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

1st Jorge Millán *Facultad de ingeniería*   
*Universidad Distrital*  
Bogotá, Colombia   
jamillang@correo.udistrital.edu.co   
 2nd Edith Aparicio *Facultad de ingeniería*   
*Universidad Distrital*  
Bogotá, Colombia   
eaparicio@udistrital.edu.co

Thu Jul 7 22:16:17 2022

En el año 2020 los casos detectados de cáncer de mama en Colombia fueron 15.509 de los cuales 4.411 terminaron en muerte. El pronóstico anticipado de esta enfermedad se ha convertido en una necesidad de investigación debido a que puede facilitar el tratamiento preventivo para evitar su letalidad en un estado avanzado. En este articulo se propone la metodología DSM-BCD (Data science methodology for breast cancer diagnosis) diseñada para agilizar el diagnostico del cáncer de mama a través de la mejora continua de las técnicas de Machine Learning y Deep Learning teniendo como base la perspicacia del especialista en oncología y la retroalimentación de conocimiento según el comportamiento de los datos en las diversas técnicas para la detección del cáncer de mama.

Metodología, Ciencia de datos, Cáncer de mama, Diagnostico, Machine Learning, Deep Learning

## 1 Introducción

En el año 2020 el numero de casos de cáncer de mama a nivel mundial ascendió a 2.261.419 en donde 684.996 de estos casos terminaron en muerte. A si mismo, en Colombia se presentaron 15.509 casos y 4.411 decesos[?], ocupando el primer puesto en la tasa de letalidad sobre los demás tipos de cáncer afectando a mujeres de todas las edades. Este tipo de cáncer se origina cuando las células mamarias comienzan a crecer sin control convirtiéndose en células malignas que normalmente forman un tumor que a menudo se puede observar en una radiografía o se puede palpar como una masa o bulto [?].

Colombia presenta limitaciones con respecto al acceso de la detección y el diagnóstico temprano del cáncer, provocado en la mayoría de los casos por factores como el estrato socio-económico, la cobertura del seguro de salud, el origen y la accesibilidad. En promedio, el tiempo de espera de un paciente es de 90 días desde la aparición de los síntomas hasta el diagnóstico de dicho cáncer. La primera acción para reducir la tasa de mortalidad por cáncer de mama debe estar enfocada en la agilidad del diagnóstico y el acceso oportuno a la atención [?].

El pronóstico anticipado de esta enfermedad se ha convertido en una necesidad de investigación debido a que puede facilitar el tratamiento preventivo para evitar su letalidad en un estado avanzado. Una alternativa para disminuir esta tasa de mortalidad es poder predecir e identificar, con base al análisis de un conjunto de datos obtenidos de exámenes realizados por diversos métodos médicos al individuo, que probabilidad tiene de contraer el cáncer de mama y cuales son las variables que más influyen en el padecimiento de esta enfermedad, y según estos resultados brindar un tratamiento preventivo que permita combatir el cáncer antes de que el mismo haga metástasis o que llegue a un estado avanzado en donde sea más difícil de tratar.

El análisis obtenido por medio de la ciencia de datos permite detectar el cáncer en un menor tiempo, debido a que los algoritmos de ML (Machine Learning) y DL (Deep Learning) impactan claramente en los estudios exploratorios que tienen como objetivo identificar los principios biológicos de la enfermedad lo que puede beneficiar a pacientes y médicos al acelerar el diagnóstico y brindar apoyo para tomar mejores y más rápidas decisiones a nivel clínico [?]. En los últimos años, el aumento de la potencia de las computadoras, junto con los avances matemáticos, ha permitido el uso de las redes neuronales complejas de múltiples capas (profundas) las cuales han mejorado el rendimiento de la interpretación automática de imágenes oncológicas altamente estandarizadas[?].

Por otra parte, la ciencia, en el lenguaje del método científico, es formular hipótesis o conjeturas, basadas en observaciones del mundo que nos rodea para validar o invalidar esas hipótesis mediante la realización de experimentos. Sin embargo, a diferencia de las ciencias puras, trabajar con datos no requiere necesariamente realizar experimentos. Más bien, muchas veces los datos ya han sido recopilados y organizados previamente. Entonces, el método científico, aplicado a los datos, se puede resumir como: *“Formular hipótesis basadas en el mundo que nos rodea y luego analizar los datos relevantes para validar o invalidar dichas hipótesis”*.

En la actualidad la ciencia de datos es utilizada por diferentes investigadores para modelar la progresión y el tratamiento de afecciones cancerosas debido a su capacidad para detectar características significativas en conjuntos de datos complejos. La medicina basada en datos tiene la capacidad no solo de mejorar la velocidad y precisión del diagnóstico de enfermedades genéticas, sino también de desbloquear la posibilidad de tratamientos médicos personalizados[?]. Aprender significa: *“Adquirir conocimientos o habilidades en algo a través de la experiencia”*. Por lo tanto, se podría enmarcar al ML cómo la manera en la cual una máquina gana o adquiere conocimiento a través de la experiencia. Pero ¿Cómo adquiere experiencia una máquina? Todas las entradas de una máquina son esencialmente cadenas binarias de 0 y 1, que en el dominio de las ciencias de la computación dichos binarios son simple y llanamente datos. Por consiguiente, el ML es realmente la forma en que una computadora adquiere conocimiento a través de los datos.

La ciencia de datos es fundamentalmente un proceso, mientras que el ML es una herramienta que puede ser inmensamente útil para llevar a cabo dicho proceso [?]. Por supuesto, esto no da ninguna idea del *como* en absoluto; simplemente se resume el proceso como algo que se hace con los datos de entrada para generar este conocimiento como salida. Para hacer una analogía matemática, el ML es una función tal que:

(1)

Hay que mencionar ademas, que una metodología es una estrategia general que sirve de guía para los procesos y actividades que están dentro de un dominio determinado. La metodología no depende de tecnologías ni herramientas específicas, ni es un conjunto de técnicas o recetas. Más bien, la metodología proporciona al científico de datos un marco sobre cómo proceder con los métodos, procesos y argumentos que se utilizarán para obtener respuestas o resultados[?] .

Así, el objetivo de este articulo es proponer una metodología en ciencia de datos para el diagnostico del cáncer de mama, esto con el propósito de aplicar el uso de esta ciencia computacional como un campo multidisciplinario que permita dar valor a los datos obtenidos por medio técnicas medicas para el diagnostico del cáncer de mama a través de un proceso que facilite el análisis y la selección de técnicas para poder tomar una decisión que valide la veracidad de una hipótesis planteada. El uso de la metodología propuesta repercute en la facilidad de la obtención de los datos y el procesamiento de la información para el diagnostico y pronostico del padecimiento de esta enfermedad.

## 2 Detección y diagnóstico del cáncer de mama

El cáncer de mama, con su causa incierta, ha capturado la atención de los cirujanos en todas las épocas. A pesar de siglos de laberintos teóricos y preguntas científicas, el cáncer de mama aún es una de las enfermedades humanas más temibles [?]. Según [?], el cáncer de mama se origina a través de tumores malignos, cuando el crecimiento de la célula se descontrola provocando que muchos tejidos grasos y fibrosos de la mama inicien un crecimiento anormal, lo cual tiene como consecuencia que las células cancerosas se diseminen por los tumores causando las diferentes etapas del cáncer. [?] exponen que existen diferentes tipos de cáncer de mama [?], que se producen cuando las células afectadas y tejidos diseminados se esparcen por todo el cuerpo. Pongamos por caso, el primer tipo de cáncer denominado *Carcinoma ductal in situ (DCIS)* el cual es un tipo de cáncer no invasivo [?], que se produce cuando las células anormales se propagan fuera de la mama. El segundo tipo de cáncer es el *Carcinoma ductal invasivo (IDC)*, también conocido como carcinoma ductal infiltrante [?]. Este tipo de cáncer ocurre cuando las células anormales de la mama se diseminan por todos los tejidos mamarios. Generalmente este tipo de cáncer se encuentra en los hombres [?]. El tercer tipo de cáncer es el cáncer de *Tumores mixtos (MTBC)* también conocido como cáncer de mama invasivo [?] causado por las células anormales de los conductos y las células lobulillares [?]. El cuarto tipo de cáncer es el cáncer de mama *Lobulillar (LBC)* [?] que ocurre dentro del lóbulo y aumenta las posibilidades de otros cánceres invasivos. El quinto tipo cáncer es el cáncer de mama *mucinoso (MBC)* o de *mama coloide* [?] que ocurre debido a las células ductales invasivas cuando los tejidos anormales se extienden alrededor del conducto [?]. El sexto y ultimo tipo de cáncer es el cáncer de mama *inflamatorio (IBC)*, el cual causa hinchazón y enrojecimiento del pecho. Este tipo de cáncer de mama es de rápido crecimiento, y comienza a aparecer cuando los vasos linfáticos se obstruyen en células rotas [?].

Según [?], en la mayoría de casos detectados la mujer descubre una tumoración en su mama. Otros signos y síntomas que se presentan menos a menudo comprenden: crecimiento o asimetría de la mama, alteraciones y retracción del pezón o telorrea, ulceración o eritema de la piel de la mama, una masa axilar y molestia musculoesquelética. Case señalar, que si se detectan alguno de los síntomas anteriores este tipo de cáncer puede ser detectado por medio de los procedimientos basados en *Exploración Física*, *Técnicas de imagen* y *Biopsias*.

A nivel de *Exploración física*, el cáncer de mama puede ser detectado por el oncólogo por medio de los métodos de *Inspección* y *Palpación*. Con estos métodos, se registran la simetría, el tamaño y la forma de la mama, así como cualquier evidencia de edema (piel de naranja), retracción del pezón o de la piel y eritema [?].

En la actualidad, muchas *Técnicas de imagen* se utilizan ampliamente para proporcionar un diagnóstico preciso de las lesiones mamarias[?], entre estas técnicas las mas relevantes son las siguientes: *La Mamografía* que hace uso de una unidad mamográfica que consta de un tubo de rayos X que encapsula un cátodo y un ánodo. La mama se coloca sobre el detector y se comprime mediante un dispositivo de placas paralelas, el cual mantiene la mama inmóvil y evita el desenfoque por movimiento, esto con el propósito de reducir el grosor del tejido que deben atravesar los rayos x [?]; *La ductografía* que identifica de lesiones en pacientes con secreción del pezón. Este método es eficaz para localizar e identificar las lesiones intraductales por medio de un examen mamográfico realizado tras el llenado retrógrado de los conductos galactóforos con material de contraste [?]; *La Ecografía* que permite obtener imágenes de alta resolución por medio de un pequeño transductor (sonda) de alta frecuencia que envía ondas sonoras inaudibles al interior de la mama y recibe el eco de las ondas procedentes de los órganos internos, los fluidos y los tejidos [?]; y *La Resonancia Magnética (MRI)*que es utilizada cuando las lesiones en la mama no se pueden evaluar fácilmente mediante otras técnicas. Para lograrlo, utiliza bobinas receptoras de radiofrecuencia (RF) para detectar una señal emitida por los tejidos tras la excitación de un campo electromagnético que obliga a los protones alinearse a la anatomía de la zona de interés en tamaño y forma [?].

Hay que mencionar, además que actualmente existen dos modalidades para obtener un diagnóstico por *Biopsia* para un paciente que presenta una anomalía mamaria. Por un lado tenemos, la biopsia para mamas con *lesiones palbables* tambien denominada *percutánea* o *mínimamente invasiva*. Estas biopsias incluyen la aspiración con aguja fina (FNA) y con con aguja gruesa (CNB). Las biopsias quirúrgicas abiertas se denominan a veces biopsias por escisión o biopsias por incisión. La biopsia por *escisión* indica la extirpación completa de la lesión, mientras que la biopsia por *incisión* indica la extirpación de parte de la lesión [?]. En el caso de las *lesiones no palpables*, las modalidades de imagen como la ecografía (US), la mamografía y la resonancia magnética (MRI) son complementos útiles para identificar y localizar la lesión de interés. La decisión de cuándo realizar una biopsia de mama depende de los antecedentes del paciente, los hallazgos de la exploración física y las imágenes radiológicas. El objetivo principal de la biopsia es obtener un diagnóstico tisular que pueda ayudar a dictar el tratamiento y la planificación preoperatoria, si está indicado. Por lo tanto, es imprescindible elegir una técnica de biopsia que optimice las posibilidades de obtener un diagnóstico preciso y que, al mismo tiempo, minimice los costes, limite las molestias del paciente y reduzca la necesidad de repetir el procedimiento [?] .

## 3 Data science methodology for breast cancer diagnosis (*DSM-BCD*)

Se realizó una revisión sistemática en la literatura científica de diversas metodologías, que aunque abarcan el proceso necesario para el análisis de datos, no generan el valor suficiente para la solución de un problema determinado. Por esta razón, autores como [?] consideran que aunque la comunidad de investigación en ciencia de datos está creciendo día a día, está explorando nuevos dominios, creando nuevos roles especializados y adicional está realizando un gran esfuerzo de investigación para desarrollar análisis avanzados, mejorar modelos y generar nuevos algoritmos apoyados de los campos de las matemáticas, la estadística y la informática, estas habilidades no son suficientes para su aplicación en proyectos reales, puesto que la mayoría de proyectos basados en datos presentan problemas organizativos y socio-técnicos, tales como: una falta de visión y claridad en los objetivos, un énfasis sesgado en cuestiones técnicas y ambigüedad de roles. Y aunque estos problemas existen en proyectos de ciencia de datos del mundo real, la comunidad no se ha preocupado demasiado por ellos y no se ha escrito lo suficiente sobre las soluciones para abordar estos problemas.

Dado los problemas presentados de los proyectos basados en datos, y el estudio de diversas metodologías propuestas por varios autores, se propone la metodología *DSM-BCD (Data science methodology for breast cancer diagnosis)*. Esta metodología tiene como base el *manifestó ágil* aplicado a un contexto de resultados basados en datos. Dado lo anterior *DSM-BCD* no se enfoca en evaluar la precisión de las técnicas de ML y DL sino su objetivo principal es generar valor a los datos en el tiempo mas corto posible para que los médicos diagnostiquen de manera ágil el cáncer de mama. Para lograrlo *DSM-BCD* integra la perspicacia medica 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. En *DSM-BCD*, se propone la conformación de un *Data Analysis Team*. Este equipo debe estar encabezado por el medico experto en oncología, al menos un ingeniero de datos y un científico de datos. Se recomienda que el equipo este conformado por un máximo de 5 personas para facilitar el trabajo en equipo y la comunicación interna. Dado lo anterior, la metodología se compone de las siguiente fases:

### Fase 1: Mapa de preguntas sobre el cáncer de mama (BCQM)

En esta fase se propone el uso de un mapa de preguntas sobre el cáncer de mama (BCQM). El proposito del *BCQM* es que el *Data Analysis Team* defina las preguntas que serán resueltas al finalizar cada *Release* y que permitirán tomar decisiones medicas con respecto al diagnostico de esta enfermedad. En la figura 1 se observa la estructura del BCQM.

Figure 1: Mapa de preguntas basados en los tipos de cáncer de mama *inflamatorio (IBC)*, *mucinoso (MBC)*, *Lobulillar (LBC)*, *Tumores mixtos (MTBC)*, *Carcinoma ductal invasivo (IDC)* y *Carcinoma ductal in situ (DCIS)*

El BCQM permite plantear preguntas relacionadas a los tipos de cáncer de mama y a las técnicas para el diagnostico de la misma. De modo que al finalizar el tiempo de cada *Release*, el cual puede variar entre 1 y 4 semanas, la pregunta sera respondida según el análisis de datos generado, y el medico podrá tomar una decisión de valor. Cabe resaltar, que es posible tener una o mas preguntas relacionadas a una técnica y a un tipo de cáncer de mama por cada *Release*, razón por la cual es posible encontrar correlaciones entre las variables características de cada tipo de cáncer encontrando así patrones ocultos en los diferentes conjuntos de datos. Adicionalmente, el BCQM permite identificar a que técnica para el diagnostico de cáncer mama esta relacionada la pregunta a resolver, lo cual de antemano hace posible conocer el tipo información (imágenes o datos) y el algoritmo ML o DL requerido para dar solución al problema. Así mismo, el BCQM permite definir desde la fase inicial el tipo de modelo predictivo o descriptivo según el enfoque analítico generado por la pregunta planteada. Sintetizando, el uso de BCQM facilita la comprensión del problema medico y permite identificar previamente la técnica, el tipo de información y enfoque que debe ser utilizado para el análisis de datos.

### Fase 2: Planeación de actividades

En esta fase el *Data Analysis Team* basado en las preguntas realizadas en el BCQM analiza todas las tareas que hay que llevar a cabo, las estiman en tiempo y las distribuyen entre las personas que las van a realizar durante el *Release*. Dado que el BCQM nos permite conocer de antemano el tipo de cáncer de mama y la técnica para el diagnostico de esta enfermedad, el científico de datos con ayuda del medico puede definir el origen de datos, lo cual va a permitir conocer el tipo, cantidad y peso de la información. Dado lo anterior, es recomendable que el equipo tenga al menos un ingeniero datos, ya que el es el encargado de tomar los datos y convertirlos en información significativa para que el científico pueda realizar el respectivo análisis.

### Fase 3: Adquisición de datos oncológicos

En esta fase, con base a la tareas realizadas en la planeación de actividades, el medico experto en oncología junto con el ingeniero y el científico de datos identifican y reúnen los recursos de datos disponibles (estructurados, no estructurados y semiestructurados) y relevantes para solucionar las preguntas planteadas en el *BCQM*. Cabe resaltar, que en la metodología *DSM-BCD* es factible tener varios conjuntos de datos o imágenes que están relacionados a un tipo de cáncer de mama y una técnica de diagnostico, por lo tanto el *Data Analysis Team* puede tener a varios científicos respondiendo preguntas diferentes en un mismo *Release*. Como consecuencia, al final se pueden obtener como resultado múltiples respuestas y una posible correlación entre las diversas variables oncológicas. Asimismo, en esta fase el *Data Analysis Team* debe definir la infraestructura de datos necesaria según la cantidad de información a procesar, lo cual permitirá proyectar la escalabilidad, alcance y distribución de dicha información.

### Fase 4: Análisis Exploratorio de datos oncológicos

En esta fase, el científico obtiene el conjunto de datos o imágenes que fueron organizados previamente por el ingeniero y realiza un *Análisis exploratorio de datos* para descubrir patrones generales en la información generada. Cabe resaltar, que en esta fase el acompañamiento del medico experto en oncología es de vital importancia, ya que los datos o imágenes que van ser explorados por el científico pueden contener variables que pueden tener o no un valor significativo para el experto, ayudando así a determinar si el análisis planteado para responder la pregunta va o no por un buen camino, de modo que es posible que se agreguen o eliminen diversas variables para lograr el resultado esperado. Adicionalmente, es necesario que los diversos análisis generados estén apoyados con gráficas que sean entendibles por todo el *Data Analysis Team*, esto con el proposito de aportar ideas, y desde esta fase ir encontrando posibles correlaciones entre las variables oncológicas.

### Fase 5: Procesamiento y transformación de datos oncológicos

En esta fase, se abarcan todas las actividades para construir el conjunto de datos o imágenes que se utilizará en la siguiente etapa de modelado y ejecución. Entre las actividades del procesamiento y transformación de datos oncológicos, están la limpieza de datos, combinar datos de múltiples fuentes y transformar los datos en variables de valor. En esta fase, es importante el trabajo en equipo y la comunicación continua entre el ingeniero y el científico de datos para tratar los valores no válidos o faltantes, eliminar duplicados, dar un formato adecuado y combinar archivos, tablas y plataformas. Adicionalmente, el medico experto en oncología deberá proporcionar un visto bueno para proceder con la siguiente fase. Esto dado que al ser experto en el tema de dominio tiene un conocimiento mas profundo de las variables o imágenes que esta observando, y si existiese información innecesaria para el diagnostico del cáncer de mama es posible depurar dicha información para que no afecte el entrenamiento y posterior ejecución del modelo de ML y DL.

### Fase 6: 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 obtenidos en la fase de procesamiento y transformación. En esta fase, el científico debe seleccionar el tipo de aprendizaje (supervisado, no supervisado y por refuerzo) y la técnica determinada (regresión, clasificación, clustering, CNN, RNN, etc.) acorde a las preguntas planteadas en el *BCQM*. Hay que mencionar que en esta fase el *Data Analysis Team* debe definir junto al medico experto en oncología la tolerancia de error permitida en el modelo, esto dado a que la sensibilidad de los análisis puede variar dependiendo del tipo de cáncer de mama y la técnica de diagnostico. Es probable que el científico de datos pruebe múltiples algoritmos con sus respectivos parámetros para encontrar el mejor modelo para las variables oncológicas disponibles. Cabe resaltar, que es de vital importancia que los modelos propuestos no tengan problemas de sobre-ajuste o infra-ajuste ya que esto puede generar resultados erróneos o poco significativos. Adicionalmente, el científico de datos en cuestión junto al *Data Analysis Team* deben definir la infraestructura a nivel de servidor necesaria para el entrenamiento y prueba del modelo según la cantidad de información a procesar, esto con el proposito de generar resultados acertados en el menor tiempo posible en pro de cumplir las tareas definidas en la fase de planeación de actividades y dar valor a los datos oncológicos una vez finalice el *Release*.

### Fase 7: Evaluación e Interpretación

En esta fase, el *Data Analysis Team* evalúa el modelo para comprender su calidad y garantizar que este aborda las preguntas generadas en el *BCQM* de manera adecuada y completa. Es necesario que para realizar la evaluación se utilicen medidas especializadas basadas en el rendimiento, sensibilidad y especificidad del modelo. Adicionalmente, los resultados obtenidos deben ser entendibles por el medico experto en oncología, en donde se garantice que dichos resultados sean interpretados correctamente y estén relacionados a la estadificación y los biomarcadores del cáncer de mama. Es importante que el medico junto al científico de datos ajusten el modelo según las necesidades. Dado que se esta trabajando con datos médicos sensibles, es necesario que al modelo final se aplique a un conjunto de validación para realizar una evaluación final. Además, el *Data Analysis Team* puede asignar al modelo pruebas de *significancia estadística* como prueba adicional para comprobar la respuesta obtenida a la pregunta generada. Esta prueba adicional es fundamental para justificar la implementación del modelo. Finalmente, dado que en el *BCQM* se pueden plantear múltiples preguntas relacionadas a diferentes tipo de cáncer de mama y técnicas de diagnostico durante el *Release*, es necesario que los científicos con ayuda del ingeniero de datos unan, si es posible, los resultados obtenidos en una matriz o mapa de calor para identificar el coeficiente de correlación entre dos o mas variables oncológicas. Esta matriz resultante debe ser analizada por el medico experto en oncología para determinar si existe una relación significativa entre los diferentes tipos de cáncer de mama.

### Fase 8: Retroalimentación medica

En esta fase, el medico experto en oncología determina si los resultados generados por el modelo de ML o DL lograron responder las preguntas planteadas en el *BCQM* y si la nueva informacion obtenida es suficiente para diagnosticar el cáncer de mama o si dichos resultados generaron información relevante para determinar la causa u origen de esta enfermedad, en pocas palabras, si los datos analizados produjeron un valor agregado al dominio medico. En el caso de que los resultados obtenidos no lograsen dar valor a los datos, el *Data Analysis Team* deberá decidir si es necesario re-plantear las preguntas o si se debe adquirir nuevos datos para ajustar el modelo generado. Ademas, el experto en compañía del científico y el ingeniero de datos, basado en su perspicacia medica, deberá ayudar a decidir cual estrategia es las mas apropiada para generar resultados significativos. De forma similar, si el resultado fue satisfactorio el medico debe emitir un dictamen del *nivel de impacto* que tuvo la información generada por los modelos al diagnosticar el padecimiento del cáncer de mama a un determinado paciente y unas vez comprobada la informacion, junto al *Data Analysis Team* alimentar un conjunto de datos con la informacion obtenida de los diagnósticos generados a cada individuo. Lo anterior con el proposito de mejorar el desempeño de los modelos existentes y aumentar su precisión. Finalmente, en cada *Release* se debe garantizar que el tiempo de diagnostico sea cada vez menor o que se genere nueva información que el medico pueda utilizar en sus funciones diarias y que ayude a determinar el origen, relación o posible tratamiento de esta enfermedad.

### Fase 9: Bitácora para el diagnostico del cáncer de mama (BCDL)

En esta fase, se propone el uso de una bitácora para el diagnostico del cáncer de mama (BCDL). El proposito del *BCDL* es almacenar la respuestas obtenidas por cada pregunta planteada en el *BCQM* y la relación de estas preguntas y respuestas con determinado modelo de ML o DL. Esta bitácora solamente debe ser alimentada cuando la informacion obtenida generó valor agregado al dominio medico. Su principal proposito es evitar la redundancia de la informacion y la duplicidad de preguntas planteadas en el *BCQM*, garantizando que en cada *Release* se genere nuevo conocimiento relacionado al cáncer de mama. Se recomienda que la bitácora sea diseñada por medio de un *modelo entidad relación (MER)* que este conformado por entidades como: modelo, tipo de cáncer de mama, técnica de diagnostico, conjunto de datos, pregunta y respuesta. Dado lo anterior, se sugiere que los diferentes conjuntos de datos o imágenes utilizados en los análisis realizados, sean almacenados en un servicio de alojamiento de informacion en la nube (Amazon Cloud, Google Drive, One Drive, etc.) y que dicho informacion este identificada con un código único que facilite su búsqueda cuando sea requerido. De igual manera, los diferentes algoritmos generados deben ser almacenados en un sistema de control de versiones (GitLab, GitHub, Bitbucket, etc.) con su respectivo *Readme* de funcionamiento y un código de identificación único para que pueda ser consultado fácilmente por base de datos. Por consiguiente, el uso del *BCDL* permite tener una trazabilidad detallada de los avances obtenidos en cada *Release* con respecto al diagnostico del cáncer de mama, esto con el proposito de que el *Data Analysis Team* tenga un punto de partida solido para innovar en nuevos modelos de ML y DL a través de la comparación y la mejora continua de modelos existentes que lograron agregar valor a los diferentes tipos de datos oncológicos.

## 4 Conclusión

Debido a que la ciencia de datos es un campo multidisciplinario el cual incorpora herramientas computacionales que permiten dar valor a los datos y aprender de los mismos para poder tomar una decisión que valide la veracidad de una hipótesis planteada, es la mejor opción para generar un diagnóstico significativo en la detección de esta enfermedad. Dado lo anterior, aunque la mayoría de los investigadores han puesto sus esfuerzos en los diagnósticos y pronósticos del cáncer de mama, cada técnica tiene una tasa de precisión diferente que varía según las diferentes situaciones, herramientas y conjuntos de datos que se utilizan. A medida que las capacidades de analítica de datos se vuelven más accesibles y prevalentes, los científicos de datos pueden encontrar un gran apoyo una el la metodología *DSM-BCD* debido a que proporciona una estrategia de orientación, independiente de las tecnologías, los volúmenes de datos o los enfoques involucrados y permite mejorar el desempeño de los modelos de ML y DL para garantizar el diagnostico agil y eficaz del cáncer de mama.

**References**