**Logotipo

Descripción generada automáticamente**

**Cáncer de mama: Un abordaje diagnostico desde la Inteligencia Artificial**

Andrés Mauricio Coneo Pretelt

Monografía presentada para optar al título de Especialista en Analítica y Ciencia de Datos

Asesor  
David Manuel Villanueva Valdes, MSc.

Universidad de Antioquia  
Facultad de Ingeniería

Especialización en Analítica y Ciencia de Datos

Medellín, Antioquia, Colombia

2023

|  |  |
| --- | --- |
| **Cita** | (Coneo Pretelt, 2024) |
| **Referencia**  **Estilo APA 7 (2020)** | Coneo Pretelt, AM (2024). *Cáncer de mama: Un abordaje diagnostico desde la Inteligencia Artificial*. Universidad de Antioquia, Medellín, Colombia. |

**** 

Especialización en Analítica y Ciencia de Datos, CohorteVI.

Centro de Investigación Ambientales y de Ingeniería (CIA).

|  |  |
| --- | --- |
|  | Diagrama  Descripción generada automáticamente con confianza media |

Centro de Documentación Ingeniería (CENDOI)

**Repositorio Institucional:** http://bibliotecadigital.udea.edu.co

Universidad de Antioquia - www.udea.edu.co

Rector: John Jairo Arboleda Céspedes.

Decano: Julio Cesar Saldarriaga Molina

Jefe departamento: Diego José Luis Botia Valderrama

El contenido de esta obra corresponde al derecho de expresión de los autores y no compromete el pensamiento institucional de la Universidad de Antioquia ni desata su responsabilidad frente a terceros. Los autores asumen la responsabilidad por los derechos de autor y conexos.

**Agradecimientos**

Agradecimientos especiales a la Universidad de Antioquia, su cuerpo docente, especial agradecimiento a Lina Maria Sepúlveda Cano, PhD y a David Manuel Villanueva Valdes, MSc., al igual que todas las demás personas que contribuyeron, apoyaron y creyeron en este trabajo.

**Tabla de contenido**

[Resumen Propuesta 8](#_Toc151300991)

[Abstract 9](#_Toc151300992)

[1. Descripción del problema 10](#_Toc151300993)

[1.1. Problema de negocio 10](#_Toc151300994)

[1.2. Aproximación desde la analítica de datos 13](#_Toc151300995)

[1.3. Origen de los datos 14](#_Toc151300996)

[1.4. Métricas de desempeño 14](#_Toc151300997)

[2. Objetivos 16](#_Toc151300998)

[2.1. Objetivo general 16](#_Toc151300999)

[2.2. Objetivos específicos 16](#_Toc151301000)

[3. Datos 17](#_Toc151301001)

[3.1. Datos originales 17](#_Toc151301002)

[3.2. Datasets 19](#_Toc151301003)

[3.3. Analítica descriptiva 20](#_Toc151301004)

[Referencias 21](#_Toc151301005)

**Lista de tablas**

[**Tabla 1**  Resultados del test PBQ-SF (Personality Belief Questionnaire Short Form) 19](#_heading=h.2u6wntf)

[**Tabla 2** Características demográficas y tipo de tratamiento de hemodiálisis y diálisis peritoneal con la adherencia (SMAQ) 20](#_heading=h.19c6y18)

[**Tabla 3** Categorías de la investigación 21](#_heading=h.3tbugp1)

**Lista de figuras**

[Figura 1 – **A**) Principales tipos de cáncer por país, número estimado de nuevos casos (Mundial) en 2020, sexo femenino, todas las edades. **B**) Principales tipos de cáncer por país, número estimado de muertes (Mundial) en 2020, sexo femenino, todas las edades. Tomado de: Globocan 2020 - IARC (http://gco.iarc.fr/today), World Health Organization. 13](#_Toc151299106)

**Siglas, acrónimos y abreviaturas**

**IA** Inteligencia Artificial

**DL** Deep Learning/ Aprendizaje Profundo

**Globocan** Observatorio Global de Cáncer

**CaM** Cáncer de Mama

**OMS** Organización Mundial de la Salud

**ML** Machine Learning

**TCIA** The Cancer Imaging Archive

**PhD** Philosophiae Doctor

**MSc.** Magister of Science

**NCI**  Instituto Nacional de Cáncer – USA

**NHI**  Instituto Nacional de Salud – USA

**MR**  Resonancia Magnética

**MRI**  Información de Resonancia Magnética

**SEG**  Segmentación

**FGT**  Tejido fibroglandular

**UdeA** Universidad de Antioquia

# Resumen Propuesta

El cáncer de mama es una de las enfermedades más prevalentes y mortales en todo el mundo, afectando a millones de mujeres y, en menor medida, a hombres. A lo largo de los años, la medicina ha avanzado en la lucha contra esta enfermedad, con un enfoque cada vez más orientado hacia la detección temprana y el tratamiento personalizado.

En la última década, la inteligencia artificial (IA) y el aprendizaje profundo (DL) han transformado radicalmente la medicina, incluyendo la lucha contra el cáncer. Deep Learning o aprendizaje profundo, una rama de la IA, se destaca por su capacidad para extraer automáticamente características clave y procesar grandes conjuntos de datos complejos.

La IA, en particular el DL, ha demostrado su capacidad para abordar desafíos médicos complejos. El DL se caracteriza por su capacidad para procesar grandes cantidades de datos y extraer patrones sutiles, lo que lo hace especialmente adecuado para la detección y diagnóstico de enfermedades, incluido el cáncer de mama. Su aplicación va desde la interpretación de imágenes médicas, como mamografías y resonancias magnéticas, hasta la identificación de biomarcadores genéticos y moleculares. Basándonos en esta tecnología y en un conjunto de datos médicos, este proyecto busca enfocarse en aplicar la IA, especialmente el DL, para mejorar el diagnóstico y tratamiento del cáncer en Colombia, en particular el cáncer de mama.

Repositorio de alojamiento de los recursos: <https://github.com/aconeop/monografia_udea>

*Palabras clave*: Cáncer de mama, Inteligencia artificial, Aprendizaje profundo, Detección temprana, Biomarcadores

# Abstract

Breast cancer is one of the most prevalent and deadly diseases in the world, affecting millions of women and, to a lesser extent, men. Over the years, medicine has made advances in the fight against this disease, with an increasingly focused approach on early detection and personalized treatment.

In the last decade, artificial intelligence (AI) and deep learning (DL) have radically transformed medicine, including the fight against cancer. Deep learning, a branch of AI, is distinguished by its ability to automatically extract key features and process large complex datasets.

AI, particularly DL, has demonstrated its ability to tackle complex medical challenges. DL is characterized by its ability to process large amounts of data and extract subtle patterns, making it particularly well-suited for the detection and diagnosis of diseases, including breast cancer. Its application ranges from the interpretation of medical images, such as mammograms and MRIs, to the identification of genetic and molecular biomarkers. Based on this technology and a medical dataset, this project seeks to focus on applying AI, particularly DL, to improve cancer diagnosis and treatment in Colombia, particularly breast cancer

*Keywords***:** Breast cancer, Artificial intelligence, Deep learning, Early detection, Biomarkers

# Descripción del problema

## Problema de negocio

El cáncer es una de las principales causas de morbilidad y mortalidad en el mundo, según el reporte del Observatorio Global de Cáncer (Globocan). Para el año 2020, había 113,221 nuevos casos de cáncer en Colombia, el 27,66% más que con respecto al año 2018, y unas 54,987 muertes por algún tipo de cáncer, un 7,88% más que con respecto al año 2018 (International Agency for Research on Cancer, 2021).

El cáncer de mama (CaM) es un problema de salud global, principal entre las mujeres debido a su alta tasa de mortalidad y morbilidad (Kashyap et al., 2022). Más de 2.3 millones de casos nuevos y 685,000 muertes por CaM ocurrieron en 2020 (Sung et al., 2021). El CaM es el tipo de cáncer más frecuente y la causa más común de muerte por cáncer en mujeres a nivel mundial(Arnold et al., 2022) (ver Fig. 1A-B).

El CaM representa 1 de cada 4 casos de cáncer y 1 de cada 6 muertes por cáncer, ocupando el primer lugar en incidencia en 159 de 185 países, y en mortalidad en 110 países (Sung et al., 2021). Se espera que las cifras aumenten un 39% para el 2040 (Organización Panamericana de la Salud, 2023), especialmente en países en transición y en poblaciones con ascendencia afro e hispanas, en donde son más frecuentemente los diagnósticos de tumores (Arnold et al., 2022; A. N. Giaquinto et al., 2022; Miller et al., 2021).

Map

Description automatically generated

Figura 1 – **A**) Principales tipos de cáncer por país, número estimado de nuevos casos (Mundial) en 2020, sexo femenino, todas las edades. **B**) Principales tipos de cáncer por país, número estimado de muertes (Mundial) en 2020, sexo femenino, todas las edades. Tomado de: Globocan 2020 - IARC (<http://gco.iarc.fr/today>), World Health Organization.

En Colombia, se han identificado tres regiones con mayor riesgo de mortalidad por CaM, que corresponden a la región Andina y algunos focos de la región Caribe y la Orinoquía (Pardo C, 2017). Estos patrones se relacionan con porcentajes significativos de población afrodescendiente, como en la costa Caribe, Cali, e incluso San Andrés y Providencia. Existe evidencia de grandes variaciones en la incidencia, mortalidad y supervivencia, así como un mal pronóstico para las mujeres afrodescendientes o con ascendencia de esta raza (Baquet, Mishra, Commiskey, Ellison, & DeShields, 2008).

De hecho, se ha reportado que las mujeres de ascendencia afro tienen una mortalidad por CaM 40% más alta que las mujeres blancas, estas desigualdades podrían mitigarse ampliando el acceso a servicios de prevención, detección temprana y tratamiento de vanguardia y calidad para todas las mujeres (Angela N. Giaquinto et al., 2022).

Para 2021 en Colombia, el 57,5% de los nuevos casos de CaM estadificado se diagnosticaron en etapa tardía, lo que sigue representando un gran reto para mejorar estrategias de tamización y detección temprana, como también la posibilidad de predicción del riesgo.

Todo lo anterior representa una carga de salud pública, pero también un sobreesfuerzo a las finanzas públicas por los elevados costos para el sistema de salud (Gamboa et al., 2016; Howard & Bland, 2012). La Organización Mundial de la Salud (OMS) ha enfatizado la importancia de ciertos aspectos prioritarios para disminuir la mortalidad relacionada con el CaM, estos incluyen asegurar un diagnóstico temprano y brindar un manejo integral del cáncer. Por lo tanto, se considera crucial la implementación de servicios descentralizados y multidisciplinarios para mejorar la detección temprana y el tratamiento de la enfermedad (Organización Panamericana de la Salud, 2023).

El abordaje actual para la detección de CaM tiene en cuenta la edad, y ayudas diagnosticas como la mamografía debido a la consulta del paciente por alguna dolencia, pero no tiene en cuenta la variación individual en el riesgo (Pashayan, Morris, Gilbert, & Pharoah, 2018). Con un agravante sí consideramos que la tasa de supervivencia a cinco años en el CaM metastásico es inferior al 30% (Kashyap et al., 2022).

Actualmente, se tiene una comprensión clara de la base genética de ciertos tipos de cáncer, algunos de los cuales pueden tener un origen heredable, lo que da lugar a varios síndromes de susceptibilidad al cáncer debido a mutaciones germinales o somáticas que agravan y aceleran la enfermedad, y de acuerdo con el gen afectado, podrían predisponer a toda la familia a ciertos tipos de cáncer. Hay más de cien genes conocidos que, cuando presentan mutaciones, aumentan el riesgo de predisposición a la enfermedad.

Estas mutaciones pueden ocurrir en genes conocidos como oncogenes, genes supresores de tumores, y genes de estabilidad (Cybulski, Nazarali, & Narod, 2014; Stuart & Sellers, 2009). La mayoría de los casos pueden estar relacionados con genes de baja penetrancia. El análisis multifactorial para enfermedades de origen diverso como el cáncer, que difiere tanto entre poblaciones, razas, etnias, estilos de vida, ofrece oportunidades terapéuticas y diagnósticas, pero es complejo y requiere herramientas avanzadas debido a la gran cantidad de datos médicos generados en la investigación de nuestra era (Zavala, Serrano-Gomez, Dutil, & Fejerman, 2019), como también representan oportunidades para mejorar los exámenes diagnósticos mediante adopción de estrategias de detección computacionales usando modelos de predicción de riesgo y evolución de la enfermedad (Antoniou et al., 2019; Pashayan et al., 2018; Roberts, 2018).

Dentro de las preguntas de investigación se plantean: ¿Cómo se logra prevenir, predecir y tomar decisiones sobre pacientes con alguna importancia oncológica, antes que se desarrolle la neoplasia? ¿Cómo se lograría un mejor pronóstico en pacientes que ya poseen la neoplasia, para la prevención de su avance a estados avanzados?

## Aproximación desde la analítica de datos

En los últimos tiempos, la investigación en datos médicos o informática medica ha cobrado mayor relevancia. Sin embargo, el análisis de conjuntos de datos de nueva generación representa un desafío debido a la complejidad de la información que se genera. Se requieren herramientas sofisticadas y complejas para analizar esta compleja red de datos.

A todo lo anterior se suma el hecho de que la detección precoz y el acceso a tratamiento efectivo para enfermedades como el cáncer siguen siendo un reto para países en desarrollo, por lo que se requieren esfuerzos de las naciones y los entes de apoyo a la investigación científica (Arnold et al., 2022). Otro de los problemas se ve representado por la baja implementación de la medicina personalizada en países en desarrollo como Colombia, la pobre sistematización de los datos de salud, y su disponibilidad o acceso para instituciones, investigación, desarrollo, y análisis epidemiológico.

La OMS estableció 13 desafíos de la salud mundial para esta década, dentro de los cuales se destaca o se incentiva el uso positivo de las nuevas tecnologías, puesto que, en esta era digital, las tecnologías de salud digital pueden contribuir a la resolución de muchos problemas actuales y planear nuevas y mejores formas de abordaje médico (Organización Naciones Unidas, 2020).

Las necesidades no satisfechas en la práctica clínica, la complejidad de las enfermedades tanto comunes como raras, y las estrategias centradas en el paciente implementadas en la mayoría de las instituciones biomédicas, indican la necesidad inminente de gestionar los datos médicos de manera más eficiente.

En la actualidad, la Inteligencia Artificial y la Minería de Datos se han convertido en herramientas fundamentales para aplicar metodologías de investigación específicas en el ámbito de la salud. Sus usos permiten optimizar las políticas sanitarias existentes y promover la transición hacia una medicina de precisión, lo que resulta esencial para mejorar el bienestar de la población y abordar los desafíos que plantea la atención médica en la actualidad (Pezoulas et al., 2021).

Entender como adicional al trabajo de diagnóstico en laboratorio clínico-patológico, las herramientas basadas en algoritmos computacionales pueden ayudar a comprender la etiología de la enfermedad, puede tener un rol vital. Se ha probado en diferentes casos que el aprendizaje automático puede mejorar el diagnóstico del cáncer, y servir de apoyo al diagnóstico y pronostico (Basavanhally et al., 2010; Couture et al., 2018; Jaber et al., 2020; Osareh, Shadgar, & Markham, 2009). También se ha usado para la terapia personalizada, por lo que la inteligencia artificial tiene un “Ahora” muy valioso al ser capaz de proporcionar información amplia que puede ser usada para el tratamiento del CaM (Whitney et al., 2018).

Este estudio busca ofrecer un enfoque innovador para implementar alternativas de apoyo al diagnóstico, con el fin de reducir la tasa de mortalidad, especialmente entre las mujeres. La inteligencia artificial presenta ventajas significativas, como su flexibilidad, escalabilidad y capacidad para analizar diversos tipos de datos, en tareas de predicción, diagnóstico, pronóstico y tratamiento de enfermedades (Ngiam & Khor, 2019). Es fundamental cerrar la brecha existente en el conocimiento y la adaptación de estas herramientas y tecnologías avanzadas a nivel nacional, y ofrecer soluciones basadas en innovación tecnológica de manera local. Así, las técnicas de inteligencia artificial, como el aprendizaje automático (ML) y el aprendizaje profundo (DL), se convierten en herramientas valiosas, cada vez más necesarias para el sistema de salud moderno (Beam & Kohane, 2018).

## Origen de los datos

Con un enfoque en la optimización de los procesos de diagnóstico, este proyecto se basa en la recopilación y procesamiento de datos médicos de acceso público, alojadas en “The Cancer Imaging Archive (TCIA)” perteneciente al Instituto Nacional de Cáncer (NCI) y el Instituto Nacional de Salud de Estados Unidos (NHI), los datos incluyen imágenes de mamografías y resonancias magnéticas, junto con historiales clínicos de pacientes.

Los datos se obtienen de TCIA, este conjunto de datos consta de imágenes de Resonancia Magnética Dinámica con Contraste Preoperatorio de mama y datos clínicos de 922 pacientes con cáncer de mama invasivo confirmado por biopsia (Saha et al.). Los datos abarcan un período de una década y se componen de varios componentes como datos demográficos, clínicos, patológicos, y de tratamiento recopilados a partir de diversas fuentes, como notas clínicas, informes radiológicos e informes patológicos. Estos datos han sido utilizados en la investigación, incluyendo publicaciones relacionadas con radiogenómica y predicción de resultados por algunos otros investigadores.

## Métricas de desempeño

Para evaluar el rendimiento del trabajo propuesto, consideraré las siguientes métricas:

**Métricas de machine learning**

* **Precisión:** Porcentaje de muestras que se clasificaron correctamente. Esta métrica es importante para mí porque mide la capacidad del modelo para detectar el cáncer de mama con precisión. Espero que el modelo tenga una precisión del 90% o superior.
* **Recall:** Porcentaje de muestras positivas que se clasificaron correctamente. Esta métrica es también importante para mí porque mide la capacidad del modelo para detectar todos los casos de cáncer de mama. Espero que el modelo tenga un recall del 85% o superior.
* **AUC:** Área bajo la curva de características operativas del receptor (ROC). Esta métrica es una medida global de la capacidad del modelo para clasificar correctamente las muestras. Espero que en lo posible el modelo tenga un AUC de 0,90 o superior.

**Métricas de negocio**

* **Tiempo de desarrollo:** El tiempo que se tarda en desarrollar el modelo y la aplicación.

Espero que el modelo y la aplicación puedan desarrollarse en un plazo de 1 año.

* **Coste de desarrollo:** El coste económico del desarrollo del modelo y la aplicación.
* **Eficiencia:** La capacidad del modelo y la aplicación para procesar datos de forma rápida y eficiente. Espero que el modelo pueda procesar 1000 imágenes por hora.
* **Usabilidad:** La facilidad con la que los usuarios pueden utilizar el modelo y la aplicación.
* **Accesibilidad:** La capacidad del modelo y la aplicación para ser utilizados por personas con discapacidades.

Espero que el modelo y la aplicación cumplan con las directrices de accesibilidad de la W3C.

Estos valores se basan en la idea de que un modelo de aprendizaje automático para detectar cáncer de mama debe ser capaz de detectar el cáncer con precisión, ser eficiente y fácil de usar, y ser accesible para todos los usuarios.

# Objetivos

## Objetivo general

* Desarrollar una herramienta de IA basada en aprendizaje profundo que utiliza datos médicos para optimizar el diagnóstico del cáncer de mama.

## Objetivos específicos

* Recolectar un conjunto de datos de imágenes de mamografías con etiquetas de    diagnóstico.
* Desarrollar un modelo de aprendizaje profundo que pueda identificar patrones en las imágenes de mamografías.
* Evaluar, refinar y optimizar el rendimiento del modelo teniendo en cuenta precisión, sensibilidad, especificidad.

# Datos

## Datos originales

Los datos de Resonancia Magnética de Mama se obtendrán de TCIA, tienen un volumen aproximado de 400GBs, a continuación de describen más detalladamente los datos:

* **Visión General del Conjunto de Datos:**
  + **Modalidades:** Imágenes por Resonancia Magnética (MR), Segmentación (SEG).
  + **Número de Participantes:** 922.
  + **Número de Estudios:** 922.
  + **Número de Series:** 5,161.
  + **Número de Imágenes:** 773,888.
  + **Tamaño de Imágenes (GB):** 368.4.
* **Población:**
  + El conjunto de datos incluye 922 pacientes diagnosticadas con cáncer de mama invasivo.
  + Datos recopilados en el Hospital Duke desde el 1 de enero de 2000 hasta el 23 de marzo de 2014.
  + Los pacientes deben haber tenido una resonancia magnética preoperatoria en el Hospital Duke.
* **Procedimiento de Anotación Radiologica:**
  + La anotación se llevó a cabo en dos fases:
    - Fase 1: 271 pacientes anotadas por un panel de 6 radiólogos.
    - Fase 2: 651 pacientes anotadas por un panel de 4 radiólogos.
    - Las anotaciones fueron realizadas por 8 radiólogos utilizando una interfaz gráfica en MATLAB.
    - La anotación incluyó tres secuencias de resonancia magnética: pre-contraste, primer post-contraste y sustracción.
* **Datos Demográficos, Clínicos, Patológicos, Genómicos, Tratamiento, Resultados y Otros:**
  + Los datos clínicos detallados están disponibles [aquí](https://doi.org/10.7937/TCIA.e3sv-re93) bajo "Acceso a Datos" usando el elemento "Tipo de Datos" "Características Clínicas y Otros" en su versión original, al igual que anidados en la carpeta [data](https://github.com/aconeop/monografia_udea/blob/1bd3c63fb22608bc159b0008784794703a35b78a/data/Clinical_and_Other_Features.xlsx).
  + Los datos incluyen información demográfica, características del tumor, hallazgos de la resonancia magnética, detalles de la cirugía, terapia de radiación, respuesta del tumor, recurrencia y seguimiento.
* **Información Técnica de Resonancia Magnética (MRI):**
  + Los detalles técnicos están disponibles [aquí](https://doi.org/10.7937/TCIA.e3sv-re93) bajo "Acceso a Datos" usando el elemento "Tipo de Datos" "Características de Imágenes" en su versión original, al igual que anidados en la carpeta [data](https://github.com/aconeop/monografia_udea/blob/6f1fae8a166cb30f56725faea74908b5ad101bc7/data/Imaging_Features.xlsx).
  + La información incluye desde el diagnóstico hasta la fecha de la MRI, detalles del fabricante, opciones de escaneo, intensidad del campo y más.
* **Características de Imágenes:**
  + Las características se extrajeron previamente de tumores y tejido fibroglandular (FGT) utilizando secuencias T1.
  + Las características se categorizan en varios grupos, incluyendo volumen, tamaño, morfología, realce y heterogeneidad.
* **Características para Analizar:**
  + Se tienen una serie de características que se pueden evaluar:
    - Evaluación de la variabilidad de las características debido a cambios en el protocolo de resonancia magnética.
    - Evaluación de la estabilidad entre lectores con un subconjunto de 50 pacientes.
    - Validación para predecir subtipo tumoral, estado de receptor y estado de Ki-67.
    - Validación para predecir niveles de puntuación de recurrencia de Oncotype DX.
    - Validación de características asociadas con la supervivencia libre de recurrencia a distancia.

## Datasets

Para construir los conjuntos de datos de entrenamiento y validación para la detección de cáncer de mama con resonancia magnética, se seguirán estos pasos:

**Selección de las características:** En primer lugar, se seleccionarán las características que se utilizarán para entrenar el modelo. Estas características deben ser relevantes para la detección de cáncer de mama, y deben ser lo suficientemente robustas para ser extraídas de las imágenes de resonancia magnética.

**Segmentación de los tumores:** Seguidamente, se deben segmentar los tumores en las imágenes de resonancia magnética. En el conjunto de datos que se describe, se dispone de las segmentaciones realizadas por radiólogos, por lo que este paso esta realizado, sin embargo, se quiero con estos datos entrenar el modelo para que sea capaz de realizar y optimizar la segmentación.

**División de los datos en conjuntos de entrenamiento y validación:** Una vez que se han seleccionado las características y se han segmentado los tumores, se deben dividir los datos en dos conjuntos: un conjunto de entrenamiento y un conjunto de validación. El conjunto de entrenamiento se utilizará para entrenar el modelo, y el conjunto de validación se utilizará para evaluar el rendimiento del modelo.

**Preprocesamiento de los datos:** Antes de entrenar el modelo completo, es necesario preprocesar los datos. Esto puede incluir tareas como la normalización de los datos, la eliminación de valores atípicos, y la reducción de la dimensionalidad, entre otros que irán surgiendo en el proceso de procesado más allá de la muestra de datos con la que hasta ahora se trabajó.

**Entrenamiento del modelo:** Una vez que los datos están preparados, se entrenara el modelo. Para ello, se pueden utilizar diferentes algoritmos de aprendizaje automático, como las redes neuronales convolucionales.

**Evaluación del modelo:** Una vez entrenado el modelo, se debe evaluar su rendimiento. Esto se puede realizar utilizando el conjunto de validación. Para el conjunto de datos, se tienen una serie de características que se pueden evaluar, como el subtipo tumoral, el estado de receptor, el estado de Ki-67, los niveles de puntuación de recurrencia de Oncotype DX, y la supervivencia libre de recurrencia a distancia.

## Analítica descriptiva

Para visualizar de forma óptima y mejor comprensión de la aproximación analítica puede consultar el repositorio destinado para tal fin, este contiene código en Python para ejecutar en colab al igual que los datos y tablas utilizadas, esto con el fin de que pueda usarse y ejecutarse por un tercero si es preciso. Consulte: <https://github.com/aconeop/monografia_udea>

# Referencias

1. Antoniou, A., Anton-Culver, H., Borowsky, A., Broeders, M., Brooks, J., Chiarelli, A., . . . Ziv, E. (2019). A response to "Personalised medicine and population health: breast and ovarian cancer". *Hum Genet, 138*(3), 287-289. doi:<https://doi.org/10.1007/s00439-019-01984-z>
2. Arnold, M., Morgan, E., Rumgay, H., Mafra, A., Singh, D., Laversanne, M., . . . Soerjomataram, I. (2022). Current and future burden of breast cancer: Global statistics for 2020 and 2040. *Breast, 66*, 15-23. doi:<https://doi.org/10.1016/j.breast.2022.08.010>
3. Baquet, C. R., Mishra, S. I., Commiskey, P., Ellison, G. L., & DeShields, M. (2008). Breast cancer epidemiology in blacks and whites: disparities in incidence, mortality, survival rates and histology. *J Natl Med Assoc, 100*(5), 480-488. doi:<https://doi.org/10.1016/s0027-9684(15)31294-3>
4. Basavanhally, A. N., Ganesan, S., Agner, S., Monaco, J. P., Feldman, M. D., Tomaszewski, J. E., . . . Madabhushi, A. (2010). Computerized image-based detection and grading of lymphocytic infiltration in HER2+ breast cancer histopathology. *IEEE Trans Biomed Eng, 57*(3), 642-653. doi:<https://doi.org/10.1109/tbme.2009.2035305>
5. Beam, A. L., & Kohane, I. S. (2018). Big Data and Machine Learning in Health Care. *JAMA, 319*(13), 1317-1318. doi:<https://doi.org/10.1001/jama.2017.18391>
6. Couture, H. D., Williams, L. A., Geradts, J., Nyante, S. J., Butler, E. N., Marron, J. S., . . . Niethammer, M. (2018). Image analysis with deep learning to predict breast cancer grade, ER status, histologic subtype, and intrinsic subtype. *NPJ Breast Cancer, 4*, 30. doi:<https://doi.org/10.1038/s41523-018-0079-1>
7. Cybulski, C., Nazarali, S., & Narod, S. A. (2014). Multiple primary cancers as a guide to heritability. *Int J Cancer, 135*(8), 1756-1763. doi:<https://doi.org/10.1002/ijc.28988>
8. Gamboa, Ó., Buitrago, L. A., Lozano, T., Dieleman, S., Gamboa, C., Guzmán, É. L., . . . Fuentes, J. (2016). Costos directos de la atención del cáncer de mama en Colombia. *Revista Colombiana de Cancerología, 20*(2), 52-60. doi:<https://doi.org/10.1016/j.rccan.2016.02.003>
9. Giaquinto, A. N., Miller, K. D., Tossas, K. Y., Winn, R. A., Jemal, A., & Siegel, R. L. (2022). Cancer statistics for African American/Black People 2022. *CA Cancer J Clin, 72*(3), 202-229. doi:<https://doi.org/10.3322/caac.21718>
10. Giaquinto, A. N., Sung, H., Miller, K. D., Kramer, J. L., Newman, L. A., Minihan, A., . . . Siegel, R. L. (2022). Breast Cancer Statistics, 2022. *CA: A Cancer Journal for Clinicians, 72*(6), 524-541. doi:<https://doi.org/10.3322/caac.21754>
11. Howard, J. H., & Bland, K. I. (2012). Current management and treatment strategies for breast cancer. *Curr Opin Obstet Gynecol, 24*(1), 44-48. doi:<https://doi.org/10.1097/GCO.0b013e32834da4b1>
12. International Agency for Research on Cancer. (2021). Cancer Today: Population Fact Sheets. *GLOBOCAN 2020.* Retrieved from <http://gco.iarc.fr/today/fact-sheets-populations>
13. Jaber, M. I., Song, B., Taylor, C., Vaske, C. J., Benz, S. C., Rabizadeh, S., . . . Szeto, C. W. (2020). A deep learning image-based intrinsic molecular subtype classifier of breast tumors reveals tumor heterogeneity that may affect survival. *Breast Cancer Res, 22*(1), 12. doi:<https://doi.org/10.1186/s13058-020-1248-3>
14. Kashyap, D., Pal, D., Sharma, R., Garg, V. K., Goel, N., Koundal, D., . . . Belay, A. (2022). Global Increase in Breast Cancer Incidence: Risk Factors and Preventive Measures. *Biomed Res Int, 2022*, 9605439. doi:<https://doi.org/10.1155/2022/9605439>
15. Miller, K. D., Ortiz, A. P., Pinheiro, P. S., Bandi, P., Minihan, A., Fuchs, H. E., . . . Siegel, R. L. (2021). Cancer statistics for the US Hispanic/Latino population, 2021. *CA Cancer J Clin, 71*(6), 466-487. doi:<https://doi.org/10.3322/caac.21695>
16. Ngiam, K. Y., & Khor, I. W. (2019). Big data and machine learning algorithms for health-care delivery. *The Lancet Oncology, 20*(5), e262-e273. doi:<https://doi.org/10.1016/S1470-2045(19)30149-4>
17. Organización Naciones Unidas. (2020). Los 13 desafíos de la salud mundial en esta década. Retrieved from <https://news.un.org/es/story/2020/01/1467872>
18. Organización Panamericana de la Salud. (2023). Cáncer de mama Retrieved from <https://www.paho.org/es/temas/cancer-mama>
19. Osareh, A., Shadgar, B., & Markham, R. (2009). A computational-intelligence-based approach for detection of exudates in diabetic retinopathy images. *IEEE Trans Inf Technol Biomed, 13*(4), 535-545. doi:<https://doi.org/10.1109/titb.2008.2007493>
20. Pardo C, d. V. E., Buitrago L, Gamboa O.,. (2017). *Atlas de mortalidad por cáncer en Colombia*(4 ed., Vol. 1, pp. 124). Retrieved from <https://www.cancer.gov.co/conozca-sobre-cancer-1/publicaciones/atlas-mortalidad-cancer-colombia-2017>
21. Pashayan, N., Morris, S., Gilbert, F. J., & Pharoah, P. D. P. (2018). Cost-effectiveness and Benefit-to-Harm Ratio of Risk-Stratified Screening for Breast Cancer: A Life-Table Model. *JAMA Oncology, 4*(11), 1504-1510. doi:<https://doi.org/10.1001/jamaoncol.2018.1901>
22. Pezoulas, V. C., Hazapis, O., Lagopati, N., Exarchos, T. P., Goules, A. V., Tzioufas, A. G., . . . Gorgoulis, V. G. (2021). Machine Learning Approaches on High Throughput NGS Data to Unveil Mechanisms of Function in Biology and Disease. *Cancer Genomics Proteomics, 18*(5), 605-626. doi:<https://doi.org/10.21873/cgp.20284>
23. Roberts, M. C. (2018). Implementation Challenges for Risk-Stratified Screening in the Era of Precision Medicine. *JAMA Oncology, 4*(11), 1484-1485. doi:<https://doi.org/10.1001/jamaoncol.2018.1940>
24. Saha, A. A.-O., Harowicz, M. R., Grimm, L. J., Kim, C. E., Ghate, S. V., Walsh, R., & Mazurowski, M. A. A machine learning approach to radiogenomics of breast cancer: a study of 922 subjects and 529 DCE-MRI features. (1532-1827 (Electronic)). doi:<https://doi.org/10.1038/s41416-018-0185-8>
25. Stuart, D., & Sellers, W. R. (2009). Linking somatic genetic alterations in cancer to therapeutics. *Current Opinion in Cell Biology, 21*(2), 304-310. doi:<https://doi.org/10.1016/j.ceb.2009.02.001>
26. Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., & Bray, F. (2021). Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J Clin, 71*(3), 209-249. doi:<https://doi.org/10.3322/caac.21660>
27. Whitney, J., Corredor, G., Janowczyk, A., Ganesan, S., Doyle, S., Tomaszewski, J., . . . Madabhushi, A. (2018). Quantitative nuclear histomorphometry predicts oncotype DX risk categories for early stage ER+ breast cancer. *BMC Cancer, 18*(1), 610. doi:<https://doi.org/10.1186/s12885-018-4448-9>
28. Zavala, V. A., Serrano-Gomez, S. J., Dutil, J., & Fejerman, L. (2019). Genetic Epidemiology of Breast Cancer in Latin America. *Genes (Basel), 10*(2). doi:<https://doi.org/10.3390/genes10020153>