

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

Presenta:

Esp. Jorge Armando Millán Gómez

Directora de la tesis:

Dra. Lilia Edith Aparicio Pico

Maestría en Ciencias de la Información y las Comunicaciones
Universidad Distrital "Francisco José de Caldas"

27 de mayo de 2023



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. 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.

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)* diseñada con el propósito 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 Específico 1

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

Ejecución 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, basados en los tipos de carcinoma ductal invasivo (IDC) y lobulillar invasivo (ILC), recopiladas del sitio publico cBioPortal para la genómica del cáncer.

Objetivos Específico 2

Proponer una metodología para el diagnóstico del cáncer de mama a partir de técnicas de Machine Learning (ML), Deep Learning (DL) e Inteligencia artificial (IA).

Resultado

Implementación de la metodología DSM-BCD con la cual fue posible extraer información significativa de muestras de tumores cancerígenos mamarios presentados en 817 pacientes recopilados por medio de las intervenciones quirúrgicas de *aspiración con aguja fina (FNA)* y *biopsia con aguja gruesa (CNB)*, a través del aprendizaje automático no supervisado basado en la técnica de agrupación y el modelo *BIRCH* en donde se logro determinar que el carcinoma lobulillar invasivo(ILC) es una enfermedad con rasgos genéticos característicos diferentes a los demás tipos de cáncer.

Objetivos Específico 3

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.

Data Science Methodology for Breast Cancer Diagnosis (DSM-BCD)



Desarrollo de la Investigación

Breast Cancer Question Map (BCQM)

The diagram illustrates the Breast Cancer Question Map (BCQM) as a grid of research questions. The columns represent different types of breast cancer, and the rows represent different diagnostic techniques. Each cell in the grid contains the word 'Pregunta' (Question).

Tipos de Cáncer de Mama (Types of Breast Cancer):

- IBC
- MBC
- ILC
- MTBC
- IDC
- DCIS

Técnicas de diagnóstico (Diagnostic Techniques):

- Mammography
- Ductography
- Ultrasound
- MRI
- FNA
- CNB

Release n...n+1

	IBC	MBC	ILC	MTBC	IDC	DCIS	
Release n...n+1	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

Desarrollo de la Investigación

Fase 1: BCQM

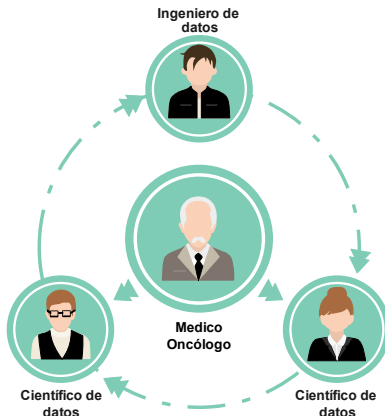
Para este caso de estudio, se plantearon las siguientes preguntas basadas en el atlas del genoma del cáncer con la finalidad de catalogar cambios moleculares de importancia biológica responsables de la aparición de esta enfermedad haciendo uso de la secuenciación genómica y la bioinformática.

	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

Desarrollo de la Investigación






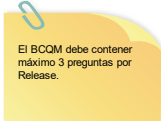
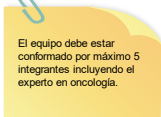
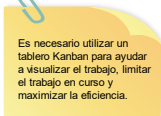
Fase 2: Planeación de actividades

En esta fase se propuso el concepto de *Data Analysis Team* y adicionalmente basados en las 3 preguntas planteadas en el BCQM, se realizó la planeación de actividades para proyectos basados en datos.



Desarrollo de la Investigación

Conformación del Data Analysis Team

División de tareas y responsabilidades	Consideraciones metodológicas	Planeación de actividades
<div>Médico</div> <div>Oncólogo: Crear BCQM con base a su experticia en el Cáncer de mama.</div> <div>Desarrolladores</div> <div>Ing de datos: Administrar, procesar y almacenar los datos para que puedan ser usados de forma accesible.</div> <div>Científico de datos: Analizar y modelar los datos para generar respuestas a las preguntas planteadas en el BCQM.</div>	<div>Comunicación: Durante el Release siempre debe existir la comunicación continua entre los desarrolladores y el experto en Oncología.</div> <div>Objetivo: Todas las acciones, propuestas y actividades deben estar orientadas a dar respuestas a las preguntas planteadas en el BCQM.</div> <div>Duración: La planeación de actividades debe durar máximo 8 horas.</div>	<div><p>El BCQM debe contener máximo 3 preguntas por Release.</p></div> <div><p>El equipo debe estar conformado por máximo 5 integrantes incluyendo el experto en oncología.</p></div> <div><p>Es necesario utilizar un tablero Kanban para ayudar a visualizar el trabajo, limitar el trabajo en curso y maximizar la eficiencia.</p></div>

Desarrollo de la Investigación

Planeación de actividades aplicada al caso de estudio



Desarrollo de la Investigación

Fase 3: Adquisición de datos oncológicos

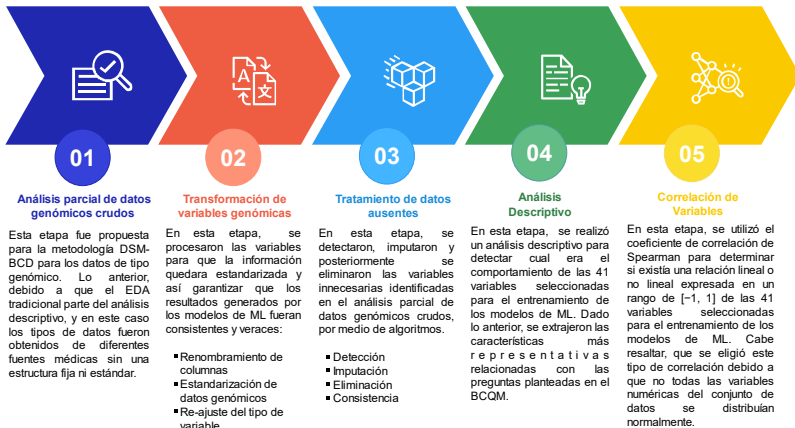
En esta fase, se utilizaron variables genéticas características de marcadores tumorales basados en los tipos de cáncer de mama carcinoma ductal invasivo (IDC) y carcinoma lobulillar invasivo (ILC). Estas variables fueron obtenidas del conjunto de datos denominado *“Breast Invasive Carcinoma”*.

N	Variable	Estadística
1	Número de variables	110
2	Variables Categóricas	95
3	Variables Numéricas	15
4	Número de filas	818
5	Celdas faltantes	37657
6	Celdas faltantes (%)	41,9 %
7	Filas duplicadas	0
8	Filas duplicadas (%)	0,0 %
9	Tamaño total en memoria	3,8mb
10	Tamaño promedio de fila en la memoria	4,8KB

Desarrollo de la Investigación

Fase 4: Análisis Exploratorio de datos oncológico

En esta fase, se realizó en análisis exploratorio de datos con los registros genéticos obtenidos del conjunto de datos "*Breast Invasive Carcinoma*".

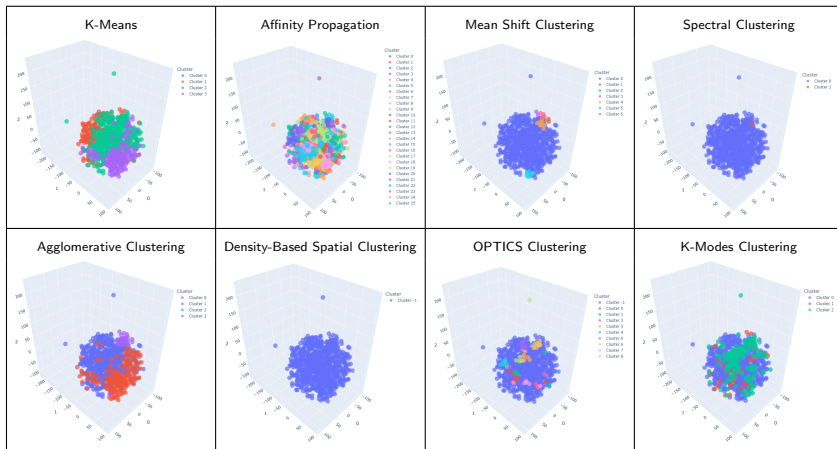


Fase 5: Modelado y Ejecución

En esta fase, se seleccionó el método de agrupamiento (*Clustering*) y se implementaron 9 modelos utilizando el 95 % de los datos para el entrenamiento y el 5 % de datos restantes para comprobar la precisión del agrupamiento y así analizar los clusters generados.

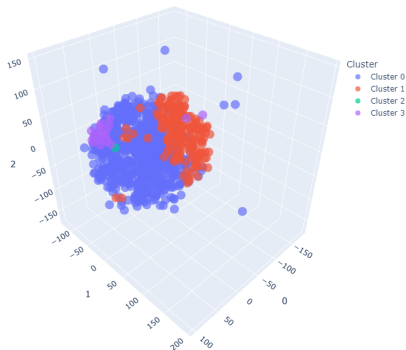
Id	Modelo Clustering	Silhouette	Davies-Bouldin	Clusters
1	K-Means Clustering	0,0826	2,5179	4
2	Affinity Propagation	0,0518	2,2123	68
3	Mean Shift Clustering	0,2790	1,2792	7
4	Spectral Clustering	0,7986	0,1419	2
5	Agglomerative Clustering	0,1034	2,0468	4
6	DB Spatial Clustering	0,0000	0,0000	-1
7	OPTICS Clustering	-0,2044	1,9565	10
8	BIRCH Clustering	0,1286	1,8703	4
9	K-Modes Clustering	0,0547	3,8189	3

Modelos Clustering aplicados al data-set “Breast Invasive Carcinoma”

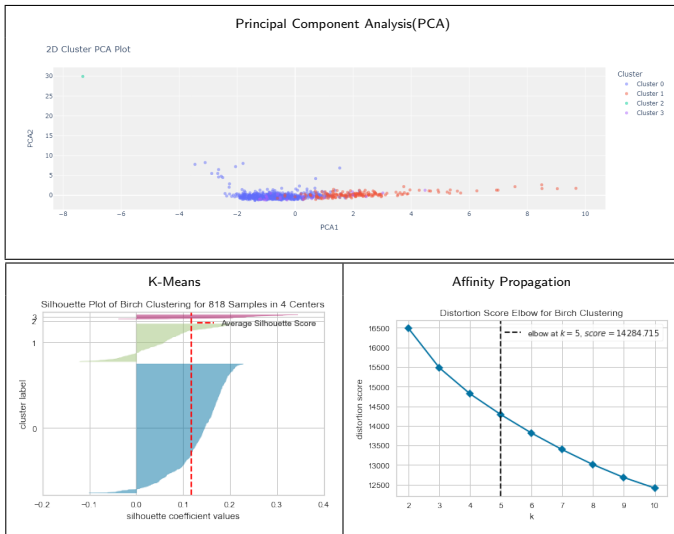


Fase 6: Evaluación e Interpretación

El modelo de ML seleccionado para analizar el comportamiento de conjunto de datos del carcinoma invasivo fue el de agrupación *BIRCH* (*Balanced, Iterative Reducing, and Clustering using Hierarchies*), debido a que los clusters generados presentaban una métrica de cohesión y separación idónea con respecto a los demás modelos.



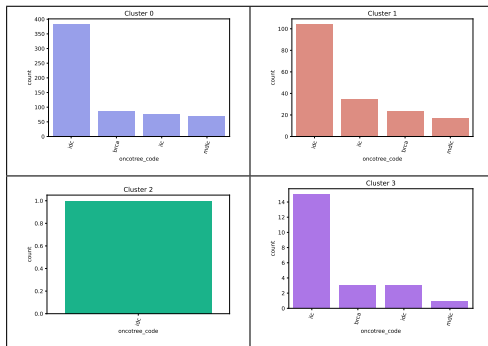
Métricas de validación internas del modelo BIRCH



Desarrollo de la Investigación

Fase 7: Retroalimentación médica

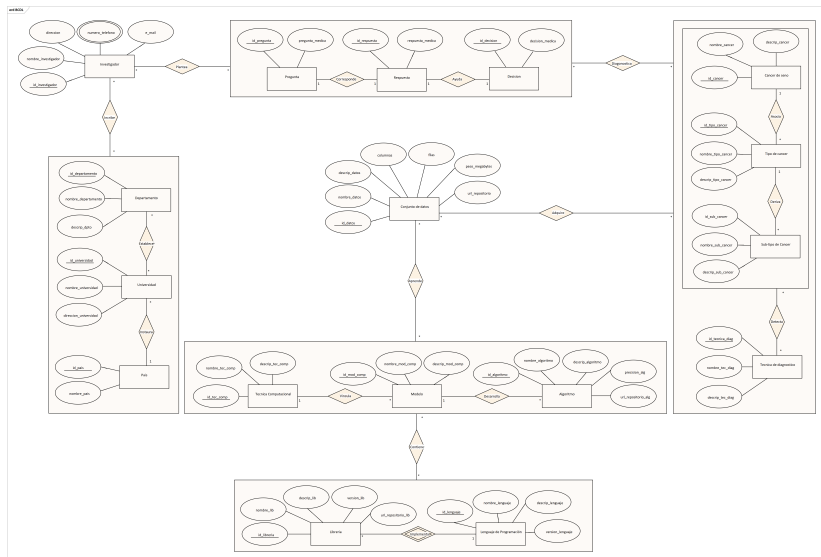
En esta fase, los resultados obtenidos por el modelo *BIRCH* fueron validados con la investigación “*Comprehensive Molecular Portraits of Invasive Lobular Breast Cancer*”, publicada por el *Ph.D Giovanni Ciriello*, en la cual se realizó un análisis exhaustivo de muestras de tumores y se determinó que el ILC es una enfermedad molecularmente distinta con rasgos genéticos característicos.



Fase 8: Bitácora para el diagnóstico del cáncer de mama (BCDL)

En esta fase, se propuso el uso de una bitácora para el diagnóstico del cáncer de mama (BCDL, por sus siglas en inglés, "*Breast Cancer Diagnostic Logbook*") basado en el desarrollo de un modelo entidad relación (MER) para facilitar el diseño de bases de datos fundamentado en la especificación de un esquema para el diagnóstico del cáncer de mama para representar una estructura lógica global que permita ver la relación entre el investigador, el tipo de cáncer de mama, la técnica de diagnóstico, la técnica computacional, el lenguaje de programación, la pregunta, la respuesta y la decisión médica.

Desarrollo de la Investigación



- Con base a las preguntas de investigación planteadas y el conjunto de datos genómicos recopilados a través de biopsias realizadas a 817 pacientes que fueron diagnosticados con cáncer de mama, se realizó la evaluación de 9 algoritmos de agrupamiento (*Clustering*). Luego, se utilizaron las métricas de validación interna basadas en el índice de *Davies-Bouldin(DB)* y el *Coeficiente de silhouette* para determinar la congruencia de los clusters entrenados. Dado lo anterior, el modelo *BIRCH* produjo un número adecuado de clusters con una estructura compacta y centros considerablemente separados los unos de los otros. De modo que la precisión del modelo *BIRCH* fue superior a la de los demás modelos de ML implementados. Por consiguiente, es plausible afirmar que el modelo *BIRCH* es el más adecuado para agrupar datos de origen genómico obtenidos por biopsias realizadas por medio de las técnicas FNA y CNB.

- Para concluir, es plausible afirmar que gracias a la aplicación de la metodología *DSM-BCD*, fue posible extraer información significativa de muestras de tumores cancerígenos mamarios recopilados por medio de las intervenciones quirúrgicas de FNA y CNB, a través del aprendizaje automático no supervisado basado en la técnica de agrupación y el modelo *BIRCH*, lo que permitió responder las preguntas planteadas en el *BCQM*, proporcionando información suficiente para diagnosticar el cáncer de mama y la identificación de rasgos genómicos característicos del IDC, ILC y MDLC, generando un valor agregado al dominio médico al confirmar que el cáncer ILC presenta características genéticas molecularmente diferentes a los demás tipos de cáncer, 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.

- Una vez la investigación fue culminada, es decir, los objetivos y resultados con base al alcance fueron solucionados satisfactoriamente, junto con la Dra.Lilia Edith Aparicio Pico se expuso el producto de la investigación en el **Segundo Congreso Interdisciplinario de Mecánica, Informática y Electricidad (IC-MECE 2022)**, realizado en *Barcelona-España*. Posteriormente, se publicó un artículo científico denominado *"Methodology for the application of data science in breast cancer diagnosis"* en la revista Turca *"Computers and Informatics"*. La ponencia realizada fue dirigida a la comunidad científica especializada en todos los aspectos de la informática, sistemas de información, aplicaciones y políticas de TI.