UNIVERSIDAD SAN FRANCISCO DE QUITO - USFQ

Colegio de Postgrados

Fundamentos de Ciencias de Datos

Proyecto 1. Detección de cáncer de mama en

Aurora Ñacato Jaya

Quito, 26 de marzo de 2025

Índice

Entendimiento del negocio/problema (*Bussines Understanding*)

El objetivo del proyecto es predecir si el cáncer de mama es maligno o benigno, en base a las características extraídas de imágenes digitalizadas de una aspiración con aguja fina PAAF de una masa mamaria. Estas imágenes describen las características de los núcleos celulares presentes en la imagen. Realizar una detección temprana de este tipo de cáncer, es clave para un adecuado tratamiento y manejo de la enfermedad. Sabiendo que el cáncer de mama está entre los tres tipos más comunes de cáncer en el 2024. (Cáncer, 2024)

Este trabajo pretende responder a la pregunta ¿Cuán efectivo puede ser un modelo predictivo para realizar la discriminación de casos malignos y benignos? En instituciones públicas donde los doctores no se dan abasto con los pacientes, este modelo, puede ser útil para agilizar el diagnóstico, reduciendo la carga de trabajo manual, que implica evaluar las imágenes.

Entendimiento de los datos (*Data Understanding*)

El dataset de este proyecto proviene del sitio web Kaggle, *"Breast Cancer Wisconsin (Diagnostic) Data Set"*. Originalmente fue recopilada por el Dr. Wolberg de sus casos clínicos desde 1989 hasta 1992, en la Universidad de Wisconsin Hospitals, Posteriormente, este conjunto de datos fue donado por el Dr. Wolberg, convirtiéndose en una valiosa fuente para investigaciones en el área de la salud. Este dataset contiene datos con 569 instancias y 33 características multivariable (*feature*s). Cabe mencionar que estos datos no tienen una actualización periódica, puesto que se tratan de un conjunto fijo de datos históricos.

El dataset describe aspectos de las imágenes de la biopsia de masa mamaria. La columna principal de diagnóstico “diagnosis”, indica si un caso es maligno (M) o benigno (B), siendo el objetivo clave del análisis. En la base los casos con diagnóstico Benigno son el 62.7% del total de la muestra; mientras que los casos Malignos representan el 37.26% de la muestra. La data ha sido tratada para que de los diez parámetros principales de la imagen se obtenga la media (mean), el error estándar (se) y el peor valor (worst), que significa que es la media de los tres valores más grandes de la muestra. Los valores que se obtiene de las imágenes computarizadas por cada núcleo celular son:

1) “*radius*”

   Radio, es la media de las distancias desde el centro a los puntos del perímetro.

   Relevancia: Un núcleo más grande puede ser una señal de crecimiento anormal, lo cual podría estar relacionado con malignidad.

2) “*texture*”

   Textura, es la desviación estándar de los valores de la escala de grises en la imagen del núcleo.

   Relevancia: Los tumores malignos a menudo tienen texturas más irregulares y heterogéneas en comparación con los benignos.

3) *“perimeter”*

   Perímetro, es la longitud total del borde del núcleo.

   Relevancia: Puede estar correlacionado con el crecimiento celular anormal, ya que un perímetro más grande podría reflejar núcleos alterados

4) *“area”*

   Área, es el tamaño total del núcleo, medido en pixeles dentro de su contorno.

Relevancia: Un área más grande podría indicar células anómalas, típicas de tumores malignos.

5) *“smoothness”*

Suavidad, es la variación local en las longitudes de los radios celulares.

Relevancia: Los tumores benignos suelen tener bordes más suaves, mientras que los malignos tienden a presentar bordes irregulares.

6) “compactness”

Compacidad, es la relación matemática entre el perímetro y el área.

Compacidad = Perimetro^2 / (area -1)

Relevancia: Tumores malignos podrían mostrar valores más altos debido a su forma irregular y menos compacta.

7) “concavity”

Concavidad, es la severidad de las porciones cóncavas del contorno del núcleo.

Relevancia: Los núcleos malignos suelen presentar concavidades más pronunciadas, reflejando alteraciones en la forma celular.

8) “concave points”

Puntos cóncavos, es el número total de las hendiduras cóncavas en el contorno celular.

Relevancia: Más puntos cóncavos pueden ser un indicador de crecimiento celular anómalo.

9) “symmetry”

Simetría, es la diferencia entre las mitades del núcleo en términos de forma y estructura.

Relevancia: Los núcleos malignos suelen ser menos simétricos debido a alteraciones celulares.

10) “fractal dimension”

Dimensión fractal, es la medida que refleja la complejidad del borde del núcleo.

Relevancia: Un valor más alto puede indicar bordes más irregulares, típicos de tumores malignos.

Las variables antes mencionadas, están directamente relacionadas con los núcleos celulares observados en las imágenes de aspiración con aguja fina (PAAF) y poseen una relación potencialmente significativa con la clasificación de cáncer de mama y se usan para entrenar el modelo predictivo. Para esclarecer de mejor manera como están representadas estas características, se detalla que mean (Promedio): representa el valor medio de la característica para el núcleo celular en la imagen. Es útil para identificar patrones generales, se (Standard Error o Error Estándar): indica la variabilidad o desviación de los valores respecto al promedio. Cuanto mayor sea el error estándar, más dispersos están los datos, lo que podría reflejar irregularidades en la muestra, worst (Mayor o "Peor"): Calcula la media de los tres valores más altos de esa característica. Es útil para capturar los casos extremos y podría ser una indicación de malignidad.

Durante la limpieza del dataset, se verificó que no existen filas duplicadas ni valores repetidos en la columna “id”. Asimismo, se llevó a cabo un análisis de valores ausentes, concluyendo que no hay columnas con datos faltantes. Sin embargo, algunos desafíos identificados podrían incluir el desbalanceo de las clases de diagnóstico. Adicional, por la forma en que está estructurado el dataset (mean, se, worst), permite hacer un análisis desde distintas perspectivas, las columnas con la clave “mean”, ayudarían a trabajar con tendencias generales. Por otro lado, las columnas con clave “se”, ayudarían a estudiar irregularidades en la data puesto que representan la dispersión de estos datos en las muestras. Las columnas con clave “worst” ayudan a observar comportamientos extremos.

Data Preparation (Preparación de los datos)

Lista las tareas de limpieza y transformación necesarias (Data Wrangling).

1) Eliminar valores duplicados y manejo de valores nulos:

   Los datos no contienen valores nulos ni duplicados.

2) Estandarización de nombres de columnas:

   Los tipos de datos no requieren realizar una transformación de tipo de datos.

   Las categorías correspondientes a los diagnósticos, benigno y maligno, han sido estandarizadas utilizando minúsculas para mantener la consistencia en el formato.

   Además, considerando que el proyecto se está desarrollando en español y será presentado a una audiencia de habla hispana, se ha realizado la traducción de los nombres de las columnas al idioma español.

3) Conversión de la columna diagnostico

   Se asigna 1 a Maligno y 0 a Benigno, esto hace que los valores sean fácilmente interpretables como etiquetas para el modelo.

4) Eliminar columnas no relevantes:

   La columna “id”, se elimina ya que no aporta información significativa para el diagnóstico de cáncer, pues su única función es servir como identificador único de cada muestra. Por lo tanto, puede excluirse del análisis sin afectar la precisión ni la interpretación del modelo predictivo.

5) Resolución del desbalanceo de clases

Resolver el desbalanceo entre las clases Maligno 212 valores y Benigno 357 valores, utilizando la estrategia SMOTE, que aumenta las muestras de la clase minoritaria maligno.

6) Normalización o escalado de características

   Se decide hacer la normalización de los datos, debido a que los valores de máximo y mínimo de las características numéricas de la data tiene valores muy distintos, y sin esta transformación, las características con valores más grandes podrían dominar el entrenamiento del modelo.

7) Revisar la relevancia de las columnas

   Los valores de las columnas con la clave “se”, presentan redundancia, pues no presentan diferencias significativas entre las clases (maligno y benigno), observados en las características de textura, radio, área, smothness, etc. no aporta información relevante para discriminar entre ellas en el modelo, por lo que se decide eliminar estas columnas con la clave “se”. Así, se simplifica el dataset, ya que no muestran patrones diferenciados.

Referencias Bibliográficas

O. L. Mangasarian and W. H. Wolberg: "Cancer diagnosis via linear programming", SIAM News, Volume 23, Number 5, September 1990, pp 1 & 18.

William H. Wolberg and O.L. Mangasarian: "Multisurface method of pattern separation for medical diagnosis applied to breast cytology", Proceedings of the National Academy of Sciences, U.S.A., Volume 87, December 1990, pp 9193-9196.

O. L. Mangasarian, R. Setiono, and W.H. Wolberg: "Pattern recognition via linear programming: Theory and application to medical diagnosis", in: "Large-scale numerical optimization", Thomas F. Coleman and Yuying Li, editors, SIAM Publications, Philadelphia 1990, pp 22-30. 4. K. P. Bennett & O. L. Mangasarian: "Robust linear programming discrimination of two linearly inseparable sets", Optimization Methods and Software 1, 1992, 23-34 (Gordon & Breach Science Publishers).

Cáncer, I. N. (9 de mayo de 2024). cancer.gov. Obtenido de NIH: https://www.cancer.gov/espanol/cancer/naturaleza/estadisticas#:~:text=Los%20tres%20c%C3%A1nceres%20m%C3%A1s%20comunes,c%C3%A1ncer%20en%20mujeres%20en%202024.