Análisis y Diagnóstico del Cáncer de Mama utilizando Machine Learning

Autor: [Ian Denis Marini]

Fecha: [10/072024]

Índice

1. Presentación del Problema.

2. Preguntas y Objetivos Investigados.

3. Narración de Datos.

4. Análisis Univariado y Bivariado.

5. Aplicación del Modelo de Aprendizaje Automático.

6. Evaluación del Modelo.

7. Validación Cruzada y Mejora del Modelo.

8. Conclusiones y Futuras Direcciones.

1. Presentación del Problema

El cáncer de mama es una de las principales causas de muerte entre las mujeres a nivel mundial. Un diagnóstico temprano y preciso es crucial para aumentar las posibilidades de tratamiento exitoso y supervivencia. El uso de técnicas de aprendizaje automático puede mejorar significativamente la precisión de los diagnósticos, ayudando a los médicos a tomar decisiones informadas y oportunas.

2. Preguntas y Objetivos Investigados

Preguntas de investigación:

1. ¿Existe una relación entre el tamaño medio de la masa mamaria y su diagnóstico de benigno o maligno?

2. ¿Las características relacionadas con la textura de la masa mamaria pueden predecir su malignidad?

3. ¿Hay alguna correlación entre la suavidad de la masa mamaria y su diagnóstico de cáncer?

4. ¿Las masas mamarias con mayor compacidad tienden a ser más malignas?

5. ¿La concavidad de la masa mamaria está relacionada con la severidad de su diagnóstico?

6. ¿Existen diferencias significativas en la simetría entre masas mamarias benignas y malignas?

7. ¿La dimensión fractal de las masas mamarias puede utilizarse como indicador de su malignidad?

8. ¿Se puede prever el diagnóstico de cáncer de mama utilizando únicamente características relacionadas con el perímetro y el área de la masa mamaria?

9. ¿La asimetría en las características de las masas mamarias puede indicar un mayor riesgo de cáncer?

10. ¿Existen patrones reconocibles en las características combinadas de las masas mamarias que puedan utilizarse para mejorar la precisión del diagnóstico de cáncer de mama?

3. Narración de Datos

Para este estudio, se utilizó un conjunto de datos que contiene diversas características que describen núcleos celulares en imágenes de aspiraciones con aguja fina. Estas características incluyen, entre otras, el radio, la textura, la suavidad, la compacidad y la concavidad de los núcleos celulares. Cada una de estas características puede desempeñar un papel crucial en la predicción de si una masa es benigna o maligna.

4. Análisis Univariado y Bivariado

En esta sección, se exploraron las distribuciones de las características individuales (análisis univariado) y las relaciones entre pares de características (análisis bivariado). Las visualizaciones de datos, como gráficos de dispersión y mapas de calor, se utilizaron para identificar patrones y correlaciones importantes.

5. Aplicación del Modelo de Aprendizaje Automático

Se aplicaron varios pasos de preprocesamiento de datos, incluyendo la codificación de la variable objetivo, la normalización de las características y la división de los datos en conjuntos de entrenamiento y prueba. Un modelo de clasificación Random Forest fue entrenado y evaluado utilizando estas características.

6. Evaluación del Modelo

El rendimiento del modelo fue evaluado utilizando métricas como la exactitud, la matriz de confusión y el informe de clasificación. También se analizaron las importancias de las características para identificar las variables más influyentes en la predicción del diagnóstico.

7. Validación Cruzada y Mejora del Modelo

Se aplicó la validación cruzada para evaluar la estabilidad del modelo, y se realizó una optimización de hiperparámetros utilizando GridSearchCV. El modelo mejorado fue entrenado y evaluado, y se comparó su rendimiento con el modelo original. También se analizó la curva ROC y el AUC para evaluar la capacidad del modelo de discriminar entre clases positivas y negativas.

8. Conclusiones y Futuras Direcciones

Este estudio demostró la eficacia del uso de técnicas de aprendizaje automático para el diagnóstico del cáncer de mama. Los modelos desarrollados mostraron una alta precisión en la clasificación de masas benignas y malignas, y las técnicas de optimización de hiperparámetros permitieron mejorar aún más el rendimiento del modelo. Futuras investigaciones podrían centrarse en la incorporación de datos adicionales, como información genética y antecedentes familiares, para mejorar aún más la precisión del diagnóstico.