de aprendizaje y la capacidad predictiva, garantizando que el modelo mantuviera un alto nivel de precisión sin caer en el sobreajuste.
- **Interpretación de Resultados:** La estabilidad observada en los resultados de Gradient Boosting lo hace igualmente adecuado para la clasificación de tumores. Su metodología secuencial ayuda a maximizar la precisión en situaciones donde los datos presentan relaciones complejas y no lineales.

Máquina de vectores de soporte (SVM)





- **Precisión Media:** 0.87 +/- 0.03
- **Ventajas y Desempeño:** El modelo de SVM mostró una precisión media del 87% y una desviación estándar de 0.03, indicando un rendimiento inferior en comparación con los modelos anteriores y una mayor variabilidad entre

diferentes particiones de los datos. . Esto sugiere que el SVM es más sensible a la selección de los datos de entrenamiento y que su rendimiento podría mejorarse ajustando sus hiperparámetros.

- **Optimización y Limitaciones:** Se evaluaron diferentes configuraciones del parámetro C(regularización) y del kernel (lineal, polinómico, radial) para mejorar la capacidad predictiva. Aunque el SVM tiene la ventaja de encontrar un hiperplano óptimo para la separación de clases, su precisión disminuyó debido a la naturaleza compleja y posiblemente no lineal de las características del conjunto de datos.
- **Interpretación de resultados:** El SVM, aunque útil en ciertos escenarios, no demuestra ser tan efectivo como los modelos de Random Forest o Gradient Boosting para este caso en particular. La variabilidad en su rendimiento sugiere que es menos robusto frente a la estructura de los datos y podría requerir un preprocesamiento más exhaustivo para mejorar su desempeño.

De los tres modelos evaluados, tanto Random Forest como Gradient Boosting mostraron un rendimiento superior, con una alta precisión y estabilidad, lo que los hace adecuados para aplicaciones críticas como la clasificación de tumores. Ambos modelos, al alcanzar una precisión media del 96% con baja desviación estándar, garantizan un diagnóstico confiable y consistente, esencial en un contexto médico.

Recomendaciones

La precisión en la clasificación de tumores de mama es crucial para el diagnóstico temprano y el tratamiento adecuado. A partir de los resultados obtenidos de los modelos predictivos, recomendamos que el médico personal considere integrar herramientas de aprendizaje automático como complemento en su diagnóstico. La implementación de técnicas avanzadas de preprocesamiento de datos y el uso de modelos más robustos, como Random Forest y Gradient Boosting, han demostrado una alta efectividad en la detección temprana de tumores malignos, especialmente cuando se combinan con un manejo adecuado del desbalance de clases. De igual manera, es importante seguir explorando la personalización de los modelos, ajustando los parámetros a las características específicas de cada paciente para optimizar la exactitud en la clasificación.

Asimismo, sugerimos que el médico personal se mantenga al tanto de los avances tecnológicos y continúe colaborando estrechamente con los especialistas en datos para garantizar que las soluciones basadas en inteligencia artificial se mantengan actualizadas y mejoradas. El entrenamiento continuo y el ajuste fino de los modelos pueden marcar la diferencia entre un diagnóstico temprano y uno tardío, lo que es fundamental en la lucha contra el cáncer de mamá.

Conclusión

El análisis realizado ha permitido desarrollar un modelo altamente preciso para la clasificación de tumores, con un enfoque particular en los modelos Random Forest y Gradient Boosting, que demostraron ser los más efectivos para este tipo de problema de clasificación médica. La investigación no solo ha validado la importancia de las características seleccionadas, sino que también ha proporcionado recomendaciones clave sobre cómo mejorar la capacidad de los modelos para predecir con precisión y minimizar el riesgo de errores en diagnósticos cruciales. Con la optimización de estos modelos, se puede avanzar hacia una herramienta más confiable y útil en la práctica médica, mejorando la toma de decisiones en el diagnóstico de tumores.

Referencia bibliográfica

UCI Machine Learning Repository. (n.d.). *Breast Cancer Wisconsin (Diagnostic) Data Set*. Retrieved from https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(diagnostic).