

# Metodología para la aplicación de la ciencia de datos en el diagnóstico del cáncer de mama

Presenta:

Esp. Jorge Armando Millán Gómez

Universidad Distrital "Francisco José de Caldas"  
Maestría en ciencias de la información y las comunicaciones

25 de mayo de 2023



# Elementos principales de la investigación

## Planteamiento del Problema

Según el informe de la organización mundial de la salud del año 2020 los casos detectados de cáncer de mama en Colombia fueron 15.509 de los cuales 4.411 casos terminaron en muerte ocupando el primer puesto de la tasa de letalidad sobre los demás tipos de cáncer[1]. Si no se tiene un diagnóstico a tiempo que detecte los aspectos más significativos que caracterizan el cáncer de mama es posible que la cifra de muertes en Colombia sea mayor en los años posteriores. En consecuencia, es necesario desarrollar una metodología que facilite la aplicación de la ciencia de datos en el diagnóstico de esta enfermedad.



## Formulación del Problema

- ¿Una metodología aplicada a técnicas en ciencias de datos para el diagnóstico de cáncer de mama mejora y facilita el análisis de patrones característicos en cada individuo para encontrar errores en el diagnóstico?



## Planteamiento de la Hipótesis

- Una metodología para comparar técnicas y grandes cantidades de datos que contienen información de resultados diagnósticos de pacientes particulares con los datos característicos de pacientes que padecen de cáncer de mama, permite hallar la similitud del comportamiento de los datos y predice de manera correcta el padecimiento de este tipo de cáncer de los pacientes particulares e identifica las variables que más influyen para contraer dicha enfermedad.



# Objetivos

## Objetivo General

Diseñar una metodología para diagnosticar el padecimiento del cáncer mama aplicando la ciencia de datos.

## Resultado

Creación de la metodología *DSM-BCD* (*Data Science Methodology for Breast Cancer Diagnosis*) con el proposito de generar valor a los datos oncológicos en el tiempo más corto posible para que los médicos diagnostiquen de manera ágil el cáncer de mama. Para lograrlo DSM-BCD integra la perspicacia médica y los resultados obtenidos por las técnicas de ML y DL en una retroalimentación continua generada en cada *Release* para producir mayor eficacia en la toma de decisiones.



# Objetivos

## Objetivos Específicos

Evaluar data-Sets con la información obtenida de técnicas médicas para la detección del cáncer de mama y realizar el Análisis exploratorio de Datos (EDA) de los mismos.

## Resultado

Implementación del Análisis Exploratorio de datos (EDA) con el data-set "*Breast Invasive Carcinoma*", el cual contiene un total de 817 muestras de tumores de mama y 110 variables genéticas características de marcadores tumorales obtenidas a partir de la Biopsia por aspiración con aguja fina (FNA) y Biopsia con aguja gruesa (CNB) del Carcinoma ductal invasivo (IDC), el Carcinoma lobulillar invasivo (ILC) y el Carcinoma de tumores mixtos (MDLC).



# Objetivos

## Objetivos Específicos

Validar la exactitud de la metodología con base en la aplicación de la ciencia de datos para el diagnóstico del cáncer de mama.

## Resultado

Comprobación de la metodología con base a la comparación del análisis descriptivo obtenido aplicando DSM-BCD y los resultados de la investigación realizada por el Ph.D Giovanni Ciriello, en donde se confirmo que el cáncer ILC presenta características genéticas molecularmente diferentes a los demás tipos de cáncer de mama, que la proteína HER2 positiva es un rasgo genético necesario para diagnosticar el cáncer IDC pero no suficiente para diagnosticar el cáncer ILC y adicional que es posible clasificar el cáncer MDLC en subgrupos de tipo LBC o IDC según sus propiedades genéticas.



# Desarrollo de la Investigación

**Bitacora de diagnóstico del cáncer de mama (BCDL)**

En esta fase, se propone el uso de la BCDL. El objetivo de la BCDL es almacenar las respuestas obtenidas para cada pregunta planteada en el BCQM y la relación de estas preguntas y respuestas con un modelo ML o DL determinado.

## Retroalimentación Médica

En esta fase, el experto en oncología médica determina si los resultados generados por el modelo ML o DL han conseguido responder las preguntas planteadas en el BCQM y si la nueva información obtenida es suficiente para diagnosticar el cáncer de mama.

### Evaluación e interpretación

En esta fase, el equipo de análisis de datos evalúa el modelo para conocer su calidad y asegurarse de que aborda de forma adecuada y completa las preguntas generadas en el BCQM. Es necesario que para realizar la evaluación se utilicen medidas especializadas basadas en el rendimiento, la sensibilidad y la especificidad del modelo.

## Modelado y Ejecución

En esta fase, el científico de datos diseña, crea o utiliza un modelo predictivo o descriptivo y lo alimenta con la versión del conjunto de datos o imágenes obtenidas. El científico debe seleccionar el tipo de aprendizaje y la técnica determinada en función de las preguntas planteadas en el RCOM.

### Mapa de preguntas sobre el cáncer de mama (BCOM)

En esta fase, se propone el uso del BCQM. El propósito del BCQM es que el Equipo de Análisis de Datos defina las preguntas que se responderán al final de cada Release y que permitirán tomar decisiones médicas sobre el diagnóstico de esta enfermedad.

### Planificación de actividades

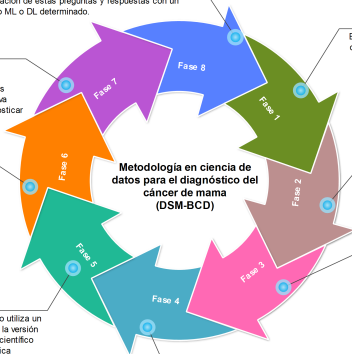
En esta fase el Equipo de Análisis de Datos, a partir de las preguntas formuladas en el BCQM, analiza todas las tareas a realizar, las estima en tiempo y las distribuye entre las personas que las llevarán a cabo durante el Release.

### Adquisición de datos oncológicos

En esta fase, basándose en las tareas realizadas en la planificación de la actividad, el médico experto en oncología junto con el ingeniero y el científico de datos identifican y recopilan los recursos de datos disponibles (estructurados, no estructurados y semiestructurados) y relevantes para resolver las preguntas planteadas en el BCOM.

### Análisis exploratorio de datos oncológicos

En esta fase, el científico de datos obtiene el conjunto de datos o imágenes previamente organizados por el ingeniero de datos y realiza un análisis exploratorio para descubrir patrones generales en la información generada. Entre las actividades se encuentran el procesamiento y transformación de datos oncológicos, en donde es necesario realizar la limpieza de datos, la combinación de datos procedentes de múltiples fuentes y la transformación de los datos en variables de valor. Al final, Las variables y patrones deben ser verificados por el médico experto en oncología.





# Desarrollo de la Investigación

Tipos de Cáncer de Mama							
Release n...n+1	IBC	MBC	ILC	MTBC	IDC	DCIS	
	Pregunta	Pregunta	Pregunta	Pregunta	Pregunta	Pregunta	Mammography
	Pregunta	Pregunta	Pregunta	Pregunta	Pregunta	Pregunta	Ductography
	Pregunta	Pregunta	Pregunta	Pregunta	Pregunta	Pregunta	Ultrasound
	Pregunta	Pregunta	Pregunta	Pregunta	Pregunta	Pregunta	MRI
	Pregunta	Pregunta	Pregunta	Pregunta	Pregunta	Pregunta	FNA
	Pregunta	Pregunta	Pregunta	Pregunta	Pregunta	Pregunta	CNB
							Técnicas de diagnóstico



# Desarrollo de la Investigación

## Fase 1: BCQM

Para este caso de estudio, se plantearon las siguientes preguntas basados en el atlas del Genoma del Cáncer con la finalidad de catalogar cambios moleculares de importancia biológica responsables de la aparición de cáncer haciendo uso de la secuenciación genómica y la bioinformática.

	ILC	IDC	MTBC	
Release 1	<i>¿Presenta el Carcinoma Lobulillar Invasivo(ILC) características genéticas molecularmente diferentes a los demás tipos de cáncer de mama?</i>	<i>¿Es la proteína HER2 positiva un rasgo genético necesario para diagnosticar el Carcinoma Ductal invasivo(IDC) pero no suficiente para diagnosticar el Carcinoma Lobulillar Invasivo(ILC)?</i>	<i>¿Es posible clasificar el Carcinoma de tumores mixtos (MDLC) en subgrupos de tipo Carcinoma Lobulillar Invasivo(LBC) o Carcinoma Ductal invasivo(IDC) según sus propiedades genéticas?</i>	FNA
				CNB



- Con respecto a los modelos de Machine Learning utilizados por diversos investigadores para predecir el Cáncer de mama se puede evidenciar que el algoritmo más utilizado es el Support Vector Machines (SVM), pero al realizar la comparación con los resultados obtenidos de la precisión de los métodos de Machine Learning observados anteriormente, el algoritmo Decision Trees es el que mejor resultado tiene, con una exactitud del 100 %, esto quiere decir que clasifico correctamente el total de las muestras. Por lo tanto según la investigación realizada se concluye que si se va a diagnosticar el Cáncer de mama los modelos más indicados para hacerlo son el Support Vector Machines(SVM) y Decision Trees.



- Según el análisis realizado con base en el diagrama de calor conformado por la correlación de las variables del Data-Set de la Universidad de Wisconsin se puede evidenciar que las variables *conca-ve\_points\_worst* y *area\_worst* generan información relevante en la realización del diagnóstico de Cáncer de mama debido a que expresan una deformidad mayor de los núcleos celulares encontrados en las masas mamarias extraídas por el método de Aspiración con Aguja Fina(FNA).





- Implementación de una capa de servicios REST basada modelos de Machine Learning para el diagnóstico de Cáncer de mama que podría ser utilizada en diferentes ámbitos en la detección y el diagnóstico de dicho Cáncer.
- Diseño y Arquitectura de un aplicativo web enfocado en el uso de modelos de Machine Learning aplicados en la rama de la Medicina especializada en Oncología.



- Creación de una aplicación web llamada OncoAnalysisApp la cual permita el diagnóstico de cualquier tipo de Cáncer teniendo como entrada Data-Sets obtenidos por diversos métodos médicos.
- Creación de una aplicación que permita el análisis de imágenes y que diagnostique el padecimiento de Cáncer de mama con base a los modelos de Deep-Learning existentes.
- Creación de una aplicación que permita crear nuevos Data-Set dinámicamente según parámetros proporcionados por el usuario.



 International Agency for Research on Cancer.  
170 Colombia fact sheets.  
*Globocan 2020*, 509:1–2, 2020.

 National Cancer Institute and National Human Genome Research  
Institute.  
The cancer genome atlas, 2023.

