Grupo de Investigación en Telemedicina GITEM++ (2021), 1–19
doi:10.1017/pasa.2022.32

CAMBRIDGE
UNIVERSITY PRESS

REVIEW

Aplicación de ciencia de datos en metodologías para el diagnóstico del cáncer de mama

Jorge Armando Millán Gómez, Lilia Edith Aparicio Pico

Grupo de Investigación en Telemedicina, Universidad Distrital "Francisco José de Caldas", Facultad de Ingeniería, Bogotá,Colombia
Author for correspondence: J. A. Millán ,L. E. Aparicio, Email: jamillang@correo.udistrital.edu.co, eaparicio@udistrital.edu.co.

(Received 20 03 2022; revised 20 03 2022; accepted 20 03 2022; first published online 19 Abril 2022)

Introducción. 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. Esta investigación se centra esencialmente en la exploración de metodologías en ciencia de datos existentes para la detección y el diagnóstico del cáncer de mama.

Objetivo. Analizar las metodologías en ciencia de datos propuestas para el diagnostico del cáncer de mama.

Métodos. Se efectuó una revisión sistemática de 20 metodologías en ciencia de datos propuestas por la comunidad científica en IEEE Access y Science Direct siguiendo las directrices PRISMA .

Resultados y conclusiones. Los resultados obtenidos permiten inferir que en las investigaciones revisadas no existe definida una metodología en ciencia de datos para el diagnóstico del cáncer de mama. Sin embargo, diversos autores proponen metodologías que permiten la aplicación de esta ciencia en proyectos reales y que pueden utilizarse en su totalidad en el dominio de las ciencias de la salud, teniendo como eje principal la comprensión de aspectos como: la comunicación constante con los interesados, el uso del enfoque ágil, la definición de roles y funciones, el valor agregado de los datos, la retroalimentación continua y la aceptación del experto en oncología para la posterior toma de decisiones que ayude a reducir de manera eficaz la morbilidad causada por esta enfermedad.

Palabras Clave. Metodología, Ciencia de datos, Cáncer de mama, Diagnostico.

1. Introducción

El Cáncer de mama ocupa el primer lugar con mayor número de muertes en Colombia ocupando el primer puesto en la tasa de letalidad sobre los demás tipos de cáncer afectando a mujeres de todas las edades. En el año 2020 los casos detectados de cáncer de mama en Colombia fueron 15.509 de los cuales 4.411 terminaron en muerte[21]. 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 [49].

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 socioeconómico, 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 [11].

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 clasificación de ML y DL 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[1].

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. Muchos in-

vestigadores 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.

La ciencia, en el lenguaje del método científico, es formular hipótesis o conjeturas sobre cómo funciona el mundo, 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[4]. Una parte fundamental de la ciencia de datos es el uso de algoritmos de Machine Learning(ML) y Deep Learning(DL).

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 [43]. 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 f tal que:

$$Conocimiento = f(Datos) \tag{1}$$

Adicionalmente, 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[27].

En otra instancia la literatura muestra que la mayoría de los casos de estudio de investigación científica y de desarrollo de aplicaciones se han dado sobre la aplicación de estas diferentes técnicas a imágenes medicas. Asimismo, otra forma de obtener información relevante es a través de técnicas de detección por Biopsia como es el caso de la aspiración por aguja Fina (FNA[a]) y aspiración por aguja gruesa (CNB[b]) y las técnicas basadas en el análisis de receptores de estrógeno en datos metabolómicos.

A medida que las capacidades de analítica de datos se vuelven más accesibles y prevalentes, los científicos de datos necesitan una metodología fundamental capaz de proporcionar una estrategia de orientación, que sea independiente de las tecnologías, los volúmenes de datos o los enfoques involucrados. 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[45] .

Así, el objetivo de esta revisión es analizar de forma sistemática las investigaciones disponibles acerca de las metodologías en ciencia de datos para el diagnostico del cáncer de mama, esto con el propósito de tener un panorama real acerca del uso de esta ciencia computacional como un campo interdisciplinario que permita dar valor a los datos obtenidos por medio técnicas medicas para el diagnostico del cáncer de mama a través de una metodología que facilite abordar el análisis y la selección de técnicas para poder tomar una decisión que valide la veracidad de una hipótesis planteada. El uso de una metodología 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 [5]. Segun [14], 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. [14] exponen que existen diferentes tipos de cáncer de mama [55], 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 [19], 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 [7]. 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 [42]. El tercer tipo de cáncer es el cáncer de *Tumores mixtos (MTBC)* también

[a] Fine Needle Aspiration
[b] Core Needle Biopsy

conocido como cáncer de mama invasivo [3] causado por las células anormales de los conductos y las células lobulillares [23]. El cuarto tipo de cáncer es el cáncer de mama *Lobulillar (LBC)* [29] 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* [31] que ocurre debido a las células ductales invasivas cuando los tejidos anormales se extienden alrededor del conducto [15]. 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 [44].

Según [6], 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 [6].

En la actualidad, muchas *Técnicas de imagen* se utilizan ampliamente para proporcionar un diagnóstico preciso de las lesiones mamarias[56], 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 [12]; *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 [18]; *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 [17]; y *La Resonancia Magnética (MRI*[c]*)* 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 [57].

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 [35]. 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 [40] .

3. Aplicaciones del ML y DL en el diagnostico del cáncer de mama

Diversas técnicas de ML y DL aplicadas en el diagnóstico del cáncer de mama han sido propuestas por una gran cantidad de investigadores. Aunque para el cribado de información se excluyeron los estudios que tienen un enfoque en técnicas y no en metodologías de ciencia de datos, se seleccionaron las siguientes investigaciones por su relevancia y aporte para el diagnostico del cáncer de mama debido a que brindan una gran cantidad de información y son un punto de partida importante para aplicar la ciencia de datos con dicho fin. Las investigaciones seleccionadas como referentes a nivel de técnicas se muestran a continuación:

[2] realizaron la identificación de subtipos moleculares sobre datos metabolómicos por medio de algoritmos de DL y ML para determinar el pronóstico del cáncer de mama y la selección terapéutica. Para ello, la investigación se basa en la clasificación de dicho cáncer en cuatro subtipos moleculares: Luminal A (*ER*+, *PR*± y *HER2*–), Luminal B (*ER*+, *PR*±, y *HER2*±), *HER2*– enriquecido (*ER*–, *PR*– y *HER2*+) y triple negativo (*ER*–, *PR*– y *HER2*–). Estos subtipos son generados a partir de las proteínas de factor de crecimiento epidérmico receptor 2 (HER2), receptor de progesterona (PR) y receptor de estrógeno (ER) encargadas de estimular el crecimiento y la diferenciación celular. Los resultados de supervivencia difieren significativamente entre estos subtipos. Los subtipos luminales A y B tienen un pronóstico benigno sin embargo, los tumores triple negativos y los tumores *HER2* tienen un pronóstico maligno. Para la investigación se utilizó un data-set conformado por 271 muestras de cáncer de mama (204 *ER*+ y 67 *ER*–) del Departamento de Patología del Hospital Charité. Posterior a eso para complementar dicho data-set se midieron un total

[c]Magnetic Resonance Imaging

de 162 metabolitos con estructura química conocida mediante cromatografía de gases para todas las muestras de tejido. Para realizar la clasificación de los datos metabólicos obtenidos se utilizó una Red Neuronal FNN[d] y los siguientes modelos de ML: RF, SVM, RPART[e], LDA[f], PAM[g] y GBM[h]. Posteriormente, se realizó una validación cruzada (Cross-validation) con el 80% de los datos de entrenamiento y se probaron los modelos con el 20% de los datos restantes. Para finalizar, se utilizaron las medidas AUC[i] y ROC[j] para evaluar el rendimiento general de los dichos modelos. En conclusión, esta investigación determino que las FNN muestran una precisión de predicción (AUC = 0, 93) mayor sobre los métodos de ML al realizar la clasificación del cáncer de mama en los subtipos moleculares basados en el receptor de estrógeno(ER).

[38] proponen de un sistema basado en métodos de DL y ML para la predicción de la proliferación tumoral a partir de imágenes WSI [k] obtenidas del recuento mitótico. La proliferación tumoral, que se correlaciona con el grado del tumor, es un biomarcador crucial que indica el pronóstico de las pacientes con cáncer de mama. El método más utilizado para predecir la velocidad de proliferación tumoral es el recuento de figuras mitóticas en los portaobjetos histológicos de hematoxilina y eosina (*H&E*). El recuento manual de mitosis consiste en identificar células tumorales implicadas en el proceso anormal de la división celular. Para la investigación se utilizó el data-set proporcionado por el desafío de evaluación de la proliferación tumoral (TUPAC16) en donde el objetivo es evaluar algoritmos que predicen las puntuaciones de proliferación tumoral a partir de imágenes WSI. Primero, al considerar el tejido epitelial como regiones de actividad de mitosis, los investigadores propusieron un método de detección de regiones de interés basado en DL para seleccionar las regiones de alta actividad de mitosis a partir de las imágenes proporcionadas en el desafío. Posteriormente, se entrenó un conjunto de redes neuronales para detectar la propagación de la mitosis en distintas áreas seleccionadas. Para finalizar, los investigadores entrenaron el modelo SVM para predecir la puntuación final de proliferación tumoral. La precisión del algoritmo entrenado logró una medida-F del 73,81% y un coeficiente kappa ponderado de 0,612, respectivamente, superando significativamente otros algoritmos que participaron en el desafío TUPAC16. Como conclusión de la investigación, los resultados experimentales demuestran que el algoritmo propuesto basado en el paradigma de DL y ML mejoró considerablemente la precisión de la predicción de la proliferación tumoral y proporciono un modelo con una precisión confiable para respaldar la toma de decisiones a nivel médico.

[30] desarrollaron un modelo de DL para identificar el padecimiento del cáncer de mama mediante imágenes mamográficas obtenidas en el Reino Unido y EE. UU. La mamografía tiene como objetivo identificar el cáncer de mama en las primeras etapas de la enfermedad, cuando el tratamiento puede tener más éxito. A pesar de la existencia de diversas técnicas de detección del cancér de mama en todo el mundo, la interpretación de las mamografías se ve afectada por altas tasas de falsos positivos y falsos negativos. El data-set del Reino Unido consistió en mamografías recopiladas de 25,856 mujeres entre 2012 y 2015 en dos centros médicos en Inglaterra, donde las mujeres se examinan cada tres años. Incluyó a 785 mujeres a las que se les realizó una biopsia y a 414 mujeres con una diagnóstico maligno del cáncer dentro de los 39 meses posteriores a la obtención de imágenes. El data-set de EE. UU consistió en mamografías que se recolectaron entre 2001 y 2018 de 3.097 mujeres en un centro médico académico. Se incluyeron imágenes de las 1511 mujeres a las que se les realizó una biopsia durante este período de tiempo y de un subconjunto aleatorio de mujeres que nunca se sometieron a una biopsia. Entre las mujeres que recibieron una biopsia, 686 fueron diagnosticadas con cáncer dentro de los 27 meses posteriores a la obtención de las imágenes. Los casos que fueron diagnosticados como positivos para cáncer fueron acompañados de un diagnóstico confirmado por biopsia dentro del período de seguimiento. una reducción absoluta del 5,7% y 1,2% (EE.UU. Y Reino Unido) en falsos positivos y del 9,4% y 2,7% en falsos negativos. En la investigación los data-set obtenidos no se utilizaron para entrenar ni ajustar el modelo basado IA. Como resultados, la investigación proporciono evidencia de la capacidad del modelo basado en IA y DL para generalizarse desde el Reino Unido a los EE.UU. En un estudio independiente de seis radiólogos, el sistema de IA superó el diagnóstico obtenido por dichos radiólogos con una precisión (AUC-ROC) mayor que la generada por el radiólogo promedio. Además, para comprobar la predicción del modelo los investigadores realizaron una simulación en la que el sistema de IA participó en el proceso de doble lectura de diagnóstico que se utiliza en el Reino Unido y se descubrió que el sistema de IA mantuvo un rendimiento elevado y redujo la carga de trabajo del segundo lector en un 88%. Como conclusión, esta investigación demostró que las técnicas de IA mejoran la precisión y la eficiencia de las técnicas tradicionales para la detección del cáncer de mama.

[36] utilizaron técnicas de DL para identificar correlaciones histológicas en biopsias de la densidad mamaria obtenidas por técnicas radiológicas para determinar el cáncer de mama entre mujeres con una densidad del tejido fibroglandular alto o bajo. La densidad de la mama es una característica radiológica que refleja el contenido de tejido fibroglandular en relación con el área o el volumen de la mama y es considerada como un factor de riesgo de padecer cáncer. En la investigación se evaluaron imágenes digitalizadas teñidas con hematoxilina y eosina (*H&E*) de biopsias de 852 pacientes. La densidad mamaria se evaluó como volumen fibroglandular global y localizado. Posteriormente, los investigadores modelaron una Red Neuronal CNN para caracterizar la composición de *H&E*. En total, se extrajeron 37 características de la salida de la red, que describen las cantidades de tejido y la estructura morfológica. Se entrenó un modelo de regresión por Bosque Aleatorio (RF) para iden-

[d] Feedforward Neural Network
[e] Recursive Partitioning
[f] Linear Discriminant Analysis
[g] Microarray Prediction Analysis
[h] Gradient Boosting Machine
[i] Area Down The Curve
[j] Receiver Operating Characteristic Curve
[k] Whole Slide Images

tificar correlaciones predictivas del volumen fibroglandular en 588 pacientes. Se evaluaron las correlaciones entre el volumen fibroglandular predicho y cuantificado radiológicamente en 264 pacientes independientes. Se entrenó un segundo clasificador de bosque aleatorio (RF) para predecir el diagnóstico invasivo vs. el diagnóstico benigno; el rendimiento de los algoritmos se evaluó utilizando el área bajo la curva (AUC). Utilizando características extraídas, los modelos de regresión predijeron el volumen fibroglandular global con una precisión de 0,94 y localizado con una precisión de 0,93 teniendo como base el contenido del estroma graso y no graso representando las correlaciones más fuertes seguidas de la cantidad de tejido epitelial. Para predecir el cáncer entre el volumen fibroglandular elevado y bajo, el clasificador logró un valor AUC de 0,92 y 0,84, respectivamente, siendo las características de tejido epitelial las más importantes. Estos resultados sugieren que el estroma no graso, la cantidad de tejido graso y el tejido epitelial predicen el volumen fibroglandular. Como conclusión, la investigación demostró que los modelos de DL y ML permiten identificar correlaciones histológicas con riesgo de cáncer en pacientes con alta y baja densidad de tejido fibroglandular.

[20] desarrollaron una técnica para el diagnóstico asistido por computadora CAD[l] bajo el paradigma de aprendizaje por transferencia profunda para realizar la diagnosis del cáncer de mama utilizando mpMRI[m].El estudio incluyó imágenes clínicas de resonancia magnética de 927 lesiones únicas de 616 mujeres, en donde se excluyeron las imágenes de lesiones que no presentaban una lesión visible, lesiones que no tenían validación de diagnóstico final, o lesiones que no pudieron ser claramente asignadas en a la categoría benigna ni maligna. Cada estudio de resonancia magnética RM[n], incluyó una secuencia con contraste dinámico DCE[o] y una imagen de RM ponderada en un radio T2w[p]. Se utilizó una Red Neuronal CNN previamente entrenada para extraer características de las imágenes DCE y T2w, y se entrenaron modelos SVM en las características de CNN para distinguir entre lesiones benignas y malignas. La información de las imágenes de resonancia magnética con contraste dinámico (DCE) y ponderada en T2w se añadió al modelo IA de tres formas diferentes: fusión de imágenes, fusión de características y fusión de clasificadores. La fusión de imágenes, se utilizó para crear una imagen compuesta RGB con base a las imágenes DCE y T2w. La fusión de características, para combinar las características generadas por las redes neuronales convolucionales con base a las imágenes DCE y T2w para posteriormente ser utilizadas como entrada en un algoritmo de máquina de vectores de soporte (SVM). La fusión clasificadora se utilizó para calcular la probabilidad de malignidad por medio de voto suave (soft-voting) con base al resultado obtenido por el algoritmo SVM al analizar las imágenes DCE y T2w. El rendimiento de los modelos de DL se evaluó utilizando la curva (ROC) y fue comparado utilizando la prueba DeLong. El método de fusión de características superó estadísticamente de manera significativa a los demás métodos. En conclusión, el método propuesto de aprendizaje de transferencia profunda CAD para mpMRI pudo mejorar el diagnóstico del cáncer de mama al reducir la tasa de falsos positivos y mejorar el valor predictivo positivo en la interpretación de imágenes obtenidas por resonancia magnética.

[54], propusieron un método de ML y diseño de una metodología para el análisis de la clasificación de tumores de mama benignos y malignos en imágenes de ultrasonido sin necesidad de un procesamiento de selección regional tumoral a priori. Las imágenes se recopilaron desde el 1 de enero de 2017 hasta el 31 de diciembre de 2018. En total, se examinaron 370 masas benignas y 418 malignas, y se estudiaron a 677 pacientes en este estudio. Los criterios de exclusión para pacientes con tumores benignos incluyeron tipos de tejido que se asociaron con las siguientes afecciones: inflamación (incluida la autoinflamación, inflamación crónica e inflamación xantogranulomatosa), abscesos y dermatitis espongiótica. Para los pacientes con tumores malignos, los criterios de exclusión incluyeron casos con tipos de tejido incompletos, clasificación de categoría BI-RADS[q] desconocida o informes de imágenes incompletas. Las edades de los pacientes oscilaron entre los 35 y los 75 años. El data-set recopilado constaba de 677 imágenes (Benignas: 312, maligno: 365) obtenidas de diferentes centros oncológicos de Estados Unidos. Una vez generado el data-set de estudio se procedió a realizar la extracción de características de las imágenes a través del método de histograma de gradientes orientados HOG[r] y el método histograma piramidal de gradientes orientados PHOG[s]. Adicionalmente, se utilizó el método FAST[t] para determinar si se podían extraer características de clasificación importantes de las imágenes preliminares. El HOG se utilizó para extraer información útil y descartar información redundante para simplificar la clasificación de imágenes calculando y contando histogramas de gradiente de áreas específicas. La distancia entre dos descriptores de imágenes PHOG refleja la medida en que las imágenes contienen formas similares y corresponde a sus diseños espaciales y el método FAST permite un rendimiento mayor que otros métodos en la detección de esquinas de los mamogramas para extraer puntos característicos y luego rastrear y mapear objetos con características relevantes en dichas imágenes. El siguiente paso fue realizar la selección de características por medio de la técnica de selección de características basada en correlación CFS[u] para evaluar los atributos importantes, y crear un subconjunto de datos considerando las habilidades predictivas junto con el grado de redundancia. La función de evaluación se utilizó para evaluar subconjuntos que contenían características que estaban altamente correlacionadas con la clase y no estaban correlacionadas entre sí. Se ignoraron las características irrelevantes y se excluyeron las características redundantes. Finalmente, se realizó la clasificación de características con

[l] Computer-Aided Diagnosis
[m] Multiparametric Magnetic Resonance Imaging
[n] Magnetic Resonance
[o] Dynamic Contrast-Enhanced
[p] T2 weighted image
[q] Breast Imaging Reporting and Data System
[r] Histogram of Oriente Gradients
[s] Pyramid Histogram of Orientad Gradients
[t] Features from the Accelerated Segment Test
[u] Correlation Based Feature Selection

base en la combinación de aprendizaje ponderado localmente LWL[v], la optimización secuencial mínima SMO[w] y el método KNN. Para la evaluación del desempeño de la clasificación, se utilizó la técnica de validación cruzada para determinar el porcentaje de error, la media, la desviación estándar y el intervalo de confianza para los algoritmos de referencia. La precisión diagnóstica se estimó comparando las curvas AUC y ROC con la prueba no paramétrica de DeLong. El rendimiento de clasificación del conjunto de datos de imágenes mostró una sensibilidad del 81,64% y una especificidad del 87,76% para imágenes malignas con AUC de 0,847. Los valores predictivos positivos y negativos fueron 84,1 y 85,8%, respectivamente. En conclusión, la comparación de los diagnósticos médicos y el diagnóstico generado por el modelo de ML propuesto arrojó resultados similares. Esto indica la posible aplicabilidad del ML en la generación de diagnósticos clínicos.

[53] desarrollaron un algoritmo de DL que puede detectar con precisión el cáncer de mama en mamografías usando un entrenamiento e2e [x] que aprovecha de manera eficiente los data-sets de imágenes mamográficas obtenidas a través de diagnósticos asistidos por ordenador CAD . En este enfoque las anotaciones para identificación de lesiones en la mama obtenidas de las imágenes mamográficas solo se utilizaron en la etapa de entrenamiento inicial, y las etapas posteriores requirieron solo las características obtenidas a nivel de imagen, eliminando la dependencia de diagnósticos obtenidos de estas anotaciones. El algoritmo generado fue testeado con dos data-set de prueba. El primer data-set fue obtenido de la base de datos de mamografías digitales del sitio web CBIS-DDSM. Este data-set almacenaba 2478 imágenes de mamografías de 1249 mujeres e incluía vistas tanto craneocaudales (CC) como oblicuas mediolaterales (MLO), además contaba con etiquetas confirmadas patológicamente en una categoría benigna o maligna para la mayoría de los exámenes. Con este data-set, el modelo logró una predicción con un valor AUC de 0,88 y el promedio de cuatro modelos entrenados posteriormente mejoró esta predicción a valor AUC de 0,91 con una sensibilidad de 86,1% y especificidad de 80,1%. El segundo data-set fue obtenido de la base de datos de mamografías digitales INbreast FFDM[y] la cual contenía información de 115 pacientes y 410 mamografías, este data-set también incluía vistas CC y MLO. Con el segundo data-set el modelo logró una predicción con valor AUC de 0,95, y el promedio de cuatro modelos entrenados mejoró el AUC a 0,98 con una sensibilidad de 86,7% y especificidad de 96,1%. La investigación demostró que el entrenamiento generado en una Red Neuronal CCN con el enfoque e2e CBIS-DDSM se puede transferir a imágenes FFDM de INbreast utilizando solo un subconjunto de datos para ajustar el modelo y sin depender de la disponibilidad de anotaciones de lesiones. Estos hallazgos demostraron que los métodos de DL se pueden utilizar para lograr una alta precisión de diagnosis en mamografía heterogéneas y es muy prometedor para mejorar las herramientas clínicas para reducir la detección de falsos positivos y falsos negativos de cáncer de mama.

[37] utilizaron el paradigma de DL con base al diagnóstico del cáncer de mama a partir del estado molecular del receptor de Estrógeno ERS[z],determinado por los patólogos a partir de la tinción inmunohistoquímica IHC[aa] que destaca la presencia de antígenos de superficie celular del tejido obtenido a través de una biopsia. Para lograrlo, se recopilaron imágenes WSI teñidas con hematoxilina y eosina (H&E) del data-set obtenido del Banco Australiano de tejidos para el cáncer de mama ABCTB[ab], que contenía 2535 imágenes de exámenes realizados a pacientes y el data-set obtenido del Atlas genómico del cáncer TCGA[ac], que contenía 1014 imágenes de 939 pacientes. Ambos conjuntos de datos informaban el estado de la presencia de receptor de Estrógeno (ER), receptor de Progesterona (PR) y el de factor de crecimiento epidérmico receptor 2 (HER2) determinado por patólogos encargados de la inspección visual de los portaobjetos teñidos mediante IHC. Los WSI se escanearon a una resolución de 20x o superior. Posteriormente, se utilizó el modelo de aprendizaje de instancias múltiples MIL[ad], el cual fue entrenado usando tinciones H&S y anotaciones IHC como etiquetas de datos de entrada. El modelo MIL se utilizó para estimar el ERS a partir de mosaicos seleccionados al azar de un WSI. Para lograr la interpretabilidad al localizar mosaicos discriminativos en imágenes de H&E e identificar características histomorfológicas que se correlacionan con el crecimiento celular impulsado por hormonas, los investigadores diseñaron una arquitectura bajo el paradigma de redes neuronales profundas llamada *ReceptorNet* basada en el proceso ejecutado por el modelo MIL. ReceptorNet aprende a asignar altos pesos a los mosaicos en la imagen de H&E que tienen la máxima capacidad discriminativa, y a asignar bajos pesos a los mosaicos que son insignificantes para el proceso de aprendizaje y diagnóstico. El análisis de los pesos asignados permitió determinar los mosaicos que se utilizaron para realizar la estimación de ERS. La arquitectura de ReceptorNet consta de tres redes neuronales interconectadas: un módulo extractor de funciones, un módulo de aprendizaje y un módulo de decisión. Adicionalmente, se comparó el método propuesto con dos métodos MIL ampliamente utilizados: Meanpool y Maxpool. En Meanpool, las representaciones de características de mosaicos se promedian para obtener una representación de características agregadas. En Maxpool, se obtiene un máximo de características para cada una de las dimensiones de características. Estos métodos se entrenaron en la misma arquitectura de modelo que ReceptorNet. El algoritmo propuesto logro una curva AUC de sensibilidad y especificidad de 0,92. En conclusión, esta investigación demostró una estimación precisa del estado del receptor de ER a partir de tinciones de H&E utilizando una Red Neuronal lo cual impacta en la disminución del trabajo al realizar procesos patológicos. En términos más

[v] Local Weighted Learning
[w] Sequential Minimal Optimization
[x] end-to-end
[y] Full-Field Digital Mammography

[z] Estrogen Receptor Status
[aa] Inspection Immunohistochemistry
[ab] Australian Breast cáncer Tissue Bank
[ac] The Cancer Genomic Atlas
[ad] Multiple Instance Learning

generales, el estudio represento una mejora de los métodos médicos para el diagnóstico del cáncer de mama y demostró el potencial del ML para mejorar el pronóstico y diagnóstico del cáncer mediante el aprovechamiento de señales biológicas imperceptibles para el ojo humano.

[46] proponen el modelo CHISEL[ae] para Segmentación y evaluación de imágenes histopatológicas asistidas por computadora. Este sistema e2e es capaz de evaluar cuantitativamente muestras digitales (benignas y malignas) con tinción nuclear inmunohistoquímica IHC de diversa intensidad y compacidad del tejido afectado por el cáncer de mama. El modelo fue capaz de procesar imágenes patológicas digitales, núcleos de microarrays de tejido TMA[af] y muestras adquiridas por cámara digital conectada a un microscopio. Una de las características principales de CHISEL es la segmentación de imágenes con base a regiones de interés ROIs[ag]. El data-set utilizado fue recopilado con base a imágenes de tejido de cáncer de mama adquiridas de diversas universidades, laboratorios, hospitales y departamentos de patología ubicados en España. Las secciones de tejido de las biopsias de mama y ganglios auxiliares se sometieron a un procesamiento histotécnico tradicional y se convirtieron en bloques de tejido embebidos en parafina y fijados con formalina. Los TMA se formaron extrayendo pequeños cilindros de tejido colocados en una matriz en un bloque de parafina. Posteriormente, se tiñeron con los anticuerpos primarios IHC indirectos contra FOXP3[ah] y los anticuerpos secundarios, incluido el bloque de peroxidasa, polímero marcado, tamponado con sustrato cromógeno DAB+ y finalmente contrastado con hematoxilina. La tinción simultánea de múltiples muestras en TMA disminuye la variabilidad intralaboratorio de la diferencia en las concentraciones de tinción entre cortes. El TMA tiene una gran ventaja para procesar imágenes digitales porque se adquieren múltiples muestras en un escaneo bajo las mismas condiciones. En general, el uso de TMA aumenta la reproducibilidad de los estudios de patrones de biomarcadores. El modelo propuesto se centró en procesar las imágenes teñidas con IHC (especialmente DAB & H) para posteriormente estimar la masa de células beta y el número de núcleos dentro del área teñida de insulina. El data-set estaba conformado por 20 imágenes de muestras de tejido teñidas con DAB y H. El modelo utilizo el paradigma de ML y el procesamiento local recursivo para eliminar los contornos distorsionados (inexactos) del conjunto de datos adquiridos. El método se validó utilizando dos data-set etiquetados que demostraron la relevancia de los resultados obtenidos. La evaluación se basó en el conjunto de datos IISPV de tejido de biopsia de pacientes con cáncer de mama, con marcadores de células T, junto con el conjunto de datos Warwick BetaCell de tejido teñido con DAB y H de pacientes con diabetes post-mórtem. Posteriormente, se utilizaron Redes neuronales Artificiales ANN[ai] para clasificar las imágenes en categorías malignas o benignas, y se optimizó el algoritmo analizando la influencia del número de capas, el número de neuronas y las funciones de activación en los resultados de la clasificación. Finalmente, la clasificación dio como resultado una precisión media del 99,2% que se probó con cinco validaciones cruzadas. Este modelo resuelve el complejo problema de la cuantificación de núcleos en imágenes digitalizadas de cortes de tejido teñidos inmunohistoquímicamente, logrando los mejores resultados para muestras de tejido de cáncer de mama teñidas con DAB y H. Para facilidad de manejo del modelo, se generó una interfaz gráfica (GUI) y se optimizó para utilizar completamente la potencia informática disponible. Con base a los resultados de los objetos detectados y clasificados con base a algoritmos de DL y ML, se llegó a la conclusión de que el método propuesto puede lograr resultados mejores o similares a los métodos de última generación.

[39] propusieron un conjunto de datos denominado SHIDC-BC-Ki-67 basado en la proteína nuclear Ki-67 y los linfocitos infiltrantes de tumores TIL [aj],los cuales son factores determinantes para predecir la progresión de un tumor cancerígeno y la respuesta probable del mismo a la quimioterapia. Teniendo en cuenta que la estimación de ambos factores depende de la observación de patólogos profesionales y que también pueden existir variaciones interindividuales, los métodos automatizados que utilizan algortimos de ML, específicamente los enfoques basados en el DL, son llamativos para realizar el diagnóstico y pronóstico del cáncer de mama. Sin embargo, los modelos de DL necesitan una cantidad considerable de datos etiquetados y este fue el motivo por lo cual se propuso la creación del data-set SHIDC-BC-Ki-67. Adicionalmente, en la investigación realizada se presenta una arquitectura en pipeline, para la estimación de la proteína Ki-67 y la determinación simultánea de la puntación de TIL intratumoral en células cancerígenas. El data-set recopilado estaba formado por imágenes digitales de muestras microscópicas con tinción nuclear inmunohistoquímica IHC obtenidas a través de biopsias realizadas con aguja gruesa(CNB) de tumores de mama malignos del tipo de carcinoma ductal invasivo. Además, se usó hematoxilina para realizar la tinción nuclear y semicuantificar el grado de inmunotinción. Las muestras de la biopsia fueron recolectadas durante un estudio clínico de 2017 a 2020. El data-set SHIDC-B-Ki-67 contiene 1656 datos de entrenamiento y 701 de prueba. Todos los pacientes que participaron en este estudio eran pacientes con un diagnóstico patológicamente confirmado de cáncer de mama cuyas biopsias fueron tomadas en los laboratorios de patología de los hospitales afiliados a la Universidad de Ciencias Médicas de Shiraz en Shiraz, Irán. Para la adquisición de las imágenes se realizaron los siguientes pasos: Primero, el patólogo identifica el tumor y define una región de interés ROI. Para cubrir todo el ROI, se capturan varias imágenes que pueden superponerse. El patólogo selecciona preferentemente imágenes del área tumoral. Luego, patólogos expertos etiquetaron las imágenes teñidas como células tumorales positivas y negativas para Ki-67 y células tumorales con linfocitos infiltrantes positivos. Finalmente, se realizó la clasificación celular y la detección de Ki-67 y TIL, para lograrlo, se usaron Redes

[ae] Computer assisted Histopathological Image Segmentation and Evaluation
[af] Tissue Micro Array
[ag] Regions of Interest
[ah] proteína P3 de la Forkhead box
[ai] Artificial Neural Network
[aj] Tumor Infiltrating Lymphocytes

Neuronales Convergentes(CNN) para extraer características relevantes y estimar mapas de densidad a partir de una imagen RGB[ak] de entrada. PathoNet primero extrae características de las imágenes de entrada y luego predice los píxeles candidatos para las células inmuno-positivas e inmuno-negativas Ki-67, y también los linfocitos con sus valores de densidad correspondientes. El backend propuesto utiliza capas convolucionales para la detección, clasificación y recuento de células en imágenes histopatológicas. Posteriormente se utilizo el algoritmo de Watershed para asignar imágenes en escala de grises a un espacio de relieve topográfico. La arquitectura en pipeline propuesta estaba conformada de una red PathoNet, posprocesamiento y el algoritmo de Watershed. En conclusión, en esta investigación se propuso un nuevo punto de referencia para la detección celular, la clasificación, la estimación del índice de proliferación y la puntuación de TIL mediante imágenes histopatológicas procesadas con modelos de DL para la extracción de características que permiten generar un data-set para la estimación intratumoral en células de cáncer de mama.

4. Métodos

En este trabajo se ha llevado a cabo una revisión sistemática de la literatura científica publicada en materia de ciencias de la computación en relación con metodologías en ciencia de datos para el diagnostico del cáncer de mama. Para su elaboración, se han seguido las directrices de la declaración PRISMA [34] para la correcta realización de revisiones sistemáticas. A continuación se detallan las etapas realizadas en el proceso.

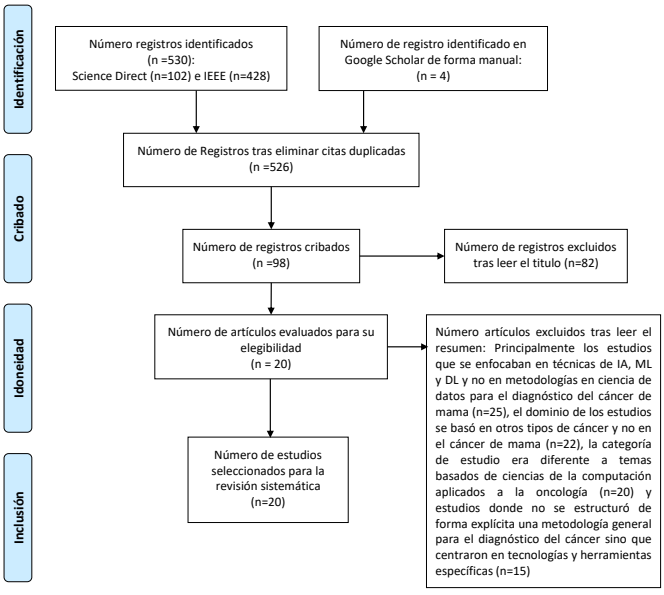

Figure 1. Diagrama de flujo PRISMA en cuatro niveles

4.1 Búsqueda inicial

Las primeras búsquedas se realizaron en abril de 2021 combinando los términos *data science, breast cancer, deep learning, machine Learning, artificial intelligence, cancer Biopsy ,Core needle biopsy, Fine needle aspiration* en las bases de datos de IEEE Access, Nature Reviews Cancer, Elsevier y Science Direct. Posteriormente, para generar una búsqueda mas especifica acerca de metodologías en ciencia de datos aplicadas al diagnostico del cáncer de mama se realizó la combinación de estos términos usando los operadores booleanos *AND* y *OR*. Los resultados obtenidos en la búsqueda arrojaron una cantidad excesiva de literatura basada en técnicas de Inteligencia Artificial, Machine Learning y Deep Learning aplicadas al diagnostico de esta enfermedad sin tener una metodología clara en ciencia de datos aplicada a esta área de la salud, por lo cual dichos resultados fueron poco útiles para la revisión. Sin embargo, las consultas realizadas permitieron tener un visión global de la complejidad del tema de investigación y los diferentes enfoques que pueden generarse a nivel de ciencia de la computación si no se limita la temática al rededor de una metodología.

Debido a que los resultados arrojados por Nature Reviews Cancer y The Lancet estaban enfocados en técnicas y no metodologías, lo cual no aportada información relevante al estudio, se decidió su eliminación de la búsqueda sistemática.

4.2 Búsqueda sistemática

La búsqueda sistemática se realizo en febrero del 2022, en IEEE Access y Science Direct, acotando los resultados de publicaciones realizadas en 2014 hasta la actualidad. La combinación de términos que arrojó mejores resultados en ambos buscadores fue la siguiente: *(methodology AND Data Science) OR (methodology AND diagnose AND cancer) OR (data science AND methodology AND health AND diagnostics) OR (data science AND methodology AND breast AND cancer)*.

Concretamente, se obtuvieron 530 resultados de los cuales 428 fueron encontrados en IEEE Access y 102 en Science Direct. Antes de realizar la selección de artículos, se definieron los siguientes criterios de inclusión y exclusión.

- ***Criterios de inclusión***
  - Tratarse de artículos de investigación, revisión sistemática, libros o manuales.
  - Que los estudios realizados estén basadas en metodologías en ciencia de datos y no en técnicas de Inteligencia Artificial, Machine Learning y Deep Learning.
  - Que la categoría de estudio este bajo la temática de ciencias de la computación aplicado a la oncología.
  - Que el dominio determinado este enfocado en el diagnostico del cáncer de mama.
  - Que los articulo de investigación realizados no dependan de tecnologías ni herramientas específicas.
  - Que se hayan publicado entre 2014 y 2022, ambos inclusive.
- ***Criterios de exclusión***
  - Se excluyen los estudios que se enfoquen en técnicas de Inteligencia Artificial, Machine Learning y Deep Learning y no en metodologías en ciencia de datos.
  - Estudios en donde el dominio se base en otros tipos de cáncer y no en el cáncer de mama.

[ak]Red, Green, Blue

- Estudios donde la categoría sea diferente a temas basados de ciencias de la computación aplicados a la oncología.
- Estudios donde no se evidencie de forma explícita una estrategia general para el diagnóstico del cáncer sino se centren en otros temas, tecnologías y herramientas específicas.

Según estos criterios, y sólo con la lectura del titulo se consideraron adecuados 98 artículos. Posteriormente, se realizo la lectura del resumen y a partir de esta lectura, se descartaron 82 artículos, principalmente los estudios que se enfocaban en técnicas de Inteligencia Artificial, Machine Learning y Deep Learning y no en metodologías en ciencia de datos para el diagnóstico del cáncer de mama (n = 25), el dominio de los estudios se basó en otros tipos de cáncer y no en el cáncer de mama (n = 22), la categoría de estudio era diferente a temas basados de ciencias de la computación aplicados a la oncología (n = 20) y estudios donde no se estructuró de forma explícita una metodología general para el diagnóstico del cáncer sino que centraron en tecnologías y herramientas específicas (n = 15). Finalmente, 16 artículos encontrados en IEEE Access y Science Direct cumplieron con los criterios de inclusión y se seleccionaron para llevar a cabo la revisión sistemática.

4.3 Búsqueda Manual

Al realizar una lectura detallada de las 16 investigaciones encontradas en la búsqueda sistemática, basándonos en sus referencias para comprobar si existía información adicional relacionada con el tema de estudio, se empleo Google Scholar con los términos de búsqueda descritos anteriormente y como resultado se incluyeron 3 capítulos de libros de ciencias de la computación que abarcaban diferentes metodologías en ciencia de datos y 1 articulo de investigación encontrado en la revista IJSR[al] que contenía información relevante para la investigación.

En definitiva, 20 investigaciones entre artículos y capítulos de libros publicados entre 2014 y 2022 fueron seleccionados para realizar la revisión sistemática de la aplicación de la ciencia de datos en metodologías para el diagnóstico del cáncer de mama (Figura 1).

5. Resultados

Teniendo en cuenta los criterios de inclusión planteados anteriormente, en la tabla 1 se encuentra una síntesis de los resultados obtenidos de cada una de las metodologías que se consideraron relevantes para la revisión sistemática. Cabe resaltar, que ninguna de las metodologías analizadas se enfoca directamente en el cáncer de mama, sin embargo hacen un gran énfasis en el entendimiento del dominio y su importancia en la definición de metas claras y alcanzables para dar valor a los datos.

[50] realizan una descripción general del enfoque de la investigación, las metodologías actuales, las mejores prácticas y las posibles brechas en la ejecución de las fases de la metodología para minería de datos *CRISP-DM* [am]. Esta metodología es una innovación cruzada entre industrias de diferentes sectores. Fue desarrollada por SPSS y Teradata en 1996 y describe un enfoque que es comúnmente utilizado por expertos en minería de datos. Esta metodología presenta un proceso iterativo estructurado, bien definido y documentado que consta de seis fases iterativas: comprensión del negocio, comprensión de datos, preparación de datos, modelado, evaluación y despliegue. Adicionalmente, esta metodología se divide en tareas genéricas con objetivos específicos y resultados concretos. Estos hitos permiten la evaluación intermedia de los resultados, una eventual re-planificación y una colaboración más fácil. Las tareas genéricas pretenden abarcar el mayor número posible de situaciones en la minería de datos, y se dividen a su vez en tareas o actividades específicas [33]. En conclusión, dado que esta esta metodología abarca desde la comprensión empresarial hasta la implementación, y ademas se compone de un proceso fácil y estructurado, confiable, de uso común e independiente de la industria, es la metodología mas popular en la práctica y en la investigación.

[47] estudiaron y evaluaron la metodología KDD [an], la cual tiene como propósito principal extraer información previamente desconocida y patrones ocultos comprensibles en los datos. Esta metodología consta de cinco etapas: selección de datos, pre-procesamiento de datos, transformación de datos , minería de datos y evaluación/interpretación. En esta metodología los datos se obtienen de múltiples fuentes, y cada una de las etapas del proceso puede ser aplicada a diferentes organizaciones. En conclusión, esta metodología se adapta al crecimiento exponencial de los datos y ayuda a que las diferentes organizaciónes generen valor a dichos datos basándose en un entendimiento previo del dominio.

[52] realizaron un estudio comparativo y una descripción puntual de los aspectos mas importantes de la metodología denominada SEMMA[ao]. Esta metodología fue creada por la organización SAS Enterprise Miner y permite comprender, organizar, desarrollar y mantener proyectos de minería de datos a través de un ciclo de cinco etapas: muestreo, exploración, modificación, modelado y evaluación de datos. En conclusión, esta metodología ofrece un ciclo de vida para la solución de los problemas típicos del ámbito empresarial al tratar de cumplir objetivos que necesiten modelos predictivos y descriptivos para el análisis de grandes volúmenes de datos.

[33] proponen una metodología denominada RAMSYS [ap], para llevar a cabo proyectos de minería de datos a distancia de manera colaborativa . Esta metodología se compone de seis fases: gestión del negocio, libertad para resolver problemas, empezar en cualquier momento, parar en cualquier momento, intercambio de conocimientos en línea y seguridad. Esta metodología profundiza las fases de CRISP-DM, y permite que el esfuerzo de minería de datos se invierta en diferentes ubicaciones que se comunican a través de una herramienta basada en la web. En conclusión, el objetivo de la metodología es permitir el trabajo colaborativo de científicos

[al] International Journal of Science and Research
[am] Cross-Industry Standard Process for Data Mining
[an] Knowledge Discovery Databases
[ao] Sample, Explore, Modify, Model, and Access
[ap] Remote Collaborative Data Mining System

de datos ubicados de forma remota, de modo que se facilite el intercambio de información a distancia, así como la libertad de experimentar con cualquier técnica de resolución de problemas [28].

[13] fundadores de Domino Data Lab en Silicon Valley, crearon la metodología Domino DS, la cual abarca el ciclo de vida de un proyecto de ciencia de datos. Esta metodología se compone de las siguientes fases: ideación, adquisición y exploración de datos, investigación y desarrollo, validación, entrega y monitoreo. Esta metodología se fundamenta en CRISP-DM, la ágilidad y las necesidades del cliente para guiar a un equipo de análisis de datos hacia un mejor desempeño. En conclusión, esta metodología integra eficazmente el proceso de la ciencia de datos, la ingeniería de software y los enfoques ágiles para generar valor a los datos a través de entregas iterativas enfocadas en primera instancia en el problema comercial y después en la implementación [28].

[22] realizaron un estudio, en donde analizaron cómo los principios y prácticas ágiles han evolucionado con la inteligencia empresarial. En este estudio los autores abordaron la metodología Agile BI Delivery Framework. Esta metodología sugiere que los desafíos a los que se enfrentan los proyectos de BI hacen que el enfoque Agile sea una respuesta atractiva debido a las semejanzas que existen entre ambas, por lo tanto sugieren que dichas metodologías pueden evolucionar en paralelo. Por el lado de la metodología BI [aq] tenemos cinco etapas: descubrimiento, diseño, desarrollo, despliegue y entrega de valor. Dado lo anterior, los autores proponen las siguientes etapas para Agile BI Delivery Framework: alcance, adquisición de datos, análisis, desarrollo de modelos, validación e implementación. En esta metodología el alcance se centra en un marco de entrega ágil que aborda la influencia de la ciencia de datos en BI. En conclusión, el propósito de la metodología es aplicar la agilidad a los problemas comunes que se encuentran en los proyectos de BI al promover la interacción y la colaboración entre las partes interesadas, debido a que esto garantiza requisitos más claros, una comprensión de los datos, una responsabilidad conjunta y la obtención de resultados de mayor calidad. Por lo tanto, se dedica menos tiempo a intentar determinar los requisitos de información y se dedica más tiempo a descubrir conocimiento oculto en los datos.

[26] plantean que el ML requiere una comprensión profunda de los dominios a los que se pueden aplicar modelos y algoritmos de aprendizaje profundo debido a que los datos determinan la funcionalidad de un sistema de información. Por lo tanto, consideran que se requiere una evaluación para determinar si los datos de entrenamiento son representativos del dominio. Dado lo anterior, afirman que el poder real de la ciencia de datos se hace evidente al aprovechar el ML en big data con el propósito de tomar decisiones que tenga un soporte solido en los datos para generar estrategias claves en la transformación digital. De lo contrario, si no se tiene una claridad absoluta del dominio podrían surgir problemas que podrían contribuir a sesgos y errores en los modelos de aprendizaje automático. Dado lo anterior, proponen una metodología de desarrollo de siete fases basada en la unión del modelado conceptual con el aprendizaje automático: Entendimiento del problema, Recopilación de datos, Ingeniería de datos, Entrenamiento del modelo, Optimización del modelo, Integración y evaluación del modelo y la Toma de decisiones analíticas. Los autores consideran que, los modelos conceptuales son un *lente* a través de la cual los seres humanos obtienen una representación mental intuitiva, fácil de entender, significativa, directa y natural de un dominio. Por el contrario, el aprendizaje automático utiliza los datos como un *lente* a través de la cual obtiene representaciones internas sobre las regularidades de los datos tomados de un dominio. En consecuencia, la unión del modelado conceptual con el ML contribuye entre sí a: mejorar la calidad de los modelos de aprendizaje automático mediante el uso de modelos conceptuales durante la ingeniería de datos, el entrenamiento de modelos y las pruebas de modelos y mejorar la interpretabilidad de los algoritmos de aprendizaje automático mediante el uso de modelos conceptuales. Los modelos prácticos de ML solo son útiles dentro de un dominio determinado, como juegos, decisiones comerciales, atención médica, política o educación. Cuando se integran en los sistemas de información, los modelos de ML deben seguir las leyes, las regulaciones, los valores sociales, la moral y la ética del dominio, y obedecer los requisitos derivados de los objetivos comerciales. Esto destaca la necesidad de un modelo conceptual para ayudar a transformar ideas comerciales en representaciones estructuradas y, a veces, incluso formales, que pueden utilizarse como pautas precisas para el desarrollo de software. Por lo tanto, ayudan a estructurar los procesos de pensamiento de los expertos en el dominio y los ingenieros de software.

[16] presentan un enfoque de una metodología de proceso moderno para el análisis de Big data (BDA[ar]) que se alinea con los nuevos cambios tecnológicos e implementa la agilidad en el ciclo de vida del análisis avanzado y el desarrollo de sistemas de ML, para minimizar el tiempo necesario para alcanzar un resultado deseado. La metodología propuesta, basada en el modelo de proceso de BDA, se llama Data Science Edge (DSE), la cual los autores consideran que junto con las metodologías ágiles sirve como un modelo de proceso para el descubrimiento de conocimientos en ciencia de datos. El ciclo de vida de DSE se compone de cinco fases: Planificar, Recopilar, Seleccionar, Analizar y Actuar. Para que DSE se alineara con una metodología ágil se proporciono una guía en cada fase sobre qué actividades serían fundamentales para comenzar y generar de forma eficiente un producto mínimo viable aunque existiera la necesidad de un refinamiento para mejorar los análisis posteriores. Por consiguiente, el propósito de la metodología propuesta fue adoptar una filosofía ágil para el desarrollo del análisis, en donde los resultados obtenidos se compartan con más frecuencia para formar un circuito de retroalimentación de la opinión de las partes interesadas y utilizar esas necesidades para validar el estado actual e influir en su evolución hacia un estado final que cumpla con los requerimientos y objetivos de un dominio especifico.

[aq] Business Intelligence

[ar] Big data analytic

[51] proponen una metodología denominada *DECIDE*[as] la cual se fundamenta en un diseño ágil basado en eventos y datos para proyectos decisionales de Big Data. El propósito de esta metodología es ayudar a las organizaciones a determinar los objetivos comerciales y analíticos deseados según la toma de decisiones con base al análisis de datos, en donde se debe: definir una estrategia de datos clara, encontrar a las personas adecuadas para llevar a cabo un cambio cultural impulsado por los datos y seguir la ética de la información. Esta metodología se basa en la metodología DSRM[at] que permite realizar investigaciones en ciencias del diseño en sistemas de información y ayuda a definir un modelo de proceso a seguir para diseñar una solución de sistemas de información, dependiendo del enfoque realizado para encontrar dicha solución. La metodología DSRM define principalmente cuatro enfoques posibles: centrado en el problema, centrado en objetivos, centrado en el diseño y desarrollo, e impulsado por cliente y contexto. En el caso de la metodología DECIDE se utiliza un enfoque centrado en el problema, que consta de las siguientes fases: identificación y motivación del problema, definición de los objetivos de la solución, diseño y desarrollo, demostración, evaluación y comunicación. Cabe resaltar que el valor de esta metodología esta basado en cinco fundamentos claves en la fase de diseño y desarrollo: Agilidad, Enfoque Bottom-up, Datos y eventos, Multi-arquitecturas y Multi-tecnologías. En conclusión, esta metodología está diseñada para respetar los conceptos y las mejores prácticas para la toma de decisiones soportada en Big data y para ser aplicada a grandes proyectos, con un tamaño de equipo significativo.

[59] aseguran que la ciencia de datos constituye la tecnología central del análisis y procesamiento de Big data que contiene un enfoque eficaz para dar valor a los datos generando oportunidades sin precedentes en al menos cuatro aspectos: la innovación en la gestión, el desarrollo industrial, el descubrimiento científico y el desarrollo de disciplinas. Ademas, consideran que en la ciencia de datos, la descripción de dichos datos necesita modelado, en donde el modelado es entendido como la formalización de objetos, objetivos y métodos de procesamiento. Por otra parte, consideran que el análisis es un proceso de juzgar las propiedades teóricas de la viabilidad, precisión, complejidad y eficiencia para realizar transformación de los datos en información, la información en conocimiento y el conocimiento en toma de decisiones, por lo tanto, para dar valor a los datos una metodología debe cumplir las siguientes fases: recopilación, convergencia, almacenamiento, administración, consulta, clasificación, extracción, análisis y la realización de descubrimientos científicos. Por consiguiente, los autores consideran que una metodología en ciencia de datos se resume como un híbrido de modelado, análisis, cálculo y aprendizaje, en donde el objetivo fundamental es realizar la cognición y el control del mundo real a través de la transformación de los datos.

[41] basados en la ciencia de datos, mejoran la fase de preparación y tratamiento de datos de una metodología para el desarrollo de aplicaciones de Minería de Datos(MD) basados en el análisis organizacional, denominada *MIDANO*. Esta metodología consta de tres fases: conocimiento de la organización, preparación y tratamiento de los datos y el desarrollo de herramientas de MD. Todas las fases y actividades de esta metodología pretenden abarcar el dominio de conocimiento que puede encontrarse en una organización, por lo que es necesario tener los datos integrados en una sola vista, la cual normalmente es conocida como *Vista Minable*, que esta compuesta por una tabla con todas las variables del proceso y los datos a considerar en el estudio de MD. Sin embargo, en las metodologías actuales no muestran cómo lograr una vista minable adecuada, es por esto que los autores para mejorar el proceso definen dos tipos de vista minable: Vista Minable Conceptual (VMC), la cual describe en detalle cada una de las variables a tomar en cuenta para la tarea de MD y una Vista Mineable Operativa (VMO) que surge como el resultado de cargar los datos del historial y de realizar la etapa de tratamiento de datos. Tanto en la VMC, como en la VMO, se identifican determinadas variables llamadas *variables objetivo* que permiten la consecución de los objetivos de MD, ya que las mismas son las que se desean predecir, clasificar, calcular, inferir según el dominio del problema de estudio. En conclusión, los autores utilizaron el conjunto de formalismos teóricos, métodos y técnicas, para el tratamiento de grandes volúmenes de datos ofrecidos por las metodologías en ciencia de datos, para detallar las variables relevantes de diversos problemas de estudio, a partir de escenarios futuros definidos en el dominio de la organización.

[9] proponen una metodología orientada a procesos para ayudar a la gestión de proyectos de ciencia de datos denominada *POST-DS*[au]. Esta metodología describe la secuencia de actividades realizadas en un proyecto de ciencia de datos y se basa en el ciclo de vida de la metodología CRISP-DM la cual proporciona un plan completo para realizar un proyecto de minería de datos basada en las siguientes fases: comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación e implementación. La metodología POST-DS está inspirada particularmente en la metodología CRISP-DM, pero con la diferencia de que permite la identificación de procesos, organización, programación y herramientas para la gestión de proyectos de ciencia de datos a través de componentes específicos para: la asignación de roles, el ajuste de expectativas, la definición del alcance del proyecto, los costos y el tiempo.Para lograrlo, esta metodología consta de las siguientes fases establecidas a través de un cronograma base: comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación y despliegue. En conclusión, esta metodología ejecuta cada una de las fases tradicionales de un proceso en ciencia de datos teniendo como eje cada una de las fases necesarias para la gestión de proyectos.

[58] aseguran que la variedad de sistemas actuales de análisis de datos comerciales y de código abierto difiere significativamente, en términos de características disponibles, funcionalidad y escalabilidad, de los sistemas de análisis de datos que respaldan el flujo de trabajo de la ciencia de datos. Los autores consid-

[as] Decisional Big Data Methodology
[at] Design Science Research Methodology

[au] Process Organization and Scheduling electing Tools for Data Science

Table 1. Características de las metodologías en ciencia de datos revisadas.

Autores	Metodología	Fases	Resultados
[50]	CRISP-DM	1. Comprensión del negocio 2. Comprensión de datos 3. Preparación de datos 4. Modelado 5. Evaluación 6. Despliegue	1. Es una de las metodologías en ciencia de datos más usada para proyectos de análisis, minería de datos y ciencia de datos. 2. Esta metodología no explica cómo deben organizarse equipos de trabajo para llevar a cabo procesos de gestión que se alineen con el software.
[33]	RAMSYS	1. Gestión del negocio 2. Libertad para resolver problemas 3. Empezar en cualquier momento 4. Parar en cualquier momento 5. Intercambio de conocimientos en línea 6. Seguridad	1. Esta metodología se basa en la metodología CRISP-DM y permite la colaboración para proyectos de minería de datos desde diferentes ubicaciones por medio de una herramienta basada en la web. 2. Esta metodología permite el trabajo colaborativo de científicos de datos ubicados de forma remota de una manera disciplinada en lo que respecta al flujo de información.
[32]	Microsoft TDSP	1. Comprensión empresarial 2. Adquisición y comprensión de datos 3. Modelado 4. Implementación 5. Aceptación del cliente	1. Esta metodología ayuda a mejorar la colaboración y el aprendizaje en equipo. 2. Esta metodología se preocupa por definir objetivos SMART (Específicos, medibles, alcanzables, relevante, con límite de tiempo). 3. Esta metodología aborda la debilidad de la falta de definición del equipo de CRISP-DM mediante la creación de roles y sus responsabilidades durante cada fase del ciclo de vida del proyecto.
[28]	Domino DS Lifecycle	1. Ideación 2. Adquisición y exploración de datos 3. Investigación y desarrollo 4. Validación 5. Entrega 6. Monitoreo	1. Esta metodología se basa en la metodología CRISP-DM y en el manifiesto ágil. 2. Esta metodología propone establecer un seguimiento para las entregas de la información y de los KPI comerciales. 3. Esta metodología hace uso de grupos de control en modelos de producción para realizar un seguimiento del desempeño del modelo y la creación de valor para la empresa.
[22]	Agile Delivery Framework	1. Alcance 2. Adquisición de datos 3. Análisis 4. Desarrollo de modelos 5. Validación 6. Implementación	1. Esta metodología está diseñada para fomentar la colaboración exitosa entre las empresas y las partes interesadas del proyecto. 2. Esta metodología separa completamente el mundo de la inteligencia empresarial y el del análisis de datos. De hecho, propone dos metodologías que evolucionan en paralelo y mediante métodos ágiles promete una colaboración eficaz entre estas dos partes.

[26]	Conceptual modeling with ML	1. Entendimiento del problema 2. Recopilación de datos 3. Ingeniería de datos 4. Entrenamiento del modelo 5. Optimización del modelo 6. Integración del modelo 7. Evaluación del modelo 8. Toma de decisiones analíticas	1. Esta metodología determina si los datos de entrenamiento de un modelo de ML son representativos del dominio. 2. Esta metodología permite que la toma de decisiones tenga un soporte solido en los datos para generar estrategias claves en la transformación digital. 3. Esta metodología considera que la unión del modelado conceptual con el ML contribuye en la mejora de la interpretabilidad de los algoritmos de aprendizaje automático mediante el uso de modelos conceptuales.
[16]	Data Science Edge (DSE)	1. Planificar 2. Recopilar 3. Seleccionar 4. Analizar 5. Actuar	1. Esta metodología se alinea con los nuevos cambios tecnológicos e implementa la agilidad en el ciclo de vida del análisis avanzado y el desarrollo de sistemas de ML. 2. Esta metodología permite la retroalimentación constante con los interesados del proyecto de ciencia de datos para validar el estado actual e influir en su evolución hacia un estado final que cumpla con los requerimientos y objetivos de un dominio específico. 3. Esta metodología proporciona una guía de las actividades que son fundamentales para generar de forma eficiente un producto mínimo viable para los interesados en un proyecto basado en ciencia de datos.
[51]	DECIDE	1. Identificación y motivación 2. Definición de los objetivos 3. Diseño y desarrollo 4. Demostración 5. Evaluación 6. Comunicación	1. Esta metodología se fundamenta en un diseño ágil basado en eventos y datos para proyectos decisionales de Big Data. 2. Esta metodología ayuda a las organizaciones a determinar los objetivos comerciales y analíticos deseados según la toma de decisiones con base al análisis de datos.
[41]	MIDANO	1. Conocimiento de la organización 2. Preparación y tratamiento de datos 3. Desarrollo de herramientas de MD	1. Esta metodología es utilizada para el desarrollo de aplicaciones de Minería de Datos (MD) basados en el análisis organizacional. 2. Esta metodología pretende abarcar el dominio de conocimiento que puede encontrarse en una organización e integrarlo una vista mineable operativa (VMO) y una vista mineable conceptual (VMC). 3. Esta metodología tiene como objetivo detallar las variables relevantes de diversos problemas de estudio, a partir de escenarios futuros definidos en el dominio de la organización.
[9]	POST-DS	1. Comprensión del negocio 2. Comprensión de los datos 3. Preparación de los datos 4. Modelado 5. Evaluación 6. Implementación	1. Esta metodología está basada en la metodología CRISP-DM, pero con la diferencia de que permite la identificación de procesos, organización, programación y herramientas para la gestión de proyectos de ciencia de datos a través de componentes específicos. 2. Esta metodología ejecuta cada una de las fases tradicionales de un proceso en ciencia de datos teniendo como eje cada una de las fases necesarias para la gestión de proyectos. 3. Esta metodología tiene como componentes claves el cumplimiento de actividades establecidas en un cronograma base generado a través del alcance y costos de un proyecto en ciencia de datos.

[58]	SANZU	1. Recopilación de datos 2. Manipulación de datos 3. Análisis estadístico 4. Retroalimentación 5. Toma de decisiones	1. Esta metodología sirve como punto de referencia para evaluar el rendimiento de las operaciones individuales que impactan en el análisis de datos. 2. Esta metodología permite representar casos de uso del mundo real para modelar las aplicaciones cuyos dominios ejecutan un flujo de trabajo específico.
[48]	SKI	1. Crear 2. Observar 3. Analizar	1. En esta metodología las etapas crear, observar y analizar garantizan que el trabajo que se requiere para la recopilación y el análisis de datos se incorpore directamente a las tareas de un equipo para una iteración de una tarea determinada. 2. Esta metodología en comparación con SCRUM, define una iteración centrada en la capacidad y no basada en el tiempo para brindarle a un equipo de análisis de datos la capacidad de ejecutar pequeñas iteraciones lógicas con una duración desconocida 3. Esta metodología proporciona una guía clara que permite a los equipos de análisis de datos aprovechar al máximo los beneficios de Kanban de manera más fácil y confiable.
[47]	KDD	1. Selección de datos 2. Pre-procesamiento de datos 3. Transformación de datos 4. Minera de datos 5. Evaluación/Interpretación	1. Esta metodología permite extracción de conocimiento oculto en una gran volumen de datos. 2. Esta metodología requiere un conocimiento previo relevante, una breve comprensión del dominio y la definición de los principales objetivos a cumplir con el análisis de datos.
[52]	SEMMA	1. Muestreo 2. Exploración 3. Modificación 4. Modelado 5. Evaluación	1. Esta metodología fue desarrollada por la compañía SAS para comprender, organizar, desarrollar y mantener proyectos de minería de datos. 2. Esta metodología proporciona las soluciones a los problemas basados en objetivos definidos en el ámbito empresarial dependiendo de su dominio.
[24]	Agile data science in healthcare	1. Definición de preguntas clínicas 2. Adquisición y validación de datos 3. Desarrollo del Modelo predictivo 4. Retroalimentación medica	1. Esta metodología fomenta el despliegue continuo de modelos predictivos en entornos clínicos durante el cual los científicos de datos pueden reunirse con los médicos y recibir comentarios sobre el rendimiento del modelo. 2. Esta metodología utiliza la perspicacia del médico unida a la ciencia de datos para determinar si los resultados del modelo son creíbles
[45]	IBM Foundational Methodology for Data Science	1. Comprensión del negocio 2. Enfoque analítico 3. Requisitos de datos 4. Recopilación de datos 5. Comprensión de datos 6. Preparación de datos 7. Modelado 8. Evaluación 9. Implementación 10. Retroalimentación	1. Esta metodología tiene algunas similitudes con las metodologías reconocidas para la minería de datos, pero pone el énfasis en varias de las nuevas prácticas en la ciencia de datos. 2. Esta metodología los modelos se mejoran y se adaptan constantemente a las condiciones cambiantes a través de retroalimentación, ajustes y reimplementaciones. De esta manera, tanto el modelo como su trabajo pueden proporcionar un valor continuo a la organización mientras la solución sea necesaria.

[10]	DataOps Manifesto	1. Ideación 2. Iniciación 3. investigación/desarrollo 4. Transición/producción 5. Retirada	1. En esta metodología las etapas son adaptables al contexto y el dominio de la organización para ayudar a los equipos de análisis de datos a ser más colaborativos en los ciclos de retroalimentación para lograr resultados más eficaces. 2. Esta metodología se basa en la experiencia laboral de varias organizaciones y los problemas tradicionales al momento de entregar tiempos de ciclo rápidos para una alta gama de análisis de datos con un manifiesto de calidad válido.
[8]	AABA	1. Catálogo de diseño conceptual 2. Arquitectura BDD 3. Implementación 4. Pruebas 5. Despliegue/retroalimentación 6. Descubrimiento del valor	1. Esta metodología ofrece una heurística que permite que la arquitectura sea exhaustiva para todas las partes interesadas , en donde el equipo de ingeniería se centre en las tareas importantes, como la validación del valor y la anticipación del cambio. 2. Esta metodología abarca aspectos importantes como: la planificación, la estimación, el coste , el calendario, el apoyo a la experimentación y el uso de la metodología DevOps para la entrega rápida y continua de un producto de valor.
[25]	KDDA	1. Formulación del problema 2. Comprensión empresarial 3. Comprensión de datos 4. Preparación de datos 5. Modelado 6. Evaluación 7. Despliegue 8. Mantenimiento	1. Esta metodología estructura su proceso en el diseño de la concha de caracol y asimila tanto las bases de conocimientos de KDD existentes como la experiencia analítica del mundo real de los investigadores. 2. Esta metodología enmarca las diferencias entre los proyectos tradicionales de minería de datos en un entorno de toma de decisiones impulsado por big data identificando los pasos que faltan en la metodología KDD. 3. Esta metodología hereda la representación del ciclo de vida del proyecto de la metodología CRISP-DM con la diferencia de que no existen secuencias estrictamente definidas entre las fases, por lo que cada fase incluye diferentes tareas, y el resultado de cada tarea determina la fase o tareas particulares a realizar en una fase determinada.

eran, que se puede utilizar un punto de referencia (benchmark) para evaluar la funcionalidad y el rendimiento de un sistema. Sin embargo afirman, que no existe un punto de referencia estándar para evaluar la capacidad de estos sistemas para hacer ciencia de datos, aunque existan categorías de benchmark que están asociadas, tales como: el benchmark de bases de datos relacionales, el benchmark de sistemas de Big data y el benchmark de análisis específicos del dominio. Por esta razón, los autores presentan una metodología denominada *Sanzu*, la cual permite evaluar sistemas que ejecuten tareas de procesamiento y análisis de datos. Esta metodología se basa en el flujo de trabajo tradicional de las ciencias de datos, el cual se compone de las siguientes fases: recopilación de datos, manipulación de datos, análisis estadístico, retroalimentación de resultados y toma de decisiones. Adicionalmente, esta metodología se compone de dos tipos de enfoques: el enfoque *micro-benchmark* que consta de seis fases: entrada y salida de archivos planos, mantenimiento de datos, estadísticas descriptivas, estadísticas de distribución e inferenciales, análisis de series de tiempo y ML. Este enfoque esta destinado a probar las tareas básicas en múltiples sistemas de análisis de datos. Dichas tareas son comunes y se utilizan todos los días para resolver problemas del mundo real y de la industria. Por otra parte, el enfoque *macro-benchmark* tiene como objetivo modelar las aplicaciones cuyos dominios ejecutan el flujo de trabajo de la ciencia de datos debido al crecimiento del volumen de datos y evalúa el desempeño de los sistemas de datos de cada una. En conclusión, la metodología Sanzu sirve como punto de referencia en ciencia de datos para evaluar el rendimiento de las operaciones individuales que impactan en el análisis de datos y para representar casos de uso del mundo real.

[32] propone una metodología llamada TDSP[av], la cuál permite ejecutar soluciones ágiles e iterativas en el análisis predictivo y el desarrollo de aplicaciones inteligentes de manera eficiente. Esta metodología incluye las mejores prácticas y estructuras de Microsoft y otros líderes de la industria para ayudar a lograr una implementación exitosa de iniciativas de ciencia de datos teniendo como base el trabajo en equipo y los criterios relevantes del manifiesto ágil. El objetivo es ayudar a que las empresas aprovechen al máximo los beneficios del análisis de datos. Dicho lo anterior, la metodología TDSP proporciona un ciclo de vida que se compone de cinco etapas principales que se ejecutan de forma iterativa: comprensión del negocio, adquisición y compresión de datos, modelado, despliegue y aceptación de cliente. Hay que mencionar, que TDSP esta sujeta al uso de la herramienta Microsoft Azure DevOps para realizar todo el proceso de ciencia de datos desde el entendimiento del negocio hasta la entrega de un producto de valor a un cliente especifico. En conclusión, el ciclo de vida de TDSP fue diseñado para proyectos destinados a generar aplicaciones inteligentes en donde el análisis predictivo se realiza a través de modelos de ML e AI. Sin embargo, puede ser utilizado para llevar a cabo proyectos en donde el propósito es tomar decisiones según el análisis obtenido en el reconocimiento de patrones de un conjunto de datos específicos.

[48] proponen un marco ágil para la ciencia de datos denominado SKI[aw]. Esta metodología adopta la filosofía de tableros Kanban para proporcionar un proceso de iteración estructurado para que los equipos de análisis de datos exploren y aprendan de manera incremental a través de pruebas de hipótesis.En general, en esta metodología la agilidad esta basada en una secuencia de ciclos iterativos de experimentación y adaptación a través de la implementación y el análisis de los resultados. Los equipos de SKI usan un tablero visual y se enfocan en trabajar en un elemento específico durante una iteración, que se basa en tareas, no en bloques de tiempo. Por lo tanto, una iteración se alinea más estrechamente con el concepto de extraer tareas de manera prioritaria, cuando el equipo tiene capacidad. Cada iteración puede verse como la validación o el rechazo de una hipótesis específica. A nivel general, una iteración se compone de tres etapas principales: Crear, Observar y Analizar. Las etapas anteriores se enfocan en garantizar que el trabajo que se requiere para la recopilación y el análisis de datos se incorpora directamente a las tareas del equipo para una iteración determinada. En conclusión, en comparación con Scrum, SKI define una iteración centrada en la capacidad y no en el tiempo, para brindarle a un equipo de análisis de datos la ejecución de pequeñas iteraciones lógicas con una duración desconocida.

[24] proponen una metodología que realiza la unión de Scrum y Kanban (Scrumban) para la investigación ágil en el cuidado de la salud. La metodología propuesta consta de cuatro fases: Definición de preguntas clínicas, adquisición y validación de datos, desarrollo del modelo predictivo y retroalimentación medica. Los autores consideran que para que se cumpla el enfoque ágil en esta metodología debe existir una colaboración continua entre científicos de datos y médicos. Ademas, debe existir almacenamiento y computación basados en la nube para proporcionar una plataforma común para acceder a los datos y modelos. Los problemas complejos se pueden dividir en tareas que pueden visualizar tanto los científicos de datos como los médicos, lo que les permite comprender mejor el trabajo que los científicos de datos deben realizar dentro de cada ciclo del Sprint. La metodología fomenta el despliegue continuo de modelos predictivos en entornos clínicos durante el cual los científicos de datos pueden reunirse con los médicos y recibir comentarios sobre el rendimiento del modelo. Ademas, la perspicacia del médico se puede aprovechar para determinar si los resultados del modelo son creíbles. En conclusion, esta metodología prescribe un proceso rápido y de mejora continua que permite a los médicos comprender el trabajo de los científicos de datos y evaluar regularmente el rendimiento de un modelo predictivo en entornos clínicos.

[45] propone la metodología IBM Foundational Methodology for Data Science, la cual tiene algunas similitudes con las metodologías mas utilizadas en minería de datos, con la diferencia que se enfatiza en las prácticas mas recientes en la ciencia de datos, como el uso de grandes volúmenes de datos, la incorporación de la analítica de texto en el modelado predictivo y la automatización de procesos. La metodología consta de

[av] Team Data Science Process

[aw] Structured Kanban Iteration

diez etapas: comprensión del negocio, enfoque analítico, requisitos de datos, recopilación de datos, comprensión de datos, preparación de datos, modelado, evaluación, implementación y retroalimentación. Estas etapas forman un proceso iterativo para el uso de datos con el propósito descubrir conocimiento oculto. En conclusion, esta metodología ilustra la naturaleza iterativa del proceso de resolución de problemas en donde, los científicos de datos vuelven frecuentemente a etapas previas para realizar ajustes a medida que van aprendiendo de los datos y el modelado.

[10] realizaron una revisión de diversas metodologías utilizadas en análisis de datos, en donde resaltan el ciclo de vida de la ciencia de datos basado en el DataOps Manifesto. Esta metodología ayuda a caracterizar qué prácticas ágiles, eventos, artefactos y roles son valiosos para agregar valor a los datos de la organización diariamente. Esta metodología se compone de cinco etapas generales: ideación, iniciación, investigación/desarrollo, transición/producción y Retirada. Estas etapas ayudan a los equipos de análisis de datos a ser más adaptables y colaborativos en los ciclos de retroalimentación para lograr resultados de manera eficiente. En conclusión, esta metodología se basa en la experiencia laboral de varias organizaciones y los problemas tradicionales al momento de entregar tiempos de ciclo rápidos para una alta gama de análisis de datos con un manifiesto de calidad válido.

[8] proponen una metodología denominada AABA[ax] para el análisis ágil de Big data centrado en la arquitectura. Esta metodología consta de seis etapas: catálogo de diseño conceptual, arquitectura BDD[ay], implementación, pruebas, despliegue/retroalimentación y descubrimiento del valor. Esta metodología se basa en el desempeño de la arquitectura de software como factor clave de agilidad. AABA proporciona una base para el descubrimiento de valor en los datos con las partes interesadas y abarca aspectos importantes como: la planificación, la estimación, el coste , el calendario, el apoyo a la experimentación y el uso de la metodología DevOps para la entrega rápida y continua de un producto de valor. En conclusión, esta metodología se centra con la colaboración estrecha entre los científicos de datos, las partes interesadas, el arquitecto de software y otros ingenieros clave como los diseñadores de bases de datos, esto con el propósito de determinar la propuesta de valor para el sistema que se está construyendo en función de de la mejora continua y el cumplimiento de los objetivos comerciales.

[25] proponen una mejora para la metodología KDD denominada KDDA[az]. Esta metodología se basa en el descubrimiento de conocimiento a través del análisis de datos y no de la minería de datos. Para lograrlo consta de un proceso de concha de caracol, el cual se compone de ochos fases: formulación del problema, comprensión empresarial, comprensión de datos, preparación de datos, modelado, evaluación ,despliegue y mantenimiento. Esta metodología hereda la representación del ciclo de vida de la metodología CRISP-DM, con la diferencia de que no tiene secuencias estrictamente definidas entre las fases. Cada fase incluye diferentes tareas, y el resultado de cada tarea determina la fase o tareas particulares de una fase a realizar, lo que permite que un resultado específico de la fase de modelado puede requerir volver a la comprensión del negocio, la comprensión de los datos, la preparación de los datos o ir directamente a la evaluación. En conclusión, esta metodología utiliza las prácticas de descubrimiento de conocimiento en el entorno analítico de una organización abarcando no solo conocimientos técnicos de TI, técnicas analíticas y algoritmos matemáticos, sino también una comprensión profunda del proceso empresarial.

[ax] Architecture-centric Agile Big data Analytics
[ay] Big Data system Design
[az] Knowledge Discovery via Data Analytics

6. Discusión

Según los resultados obtenidos, se infiere que en la información revisada no existe una metodología en ciencia de datos definida para el diagnóstico del cáncer de mama. En efecto, la mayoría de la literatura científica apunta directamente al uso de técnicas de ML y DL para el diagnóstico o pronostico del cáncer de mama exponiendo el nivel de precisión, cantidad de falsos positivos, gasto computacional y modelos algorítmicos utilizados para determinar el posible padecimiento de esta enfermedad. Y aunque estas investigaciones, brindan información valiosa para mejorar la precisión, sensibilidad y especificidad de las técnicas de ML y DL en el diagnóstico del cáncer, carecen de una metodología clara en donde la idea principal gire entorno de la comprensión del dominio y la toma de decisiones por parte de los oncólogos. En particular, la mayoría de investigaciones llegan a resultados en términos de precisión y exactitud, pero no profundizan en el valor real que el especialista en oncología atribuye a los datos para tomar una decisión y el impacto que dicha decisión tiene en la usabilidad del modelo generado, a sabiendas que el experto a través de su perspicacia medica es quien finalmente evalúa si los resultados obtenidos por los algoritmos son veraces y permiten diagnosticar el cáncer de forma ágil, generando un valor agregado que cumpla con los objetivos de las ciencias de la salud. Por consiguiente, aunque la comunidad de investigación en ciencia de datos este en crecimiento constante, esté explorando nuevos dominios, creando nuevos roles especializados y este realizando un gran esfuerzo de investigación para desarrollar análisis avanzados, mejorar modelos de datos y generar nuevos algoritmos apoyados de los campos de las matemáticas, la estadística y la informática, estas habilidades no son suficientes para la aplicación de la ciencia de datos en proyectos reales [28], 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 los roles. Dicho lo anterior, aunque las metodologías seleccionadas en la revisión sistemática no giren en torno al cáncer, proponen etapas que abarcan aspectos relevantes en la organización, a nivel de planteamiento y ejecución de un proyecto en ciencia de datos que tiene como eje el dominio basado en la prevención, el diagnostico y el tratamiento del cáncer de mama. Cabe resaltar que las metodologías CRISP-DM y KDD fueron selec-

cionadas debido a que comprenden todas las fases básicas que debe cumplir cualquier proyecto que tenga en su contexto el análisis de datos, ademas de que son bastante utilizadas por una cantidad considerable de investigadores, sin embargo aunque sean las metodologías más utilizadas, carecen de lineamientos que profundicen en la organización de los equipos de trabajo para llevar a cabo procesos de gestión que se alienen con el software y las metodologías de desarrollo ágiles. Ademas, autores como [28] sugieren que para ofrecer una solución integral para la ejecución exitosa de una metodología en ciencia de datos se deben cubrir estrictamente tres áreas: gestión de proyectos, gestión de equipos y gestión de datos e información. Por consiguiente, tras el esfuerzo por integrar los resultados analizados en este trabajo, parece bastante plausible afirmar que no existe una metodología puntual en ciencia de datos que se enfoque en el diagnóstico del cáncer de mama, sin embargo, diversos autores proponen metodologías que permiten la aplicación de esta ciencia en el dominio de las ciencias de la salud, teniendo como eje principal la comprensión de aspectos como: la comunicación constante con los interesados, el uso del enfoque ágil, la definición de roles y funciones, el valor agregado de los datos, la retroalimentación continua y la aceptación del experto en oncología para la posterior toma de decisiones que ayude a reducir de manera eficaz la morbilidad causada por esta enfermedad.

Referencias

[1] 2020, Nature Reviews Cancer, 1, 137

[2] Alakwaa, F. M., Chaudhary, K., & Garmire, L. X. 2018, Journal of Proteome Research, 17, 337

[3] B., T. A., O'Malley, F. P., Singhal, H., & Tonkin, K. S. 1997, Arch Pathol Lab Med .

[4] Baker, D. J. 2018, Artificial Intelligence: The Future Landscape of Genomic Medical Diagnosis: Dataset, In Silico Artificial Intelligent Clinical Information, and Machine Learning Systems (Elsevier Inc.), 223–267, doi:10.1016/B978-0-12-809414-3.00011-5

[5] Bland, K. I., & Copeland, E. M. 2009, The Breast : comprehensive management of benign and malignant diseases (Saunders/Elsevier), 62

[6] Brunicardi, C. F. 2010, Schwartz- Principios de Cirugía, novena edn., ed. D. K. Andersen, T. R. Billiar, D. L. Dunn, J. G. Hunter, J. B. Matthews, & R. E. Pollock, Vol. 9 (McGRAW-HILL INTERAMERICANA EDITORES, S. A. de C. V), 424–469

[7] Chaudhury, A. R., Iyer, R., Iychettira, K. K., & Sreedevi, A. 2011, ICIIP 2011 - Proceedings: 2011 International Conference on Image Information Processing, doi:10.1109/ICIIP.2011.6108877

[8] Chen, H. M., Kazman, R., & Haziyev, S. 2016, Proceedings of the Annual Hawaii International Conference on System Sciences, 2016-March, 5378

[9] Costa, C. J., & Aparicio, J. T. 2020, Iberian Conference on Information Systems and Technologies, CISTI, 2020-June, 24

[10] Dastgerdi, A. K., & Gandomani, T. J. 2021, 2021 International Conference on Information Technology, ICIT 2021 - Proceedings, 667

[11] Duarte, C., Salazar, A., Strasser-Weippl, K., et al. 2021, Breast Cancer Research and Treatment, 186, 15

[12] Ebrahimi, M. 2019, Encyclopedia of Biomedical Engineering, 1-3, 501

[13] Elprin, N., Yang, C., Gleason, T., & Tacelli, K. 2022, Domino Data Lab

[14] Fatima, N., Liu, L., Hong, S., & Ahmed, H. 2020, IEEE Access, 8, 150360

[15] Gradilone, A., Naso, G., Raimondi, C., et al. 2011, Annals of Oncology, 22, 86

[16] Grady, N. W., Payne, J. A., & Parker, H. 2017, Proceedings - 2017 IEEE International Conference on Big Data, Big Data 2017, 2018-January, 2331

[17] Hasan, M. K., & Ara, S. R. 2019, Encyclopedia of Biomedical Engineering, 1-3, 331

[18] Hirose, M., Nobusawa, H., & Gokan, T. 2007, Radiographics, 27, doi:10.1148/RG.27SI075501

[19] Hou, R., Mazurowski, M. A., Grimm, L. J., et al. 2020, IEEE Transactions on Biomedical Engineering, 67, 1565

[20] Hu, Q., Whitney, H. M., & Giger, M. L. 2020, Nature Scientific Reports, 10, 1

[21] International Agency for Research on Cancer. 2020, Globocan 2020, 509, 1

[22] Larson, D., & Chang, V. 2016, International Journal of Information Management, 36, 700

[23] Lee, B., Kim, K., Choi, J. Y., et al. 2017, Medicine (United States), 96, doi:10.1097/MD.0000000000008089

[24] Lei, H., O'Connell, R., Ehwerhemuepha, L., et al. 2020, Intelligence-Based Medicine, 3-4, doi:10.1016/j.ibmed.2020.100009

[25] Li, Y., Thomas, M. A., & Osei-Bryson, K. M. 2016, Decision Support Systems, 91, 1

[26] Maass, W., & Storey, V. C. 2021, Data and Knowledge Engineering, 134, 1

[27] Mann, R. M., Hooley, R., Barr, R. G., & Moy, L. 2020, Radiology, 297, 266

[28] Martinez, I., Viles, E., & G. Olaizola, I. 2021, Big Data Research, 24, 100183

[29] Masciari, S., Larsson, N., Senz, J., et al. 2007, Journal of Medical Genetics, 44, 726

[30] McKinney, S. M., Sieniek, M., Godbole, V., et al. 2020, Nature, 577, 89

[31] Memis, A., Ozdemir, N., Parildar, M., Ustun, E. E., & Erhan, Y. 2000, European Journal of Radiology, 35, 39

[32] Microsoft. 2022, Azure Architecture Center, 955

[33] Mladenic, D., Lavrač, N., Bohanec, M., & Moyle, S. 2012, Data Mining and Decision Support: Integration and Collaboration, The Springer International Series in Engineering and Computer Science (Springer US)

[34] Moher, D., Liberati, A., Tetzlaff, J., et al. 2009, PLoS Medicine, 6, doi:10.1371/journal.pmed.1000097

[35] Mulholland, M., Lillemoe, K., Doherty, G., et al. 2012, Greenfield's Surgery: Scientific Principles & Practice (Wolters Kluwer Health)

[36] Mullooly, M., Ehteshami Bejnordi, B., Pfeiffer, R. M., et al. 2019, Nature npj Breast Cancer, 5, doi:10.1038/s41523-019-0134-6

[37] Naik, N., Madani, A., Esteva, A., et al. 2020, Nature Communications, 11, doi:10.1038/s41467-020-19334-3

[38] Nateghi, R., Danyali, H., & Helfroush, M. S. 2021, Artificial Intelligence in Medicine, 114, 102048

[39] Negahbani, F., Sabzi, R., Pakniyat Jahromi, B., et al. 2021, Nature Scientific Reports, 11, 1

[40] Obeng-Gyasi, S., Grimm, L. J., Hwang, E. S., Klimberg, V. S., & Bland, K. I. 2018, The Breast: Comprehensive Management of Benign and Malignant Diseases, 377

[41] Pacheco, F., Rangel, C., Aguilar, J., Cerrada, M., & Altamiranda, J. 2014

[42] Page, D. L., Dupont, W. D., Rogers, L. W., & Landenberger, M. 1982, American Cancer Society, doi:10.1002/1097-0142

[43] Pillai, N. 2020, in MinTIC- (Bogota: Correlation One)

[44] Robertson, F. M., Bondy, M., Yang, W., et al. 2010, CA: A Cancer Journal for Clinicians, 60, 351

[45] Rollins, J. 2015, IBM Analytics, 1

[46] Roszkowiak, L., Korzynska, A., Siemion, K., et al. 2021, Nature Scientific Reports, 1

[47] Safhi, H. M., Frikh, B., & Ouhbi, B. 2019, in (Elsevier B.V.), 30–36

[48] Saltz, J., & Suthrland, A. 2019, IEEE

[49] Sauer, A. G., Jemal, A., Siegel, R. L., & Miller, K. D. 2019, Breast Cancer Facts and Figures 2019-2020

[50] Schröer, C., Kruse, F., & Gómez, J. M. 2021, Procedia Computer Science, 181, 526

[51] Sfaxi, L., & Ben Aissa, M. M. 2020, Data and Knowledge Engineering, 130, 101862

[52] Shafique, U., & Qaiser, H. 2014, A Comparative Study of Data Mining Process Models (KDD, CRISP-DM and SEMMA)

[53] Shen, L., Margolies, L. R., Rothstein, J. H., et al. 2019, Nature Scientific Reports, 1

[54] Shia, W. C., Lin, L. S., & Chen, D. R. 2021, Nature Scientific Reports, 11, 1

[55] Sun, Y. S., Zhao, Z., Yang, Z. N., et al. 2017, International Journal of Biological Sciences, 13, 1387

[56] Tamam, N., Salah, H., Rabbaa, M., et al. 2021, Radiation Physics and Chemistry, 188, doi:10.1016/J.RADPHYSCHEM.2021.109680

[57] Tse, G. M., Yeung, D. K., & Chu, W. C. 2014, Comprehensive Biomedical Physics, 3, 205

[58] Watson, A., Babu, D. S. V., & Ray, S. 2017, Proceedings - 2017 IEEE International Conference on Big Data, Big Data 2017, 2018-Janua, 263

[59] Xu, Z., Tang, N., Xu, C., & Cheng, X. 2021, Data Science and Management, 1, 32