Modelo Predictivo para la Clasificación de Tumores Mamarios relacionados con el cancer de mamá

Frida Bahena Frutos

10/06/24

Abstract

Este documento presenta un análisis detallado de las características de los tumores de cáncer de mama. Se utilizan técnicas de ciencia de datos para predecir la malignidad de los tumores. Los métodos incluyen la limpieza de datos, la normalización y el uso de modelos de clasificación como Random Forest. Los resultados muestran una alta precisión en la predicción de la malignidad de los tumores, lo que sugiere la eficacia del enfoque propuesto.

1 Introducción

Según la Organización Mundial de la Salud (OMS), en el año 2022 se diagnosticaron 2.3 millones de casos de cáncer de mama en mujeres a nivel mundial, y se registraron 670,000 defunciones debido a esta enfermedad[1]. La detección temprana y el diagnóstico preciso son esenciales para mejorar las tasas de supervivencia y la eficacia del tratamiento. Sin embargo, los métodos tradicionales de diagnóstico pueden ser invasivos, costosos y a menudo sujetos a errores humanos.

En este contexto, es crucial desarrollar métodos más rápidos y eficientes, y la ciencia de datos puede desempeñar un papel significativo en la mejora de estas tasas de supervivencia. Actualmente, se están realizando investigaciones para el diagnóstico del cáncer en humanos mediante resultados imageneológicos, como las imágenes histopatológicas,de TAC, digitalizadas de x-ray, las imágenes digitalizadas obtenidas por técnicas como la punción aspirativa con aguja fina (PAAF), etc. La PAAF consiste en extraer una muestra con una aguja fina que luego se estudia microscópicamente.

Estas imágenes se procesan para extraer características como tamaño, contraste y textura. Junto con otros datos derivados de la investigación, como los datos genómicos, proteómicos y transcriptómicos, se puede crear un modelo integrativo para predecir la supervivencia del pacienten en ciertos tipos de cáncer, mediante bosques aleatorios u otras tecnicas.

Los bosques aleatorios son algoritmos de aprendizaje automatico, que toma el resultado de multiples arboles de decisiones para obtener un resultado. Es optimo para problemas de clasificación como regresión. Los árboles de decisión se estructuran similarmente a un árbol y son jerárquicos. Es una representación de decisiones resultado de preguntas y posibles consecuencias o respuestas[2][3].

En este proyecto, se utilizan datos de tumores en la región mamaria, específicamente imágenes digitalizadas obtenidas mediante PAAF (Fig. 1).

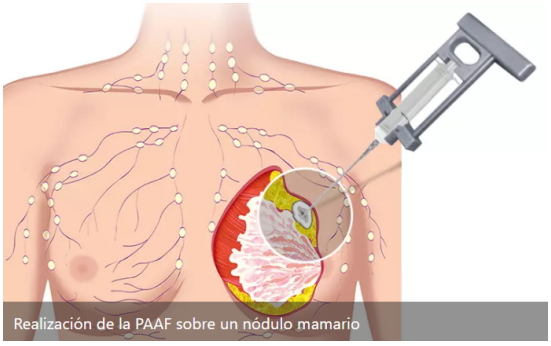

Fig.1 Ejemplo del metodo de PAAF,

técnicas de diagnostico por punción aspirativa con aguja fina.

1.1 Datos

Características de los Datos Utilizados.

El conjunto de datos se obtuvo de la plataforma kaggle, los datos son recuperados de la universidad de Wisconsin mediante la digitalizacion de imagenes obtenidas con la tecnica de diagnostico PAAF.(URL en el anexo) El conjunto de datos incluye información detallada sobre los tumores, tales como:

- Identificación del paciente
- Diagnóstico (Benigno = B o Maligno = M)
- Radio
- Textura
- Perímetro
- Área
- Suavidad (variación local en la longitud del radio)
- Compacidad (perímetro^2 / área - 1,0)
- Concavidad (severidad de las porciones cóncavas del contorno)
- Puntos cóncavos (número de porciones cóncavas del contorno)
- Simetría
- Dimensión fractal ("aproximación de la línea costera" - 1)

Para cada imagén, se calcularon la media, el error estándar y el valor más alto ("peor" o mayor, definido como la media de los tres valores más grandes), resultando en un total de 30 características por imágen.

Resultados del Conjunto de Datos: El conjunto de datos consta de un total de 569 casos, de los cuales 357 son benignos y 212 son malignos.

2 Objetivo

Desarrollar un modelo predictivo basado en técnicas de ciencia de datos que permita clasificar de manera precisa y eficiente los tumores de cáncer de mama como benignos o malignos, utilizando un conjunto de datos con características detalladas de los tumores,contribuyendo así a la detección temprana y al tratamiento eficaz del cáncer de mama.

3 Desarrollo

El desarrollo del proyecto se llevaran a cabo a en Python 3. Las librerias que se utilizaron fueron pandas para la visualización y manejo de los DataFrames, numpy para realizar operaciones matemáticas mas especificas, seaborn para facilitar la creación de gráficos al igual que matplotlib. Además de librelias de sklearn de aprendizaje automático que nos ayudara para el modelado de los datos.

3.1 Entendimiento del Problema - EDA

Con los datos en formato CSV listos para ser leídos, se utilizó la librería Pandas para visualizarlos en una tabla. A continuación, se empleó el comando info() para observar el tipo de datos, las columnas y el número de datos. Posteriormente, se realizó un análisis descriptivo de los datos y se visualizaron mediante histogramas y diagramas de caja para obtener una mejor comprensión de los mismos, para ello antes se definieron las variables numericas y las categoricas(Fig.2 y Fig. 3). En este punto, se observó que la última columna de los datos contenía únicamente valores nulos. Además, se detectó un número significativo de valores atípicos en casi todas las columnas, lo cual indica la heterogeneidad de los datos.

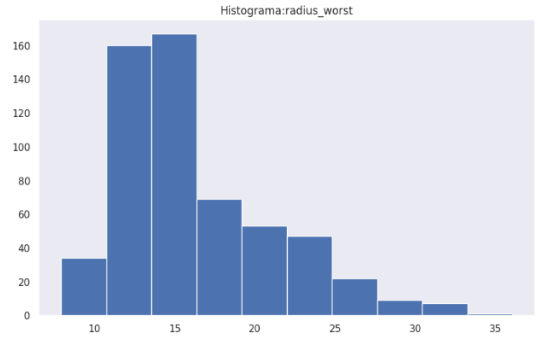

Fig.2 Ejemplo de visualizacion grafica de los datos. Histograma de la columna "radius_worst"

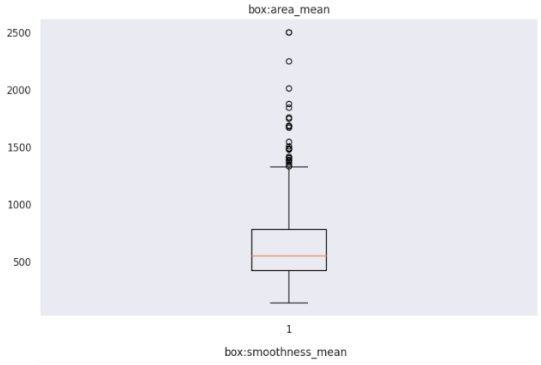

Fig.3 Ejemplo de visualizacion grafica de los datos. Diagrama de cajas de la columna "area_mean"

3.2 Procesamiento de datos

- Manejo de valores nulos.
  Como se observó, la última columna de los datos contenía únicamente valores nulos. Por lo tanto, se procedió a eliminar esta columna, denominada "Unnamed: 32", utilizando el comando .dropna(axis=1). Se denomino al DataFrame data2. Ademas, se realizo la busqueda de los valores duplicados, los cuales no se encontraron en el DataFrame.

- Estandarización de datos
  Con el fin de tener el Dataframe con mayor heterogeneidad, los datos sean mas interpretables y eliminar sesgos, se estandarizan los datos para que tengan media de 0 y desviacion estandar de 1. Para la estandarizacion de los datos se utilizaron solo las columnas numericas, por lo que se definio un avariable llamada "Columnas numericas" excluyendo la columna 'diagnosis' que es la unica columna categorica.

- Tratamiento de Outliers
  Se observo anteriormente que se tienen muchos valores atipicos en cada columna. Con el fin de tener una mayor precisión con la normalidad de los datos, evitar errores y obtener una mejor interpretacion y precisión del modelo se hizo el tratamiento de los datos. Para ello se se crea un escalador, esto con un modulo "StandardScaler" de "sklearn.preprocessing", el cual calulara la media y desviacion para luego ajustar y transformar los datos.

  Después del tratamiento de los datos se realizó una vizualización mediante diagramas de cajas y se observo todas las columnas sin datos atipicos.

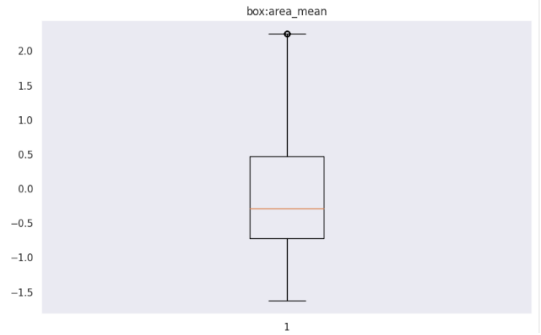

Fig.3 Ejemplo de visualizacion grafica de los datos. Diagrama de cajas de la columna "area_mean" con tratamiento de Outliers

- Cardinalidad de datos
  Para este caso, la unica columna categoricas es "Diagnosis" con los estadios del tumor M= maligno, B= benigno. Por lo que no se palico ninguna reduccion de dimencionalidad.

- Redundancia columnar
  Se observa en el heat map antes desarrollado con lamatriz de correlacion que no existen caracteristicas altamente correlacionadas. Sin embargo, para buscara columnas altamente correlacionadas en el DataFrame, eliminarlas y evitar problemas de multicolinealidad. Se calcula la matriz de correlacion de las columnas y seleccione el triángulo superior para reducir la duplicación. Luego, se identifico las columnas con una potencia superior al umbral definido y se eliminaron del

DataFrame. En este punto la data limpia se encuentra con el nombre de "data3"(Fig. 4).

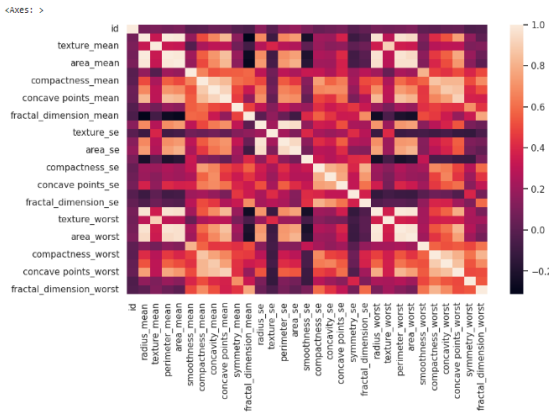

Fig. 4 Ejemplo de visualizacion grafica de los datos. Heatmap de los datos

3.3 Preparacion de datos- Parte 2

Para este proyecto con baja dimensionalidad y debido a la naturaleza de las caracteristicas. No es necesario plicar ninguna tecnica de reduccion de cardinalidad pues para las caracteristicas es crucial asegurarse que se conserven lo mejor posible para tener un modelo predictorio con alta prediccion.

3.4 Modelado de datos

Random Forest

En este proyecto se emplearon árboles de decisiones para abordar el problema, dada su idoneidad para problemas de clasificación. La tarea consistía en discernir la naturaleza benigna o maligna de los tumores en función de sus características. El uso de Random Forests en este contexto presenta varias ventajas destacadas: además de su notable facilidad de interpretación, requiere una preparación mínima de los datos y exhibe una resistencia robusta ante la presencia de valores faltantes. La utilización de árboles de decisión contribuye a mitigar el riesgo de sobreajuste, lo que puede resultar fundamental en el análisis de datos biomédicos. En particular, para este proyecto, los Random Forests ofrecen la posibilidad de obtener una precisión elevada en la clasificación de tumores.

4 Propuesta de solucion

Para utilizar los árboles de decisiones y random tree en python se utilizan las paqueterias de sklearn(tree), sklearn. ensemble (modulo = RandomForestClassifier), sklearn.model_selection(modulo = train_test_split), sklearn.metrics (modulo = classification_report, accueracy_ score).

Se comienza con dividir las columnas en dos, en este caso x e y (columnas numericas y cateórica respectivamente). Para crear los conjuntos de entrenamiento y prueba se definen variables para x de prueba y de entrenamiento, lo mismo para y. Para ello se uso la función train_test_split.

Se sigue por crear los arboles de decisión, para ellos se llamo a la funcióin DecisionTreeClassifier como ad. Para el bosque aleatorio se llamó rf y se uso la función RandomForestClassifier.

Se siguió por entrenar el modelo con las funciones .fit y los conjuntos de entrenamiento y prueba.

Para la predicción de las variables de prueba se uso la fucnion .predict() con los conjuntos te pureba y por ultimo se estimo la presicion del modelo con la función accuracy_score[4].

5 Resultados

Al aplicar este proceso se obtuvo una presición del modelo de Random Forest y árbol de decisión de de 0.96 y 0.93 respectivamente. Lo cual son resultados muy buenos por lo que el modelo predictivo presenta un alta presición de predicción (Fig. 5).

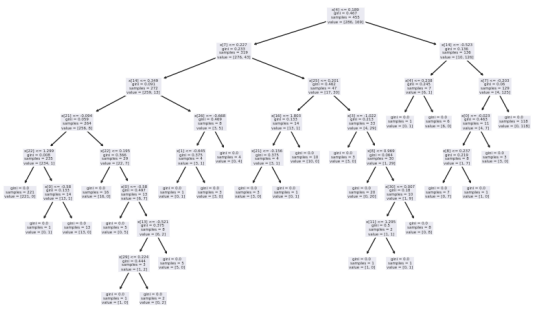

Fig. 5 Ejemplo de visualizacion grafica el arbol de decision

6 Evaluación

Para evaluar la precsion y rendimiento detallado de los resultados se utilizo la funcion classification_report con los datos de prueba de las etiquetas y los datos predecidos de las etiquetas. Esta funcion nos devuelve metricas como la precision que indica cuantas de las predicciones positivas fueron correctas, el recall que indica cuantos de los casos positivos fueron identificados correctamente , F1-Score se puede interpretar como una media armónica de la precisión y la recuperación, donde una puntuación F1 alcanza su mejor valor en 1 y su peor puntuación en 0. Support que es el numero de ocurrencias de cada clase en el conjunto prueba (Fig. 6).

```
[275] #evaluacion de la precision de la prediccion
      report_rf = classification_report(y_test, y_pred_rf)
      print(report_rf)
```

	precision	recall	f1-score	support
B	0.96	0.99	0.97	71
M	0.98	0.93	0.95	43
accuracy			0.96	114
macro avg	0.97	0.96	0.96	114
weighted avg	0.97	0.96	0.96	114

```
[276] report_ad = classification_report(y_test, y_pred_ad)
      print(report_ad)
```

	precision	recall	f1-score	support
B	0.96	0.94	0.95	71
M	0.91	0.93	0.92	43
accuracy			0.94	114
macro avg	0.93	0.94	0.93	114
weighted avg	0.94	0.94	0.94	114

Fig. 6 evaluacion de la precision y rendimiento detallado de los resultados

Con los resultados obtenidos, obserbamos un desempeño muy bueno tanto del modelo de arbol de decicion como el bosque aleatorio.

7 Conclusión

- Alta precisión y un elevado recall, lo que indica que el modelo es efectivo en la detección de tumores malignos.

- En el ámbito biomédico es significativa, ya que puede ayudar a mejorar la detección temprana y el diagnóstico preciso del cáncer de mama.

8 Tabajo a futuro

- Desarrollar un modelo predictivo de tumores enfocado en el cáncer de mama a partir de imágenes histopatológicas y crear sistemas que agilicen el diagnóstico basado en dicho modelo.

9 Datos y Codigo QR al repositorio

- https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data.
- Codigo QR

10 Referencias

1 Organizacion mundial de la salud. (2024). Cancer de mama. Recuperado de OMG: https://www.who.int/es/news-room/fact-sheets/detail/breast-cancer

2 IBM.(2024).¿Qué es el bosque aleatorio?. IBM:https://www.ibm.com/topics/random-forest

3 DataScientest.(2024).Random Forest: Bosque aleatorio. Definición y funcionamiento. Recuperado de : https://datascientest.com/es/random-forest-bosque-aleatorio-definicion-y-funcionamiento

4 scikitlearn(2024).precision_recall_fscore_support.Rec de: https://scikit-learn.org/stable/modules/generat