EDUCACIÓN
SECRETARÍA DE EDUCACIÓN PÚBLICA

TECNOLÓGICO NACIONAL DE MÉXICO

Tecnológico Nacional de México

Centro Nacional de Investigación y Desarrollo Tecnológico

Tesis de Maestría

Revisión sistemática de publicaciones de ciencia de datos

presentada por
Jorge Francisco Ruiz Lopez

como requisito para la obtención del grado de
Maestro en Ciencias de la Computación

Director de tesis
Dr. Joaquín Pérez Ortega

Codirector de tesis
Dr. Javier Ortiz Hernández

Cuernavaca, Morelos, México. Julio de 2021.

cenidet
Centro Nacional de Investigación y Desarrollo Tecnológico

Oficio de aceptación de documento de tesis

Dedicatoria

A Jesucristo,
por su gracia y sabiduría
para enfrentar los retos de este proceso

A mi mamá Gina Lopez,
por su consejo y apoyo para realizar mis estudios.

A mi papá Francisco Ruiz,
por su apoyo incondicional en la vida.

A la memoria de mis abuelos, Juanita y Francisco,
por sus consejos y ejemplos para mi vida.

A toda mi familia,
en especial a mis tios Lupita y Jaime,
por su comprensión y consejo en diferentes etapas de mi vida.

A mi novia Sylvie,
por tu apoyo y ayuda en esta etapa.

Agradecimientos

Agradezco a Dios por la vida y por la salud. Por la oportunidad de alcanzar una meta más. A CONACYT por el apoyo económico que hizo posible la realización de la maestría. Al Tecnológico Nacional de México y al Centro Nacional de Investigación y Desarrollo Tecnológico, por ser las instituciones que hicieron posible este logro.

Al Dr. Joaquín Pérez Ortega por ser el director de esta tesis y por su apoyo durante este proceso de formación profesional y guiarme a lo largo de mis estudios de maestría.

Al Dr. Javier Ortíz Hernández por ser el codirector de esta tesis y por su tiempo y disposición para contribuir con la presente investigación.

Al Dr. José Crispín Zavala Díaz y la Dra. María Yasmín Hernández Pérez por ser mi comité revisor y por su disposición y aportaciones a la presente investigación.

A mis padres, José Francisco Ruiz Lopez y Cornelia Georgina Lopez Ortíz por su apoyo en la vida y en este proceso de formación.

A mi tío, Dr. Jaime Arau Roffiel por su disposición y consejo en este proceso de formación profesional.

A mi novia, Sylvie Caroy por su ayuda, motivación y compañía en esta etapa de mi vida…

A mis compañeros de generación y laboratorio, Carlos Moreno Calderón y Dianely Aparicio García, Erick Infante Covarrubias.

Resumen

En la presente tesis, se realizó una la revisión sistemática de metodologías de Ciencia de Datos. Para realizar la revisión sistemática, se utilizó la metodología que fue adaptada por la Dra. Bárbara Kichenham a las ciencias computacionales, la cual consta de 3 fases: 1) fase de planeación, 2) fase conductora y 3) elaboración del reporte. Como resultado de la aplicación de la metodología de revisión sistemática propuesta, se recuperaron 3,451 publicaciones, sobre las cuales se aplicaron criterios de inclusión y exclusión, quedando un total de 24 publicaciones. De los 24 artículos se identificaron: 7 metodologías propuestas por la comunidad científica y 2 propuestas por la industria; 4 metodologías para gestión de proyectos de Ciencia de Datos; 2 metodologías de flujos de trabajo de Ciencia de Datos y, por último, 2 metodologías de investigación de Ciencia de Datos.

Asimismo, se realizó un análisis mixto de los resultados de la revisión sistemática teniendo, por un lado, el análisis cuantitativo y por el otro, el análisis cualitativo. En el análisis cuantitativo, se presentan las publicaciones por año, publicaciones por buscador, publicaciones por repositorio, publicaciones más citadas, palabras clave y cadenas de búsqueda. En el análisis cualitativo, se presentan las palabras clave de publicaciones, panorama general de metodologías de Ciencia de Datos. Asimismo, dentro de los resultados se presenta la clasificación de metodologías de Ciencia de Datos y las perspectivas de dicha temática.

En la presente investigación se proponen dos clasificaciones: en la primera, se clasifican las publicaciones relevantes identificadas de acuerdo con su enfoque y en la segunda, se crea otra clasificación con cuatro categorías para clasificar las metodologías de Ciencia de Datos (metodologías para el proceso de los datos, metodologías de gestión de proyectos, metodologías de investigación y metodologías de flujo de trabajo) de acuerdo con su objetivo.

Abstract

In this thesis, a systematic review of Data Science methodologies was carried out. To carry out the systematic review, the methodology adapted by Dr. Barbara Kichenham to computational sciences was used, which consists of 3 phases: 1) planning phase, 2) conducting phase and 3) report elaboration. As a result of the application of the proposed systematic review methodology, 3,451 publications were retrieved, to which inclusion and exclusion criteria were applied, leaving a total of 24 publications. Of the 24 articles, the following were identified: 7 methodologies proposed by the scientific community and 2 proposed by the industry; 4 methodologies for Data Science project management; 2 methodologies for Data Science workflows and, finally, 2 methodologies for Data Science research.

Likewise, a mixed analysis of the results of the systematic review was performed, having, on the one hand, the quantitative analysis and, on the other hand, the qualitative analysis. In the quantitative analysis, publications by year, publications by search engine, publications by repository, most cited publications, keywords and search strings are presented. In the qualitative analysis, the key words of publications, general overview of Data Science methodologies are presented. Also, within the results, the classification of Data Science methodologies and the perspectives of this subject are presented.

In this research, two classifications are proposed: in the first one, the relevant publications identified are classified according to their focus and in the second one, another classification is created with four categories to classify Data Science methodologies (data processing methodologies, project management methodologies, research methodologies and workflow methodologies) according to their objective.

Tabla de contenido Página

Lista de figuras

Lista de tablas

Capítulo 1

Introducción

1 Introducción

Este capítulo introductorio está organizado de la siguiente manera, en la Sección 1.1; se describe el contexto en el que se encuentra enmarcado la investigación; en la Sección 1.2; se presentan las bases y justificación del problema que se pretende resolver y que es la principal motivación de la presente investigación; las secciones 1.3 y 1.4, describen los objetivos general y específicos que fueron planteados en este trabajo para atender la problemática a resolver en esta investigación; en la Sección 1.5; se describen los alcances y limitaciones de la tesis a partir de las facilidades con las que se contó para este estudio; en la Sección 1.6 se muestra la organización del documento.

1.1. Contexto de la investigación

Actualmente, la Ciencia de Datos es emergente lo cual ha provocado un interés destacado de la comunidad científica dedicada al análisis de datos. La Ciencia de Datos, surge en 1967 [1], como una propuesta de evolución a partir del análisis de datos, debido a que esta cumplía con los criterios para ser considerada como ciencia. Dando origen a la Ciencia de Datos.

Día con día, se ha incrementado de manera significativa la cantidad de publicaciones sobre dicha ciencia, ya sea, porque es una herramienta útil para la toma de decisiones o simplemente porque la comunidad científica y la industria identificaron una ciencia contemporánea capaz de cambiar la forma en que vemos el mundo. Algunos de los campos del sector empresarial en los cuales se utiliza la Ciencia de Datos son: finanzas, marketing, recursos humanos, desarrollo tecnológico e inteligencia de negocios entre otros. Por lo anterior, en los últimos años se ha incrementado exponencialmente la cantidad de publicaciones sobre Ciencia de Datos.

En el Departamento de Ciencias Computacionales del Centro Nacional de Investigación y Desarrollo Tecnológico (CENIDET), , específicamente dentro del grupo de investigación sobre Cómputo Inteligente, desde 2018 se inició un primer trabajo de investigación en el campo de la Ciencia de Datos con la tesis de maestría denominada: “Desarrollo de una aplicación de la Ciencia de Datos” [2].

1.2. Descripción del problema

Desde su creación en 1972, la Ciencia de Datos ha sido ampliamente utilizada en diferentes campos del conocimiento debido a que profundiza más que otras disciplinas para analizar los datos con los cuales se interactúa. Los sectores empresariales en los cuales se utiliza la Ciencia de Datos son: finanzas, marketing, recursos humanos, desarrollo tecnológico e inteligencia de negocios entre otros. Por lo anterior, en los últimos años se ha incrementado exponencialmente la cantidad de publicaciones sobre Ciencia de Datos.

Sin embargo, se identificó que dicha literatura no está suficientemente clasificada y organizada para facilitar su consulta. Ya que, la búsqueda e identificación de publicaciones relevantes sobre metodologías de Ciencia de Datos requiere de una gran inversión de tiempo para los investigadores y personas interesadas en ampliar su conocimiento.

Como sucede con frecuencia, la búsqueda de publicaciones relevantes sobre un tópico específico, no es una tarea fácil, y se dificulta más aún cuando dichas publicaciones no están clasificadas adecuadamente. De esta manera, analizar el estado del arte mediante el uso de una metodología facilitará la investigación. En esta tesis, se aborda el problema de seleccionar y clasificar las publicaciones relevantes sobre metodologías de Ciencia de Datos reportadas en la literatura.

1.3. Objetivo general

Aplicar una revisión sistemática de metodologías de Ciencia de Datos con base en las publicaciones reportadas en la literatura sobre dicha temática, que permita seleccionar y clasificar las metodologías de acuerdo con su enfoque.

1.4. Objetivos específicos

1.- Identificar las publicaciones de metodologías de Ciencia de Datos.

2.- Caracterizar las publicaciones de metodologías de Ciencia de Datos

3.- Clasificar las publicaciones relevantes de metodologías de Ciencia de Datos.

1.5. Alcances y limitaciones de la investigación

Los alcances son los siguientes:

- La revisión sistemática será aplicada a las metodologías de Ciencia de Datos publicadas en los repositorios científicos.
- El presente trabajo no solo se limita a identificar las publicaciones relevantes, sino que también propone dos clasificaciones una de las publicaciones identificadas y otra de las metodologías de Ciencia de Datos.

Las limitaciones son las siguientes:

- Para propósitos del estudio solo se recopilará el material bibliográfico de las bases de datos a las cuales tenga acceso el CENIDET, los artículos que ya se tienen en el grupo de investigación, así mismo, se buscará acceso a bases de datos de otras instituciones.
- Las búsquedas se realizarán en las bases de datos a las cuales el CENIDET, tiene acceso.

1.6. Organización del documento

Esta tesis consta de 7 Capítulos, en el Capítulo 2, se elabora una introducción teórica en relación con el campo de Ciencia de Datos; en el Capítulo 3, se describen los trabajos relacionados que fueron identificados para la presente investigación; en el Capítulo 4, se describe la propuesta de solución al problema abordado en esta tesis; en el Capítulo 5, se describen las actividades realizadas para desarrollar la revisión sistemática; en el capítulo 6, se describen los resultados de la revisión sistemática y por último, en el Capítulo 7, se describen las conclusiones de la tesis y los trabajos futuros planteados.

Capítulo 2

Marco Teórico

2 Marco teórico

En este capítulo, se explicarán algunos de los conceptos esenciales que permitirán la mejor comprensión de la terminología empleada en la presente tesis, para lo cual en la Sección 2.1, se presenta el concepto de Ciencia de Datos; en la Sección 2.2, se describe el concepto de metodología de Ciencia de Datos; en la sección 2.3, se describe el concepto de revisión sistemática y por último, en la sección 2.4, se describe la metodología empleada en esta investigación.

2.1. Ciencia de Datos

2.1.1 Definición de Ciencia de Datos

En [3], se menciona que "*La Ciencia de Datos es un nuevo campo interdisciplinario que sintetiza y se basa en estadísticas, informática, computación, comunicación, gestión y sociología para estudiar los datos y sus entornos (dominios y otros aspectos contextuales, como los aspectos organizativos y sociales) para transformar los datos en percepciones y decisiones siguiendo un pensamiento y una metodología"*, lo que nos hace reflexionar sobre cómo la Ciencia de Datos, seguirá añadiendo cada vez más áreas que ayuden a tomar mejores decisiones basadas en datos y poder construir una ciencia más completa y ágil que considere diferentes perspectivas para su aplicación.

De manera general, la Ciencia de Datos, surge como un proceso para identificar patrones en los datos que ayuden a la toma de decisiones, sin embargo, la Ciencia de Datos necesita de una metodología que garantice la reproducibilidad del proceso.

2.1.2 Historia de la Ciencia de Datos

De acuerdo con la literatura revisada, se identificó que la primera aproximación a este campo fue desarrollada por Turkey en 1962, la cual surgió como una propuesta con el propósito de evolucionar el análisis de datos. En esta propuesta, se justificaba la razón por la cual el análisis de datos debía ser considerado como una ciencia, ya que el autor demostró que cumplía con todos los requisitos para ser considerada como tal. En 1974, 12 años después Peter Neur propone utilizar el término "*Data Science*", como una nueva ciencia emergente. En 1977, se funda la Asociación Internacional de Computación Estadística

teniendo como objetivo fusionar la estadística, la informática moderna y el conocimiento de los expertos para convertir los grandes volúmenes de información en conocimiento. En 1989, Gregory Piatetsky organiza el primer taller de descubrimiento de conocimiento en bases de datos (KDD por sus siglas en inglés). En 1996, se emplea por primera vez el término "Ciencia de Datos" en el título de una conferencia (*Ciencia de Datos, clasificación y métodos relacionados*).

En 1999, Jacob Zahavi, plantea la problemática de las grandes bases de datos con altos volúmenes de información y las limitaciones que la minería de datos supone. En 2001, William S. Cleveland publica "*Ciencia de Datos: un plan de acción para expandir las áreas técnicas del campo de la estadística*", publicación que plantea un plan para ampliar las principales áreas del trabajo técnico de la estadística clásica. En 2002, se lanza el "*Data Science Journal*", revista que está enfocada en publicar artículos sobre gestión de datos, minería de datos, ciencia y tecnología.

En 2007, se establece el Centro de Investigacion de Datalogy y Data Science (CIDD) en China. En 2009, Mike Driscoll publica "*The three sexy skills of data geeks*", mencionando que la era de los datos ha llegado y aquellos que tengan las capacidades para modelar, combinar y presentar adecuadamente los datos serán llamados estadísticos de datos. En 2012, Tom Daverport y D.J Patil publican el artículo "*Data Scientist: the sexiest jobo f the 21st Century*" en la revista de Harvard: Business Review.

En 2017, Lombing Cao, publica "Data Science Overview" publicacion enfocada a dar un panorama general de Ciencia de Datos y en el mismo año, David Donoho publica "50 year of data Science", publicación enfocada a las discusiones de estadísticos sobre Ciencia de Datos.

2.1.3 Conceptos relacionados con Ciencia de Datos

1) Descubrimiento de conocimiento

El proceso de descubrimiento de conocimiento o *KDD* por sus siglas en inglés, ha sido un proceso mediante el cual a partir de su creación fue la base para la minería de datos. En [62], se define como "*la extracción no trivial de información potencialmente útil a partir de un gran volumen de datos, en el cual la información está implícita, donde se trata de*

interpretar grandes cantidades de datos y encontrar relaciones o patrones, para conseguirlo harán falta técnicas de aprendizaje, estadística y bases de datos".

Desde esta perspectiva, el descubrimiento del conocimiento dio inicio a la obtención de conocimiento a partir de los datos, sin embargo, aún en sus inicios no contemplaba grandes volúmenes de información.

2) Minería de Datos

En [62], se define como: "*el proceso de extraer conocimiento de bases de datos. Su objetivo es descubrir situaciones anómalas y/o interesantes, tendencias, padrones y secuencias en los datos*". Como evolución del concepto descubrimiento de conocimiento, surge la minería de datos que es el área de aplicación donde ahora se aplicará *KDD* para extraer conocimiento de las bases de datos, sin embargo, se en el momento del surgimiento de la minería de datos se tenían limitantes en las herramientas para procesar grandes volúmenes de información.

2.2 Metodología de Ciencia de Datos

2.2.1 Definición de metodología de Ciencia de Datos

En [60], se define una metodología de Ciencia de Datos como: "*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*".

Dada esta definición, se puede entender a una metodología de Ciencia de Datos, como un conjunto de pasos sistematizados que utilizan grandes volúmenes de información como materia prima para aplicar técnicas y herramientas de diferentes áreas del conocimiento que permita identificar patrones y obtener respuestas a una pregunta de investigación que ayude a la toma de decisiones.

2.3 Revisión sistemática

2.3.1 Definición de revisión sistemática

Una tarea importante en el campo de la investigación científica es la revisión sistemática de la literatura generada acerca de determinado tema. En [5], se explica que "*Las revisiones sistemáticas son investigaciones en las cuales la unidad de análisis son los estudios originales primarios. Constituyen una herramienta esencial para sintetizar la información científica disponible, incrementar la validez de las conclusiones de estudios individuales e identificar áreas de oportunidad donde sea necesario realizar investigación*". Es un concepto que en los últimos años solo se había utilizado en el área de ciencias de la salud y con el tiempo la comunidad científica comenzó la adaptación de dicha metodología de revisión de la literatura a otros dominios.

De manera general, una metodología de revisión sistemática es el proceso sistematizado de revisión de la literatura mediante un conjunto de pasos consecutivos que garantizan la reproducibilidad de una investigación y elimina la selección de trabajos de acuerdo con los criterios propios del autor y que, además, ayuda a la identificación de áreas de oportunidad.

2.3.2 Historia de la revisión sistemática

En 1979, el epidemiólogo Archie Cochrane identificó que no se tenían registros adecuadamente sistematizados de la efectividad de los tratamientos a sus pacientes. En [6], se explica que, como metodología de investigación, la revisión sistemática nació en el campo de la medicina, cuando por parte de los investigadores surgió el interés por conocer los resultados de las intervenciones médicas. No obstante, aunque en estos momentos ya se tenía un gran número de investigaciones y de resultados, éstos no se encontraban debidamente organizados de acuerdo con una metodología específica.

En 1993, se funda la colaboración Cochrane con la intención de crear un organismo sin fines de lucro que facilite las revisiones sistemáticas y a partir de ese año se comenzó a publicar el manual Cochrane el cual es mejorado cada año con el objetivo de tener una mejor calidad de revisiones sistemáticas. Asimismo, en el mismo año se crea el repositorio de revisiones sistemáticas para administrar todas las revisiones sistemáticas realizadas año con año.

En 2006, la Dra. Barbara Kichenham adapta el manual Cochrane al área de las Ciencias computacionales y crea “Guidelines for Performing Systematic Literature Reviews in Software Engineering”.

2.4 Metodología de revisión sistemática de la Dra. Barbara Kichenham.

2.4.1 Descripción general de metodología

Con la intención de brindar una explicación de la metodología de revisión sistemática de la Dra. Bárbara Kichenham en [7], se brindará información en este apartado, sin embargo, la metodología será empleada y analizada a profundidad en el capítulo 4 “Propuesta de solución”. A continuación, se describen de manera general las fases que componen la metodología de revisión sistemática utilizada en la presente tesis.

1. Fase de planeación de la revisión.

En esta fase, se contemplan todos aquellos factores que se deben considerar antes de iniciar con la revisión sistemática. Se refieren a definir una justificación, un enfoque, las preguntas de investigación y los criterios, por mencionar algunos. Esta fase está compuesta por las siguientes partes:

A) Identificación de la necesidad de la revisión sistemática
B) Desarrollo de un protocolo de revisión

2. Fase conductora

En esta fase se utiliza todo lo especificado en el protocolo de revisión sistemática. El protocolo debe pasar por un proceso de refinamiento y revisión antes de iniciar con esta fase, la cual está compuesta por las siguientes partes:

A) Identificación de búsqueda
B) Selección de estudios primarios
C) Asignación de calidad de estudios primarios
D) Síntesis y extracción de datos

3. Fase de elaboración del reporte.

En esta Fase, se reportan los resúmenes de las publicaciones relevantes identificadas en la fase anterior. Adicionalmente, en esta fase se realiza un análisis cuantitativo y cualitativo. Dicha fase está compuesta de las siguientes partes:

A) Resúmenes de publicaciones seleccionadas
B) Perspectivas de artículos involucrados

Capítulo 3

Trabajos relacionados

3 Trabajos relacionados

En este capítulo, se describen los trabajos relacionados que fueron identificados para la presente investigación. En la Sección 3.1, se describen los antecedentes dentro del CENIDET; en la Sección 3.2, se presenta el análisis de las publicaciones sobre Ciencia de Datos, revisión sistemática y metodologías de revisión sistemática; en la Sección 3.3, se presentan las observaciones de las publicaciones de Ciencia de Datos, revisión sistemática y metodologías de revisión sistemática y finalmente en la Sección 3.4, se presenta la comparación entre la propuesta actual y los trabajos identificados.

3.1 Antecedentes dentro del CENIDET

La tesis de maestría “Revisión del estado del arte de los algoritmos *K-means* y sus mejoras”[8], desarrollada dentro del CENIDET por A. Villalobos AV. Vega, aborda la problemática de la clasificación de las publicaciones que describen las mejoras más relevantes del algoritmo de agrupamiento *K-means* de acuerdo con sus etapas. Para esta clasificación, se aplicó la metodología de revisión sistemática en la cual se propuso una taxonomía organizada de acuerdo con distintas etapas del algoritmo *K-means*. Como resultado, se agruparon 1,125 documentos, sobre los cuales se aplicaron criterios de inclusión y exclusión, quedando categorizados en total 79 estudios, de los cuales 38 se encuentran en la fase de inicialización; 33 para la combinación entre las fases de clasificación y cálculo de centroides; la etapa de convergencia contiene cuatro publicaciones, dos para la combinación de convergencia-inicialización y dos para convergencia-clasificación.

Dicha tesis, provocó el interés por realizar revisiones del estado del arte mediante el uso de una metodología que permita garantizar la reproducibilidad del proceso e identificar las publicaciones relevantes eliminando los criterios del autor para su selección. Asimismo, los resultados de esta tesis fueron muy satisfactorios ya que generaron un artículo y un capítulo de libro.

3.2 Análisis de publicaciones relevantes

En esta Sección, se presenta el análisis de los artículos de Ciencia de Datos, revisión sistemática, y metodologías de revisión sistemática identificados como relevantes para esta investigación.

3.2.1 Ciencia de datos

Se analizaron las publicaciones acerca de esta temática con el objeto de identificar las aportaciones importantes para esta investigación. Para este análisis, se tomaron en cuenta los siguientes criterios: definición y origen de ciencia de datos; si son una revisión sistemática; si utilizan de manera explícita una metodología de revisión sistemática; lugar de publicación; número de citas; y autores.

1) *Data Science: A comprehensive overview*

En [9], se proporciona una descripción general sobre Ciencia de Datos, tal como, estructuras, áreas de aplicación, habilidades, áreas de oportunidad, universidades donde se imparte la carrera de científico de datos, sitios *web* donde se imparten cursos sobre Ciencia de Datos e historia. Es importante resaltar, que el autor menciona cómo ha impactado la Ciencia de Datos en las diferentes áreas de aplicación.

Al analizar las partes fundamentales para la presente investigación, se identificó que no se expresa de manera explícita una metodología para realizar la revisión de la literatura. El autor reconoce que el artículo fue construido por las observaciones y experiencias obtenidas dentro de la industria y organizaciones gubernamentales. Asimismo, el autor define a la Ciencia de Datos como: *Un nuevo campo interdisciplinario que sintetiza y se basa en la estadística, informática, computación, comunicación, gestión y sociología para estudiar los datos y su medio, para la transformación de los datos en ideas y decisiones siguiendo un pensamiento y una metodología*. Por otro lado, el autor atribuye el origen de la ciencia de datos a los trabajos de Turkey [1] y Neur [4].

2) *50 years of data science*

En [3], se reportan discusiones realizadas en el *Turkey centennial workshop,* con las cuales se busca precisar qué es la Ciencia de Datos. Asimismo, se explica lo siguiente: ¿Cuál es el verdadero trabajo de un científico de datos?, ¿Por qué los egresados de dicha carrera no pueden considerarse científicos de datos? Finalmente, identifica seis habilidades que deben de cumplirse antes de considerarse científicos de datos. Tales habilidades son: preparación y exploración de los datos; transformación y representación de los datos; cómputo de datos; modelado de datos: visualización y presentación de datos.

Al aplicar los criterios de análisis ya mencionados a esta publicación, se identificó que esta publicación fue construida a partir de las discusiones realizadas en el *Turkey centennial workshop* y también con las investigaciones de Turkey en las cuales aborda el análisis de datos y la causa por la cual debe ser una ciencia. Es importante mencionar que no brinda una definición de Ciencia de Datos. Sin embargo, presenta una visión de la Ciencia de Datos basada en las actividades que realizan las personas que están aprendiendo de los datos y describe un campo académico dedicado al mejoramiento de la Ciencia de Datos, basado en evidencias.

Continuando el análisis del artículo, el autor no menciona en qué año se reportó la aparición del término *Ciencia de Datos*. Sin embargo, se menciona la investigación que dio origen a la Ciencia de Datos, el artículo : "*The Future of Data Analysis*" publicado por John Turkey en 1962, en el cual afirma que el análisis de datos debe evolucionar, ya que desde su punto de vista cumple con todos los requisitos para ser considerado una ciencia.

3) *Data Science Methodologies: Current Challenges and Future Approaches*

En [10], se realizó una revisión crítica de las metodologías que ayudan a la gestión de proyectos de Ciencia de Datos. Esta investigación tiene dos enfoques, el primero de ellos es clasificar las metodologías de proyectos de Ciencia de Datos y el segundo motivar al debate de la importancia de la formulación y el uso de un método científico para las actividades de investigación de Ciencia de Datos en la industria y en proyectos de negocio. En este artículo se propone un cuadro de trabajo que contiene características que una

metodología para la gestión de proyectos de Ciencia de Datos debe contener considerando un punto de vista holístico.

Al aplicar los criterios de análisis ya mencionados a esta publicación, se identificó que el autor realiza una revisión crítica de la literatura para identificar las metodologías de gestión de proyectos de Ciencia de Datos. Asimismo, el autor realiza la revisión crítica de la literatura como una "*combinación de nuestro conocimiento y comprensión de lo que se ha escrito, nuestras habilidades de evaluación y juicio y nuestra capacidad para estructurarlos de forma clara y lógica por escrito*". El autor no explica de manera explícita cuáles fueron los pasos que siguió para realizar la revisión crítica, sin embargo, menciona los siguientes detalles:

1. Establecimiento de los criterios de selección y exclusión
2. Extracción de elementos importantes de los artículos (detalles del artículo, ideas principales, perspectivas y hallazgos).
3. Tabla comparativa de las metodologías para identificar cual cumple con los retos y desafíos identificados.

3.2.2 Revisión Sistemática

Se realizó un análisis de revisiones sistemáticas aplicadas en otros campos de la investigación. En este análisis únicamente se buscó identificar lo siguiente: metodología, definición, año de publicación, revista de publicación y autores.

1) *A Systematic Review for Smart City data analytics*

En [11], el autor utilizó una metodología de revisión sistemática tomada de [7], para identificar las publicaciones relevantes sobre ciudades inteligentes y el análisis de datos. El autor define la revisión sistemática como: *Las revisiones sistemáticas de la literatura son estudios de un nivel secundario y la calidad de sus hallazgos depende significativamente de la calidad de los estudios primarios que se utilizaron*. Para el proceso de implementación de la revisión sistemática, el autor identifica las siguientes fases:

1. Fase de planeación de revisión.
    a) Se identifica la necesidad de la revisión sistemática.
    b) Se diseña un esquema para el protocolo de revisión.

2. Fase conductora: se debe establecer un flujo, el cual es posible iterar porque muchas actividades que se inician en esta fase, deben ser refinadas (se crea un flujo iterativo de revisión). Esta fase se compone de dos pasos:
    a) Se identifica la investigación.
        - Pregunta de investigación.
    b) Se seleccionan los estudios primarios.
        - Criterios de inclusión y exclusión.
    c) Se caracteriza la calidad de los artículos seleccionados.
    d) Se extraen datos y se realiza una síntesis.

3. Fase de elaboración del reporte.
    a) Se resumen los artículos seleccionados y se responden algunas preguntas de investigación.
    b) Se ofrece una perspectiva de los artículos involucrados.

2) *Big Data analytics and Big Data Science: a survey*

En [12], se utiliza una metodología de revisión sistemática propia de los autores. Esta revisión está enfocada en seleccionar las publicaciones más relevantes sobre *Big Data*, con las cuales tiene la intención de brindar una visión general del concepto *Big Data*, ya sea para profesionales o para estudiantes. Los autores describen la metodología empleada en la revisión sistemática aplicando los siguientes pasos:

1. Se utilizó como palabra clave o *keyword,* la palabra *Big Data* y se filtraron los trabajos publicados entre 2011 y 2015 en SCI (*Science Citation Index* por sus siglas en inglés **)** y SSCI (*Social Sciences Citation Index* por sus siglas en inglés).
2. Se identificaron 186 artículos para seleccionarlos como muestra.
3. Se clasificó por número de artículos publicados y por fecha, con el fin de destacar la tendencia y el incremento de interés en el tema.
4. Se clasificaron los artículos seleccionados por revista y posteriormente se agruparon, con la finalidad de identificar qué revista tenía la mayoría de los artículos publicados durante los años seleccionados para su investigación.
5. Se seleccionaron las revistas que contienen mayor número de artículos relevantes. Cada uno de los artículos fue revisado por los buscadores.
6. Se revisaron dos veces cada uno de los artículos, con la finalidad de categorizarlos correctamente, de la siguiente manera:
    a) En la primera corrida, los artículos fueron revisados, leídos y categorizados por cada uno de los buscadores,
    b) En la segunda corrida se leyeron aquellos artículos en los cuales los buscadores no estuvieron de acuerdo, para volver a categorizarlos.
7. Se categorizaron los artículos en cada una de las categorías y subcategorías que establecieron anteriormente.
8. Se discutieron las categorías en donde fueron contenidos los artículos, de acuerdo con características.
9. Se colocaron cada uno de los artículos dentro de cada una de las categorías y subcategorías. También, se procedió a explicar la razón por la cual se posicionó cada uno de los artículos en las diferentes categorías.

3.2.3 Metodologías de revisión sistemática

Se identificó que en la literatura están reportadas metodologías de revisión sistemática que son aplicadas a la ciencia de la salud. Dichas metodologías relevantes son las siguientes:

1) ***Manual Cochrane de revisiones sistemáticas de intervenciones***

En [13] se describe una metodología para realizar una revisión sistemática en el área de ciencias de la salud que pretende brindar resultados del uso de tratamientos médicos. Esta metodología consiste en los siguientes pasos:

1. Se formula la pregunta para la revisión y se desarrollan los criterios para incluir los estudios,
2. Se buscan los estudios,
3. Se seleccionan los estudios y se obtienen los datos,
4. Se evalúa el riesgo de sesgo en los estudios incluidos,
5. Se analizan los datos y se realiza un meta-análisis,
6. Se analiza el sesgo del informe,
7. Se presentan los resultados y las tablas *Resumen de los Resultados*,
8. Se interpretan los resultados y se elaboran las conclusiones.

2) ***Five steps to conducting a systematic review***

En [14] se provee una metodología para los profesionales en el área de la salud que ayuda a realizar revisiones sistemáticas en este campo de investigación. Esta metodología consiste en 5 pasos:

1. Enmarcar las preguntas de la revisión sistemática,
2. Identificar los trabajos relevantes,
3. Evaluar la calidad de los estudios,
4. Resumir la evidencia,
5. Interpretar los descubrimientos.

3.3 Observaciones de trabajos relacionados

En esta Sección se presentan los hallazgos más relevantes identificados en las publicaciones de Ciencia de Datos y en la revisión sistemática. Los hallazgos se enlistan a continuación de manera puntual con la finalidad de denotar las partes de mayor interés.

3.3.1 Ciencia de datos

Las observaciones relevantes identificadas sobre Ciencia de Datos son :

- Las publicaciones [3] y [9] no describen de manera explícita una metodología de revisión del estado del arte. Por otro lado, en [10] se describe una metodología de revisión de la literatura, sin embargo, no se detallan los pasos para realizarla.
- La definición del concepto Ciencia de Datos aún no se presenta como una definición precisa, incluso los autores la definen desde sus propios puntos de vista.
- La cantidad de publicaciones relevantes referenciadas en los estudios sobre Ciencia de Datos presentan un incremento exponencial a partir del 2014, con lo cual, se comprueba que dicho tema es de gran interés.
- La cantidad de artículos referenciados en las publicaciones de investigación oscila entre 30% y 50%.
- El enfoque de los trabajos recientes es diferente: en [9] se brinda una visión general sobre la Ciencia de Datos; en [3] el objetivo es reportar las discusiones entre estadísticos y científicos de datos en relación con el concepto de Ciencia de Datos, tomando en cuenta diferentes aspectos. En [10], se clasifican y evalúan las metodologías de Ciencia de Datos para la gestión de proyectos.
- La accesibilidad a las referencias reportadas en los trabajos analizados oscila entre 88% y 93%.
- En [10] se identificó que se mezclan las metodologías para la gestión de proyectos con las metodologías de gestión de los datos.

3.3.2 Revisión Sistemática

De las publicaciones seleccionadas que aplican una revisión sistemática y las metodologías encontradas se identificó lo siguiente:

- En [12], la metodología no fue referenciada por el autor. Por otro lado en [11], la metodología de revisión sistemática empleada corresponde a otra investigación a la cual hace referencia el autor.
- El concepto de revisión sistemática únicamente fue explicado en [11].
- El trabajo más reciente corresponde al año 2018 [11].
- La metodología empleada por [11] consta de 4,079 citas.

3.4 Comparación de trabajos relacionados

En esta Sección, se presentan tres comparaciones de los trabajos relacionados con la investigación actual. En la primera se comparan los trabajos de Ciencia de Datos, en la segunda se comparan los trabajos de revisión sistemática y en la tercera se comparan las metodologías de revisión sistemática.

3.4.1 Ciencia de Datos

En la Tabla 1 se presenta una comparación entre los trabajos identificados sobre Ciencia de Datos y la presente investigación. Con la finalidad de identificar las partes relevantes de las publicaciones de Ciencia de Datos, en esta comparación se establecieron los siguientes criterios:

1. Describe una metodología de revisión de estado del arte
2. Contexto teórico (aborda los aspectos necesarios para entender el uso de la Ciencia de Datos)
3. Contexto histórico (detalla las investigaciones que dieron origen a la Ciencia de Datos)
4. Aborda investigaciones sobre Ciencia de Datos.
5. Implementa una clasificación de publicaciones

Tabla 1. Trabajos relacionados de Ciencia de Datos

Estudio	1	2	3	4	5
50 years of data science (2015)		●		●	
Data Science: A comprenhensive OverView (2017)		●	●	●	
Data science methodologies: Current challenges and future	●	●		●	
Presente investigación	●	●	●	●	●

3.4.2 Revisión sistemática

En la Tabla 2 se presenta una comparación entre los trabajos identificados sobre revisión sistemática y la presente investigación. Para identificar las partes relevantes de las publicaciones de revisión sistemática, en esta comparación se establecieron los siguientes criterios:

1. Describe de manera explícita la metodología
2. Contexto teórico (detalla los conceptos de revisión sistemática)
3. Cita el autor de la metodología de revisión sistemática implementada
4. Hace referencia de la publicación que ayudó a realizar la investigación en las fuentes consultadas

Tabla 2. Trabajos relacionados de revisión sistemática

Estudio	1	2	3	4
Chen - Big data analytics and Big data science: a survey	●			
Moustaka - A systematic review for Smart city data analytics	●	●	●	●
Presente investigación	●	●	●	●

3.4.3 Metodologías de revisión sistemática

En este apartado, se presenta una comparación entre los trabajos identificados sobre metodologías de revisión sistemática que se muestran en la Tabla 3. Con la finalidad de identificar las partes relevantes de las publicaciones de metodologías de revisión sistemática se establecieron los siguientes criterios:

1. Describe de manera explícita la metodología
2. Está fundamentada en el manual Cochrane de revisiones sistemáticas
3. Separa por fases la aplicación de la metodología de revisión sistemática
4. Fue implementada en el área de Ciencia Computacionales

Tabla 3. Trabajos relacionados de metodologías de revisión sistemática

Estudio	1	2	3	4
Manual Cochrane de revisiones sistemáticas de intervenciones	●	●	●	
Khan - Five steps to conducting a systematic review	●		●	
Kitchenham- Lessons from applying the systematic literature review process within the software engineering domain	●	●	●	●

Capítulo 4

Propuesta de Solución

4 Propuesta de solución

En este capítulo se describe la propuesta de solución a la problemática descrita en el presente documento de tesis. En la Sección 4.1 se describen los elementos que componen una revisión sistemática y en la Sección 4.2 se presenta el esquema de solución general y las etapas que la componen.

4.1 Elementos para realizar una revisión sistemática

Los elementos necesarios identificados para realizar una revisión sistemática son los siguientes: buscadores científicos de información, bases de datos de artículos (repositorios), publicaciones científicas y tecnológicas, filtros de búsqueda (que permiten realizar las búsquedas más especializadas), selección y agrupamiento de publicaciones. En el Figura 1, se muestran los elementos identificados para realizar una revisión sistemática.

Figura 1. Elementos identificados de una revisión sistemática

4.2 Etapas de la propuesta de solución

La propuesta de solución para la presente investigación está dividida en 4 etapas. En la etapa 1, se seleccionan las bases de datos en las cuales se buscarán aquellos trabajos que buscan solucionar la problemática planteada en esta tesis con el objetivo de construir el estado del arte; en la etapa 2, se identifican cuáles son las metodologías de revisión sistemática aplicadas a otros campos de la investigación y cuáles son las más relevantes, asimismo, también se realiza una comparación de ellas para identificar sus ventajas y desventajas; en la etapa 3, se selecciona una metodología de revisión sistemática más adecuada para la presente investigación y se comienza con su implementación; por último en la etapa 4, se realiza el análisis de resultados y se implementa una clasificación de las publicaciones. De manera general el proceso se encuentra descrito en la Figura 2.

Figura 2. Proceso de solución

4.2.1 Etapa 1

En esta etapa se seleccionaron los buscadores y las bases de datos para realizar la búsqueda de los artículos relevantes. Dicho esto, en el primer nivel las publicaciones sobre ciencia de datos relevantes; en el segundo nivel las publicaciones sobre revisión sistemática aplicada

a otras áreas de investigación y, por último, en el tercer nivel las metodologías de revisión sistemática reportadas en la literatura. Este proceso se muestra en la Figura 3.

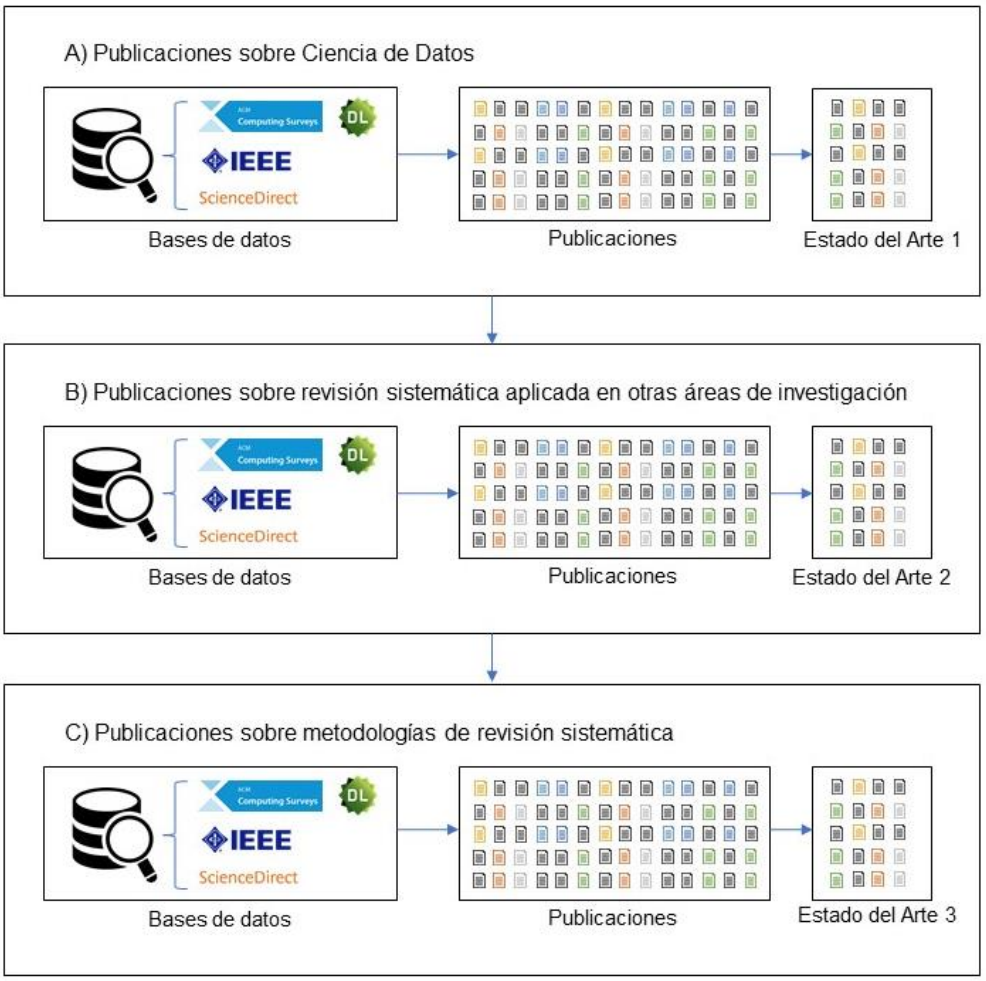

Figura 3. Proceso de la etapa 1

Adicionalmente, se realizó una búsqueda de referencias cruzadas de forma manual, recuperando documentos de interés principalmente por medio del buscador de *Google académico*, así como también se tomaron como base inicial, los artículos, libros y documentos relacionados recuperados en el grupo de investigación.

4.2.2 Etapa 2

En la etapa 2, se realizó una comparación de las metodologías de revisión sistemática que son utilizadas en diferentes campos de investigación. La finalidad de esta comparación es la selección de la metodología más adecuada para la presente investigación. La Figura 4, describe el proceso de la etapa 2.

Figura 4. Proceso de la etapa 2.

4.2.3 Etapa 3

En la etapa 3, se seleccionó una metodología de revisión sistemática basándonos en los resultados del análisis de la etapa anterior. Asimismo, se comenzó con la implementación de la metodología seleccionada. La metodología seleccionada [7], se describe en la Figura 5.

Figura 5. Metodología de revisión sistemática

4.2.4 Etapa 4

En la etapa 4, como parte del análisis de los resultados se realizó un análisis cuantitativo y cualitativo de los resultados obtenidos. Asimismo, se realizó un análisis de las publicaciones para poder agruparlas de acuerdo con su enfoque.

Capítulo 5

Revisión sistemática

5 Revisión sistemática

El interés de esta investigación, consiste en identificar publicaciones sobre metodologías de Ciencia de Datos, debido a que estas metodologías son la base para el desarrollo de proyectos en este campo. Este interés está motivado por la falta de especificidad de las metodologías de búsqueda hasta ahora reportadas en la literatura.

La revisión sistemática implementada en la presente investigación es la propuesta por [7], la cual se compone por tres fases principales:

- Fase 1: Planeación de la revisión.
- Fase 2: Implementación de la revisión.
- Fase 3: Reporte de revisión sistemática.

5.1 Identificación de la necesidad de revisión sistemática

Se identificó que el crecimiento exponencial en la cantidad de publicaciones actuales y el tiempo invertido para identificar las más relevantes, son las principales razones para realizar una revisión sistemática. Asimismo, se identificó que las metodologías de Ciencia de Datos son de manera general importantes para su aplicación por lo cual se considera relevante aplicar la revisión sistemática a las metodologías de Ciencia de Datos.

5.2 Desarrollo del protocolo de revisión

Se realizó un esquema del protocolo de revisión sistemática, el cual se sometió a análisis, para posteriormente continuar con la construcción del protocolo de revisión sistemática. El protocolo está compuesto por las siguientes partes: antecedentes, preguntas de investigación, estrategia de selección de estudios primarios, criterios de selección, criterios para la evaluación de la calidad, extracción de datos, síntesis, estrategia de diseminación y calendario de actividades.

5.3 Identificación de la búsqueda

Para este apartado se formuló la pregunta de investigación, la cual tiene la intención de ayudar a identificar las publicaciones relevantes del tema metodológico de Ciencia de Datos. Para la formulación de la pregunta, se tomó en consideración la gran importancia de seguir un procedimiento o metodología cuando se utiliza la Ciencia de Datos.

5.4 Selección de estudios primarios

En este apartado, se buscaran los estudios primarios, es decir, las publicaciones relevantes que responderán la pregunta de investigación planteada en el protocolo de revisión sistemática. Para la selección de estudios, se siguió la estrategia de búsqueda diseñada en el protocolo de revisión sistemática. La estrategia consiste en la búsqueda de publicaciones mediante el uso de cadenas de búsqueda en los buscadores y repositorios a los que el CENIDET tiene acceso. Para las cadenas de búsquedas se identificaron "*Methodology AND Data Science*". Posteriormente se reconocieron sinónimos de la palabra "*Methodology*": "Method", "Process" y "Procedure" en idioma español e inglés para sustituirlo en la cadena de búsqueda. Asimismo, en esta actividad se determinaron los valores de los criterios de búsqueda. Dichos criterios son: título, año de publicación, número de citas e idioma. En la Figura 6 se muestra de manera general el proceso de selección de estudios primarios en donde se muestra la cantidad de artículos encontrados en los repositorios, así como los artículos que se fueron filtrando después de aplicar los criterios de inclusión y exclusión con la finalidad de seleccionar las publicaciones de mejor calidad.

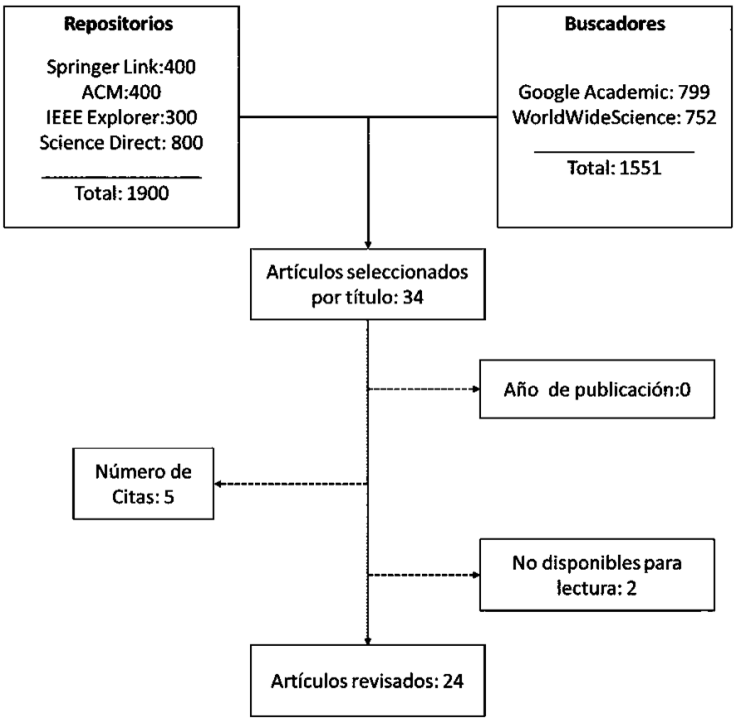

Figura 6. Selección de estudios primarios.

5.5 Caracterización de calidad de las publicaciones

Para realizar la caracterización, se extrajeron aquellos elementos que permitirán determinar la calidad de cada una de las publicaciones identificadas en la búsqueda. Por calidad entiéndase aquel conjunto de elementos que posicionan a una publicación respecto a las demás. Para dicho proceso fue necesario extraer los siguientes elementos: nombre de artículo, número de páginas, tema, número de citas, año de publicación, citas por año, link, año de búsqueda y palabras clave. Asimismo, en esta actividad se aplicaron los criterios de inclusión y exclusión a manera de filtro para las publicaciones identificadas en la búsqueda.

5.6 Extracción de datos y síntesis

Se realizó el análisis y extracción de datos de las publicaciones seleccionadas como las más relevantes de acuerdo con la pregunta de investigación. En el análisis se contempló leer el resumen, contenido y conclusiones de cada publicación. Una vez terminado este proceso, se elaboró una síntesis de las publicaciones.

5.7 Resúmenes de publicaciones

24 resúmenes correspondientes a cada una de las publicaciones seleccionadas como relevantes. Estas publicaciones se enlistan a continuación y se presenta un resumen de su contenido:

5.7.1 *Reimaging Research Methodology as Data Science*

En [27] se extraen ideas de un proyecto de investigación a gran escala que examina las concepciones y prácticas actuales de académicos (n = 144) involucrados en la enseñanza de métodos de investigación en universidades de investigación intensiva en 17 países. Los datos se obtuvieron mediante un cuestionario en línea.

También, señala que las universidades enfocan la enseñanza de los cursos de métodos de investigación principalmente como un "servicio" para los estudiantes y departamentos, no como parte del plan de estudios básico. Para hacer frente a los crecientes cambios en las estructuras de datos y el entorno de investigación impulsado por la tecnología, el estudio recomienda a las instituciones que vuelvan a crear imágenes de los programas de

metodología de la investigación para permitir que los estudiantes desarrollen las competencias adecuadas para hacer frente a los desafíos de trabajar con grandes cantidades de datos y análisis asociados.

5.7.2 *Comparing Data Science Project Management Methodologies via a Controlled Experiment*

En [28] se informan los resultados de un experimento que compara cuatro metodologías diferentes para administrar y coordinar un proyecto de ciencia de datos. Dichas metodologías son: *Agile Scrum*, *Agile Kanban*, *CRISP-DM* y las metodologías propias. Primero, se presenta un modelo para comparar diferentes metodologías de gestión de proyectos y luego informa sobre los resultados del experimento.

Así mismo, muestra los resultados del experimento, los cuales muestran que existen diferencias significativas en función de la metodología utilizada, siendo la metodología *Agile Kanban* la más efectiva y sorprendentemente, la metodología *Agile Scrum* siendo la menos efectiva. No obstante, el autor menciona que el enfoque de gestión de proyectos con ciencia de datos, ayudará a los estudiantes a resolver problemas de la sociedad digital. Dicha investigación, tiene como objetivo crear conciencia sobre volver a estructurar las metodologías de investigación debido al crecimiento excesivo de datos.

5.7.3 *Business Information Modeling: A Methodology for Data-Intensive Projects, Data Science and Big Data Governance*

En [29] se propone una metodología *Business Information Modeling* (BIM) integrada para estructurar y formalizar los requisitos comerciales en grandes proyectos con uso intensivo de datos. También, propone un flujo de trabajo con Ciencia de Datos en el cual considera 4 fases: planeación, análisis, diseño y construcción. Además, muestra que el enfoque es adecuado más allá de los entornos de almacenamiento de datos tradicionales, aplicándolo también a entornos de *Big data* e iniciativas de Ciencia de Datos, donde el análisis de requisitos comerciales a menudo se descuida.

Dado que la compatibilidad adecuada con las herramientas ha resultado ser inevitable en muchos entornos del mundo real, también discute los requisitos de software y su implementación en la herramienta *Accurity Glossary*. El enfoque se evalúa en función de

un gran proyecto de almacenamiento de datos bancarios en el que los autores participan actualmente.

5.7.4 *A Data Science Methodology Based on Machine Learning Algorithms for Flood Severity Prediction*

En [30] se presenta una metodología de Ciencia de Datos utilizando algoritmos de *machine learning* que incluyen la arquitectura de redes neuronales para la predicción de la gravedad de las inundaciones. Este documento intenta abordar el tema de la mitigación de inundaciones a través de la presentación de un nuevo conjunto de datos de inundaciones, que comprende 2000 eventos de inundación anotados, donde la gravedad del resultado se clasifica de acuerdo con 3 clases objetivas, lo que demuestra la gravedad respectiva de las inundaciones.

Asimismo, el documento presenta varios tipos de algoritmos de *machine learning*. Dichos algoritmos presentados son: *Random Forest Classifier, Support vector machines (SVM), Levenberg-Marquartdt training algorithm* y *Baseline algorithms*. Estos algoritmos, son útiles para esta investigación ya que por su utilidad ayudan a predecir la gravedad de las inundaciones y clasificación de los resultados en tres clases: inundaciones normales, anormales y de alto riesgo.

La metodología propuesta es la siguiente:

1. Preprocesamiento de los datos con herramientas,
2. Preparación de datos: reducción de dimensionalidad, selección de características y extracción de características,
3. División de conjunto de datos: conjunto de datos de entrenamiento, conjunto de datos de validación y conjunto de datos de prueba,
4. Construcción de modelos: algoritmos de *machine learning*,
5. Pruebas de modelos: iterar el mejor modelo y el uso de la evaluación del desempeño,
6. Selección de modelo: aplicación del modelo a los datos y visualización de los datos.

5.7.5 Data Science Methodology for Cybersecurity Projects

En [31] se presenta las metodologías de Ciencia de Datos populares *CRISP-DM*, *FMDS (IBM)*, *TDSP (Microsoft)*, *CRISP-DM* y *SAS SEMMA*. Algunas de las anteriores son comparadas de acuerdo con los desafíos de ciberseguridad. También se realiza una comparativa para explicar las fortalezas y debilidades de las metodologías en el caso de proyectos de ciberseguridad. Como resultado de la evaluación de las metodologías de Ciencia de Datos, el autor considera que existen 4 actividades que son las generales que actualmente se encuentran en cada una de las metodologías evaluadas y que además se encuentran en un ciclo. Estas actividades son: definición del problema/formulación, recopilación de datos, modelado de datos y producción de datos.

El autor aborda la ciberseguridad con el objetivo de identificar que los métodos tradicionales para la identificación de ciberataques tienen limitaciones. Por lo anterior, el autor considera que el uso de la Ciencia de Datos ayuda a cruzar esas limitaciones. Adicionalmente, el autor propone el uso de la metodología *FMDS* (IBM) para aplicarlo en el campo de la ciberseguridad agrupando cada una de las actividades en su propuesta general de proceso de Ciencia de Datos.

Metodología *FMDS* para ciberseguridad:

1 formulación y definición del problema

1.1 Entendimiento del negocio

1.2 Enfoque analítico

2 Recopilación de datos

2.1 Requerimientos de datos

2.2 Recolección de datos

2.3 Entendimiento de los datos

2.4 Preparación de los datos

3 Análisis de datos

3.1 Modelado

3.2 Evaluación

4) Producción

4.1 Despliegue

4.2 Retroalimentación

5.7.6 *Adoption-Driven Data Science for Transportation Planning: Methodology, Case Study, and Lessons Learned*

En [32] se propone una metodología para unir dos disciplinas, la ciencia de datos y el transporte, para identificar, comprender y resolver problemas de planificación del transporte con soluciones basadas en datos que sean adecuadas para su adopción por parte de los planificadores urbanos y los responsables políticos.

La metodología se define en cuatro pasos, donde personas de ambas disciplinas van desde la definición de algoritmos y modelos hasta el desarrollo de una solución potencialmente adoptable con resultados evaluados. También, se describe cómo se aplicó esta metodología para definir un modelo para inferir viajes diarios con el modo de transporte a partir de los datos del teléfono móvil, e informamos las lecciones aprendidas durante el proceso. El autor propone la siguiente metodología en la que interactúan los actores del dominio y los actores del dominio. El objetivo de la metodología es integrar un equipo multidisciplinario que consiga la mejor solución. Dicha metodología se describe a continuación;

1. Diseño de algoritmos: para procesar tanto los requisitos del dominio como los conjuntos de datos no tradicionales para llegar a una primera versión de la solución.
2. Evaluación basada en datos: para evaluar el rendimiento de la solución en comparación con un conjunto de datos de referencia de dominio que se utilizará para iterar la solución.
3. Evaluación interactiva y colaborativa: para evaluar si la solución es lo suficientemente útil para ser adoptada en la planificación y formulación de políticas. Al igual que en el paso anterior, su salida se utiliza para iterar la solución.
4. Consolidación del conocimiento: congregar el conocimiento y los sistemas generados, evaluando los actores los requisitos para que la solución sea adoptada en los escenarios del mundo real. Similar a pasos anteriores, su salida se utiliza para iterar la solución.

5.7.7 *Which method to use? An assessment of data mining methods in Environmental Data Science*

En [34] se proporciona una conceptualización de los sistemas ambientales y una conceptualización de los métodos de minería de datos, que se encuentran en el paso central del proceso de ciencia de datos. Estos dos elementos definen un marco conceptual que se basa en una nueva metodología propuesta para relacionar las características de un determinado problema ambiental con una familia de métodos de minería de datos.

El artículo ofrece una visión general y pautas de las técnicas de minería de datos a un usuario no experto, que puede decidir con este soporte cuál es la técnica más adecuada para resolver su problema en cuestión. Además, se presenta y discute cada par de métodos de minería de datos del sistema ambiental. También, se presentan algunos ejemplos de cómo se utiliza la metodología propuesta para respaldar la selección del método de minería de datos, y se identifican desafíos y tendencias futuras.

5.7.8 *A Data Science Methodology for Internet-of-Things*

En [37] se discuten las oportunidades y preocupaciones de la analítica de *IoT*. Además, proponemos una metodología de ciencia de datos genérica para el análisis de datos de *IoT* denominada *Plan, Collect and Analytics for Internet-of-Things* (PCA-IoT). La metodología propuesta podría aplicarse en escenarios de *IoT* para realizar análisis de datos para una toma de decisiones efectiva y eficiente. La metodología planteada es la siguiente: 1. Planeación,2. Recolección y 3. Análisis. Para ilustrar esto, se muestra en el Figura 7.

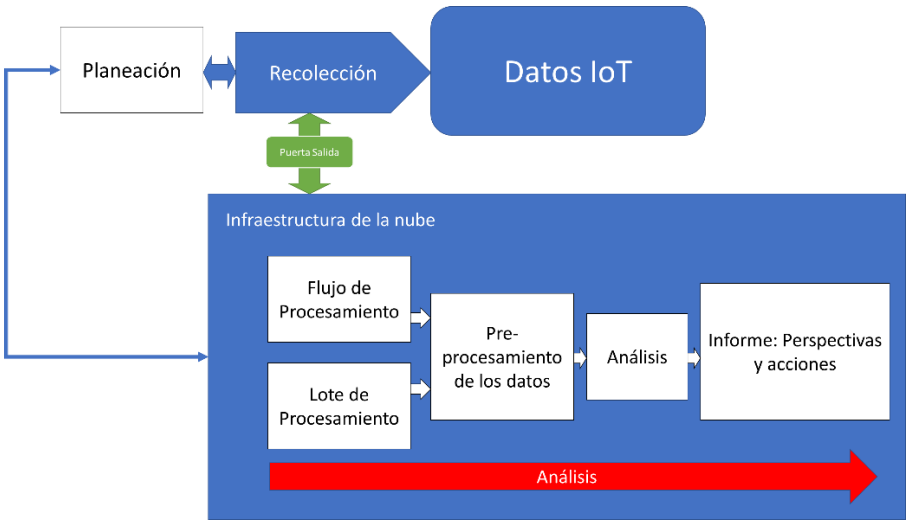

Figura 7. Metodología para internet de las cosas.[1]

5.7.9 *Methodological Framework for Data Processing based on the Data Science Paradigm*

En [38] se propone el uso de la ciencia de datos en la fase dos de una metodología para el desarrollo de aplicaciones de minería de datos llamada *MIDANO* (Metodología para el desarrollo de aplicaciones de Minería de Datos basados en el análisis organizacional).

La metodología *MIDANO* consta de las siguientes partes: 1. Identificación de las fuentes para la extracción de conocimiento en una organización y de los problemas que se pueden resolver con ellas en dicha organización, 2. Preparación y tratamiento de los Datos, 3. Desarrollo de herramientas de minería de datos. La metodología se describe en Figura 8.

Figura 8. Metodología MIDANO.[2]

[1] S.N. Brohi, M. Marjani, I.A.T. Hashem, T.R. Pillai, S. Kaur and S. Amalathas, "A Data Science Methodology for Internet-of-Things*", presented at the Int. Conf. for Emerg, Tech. in Comp.* London, UK., Aug. 19-20, 2019, pp. 184.

[2] P. F. Rangel, C. Aguilar, J. Cerrada M. and A. J. "Methodological Framework for Data Processing based on the Data Science Paradigm", *presented at the XL Latin American Comp*. Conf., Uruguay, pp. 3 , Nov. 15-19, 2014.

5.7.10 *Artificial Intelligence Data Science Methodology for Earth Observation*

En [39] se describe un demostrador a pequeña escala de capas intermedias de *Copernicus Access Platform*, que es una plataforma general para el manejo, análisis e interpretación de imágenes de satélite de observación de la Tierra, que explota principalmente los macrodatos del Programa Europeo *Copernicus* mediante métodos de inteligencia artificial (IA). A partir de 2020, la plataforma se aplicará a nivel regional y nacional para diversos usos.

Casos como expansión urbana, salud forestal y desastres naturales. Sus flujos de trabajo permiten la selección de imágenes de satélite de archivos de datos, la extracción de información útil de los metadatos, la generación de descriptores para cada imagen individual, la ingestión de imágenes y datos de descriptores en una base de datos común, la asignación de etiquetas de contenido semántico a la imagen y la posibilidad de buscar y recuperar capturas de imágenes similares relacionados con el contenido. Los dos componentes principales, a saber, minería de datos y fusión de datos, están detallados y validados.

5.7.11 *The Big Three: A Methodology to Increase Data Science ROI by Answering the Questions Companies*

En [40] se propone una metodología para categorizar y responder a las preguntas de 'Los tres grandes' (qué está sucediendo, qué lo está causando y qué acciones puedo tomar para optimizar lo que me importa) utilizando la ciencia de datos.

Las aplicaciones de la Ciencia de Datos parecen ser casi infinitas en el panorama moderno de hoy, cada empresa compite por posicionarse en la nueva economía de datos y conocimientos. Sin embargo, los científicos de datos parecen centrarse únicamente en el uso de métodos de clasificación, regresión y agrupación para responder a la pregunta "qué está pasando".

Responder preguntas sobre por qué están sucediendo las cosas o cómo tomar las acciones óptimas para mejorar las métricas se relegan a campos de investigación especializados y, en general, se descuidan en el análisis de la ciencia de datos de la industria.

5.7.12 *Big data and data science methods for management research*

En [41] se proporciona un manual básico o un "kit de inicio" para posibles aplicaciones de la Ciencia de datos en la gestión de la investigación. Así mismo, se advierte que los campos emergentes son obsoletos y mejoran las metodologías, mientras que a menudo son reemplazados por nuevas aplicaciones. Sin embargo, este manual puede guiar a los académicos de la administración que deseen utilizar técnicas de ciencia de datos para obtener mejores respuestas a preguntas existentes o explorar preguntas de investigación completamente nuevas.

5.7.13 *Processes Meet Big Data: Connecting Data Science with Process Science*

En [42] el autor destaca que actualmente la mayoría de los proyectos de *Big data* están enfocados a su almacenamiento, procesamiento y tareas analíticas bastantes simples. Debido a esto, surge la necesidad de cambiar dicho enfoque por uno en el cual se haga un análisis más profundo de los datos, es decir, un enfoque de ciencia de datos.

Las causas o iniciativas de los proyectos de *Big data* no contemplan las mejoras del proceso de inicio a fin. El autor considera que es necesario una integración de la ciencia de datos, la tecnología de datos y la ciencia de procesos. Este nuevo enfoque, ayudará a impulsar dichos proyectos por modelos sin considerar la evidencia oculta en los datos. Por ello, la minería de procesos tiene como objetivo reducir el sesgo presente entre dichos conceptos (Ciencia de datos, tecnología de datos y ciencia de procesos).

5.7.14 Process-Structure Linkages Using a Data Science Approach: Application to Simulated Additive Manufacturing Data

En [43] se muestra el flujo de trabajo basado en datos se aplica a un conjunto de microestructuras sintéticas de fabricación aditiva obtenidas mediante el enfoque de *Potts-kinetic Monte Carlo* (kMC por sus siglas en inglés).

Las técnicas de fabricación aditiva, producen de forma inherente microestructuras que pueden variar significativamente con las condiciones de procesamiento. Utilizando el flujo de trabajo desarrollado, se establece un modelo basado en datos de baja dimensión para correlacionar los parámetros del proceso con la microestructura final prevista.

5.7.15 *Exploring the Process of Doing Data Science Via an Ethnographic Study of a Media Advertising Company*

En [44] se presenta los resultados de un estudio etnográfico centrado en cómo se llevaron a cabo los proyectos de ciencia de datos dentro de una empresa global de publicidad en medios. Se documentan las observaciones, mediante la incorporación de un investigador dentro del equipo, así como entrevistas y encuestas más estructuradas. También, se discuten recomendaciones para mejorar la metodología actual de ciencia de datos dentro de la empresa.

El autor también menciona que, se le ha prestado poca atención a la metodología de proceso del equipo y las mejoras de proceso sugeridas darían como resultado que los proyectos de ciencia de datos de la empresa tuvieran menos riesgos y plazos más cortos. Otros equipos de *Big data*, también podrían beneficiarse de la revisión y el perfeccionamiento de sus procesos de trabajo, pero es necesario trabajar más para validar esta suposición. La metodología propuesta se muestra en el Figura 9.

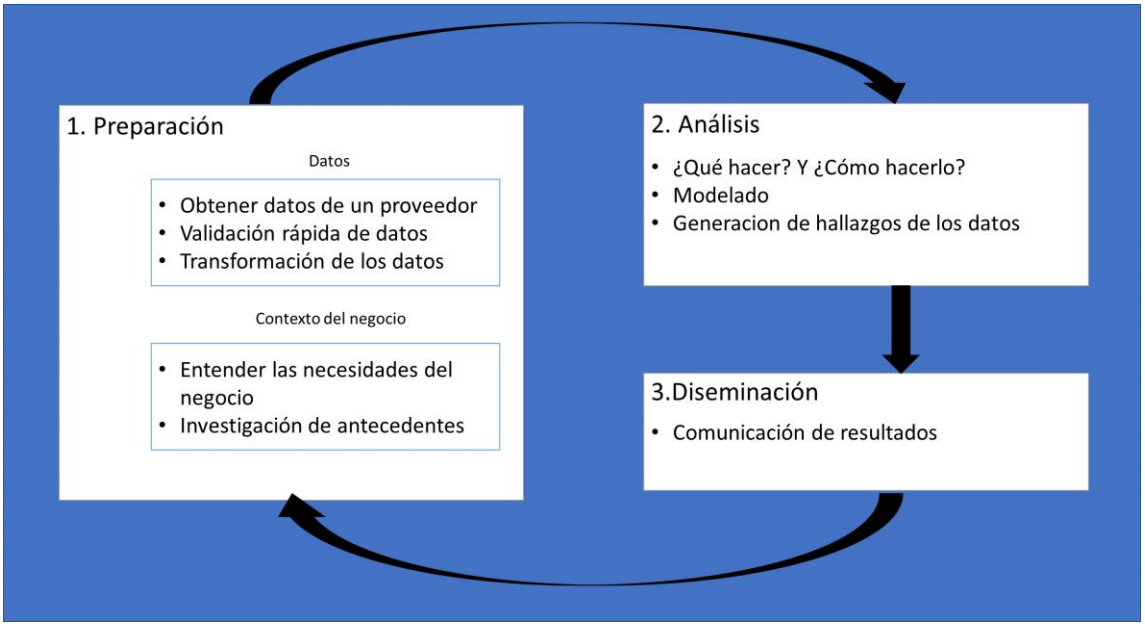

Figura 9. Metodología mejorada en la empresa.[3]

[3] J.S. Saltz and I. Shamshurin, "Exploring the Process of Doing Data Science Via an Ethnographic Study of a Media Advertising Company", *in 2015 IEEE Int. Conf. on Big Data, Santa Clara*, CA, 2015, pp. 2100.

5.7.16 Exploring the Relationship Between Data Science and Circular Economy: Ann Enhanced CRISP-DM Process Model

En [45] se aborda cómo las organizaciones pueden estructurar mejor su comprensión y preparación de datos para alinearse con los objetivos generales de negocios y economía circular. Por lo tanto, con base en la literatura y un estudio de caso, se explora la relación entre la ciencia de datos y la economía circular, y se propone un modelo de proceso genérico. El modelo de proceso propuesto amplía el Proceso estándar de la industria para la minería de datos (*CRISP-DM*) con una fase adicional de validación de datos e integra el concepto de perfiles analíticos. Demostramos su aplicación para el caso de estudio de una empresa de fabricación que busca implementar la estrategia circular inteligente: mantenimiento predictivo. El proceso mejorado de *CRISP-DM*, en el cual pretende adicional la validación de los datos se ilustra gráficamente en el Figura 10.

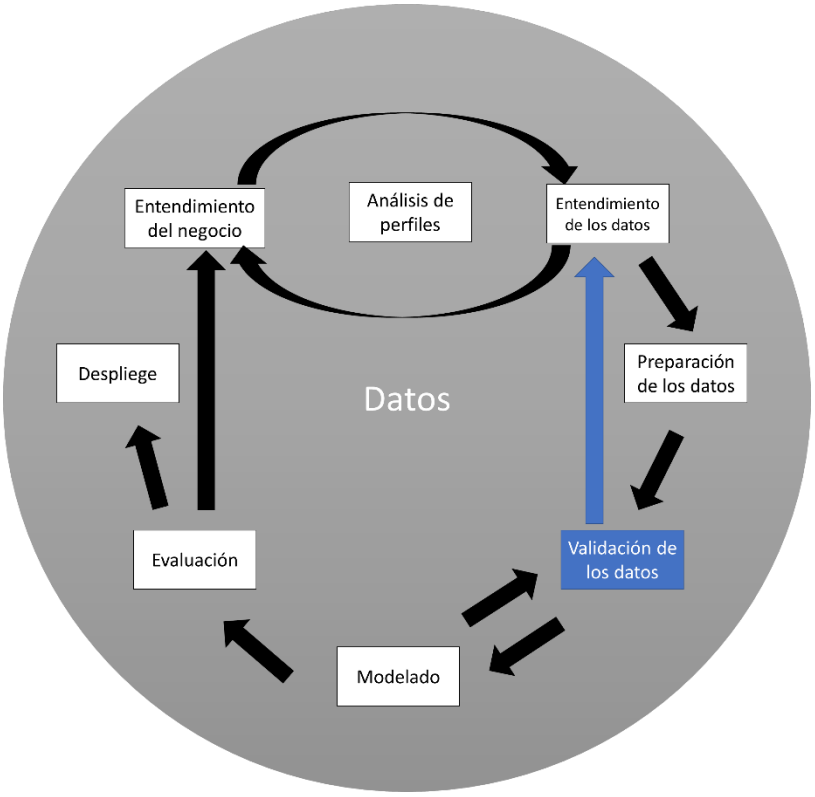

Figura 10. Proceso de Minería de datos mejorado.[4]

.

[4] Tomada de: E. Kristoffersen, O.O. Aremu, F. Blomsma, P. Mikalef and J. Li, "Exploring the Relationship Between Data Science and Circular Economy: An Enhanced CRISP-DM Process Model", presented ath the 2019 Conf. on e-Business, e-Services and e-Society: Digital Tf. for a Susble. Society in the 21st Century, Trondheim, Norway, 2019, pp.183.

5.7.17 *POST-DS: A Methodology to Boost Data Science*

En [46] se analizan las metodologías y modelos para el proceso de datos tales como: *KDD*, *CRISP-DM*, *SEMMA*, *ASUM* y *TDSP*. Dichos modelos y metodologías según la perspectiva del autor no se encuentran completas por lo que se propone una metodología llamada POST-DS (Organización de procesos y programación de herramientas de elección para ciencia de datos). Esta es una metodología orientada a procesos para ayudar a la gestión de proyectos de ciencia de datos. Este enfoque no solo se admite en el proceso, sino también en la programación de la organización y la selección de herramientas.

Además, el autor identifica los siguientes cuatro bloques que hacen frente a las problemáticas de la ciencia de datos: organización, calendarización, proceso y herramientas. En la siguiente lista se describe el proceso de la metodología:

1. Proceso
2. Organización
3. Calendarización
4. Herramientas

Este enfoque permite la identificación de procesos, organización, programación y herramientas. Este enfoque está inspirado particularmente en el Proceso Estándar Interindustrial para Minería de Datos, pero tiene la intención de brindar pautas adicionales. Esta metodología se aplicó en un proyecto específico de ciencia de datos. La aplicación permitió concluir que este POST-DS puede contribuir a una mejor alineación de la gestión general del proyecto.

5.7.18 *CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories*

En [47] se busca argumentar porque sí y en qué contextos, CRISP-DM sigue siendo apto para proyectos de ciencia de datos. El autor argumenta que, si el proyecto está dirigido a objetivos e impulsado por procesos, la visión del modelo de proceso todavía se mantiene en gran medida. Por otro lado, cuando los proyectos de ciencia de datos se vuelven más exploratorios, los caminos que puede tomar el proyecto se vuelven más variados y se requiere un modelo más flexible.

Además, el autor sugiere cómo se verían los contornos de dicho modelo basado en trayectorias y cómo se puede utilizar para categorizar proyectos de ciencia de datos (dirigido a objetivos, exploratorio o gestión de datos). No obstante, el autor realiza un análisis de siete ejemplos de la vida real en los que las actividades exploratorias juegan un papel importante y los compara con 51 casos de uso extraídos del *Grupo de Trabajo Público de Big Data* del NIST. El autor reconoce la evolución de los modelos y metodologías de *Data Mining* y *Data Science* más relevantes. *KDD* y *CRISP-DM* son las metodologías "canónicas".

5.7.19 *Citizen Data Scientist: A Design Science Research Method for the Conduct of Data Science Projects*

En [50] el autor menciona que actualmente las empresas buscan identificar y obtener mayor conocimiento de los datos recopilados. Como consecuencia de esta acción, se reconoce que para implementar de manera efectiva y eficiente los recursos del proyecto de ciencia de datos, las empresas descubrieron que es necesario tener un cambio de enfoque en los proyectos de Ciencia de Datos.

El enfoque descubierto lleva por nombre diseño de ciencia de investigación (DSR), el cual busca extenderse a la ciencia de datos a través de una evaluación iterativa y evaluativa. No obstante, *DSR* por sus siglas en inglés, también proporciona un paradigma de búsqueda emergente guiada que puede ser integral para encontrar conocimientos ocultos.

Este artículo además examina un caso para utilizar el método de investigación de diseño de acción elaborado (eADR) para informar el enfoque de gestión de proyectos (PM) de ciencia de datos in situ con un fabricante *global de Fortune 100*. El enfoque innovador de ciencia de datos con gestión de proyectos dio como resultado múltiples artefactos innovadores de soluciones de ciencia de datos construidos y evaluados por una docena de equipos de ciencia de datos con gestión de proyectos en la empresa durante los primeros dos años de su implementación.

5.7.20 *The Importance of Multidisciplinary in Data Science: Application of the Method in Health Sector to Telecommunication Sector*

En [52] se revela la importancia de la perspectiva multidisciplinaria en la ciencia de datos. Para ello, los autores aplicaron el método de análisis de supervivencia desarrollado en el sector salud a los datos del sector de telecomunicaciones.

Como resultado de la insuficiencia de los métodos tradicionales de ciencia de datos sobre los datos obtenidos, los autores han buscado diferentes métodos y los autores que tienen una maestría en estudios biomédicos encontraron similitud entre abandono y supervivencia. Con este artículo se trata de demostrar el beneficio de ser multidisciplinarios que quieran estudiar ciencia de datos.

5.7.21 *Applying Data Science Methods for Early Prediction of Undergraduate Student Retention*

En [53] se presenta un estudio de caso de aplicación de los métodos de ciencia de datos a una gran cantidad de datos educativos recopilados en una universidad durante 7 años. El objetivo del estudio es comprender las características importantes y derivar modelos para predecir la retención de estudiantes. Se discutieron cuestiones relacionadas con los datos del mundo real, por ejemplo, definición de variables, manejo de datos faltantes y limpieza de datos.

Se desarrolló un nuevo método de selección de características basado en la eliminación de características recursivas. Este estudio derivó características y modelos para cuatro grupos de estudiantes diferentes, los estudiantes de primera generación, los estudiantes afroamericanos, los estudiantes hispanos y los estudiantes discapacitados. Las características identificadas y los modelos predictivos construidos se compararon cruzando los cuatro grupos. (1) preprocesamiento de datos, (2) selección de características y (3) construcción de modelos predictivos.

5.7.22 Ciencia de datos y estudios globales: aportaciones y desafíos metodológicos

En [54] se parte del análisis crítico de información sobre los fundamentos conceptuales de la ciencia de datos y lo contrasta con sus posibles aplicaciones en la disciplina de los estudios globales. De esta manera, se examinan los elementos conceptuales que definen la

ciencia de datos, se establecen las concepciones paradigmáticas de los estudios globales, se analizan los vínculos posibles entre la ciencia de datos y los estudios globales y se discuten los límites y alcances que la ciencia de datos puede aportar como enfoque metodológico.

El autor destaca que la aplicación de una metodología de ciencia de datos al campo de las ciencias sociales tiene un gran impacto. La razón de esto está fundada en que normalmente las ciencias sociales realizan un análisis del comportamiento humano basado en teorías. Por otro lado, la ciencia de datos enfocada a esta área permite reconocer patrones e identificar comportamientos atípicos de un sector de la población y como es fácil darse cuenta esta puede incrementar su validez.

5.7.23 Elaboración e implementación de una propuesta metodológica para la evaluación y gestión de la calidad del aire mediante el enfoque de la ciencia de datos

En [56] se propone una metodología para medir la calidad del aire utilizando un enfoque de ciencia de datos. También, hace una comparación entre los modelos y metodologías tradicionales para medir la calidad del aire.

Esta metodología se aplica a 3 casos de estudio: el primero en la Comunidad Autónoma de Valencia (España) que cuenta con una extensa red de monitoreo de la calidad del aire; el segundo en la ciudad de Buenos Aires (Argentina) con tres estaciones de monitoreo en su zona central y, por último, el tercero en la ciudad de La Plata (Argentina).

La metodología propuesta consta de 3 etapas y es descrita de manera gráfica en el Figura 11.

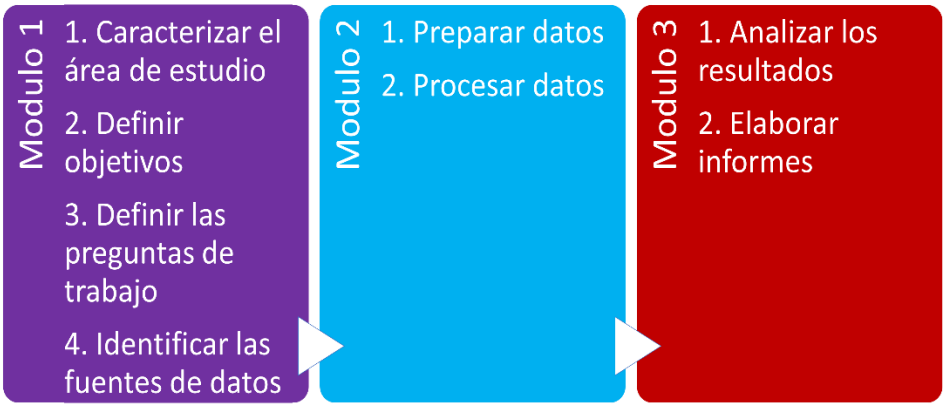

Figura 11. Metodología de gestión de la calidad del aire.[5]

5.7.24 Metodología para explorar datos abiertos de accidentalidad vial usando Ciencia de Datos: Caso Medellín

En [57] se propone una metodología para estudiar datos abiertos sobre accidentalidad vial mediante el uso de Ciencia de datos. Dicha metodología, se propone como cuatro grandes macro procesos: 1. Planificación, 2. Preparación de los datos, 3. Análisis automático y 4. Visualización de datos. El objetivo de la implementación de dicha metodología, es proveer de forma automática información estadística (exploratoria) sobre métricas de accidentalidad vial desde un enfoque de ciencia de datos. La metodología se describe a continuación:

1. Macroproceso de planificación del estudio.
2. Macroproceso de preparación de datos.
    2.1 Lectura, limpieza y procesado
    2.2 Integración de datos
3. Macroproceso de análisis automático

3.1 – Etapa descriptiva

- 3.1.1 Gravedad de accidentes
- 3.1.2 Segmentación de accidentes
- 3.1.3 Mapa de accidentalidad
- 3.1.4 Calles y carreras de accidentalidad
- 3.1.5 Evolución gráfica de la accidentalidad
- 3.1.6 Evolución tabular de la accidentalidad

[5] N.S. Represa, "Elaboración e implementación de una propuesta metodológica para la evaluación y gestión de la calidad del aire mediante el enfoque de la ciencia de datos", Ph.D., dissertation, *Univ. Valencia and Univ. De la Plata*, pp. 46, 2020. Avaliable: https://riunet.upv.es/handle/10251/144645.

3.2 – Etapa relacional

3.2.1 Métricas sobre columnas

3.2.2 Relaciones bivariadas

3.2.3 Patrones de agrupación de comunas

3.2.4 Indicadores latentes de los tipos de población

3.2.5 Factores latentes

3.2.6 Jornada y horarios

4.- Macroproceso de visualización

4.1 Diseño

4.2 Desarrollo

4.3 Validación

5.8 Perspectivas de los artículos

Se realizó un análisis cuantitativo y cualitativo de las publicaciones relevantes de metodologías de Ciencia de Datos. En el análisis cuantitativo se identificaron 3 aspectos relevantes: 1) el número de publicaciones relevantes de metodologías de Ciencia de datos, 2) la proporción de publicaciones por buscador *WEB* y 3) el número de publicaciones por repositorio.

Por otro lado, el análisis cualitativo consistió en identificar que fueron analizados un total de 3,451 artículos por título de los cuales únicamente 34 de ellos pasaron el primer filtro de palabras clave. De los 34 artículos, únicamente 24 de ellos fueron considerados como relevantes después de aplicar los criterios de inclusión y exclusión con el objeto de seleccionar las publicaciones de mejor calidad.

De los 24 artículos, se realizó una clasificación por su contenido identificando un el tipo de metodología de Ciencia de Datos que actualmente se encuentran reportadas en la literatura. Asimismo, dentro de esta clasificación se consideraron las siguientes categorías:

- Metodologías que nacieron en minería de datos, pero se han utilizado en Ciencia de Datos
- Metodologías que son adecuaciones de la minería de datos
- Metodologías utilizadas para dar solución a un proyecto de Ciencia de Datos
- Metodología propuesta por la comunidad científica

- Metodologías de investigación con un enfoque a Ciencia de Datos
- Metodologías para la gestión de proyectos con un enfoque de Ciencia de Datos
- Propuestas de flujo de trabajo para la Ciencia de Datos
- Publicaciones que destacan el uso de una determinada metodología de Ciencia de Datos
- Contienen las palabras clave de la cadena de búsqueda, pero no proponen una metodología, ofrecen diversas comparativas e información relacionadas con las metodologías disponibles en ciencia de datos.

Capítulo 6

Resultados

6 Resultados

Para presentar los resultados de la investigación, se consideró un análisis mixto el cual contempla el análisis cuantitativo y cualitativo de las publicaciones. Ambos enfoques permitirán identificar cuáles son las publicaciones relevantes de metodologías de Ciencia de Datos y sus tendencias actuales.

Para identificar las publicaciones analizadas en este apartado, se utilizaron las cadenas de búsqueda (*Data Science and Methodology*, *Data Science and Method*, *Data Science and Process* y *Data Science and Procedure*) generadas a partir de la pregunta de investigación *"¿Cuáles son las publicaciones relevantes de metodologías de Ciencia de Datos reportadas en la literatura?*". En consecuencia, se generaron un total de 3,451 artículos a los cuales se aplicaron criterios de inclusión y exclusión, quedando únicamente un total de 24 publicaciones. Dichos criterios son los siguientes: título, año de publicación, número de citas e idioma.

6.1 Análisis cuantitativo

6.1.1 Publicaciones por año

Con el análisis cuantitativo de este apartado, se logró identificar en que año se realizaron la mayor cantidad de publicaciones acerca de metodologías de Ciencia de Datos. Esta información es relevante, ya que confirma que actualmente existe un crecimiento en la cantidad de publicaciones de metodologías de Ciencia de Datos, sin embargo, se identifica como una cantidad reducida en comparación con las publicaciones de Ciencia de Datos en general.

En la Figura 12, se describe que el año con mayor cantidad de publicaciones fue el 2019, con un total de 8 publicaciones sobre metodologías de Ciencia de Datos y los menores fueron el 2014 y 2016 con solo una publicación respectivamente.

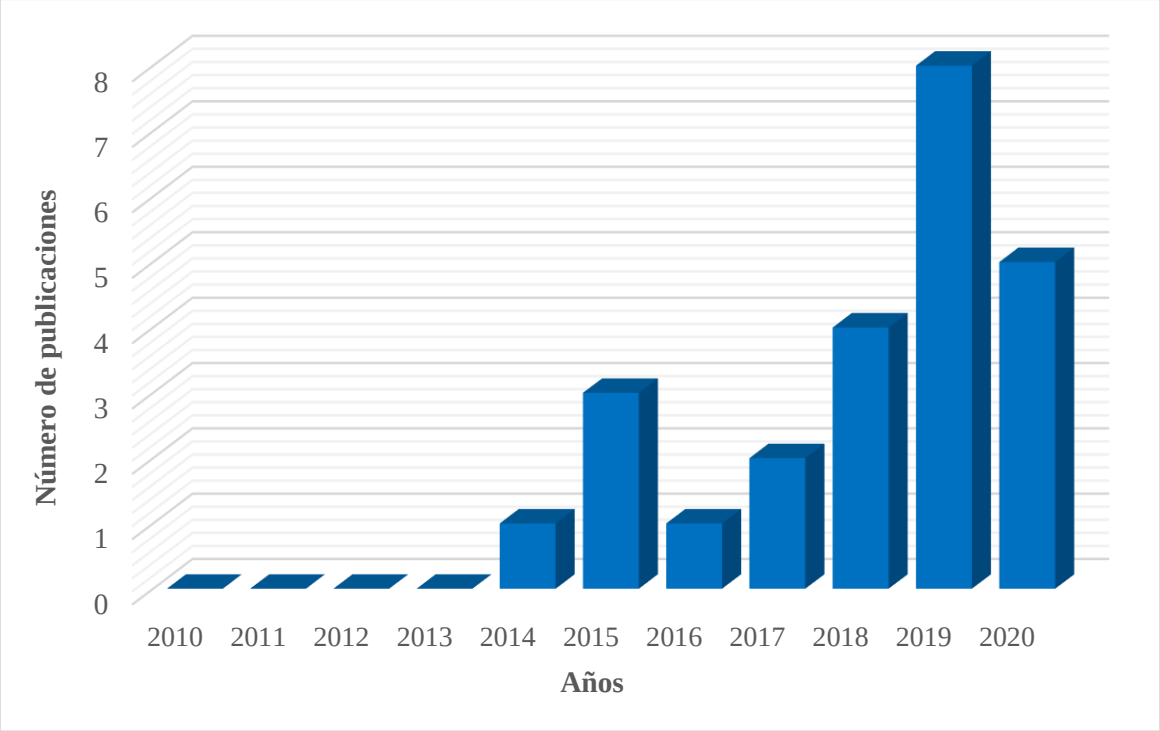

Figura 12. Número de publicaciones de metodologías de Ciencia de Datos por año

6.1.2 Publicaciones por buscador

Para la selección de buscadores de esta investigación, se consideraron aquellas bibliotecas digitales y sistemas de indización que presentan una alta adopción por parte de la comunidad científica. Alice en [58] y Daren en [59], reconocen que *Google Scholar*, *Scopus* y *IEEE Xplore* son los buscadores más utilizados por la comunidad científica. Asimismo, también consideraron los buscadores seleccionados por Moustaka *and et al* en [11] para la revisión sistemática enfocada a las ciudades inteligentes. Para la presente investigación el buscador *Scopus* fue descartado debido a que el Centro Nacional de Investigación y Desarrollo tecnológico (CENIDET), no cuenta con acceso para realizar las búsquedas. Con la finalidad de reemplazar este buscador se identificó que *WorldWideScience* es otro buscador ampliamente aceptado por la comunidad científica, además de ser el único que está potenciado por *Deep Technologies* compañía especializada en la extracción de conocimiento de la *Deep Web*.

La Figura 13, está dividida en círculos que representan los tres buscadores utilizados en la presente investigación: Google Académico, *IEEE Xplore* del *Institute of Electrical and Electronics Engineers* y el *WorldWideScience*. De color azul, es representado el buscador

Google Académico el cual identificó 26 publicaciones; de color naranja, es representado el buscador *IEEE xplore* el cual identificó seis publicaciones y, por último, de color rojo, es representado el buscador WorldWideScience el cual identificó dos publicaciones.

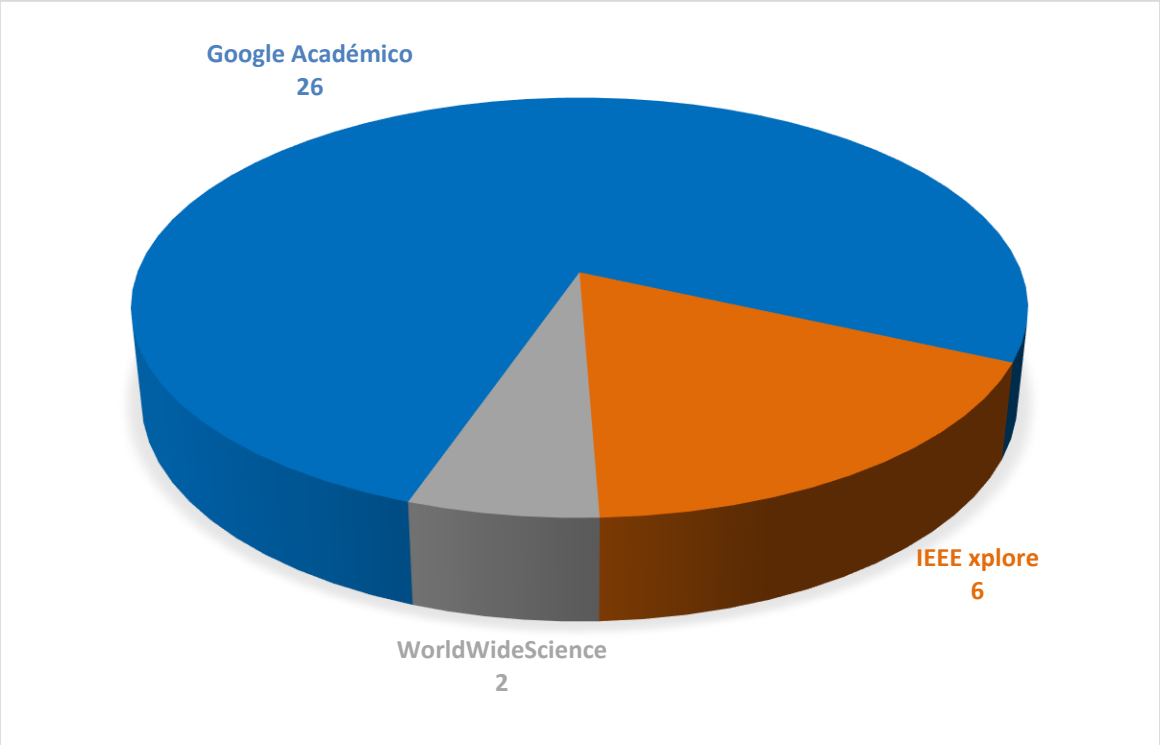

Figura 13. Distribución de publicaciones de metodologías de Ciencia de Datos por buscador

6.1.3 Publicaciones por repositorio

Los repositorios a los cuales se tiene acceso por parte del CENIDET son: *IEEE, Springer Link, ACM digital library y Science Direct.* Sin embargo, se identificaron otros repositorios donde se encontró al menos una publicación sobre metodologías de Ciencia de Datos. El gráfico de barras descrito en la Figura 14, describe el número de publicaciones identificadas por repositorio en esta investigación. Los repositorios con mayor número de publicaciones identificadas fueron *IEEE* y *Springer*. Al utilizar las cadenas de búsqueda, se identificó que el repositorio *ACM digital library*, no generó publicaciones en las cuales apareciera al mismo tiempo en el título las palabras clave "Ciencia de Datos" y "Metodología", sin embargo, dichas cadenas de búsqueda generaron publicaciones sobre proyectos de Ciencia de Datos

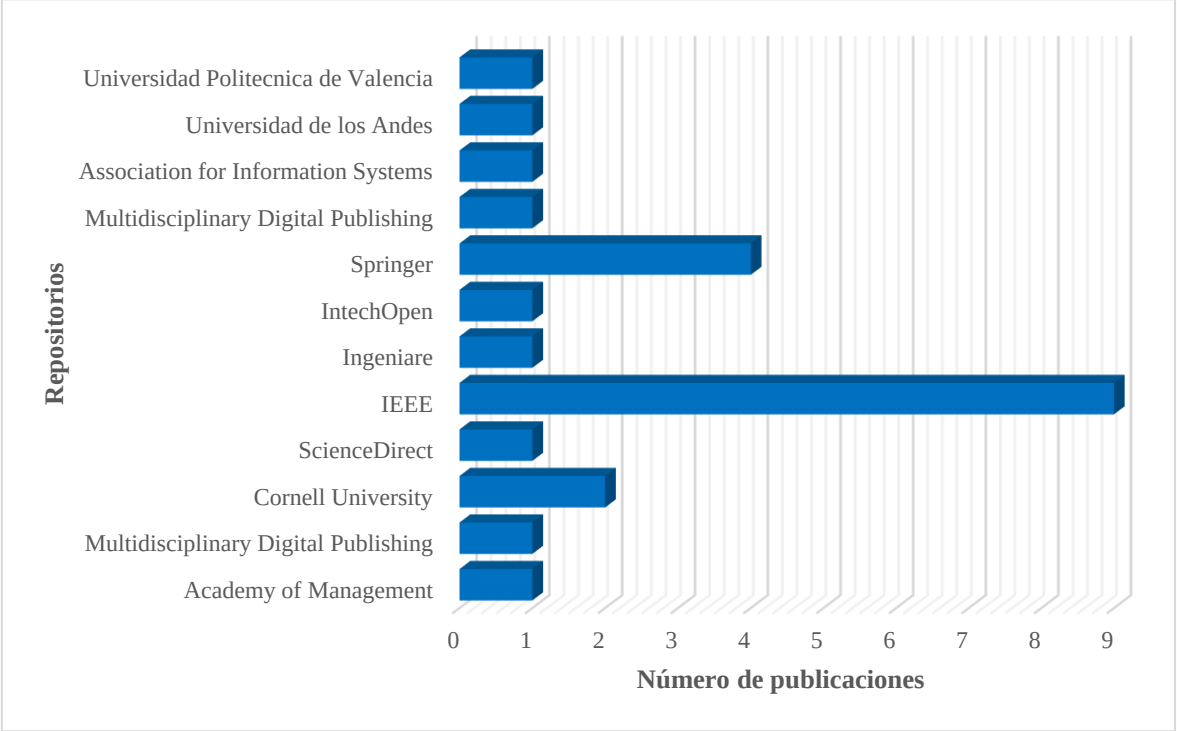

Figura 14. Número de publicaciones de metodologías de Ciencia de Datos por repositorio

6.1.4 Publicaciones más citadas

El análisis cuantitativo de las publicaciones más citadas, ayudó a identificar a los artículos más destacados de metodologías de Ciencia de Datos. Asimismo, se identificó que Jeffrey S. Saltz es el autor que más publicaciones aportó a esta investigación con un total de seis identificadas, de las cuales al aplicar criterios de inclusión y exclusión únicamente dos pasaron dicho filtrado.

En la Figura 15, se describen las publicaciones más citadas de cada autor. De color naranja observamos que la publicación con mayor número de citas fue [41] con 50 citas. Adicionalmente, se identificó que el autor Jeffrey S. Saltz tiene dos publicaciones relevantes.

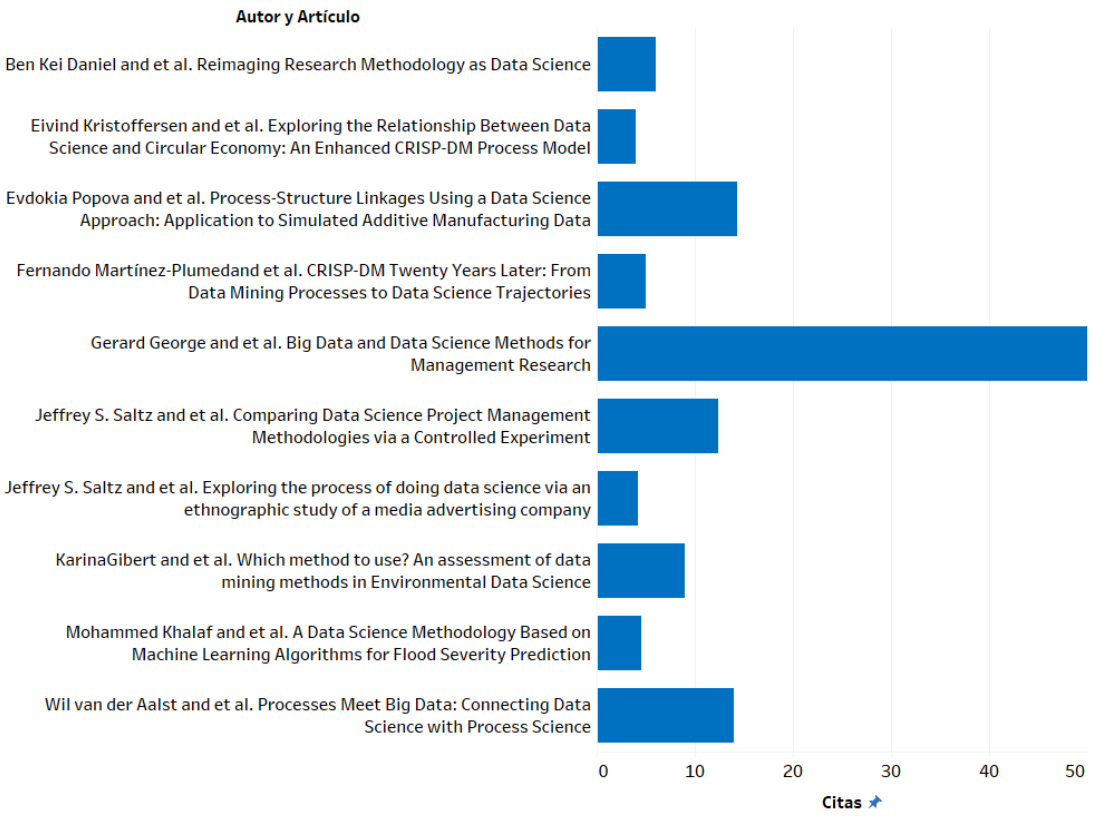

Figura 15. Publicaciones de metodologías de Ciencia de Datos con más citas

6.1.5 Palabras clave

El análisis de las palabras clave de las publicaciones sobre metodologías de Ciencia de Datos es muy útil para identificar las tendencias actuales de dicha temática. Se identificó que la comunidad científica comenzó la incorporación de la Ciencia de Datos a dominios en los que la toma de decisiones es fundamental. Esta incorporación dio origen a la generación de metodologías específicas y generales de Ciencia de Datos, ya que para aplicar la Ciencia de Datos a cierto dominio es necesario seguir una metodología que garantice la reproducibilidad del proceso de los datos.

La Figura 16, mediante un gráfico de barras describe las 19 palabras clave más repetidas en esta investigación. La palabra *Data Science* es la que más se repite como *index term* o como palabra clave con un total de 20 veces. Se identificó que las palabras *Big Data*, *Data*, *Data Mining* y *Machine Learning* son palabras presentes en las publicaciones de metodologías de Ciencia de Datos.

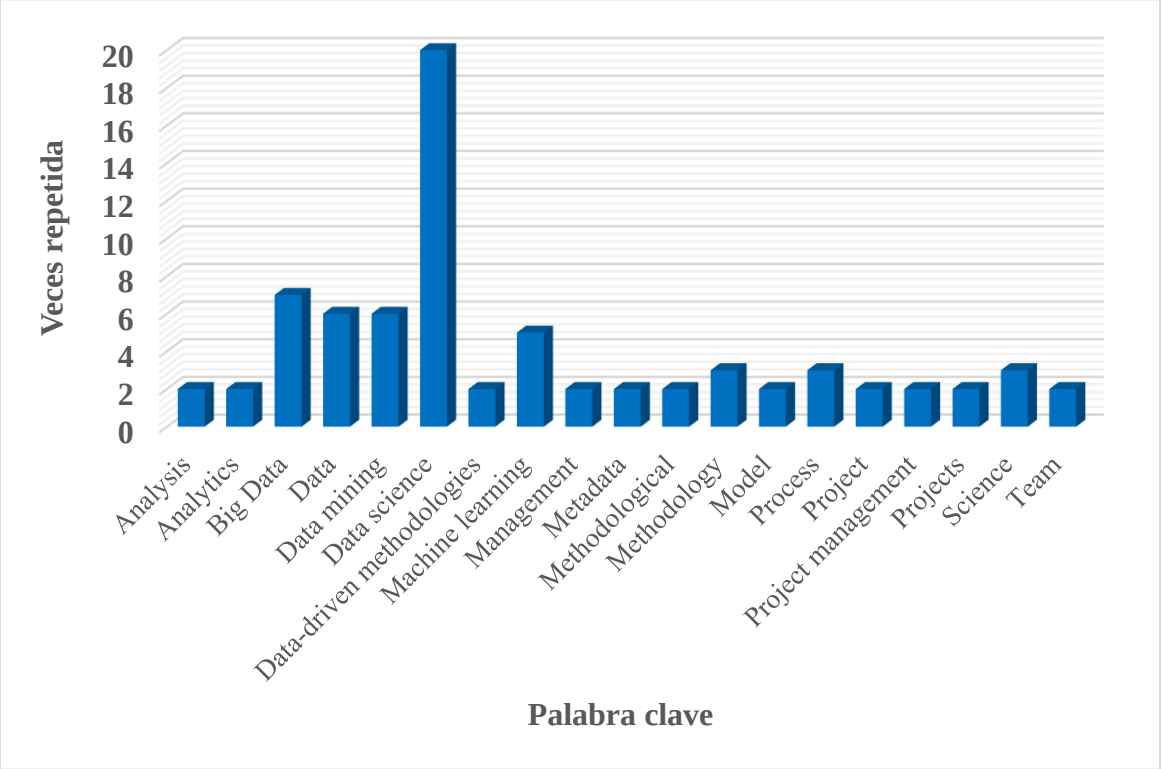

Figura 16. Frecuencia de aparición de palabras claves en los artículos seleccionados de metodologías de Ciencia de Datos.

6.1.6 Cadenas de búsqueda

Es importante señalar que debido a que la Ciencia de datos es emergente a pesar de tener más de 50 años, el número de publicaciones encontradas claramente son en el idioma inglés, por lo que se recomienda el uso de cadenas de búsquedas en este idioma. No obstante, en español se identificaron cuatro publicaciones relevantes con la cadena de búsqueda "Ciencia de Datos" y "Metodología". Esto nos indica que pese a que la comunidad científica de habla hispana ha comenzado con las investigaciones sobre metodologías de dicha ciencia, la comunidad científica de habla inglesa presenta mayor contenido, ya que las cuatro cadenas de búsquedas utilizadas en este idioma identificaron al menos una publicación relevante.

En la Figura 17 se describe cuál cadena de búsqueda dio origen a encontrar más publicaciones sobre metodologías de Ciencia de Datos. La cadena de búsqueda que presentó mayor efectividad fue "*Data Science and Methodology*" la cual ayudó a encontrar un total de 17 publicaciones.

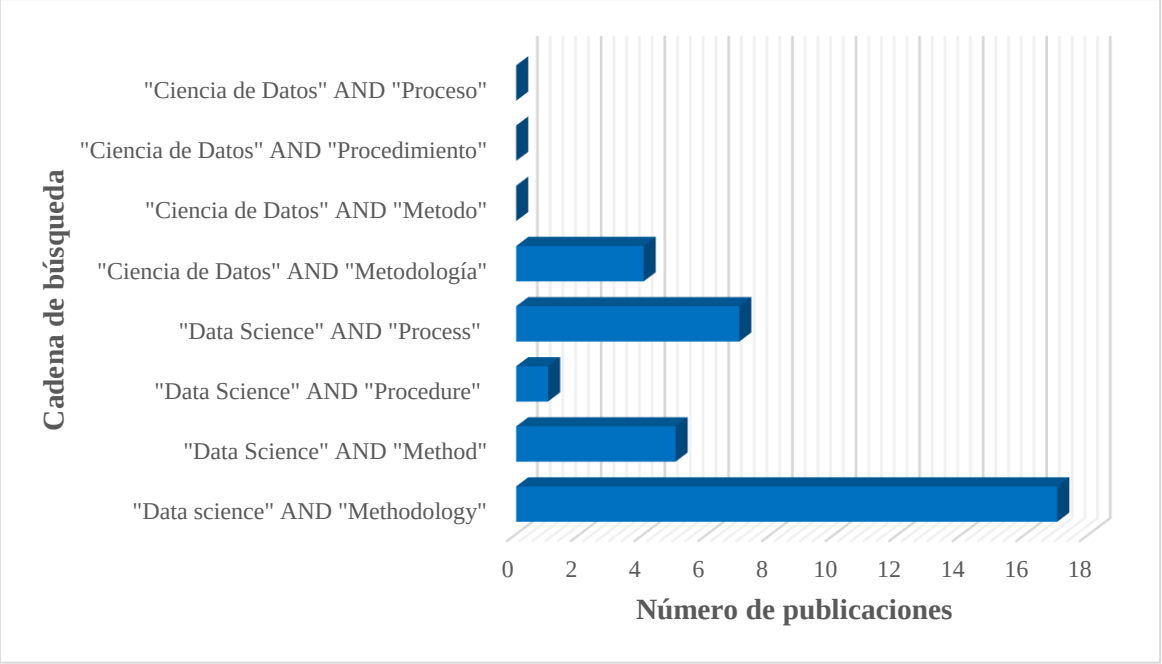

Figura 17. Efectividad de cadenas de búsqueda utilizadas en la búsqueda de metodologías de Ciencia de Datos

6.2 Análisis cualitativo

El presente análisis cualitativo involucró a las 24 publicaciones que pasaron los criterios de inclusión y exclusión. En este análisis se identificó lo siguiente: palabras clave de publicaciones, panorama general de metodologías de Ciencia de Datos y la clasificación de publicaciones de metodologías.

6.2.1 Palabras clave de publicaciones

En la Figura 18 se presenta mediante un diagrama de nube las palabras claves que están contenidas en los 24 artículos identificados como relevantes. Se muestra con mayor tamaño aquellas palabras que se repitieron más y con menor tamaño aquellas que no fueron tan repetidas.

Figura 18. Palabras clave de publicaciones

6.2.2 Panorama general de metodologías de Ciencia de Datos

Después de dar lectura y analizar cada una de las publicaciones, se identificaron los tipos de metodologías actualmente reportadas en la literatura. Estas metodologías son propuestas por la comunidad científica y por la industria. Asimismo, tenemos propuestas por campos del conocimiento como lo son la gestión de proyectos y flujos de trabajo. Por último, se identificó que en [47] se mencionan como complementos a las metodologías de Ciencia de Datos a *Cross Industry Standard Process for Data Mining* (*CRISP-DM*), *Knowledge Discorvery in Data Base* (*KDD*) y *Sampling, Exploring, Modifying, Modeling, and Assessing* (*SAS SEMMA*). En la Figura 19, se describe de manera general la clasificación de las metodologías identificadas en los artículos identificados como relevantes de metodologías de Ciencia de Datos.

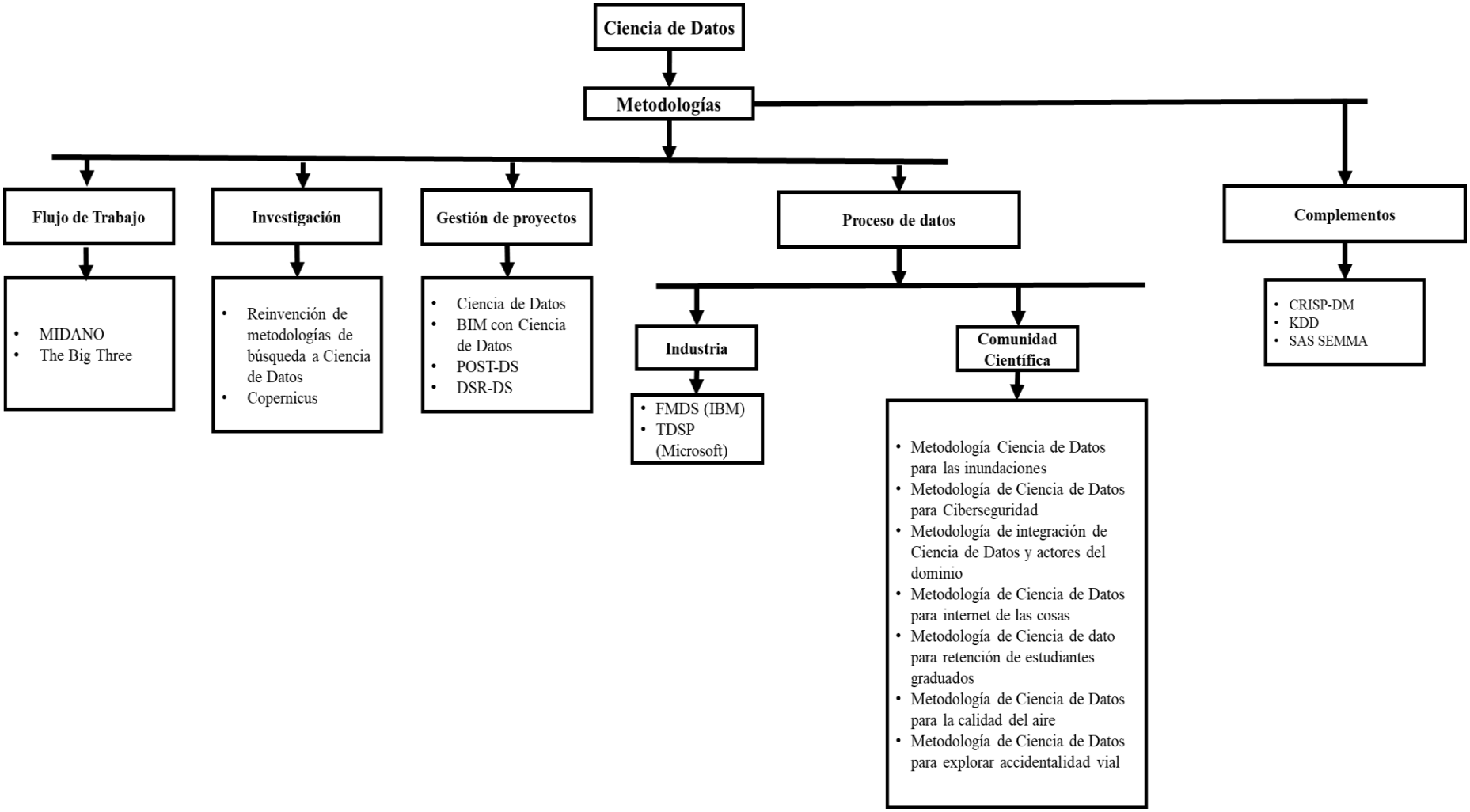

Figura 19. Metodologías de Ciencia de Datos identificadas

1) Metodologías de Ciencia de Datos propuestas por la comunidad científica

Esta categoría presenta las metodologías de Ciencia de Datos que fueron propuestas por la comunidad científica para aplicarlas a un dominio específico. En este caso, los dominios donde se ha evidenciado la aplicación de la Ciencia de Datos son: meteorológico (inundaciones), ciberseguridad, *Internet of the things* (IoT), calidad del aire y accidentalidad vial. También se presenta una propuesta de integración entre actores del dominio con Ciencia de Datos. De manera general las metodologías propuestas por la comunidad científica son: a) Metodología de Ciencia de Datos para las inundaciones; b) Metodología de Ciencia de Datos para ciberseguridad; c) Metodología de integración de Ciencia de Datos y actores del dominio; d) Metodología de Ciencia de Datos para internet de las cosas; e) Metodología de Ciencia de Datos para retención de estudiantes graduados Metodología de Ciencia de Datos para la calidad del aire y f) Metodología de Ciencia de Datos para explorar accidentalidad vial.

En la Tabla 4, se describen los pasos identificados en cada una de las metodologías propuestas por la comunidad científica para el proceso de datos. La metodología de Ciencia de Datos de ciberseguridad es la única metodología que toma como base la propuesta de IBM (FMDS), el resto no presenta de manera explícita cuál fue la base para su generación.

Tabla 4. Comparación del proceso de datos propuesto por la comunidad científica

	Metodología de Ciencia de Datos para inundaciones	**Metodología de Ciencia de Datos para Ciberseguridad**	**Metodología de integración de Ciencia de Datos y actores del dominio**	**Metodología de Ciencia de Datos para internet de las cosas**	**Metodología de Ciencia de Datos para retención de estudiantes graduados**	**Metodología de Ciencia de Datos para medir la calidad del aire**	**Metodología de Ciencia de Datos para explorar la accidentalidad Vial**
1	Preprocesamiento de los datos con herramientas	Entendimiento del negocio	Diseño de algoritmos	Análisis de los requerimientos	Compilación de tabla de definición de variables	Caracterización del área de estudio	Planificación del estudio
2	Preparación de datos	Enfoque analítico	Evaluación basada en datos	Entendimiento de requerimientos	Construcción de base de datos relacional	Definir los objetivos	Lectura, limpieza y procesado de datos
3	División de conjunto de datos	Requerimientos de datos	Evaluación interactiva y colaborativa	Formulación de objetivos analíticos	Limpieza de datos	Definir las preguntas de trabajo	Integración de datos
4	Construcción de modelos	Recopilación de Datos	Consolidación del conocimiento	Establecimiento de relación entre variables	Caracterización de datos de retención	Identificar fuentes de datos	Análisis de datos
5	Pruebas de modelos	Entendimiento de los datos	-	Recopilación de los datos	Selección de características basadas en la eliminación recursiva	Preparar datos	Análisis relacional
6	Selección de modelo	Preparación de los datos	-	Limpieza de datos	Resultados de selección de características	Procesar datos	Visualización de datos
7	-	Modelado	-	Análisis de datos	Clasificación de modelos	Analizar resultados	Validación
8	-	Evaluación del modelo	-	Procesamiento de datos	Evaluación de modelos	Elaborar informes	-
9	-	Despliegue	-	Visualización de datos	-	-	-
10	-	Retroalimentación	-	-	-	-	-

2) Metodologías de Ciencia de Datos propuestas por la industria

En esta categoría se presentan las dos metodologías de Ciencia de Datos propuestas por la industria son *Foundational Methodology for Data Science* (FMDS) [60] y *Team Data Science Process* (TDSP) [61]. Estas metodologías son de las compañías de IBM y Microsoft respectivamente.

En la Tabla 5 se describen los pasos identificados en cada una de las metodologías propuestas por la industria para el proceso de datos. La metodología de Ciencia de Datos de IBM (FMDS) es de uso general para cualquier campo de investigación. Por otro lado, la metodología de Microsoft (TDSP) también es de uso general, pero está limitada al uso del paquete de programas implementados por Microsoft.

Tabla 5. Comparación del proceso de datos propuesto por la industria

n.	**Metodología fundamental para Ciencia de Datos**	**Proceso de Ciencia de Datos en equipo**
1	Comprensión del negocio	Definición de objetivos
2	Enfoque analítico	Identificación de fuentes de datos
3	Requisitos de datos	Adquisición de datos
4	Recopilación de datos	Entendimiento de los datos
5	Comprensión de datos	Generación de modelo
6	Preparación de datos	Entrenamiento del modelo
7	Modelado	Implementación del modelo
8	Evaluación	Validación del modelo
9	Implementación	-
10	Retroalimentación	-

3) Metodologías de gestión de proyectos para Ciencia de Datos

Se identificó que se ha comenzado a incorporar a la Ciencia de Datos para la gestión de proyectos. De manera general, el interés de gestionar los proyectos utilizando la Ciencia de Datos ha tomado relevancia por la gran ayuda para la toma de decisiones basadas en datos. Las metodologías de gestión de proyectos de Ciencia de Datos fueron: a) Metodología de gestión de proyectos de Ciencia de Datos; b) Metodología BIM de gestión de proyectos de Ciencia de Datos; c) Metodología de gestión de proyectos POST-DS y d) Metodología DRS para proyectos de Ciencia de Datos.

En la Tabla 3 se describen las metodologías identificadas para gestión de proyectos con Ciencia de Datos. Es importante destacar que estas metodologías están enfocadas en la administración de proyectos utilizando un enfoque de Ciencia de Datos. En todos los casos, las metodologías siempre plantean el trabajo multidisciplinario que en estos proyectos interactúan.

Tabla 6. Comparación de metodologías de gestión de proyectos para Ciencia de Datos

n.	**Metodología de gestión de proyectos de Ciencia de Datos**	**Metodología BIM de gestión de proyectos de Ciencia de Datos**	**Metodología de gestión de proyectos POST-DS**	**Metodología DSR para proyectos de Ciencia de Datos**
1	Asignación de roles de proyecto	Representación del glosario del negocio	Análisis del proyecto	Diagnóstico del proyecto
2	Documentación del proceso	Ligamiento del negocio con los requerimientos de datos	Asignación de roles	Diseño del proyecto
3	Preparación de los datos	Catalogar representaciones de datos físicos	Ejecución del proceso de datos	Implementación del proyecto
4	Análisis de los datos	Mapeo detallado de representaciones de datos físicos	Evaluación de resultados	Seguimiento del proyecto
5	Diseminación de resultados	Ligamiento de los datos	Diseminación de resultados	Diseminación de resultados

4) Metodologías de flujo de trabajo para Ciencia de Datos

En esta categoría se identificaron dos metodologías de flujo de trabajo para Ciencia de Datos reportadas en la literatura, por un lado, la metodología de flujo de trabajo *MIDANO* y por el otro, la metodologia de flujo de trabajo "*The Big Three*"

En la Tabla 7 se describen las metodologías de flujo de trabajo para Ciencia de Datos reportadas en la literatura. Ambas metodologías están enfocadas al flujo de trabajo dentro de las organizaciones.

Tabla 7. Comparativa de metodologías de flujo de trabajo

n.	**Metodología de flujo de trabajo MIDANO**	**Metodología de flujo de trabajo "The Big Three"**
1	Identificación de las fuentes de extracción del conocimiento en la organización	Identificación del problema
2	Preparación y tratamiento de los datos	Análisis de causas del problema
3	Desarrollo de herramientas de minería de datos	Planteamiento de solución

5) Metodologías de Investigación con Ciencia de Datos

En esta categoría se identificaron dos metodologías de investigación con Ciencia de Datos reportadas en la literatura; en primer lugar, Reinvención de metodologías de investigación a Ciencia de Datos y en segundo, la metodología *Copernicus* para investigación.

La Tabla 8 describe las 2 metodologías de investigación con Ciencia de Datos. La metodología de reinvención utiliza su propuesta metodológica en la cual aplica Ciencia de Datos para comparar las metodologías de investigación de 144 instituciones. Por otro lado, la metodología *Copernicus* está enfocada en la observación del planeta tierra utilizando la Ciencia de Datos.

Tabla 8. Comparativa de metodologías de investigación

n.	**Reinvención de metodologías de investigación**	**Metodologías Copernicus para investigación**
1	Estrategia de muestreo	Análisis de las fuentes de datos
2	Segmentación de participantes	Establecimiento del proceso de datos
3	Análisis de datos	Implementación lógica de la arquitectura de la investigación
4	Hallazgos y discusiones	Análisis de resultados
5	-	Conclusiones de la investigación

6) Complementos de las metodologías de Ciencia de Datos.

Se identificaron las metodologías que son complementos utilizados para realizar Ciencia de Datos (CRISP-DM, KDD y SAS-SEMMA). La identificación de complementos se basa en que estas metodologías no fueron creadas para Ciencia de Datos.

6.3 Clasificación de metodologías de Ciencia de Datos

En esta sección, se muestra la clasificación propuesta al realizar el análisis de las publicaciones seleccionadas con la revisión sistemática del tema "Metodologías de Ciencia de Datos". La clasificación propuesta es la siguiente:

- Metodologías que nacieron en minería de datos, pero se han utilizado en Ciencia de Datos
- Metodologías que son adecuaciones de la minería de datos
- Metodologías utilizadas para dar solución a un proyecto de Ciencia de Datos
- Metodología propuesta por la comunidad científica
- Metodologías de investigación con un enfoque a Ciencia de Datos
- Metodologías para la gestión de proyectos con un enfoque de Ciencia de Datos
- Propuestas de flujo de trabajo para la Ciencia de Datos
- Publicaciones que destacan el uso de una determinada metodología de Ciencia de Datos
- Contienen las palabras clave de la cadena de búsqueda, pero no proponen una metodología, ofrecen diversas comparativas e información relacionadas con las metodologías disponibles en ciencia de datos.

Es importante destacar que en la presente clasificación se consideró la información extraída durante la revisión sistemática. Esta información fue útil para clasificar las publicaciones de acuerdo con su contenido. En la Tabla 9, se muestra la clasificación propuesta que incluye las publicaciones clasificadas de acuerdo con su categoría.

Tabla 9. Clasificación de metodologías de Ciencia de Datos

N.	Clasificación	Artículos
1	Metodologías que nacieron en minería de datos, pero se han utilizado en Ciencia de datos	[47]
2	Metodologías que son adecuaciones de la minería de datos	[45]
3	Uso de metodología para dar solución a un proyecto de Ciencia de Datos	[31],[44],[53]
4	Propuesta metodológica de la comunidad científica	[30],[32],[37],[46],[55],[56]
5	Proponen metodologías de investigación con un enfoque a Ciencia de Datos	[27],[41],[50]
6	Proponen el uso de una metodología para la gestión de proyectos con un enfoque de Ciencia de Datos	[28]
7	Proponen un flujo de trabajo de Ciencia de Datos	[29],[38],[43]
8	Destacan el uso de una metodología de Ciencia de Datos	[54]
9	Contienen las palabras clave de la cadena de búsqueda, pero no proponen una metodología, ofrecen diversas comparativas e información relacionadas con las metodologías disponibles en ciencia de datos.	[34],[39],[40],[42],[52],[10]

Capítulo 7

Conclusiones y Trabajo futuro

7 Conclusiones y trabajo futuro

En este capítulo, se brindan las conclusiones de la investigación sobre metodologías de Ciencia de Datos. En la Sección 7.1, se describe la clasificación de las metodologías de Ciencia de Datos y los hallazgos identificados el análisis cuantitativo y cualitativo; en la Sección 7.2, se describen el trabajo futuro generado a partir de la presente investigación.

7.1 Conclusiones

Con el uso de una metodología de revisión de la literatura en este caso la revisión sistemática, la cual es una herramienta fundamental para la identificación y selección de publicaciones relevantes. Para realizar nuestra investigación se seleccionó la metodología de la Dra. Barbara Kichenham [7].

Se identificó que existen diversos tipos de metodologías en el ámbito de la Ciencia de Datos, por ejemplo, para el proceso de datos, para la gestión de proyectos, para realizar investigaciones específicas y para organizar el flujo de trabajo en ciencia de datos. Asimismo, se identificó que los dos principales proveedores de metodologías en la Ciencia de Datos son la comunidad científica la cual publica en repositorios y/o bases de datos reconocidos por esta comunidad y por el otro lado la industria, la cual publica en sus propias páginas bajo los llamados *white papers*. Para ilustrar esto, en la Figura 20 se describe a los dos proveedores de metodologías de Ciencia de Datos y la clasificación propuesta para las metodologías de Ciencia de Datos.

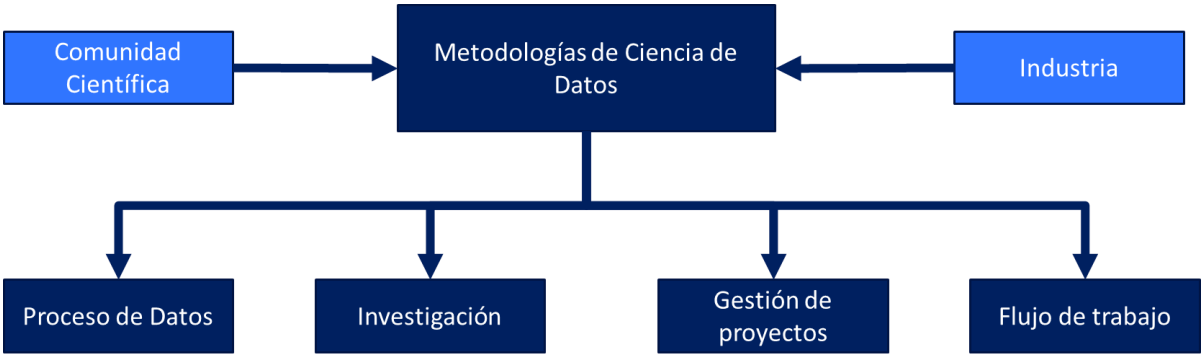

Figura 20. Clasificación de metodologías de Ciencia de Datos.

Las metodologías para el proceso de Ciencia de datos están enfocadas en el área técnica de Ciencia de Datos, es decir, en ellas se presenta de manera explícita cómo debe ser el proceso de gestión de los datos para obtener una respuesta que sea de ayuda para la toma de decisiones en un dominio específico. Asimismo, se identificó que las metodologías propuestas por la industria tienen un enfoque generalista aplicable a cualquier dominio, mientras que las metodologías propuestas por la comunidad científica son propuestas que son creadas para un dominio específico.

Las metodologías para la gestión de proyectos de Ciencia de Datos están enfocadas en el área de la administración de proyectos de Ciencia de Datos, los cuales tienen la característica de integrar actores del dominio de aplicación y científicos de datos en proyectos que involucran grandes volúmenes de información. En ellas, podemos identificar cómo deben interactuar las personas en un proyecto con un enfoque en datos. Este tipo de metodologías son propuestas generalmente por la comunidad científica. Por parte de la industria no se identificaron propuestas metodológicas para la gestión de proyectos de Ciencia de Datos.

Las metodologías para la investigación con Ciencia de Datos están enfocadas para investigaciones que utilizan como medio a los datos. En ellas podemos encontrar cómo debe de gestionarse una investigación cuando se manejan grandes volúmenes de información con el objetivo de encontrar conocimiento en los datos. Este tipo de metodologías son propuestas por la comunidad científica.

Las metodologías para crear un flujo de trabajo de Ciencia de Datos están enfocadas en la generación de flujos de trabajo para los proyectos de Ciencia de Datos. Estas metodologías se enfocan en estructurar y formalizar los requisitos comerciales en grandes proyectos con uso intensivo de datos.

Los complementos de metodologías de Ciencia de Datos, son todas aquellas metodologías para el proceso de gestión de datos que fueron originadas en otros campos del conocimiento, pero se utilizan en la Ciencia de Datos. Estas metodologías en algunos casos también se reconocen como “canónicas” ya que fueron las primeras en gestionar grandes volúmenes de información. Algunas de estas metodologías complementarias son: *KDD*, *CRISP-DM* y *SAS SEMMA*.

7.1.1 Observaciones durante el proceso de investigación

1) El año 2018 es el año con mayor número de publicaciones sobre metodologías de Ciencia de Datos con un total de ocho. Asimismo, se identificó que el interés de la comunidad científica por las metodologías surge en 2014.
2) El buscador Google académico, identificó un total de 26 publicaciones con el uso de las cadenas de búsqueda propuestas. Por su parte, *IEEE* y *WorldWideScience* también identificaron publicaciones con un total de seis y dos respectivamente.
3) Los repositorios de *IEEE* y *Springer* presentan una clara ventaja de publicaciones respecto a *ACM digital library*.
4) En el repositorio *ACM digital library*, pese a ser uno de los más reconocidos por la comunidad científica en el área de ciencias de la computación, no se identificaron publicaciones en las cuales por título contengan las palabras clave empleadas en otros repositorios.
5) Jeffrey S. Saltz es el autor con más publicaciones de metodologías de Ciencia de Datos con un total de 6 identificadas y únicamente 2 relevantes considerando los criterios de calidad definidos en la presente investigación.
6) La publicación más citada es *Big Data and Data Science Methods for Management Research* con un total de 50 citas.
7) Se identificó que la comunidad científica comenzó la incorporación de la Ciencia de Datos a dominios en los que la toma de decisiones es fundamental. Esta incorporación dio origen a la generación de metodologías específicas y generales de Ciencia de Datos, ya que para aplicar la Ciencia de Datos a cierto dominio es necesario seguir una metodología que garantice la reproducibilidad del proceso de los datos.
8) El idioma dominante de las publicaciones de metodologías de Ciencia de Datos es el inglés.
9) La cadena de búsqueda más efectiva fue *Data Science and Methodology* la cual identificó un total de 17 publicaciones.
10) La metodología de Ciencia de Datos de ciberseguridad, es la única metodología que toma como base la propuesta de IBM (FMDS), el resto no presenta de manera explícita cual fue la base para su generación.

11) Todas las metodologías propuestas por la comunidad científica, de manera general convergen en el proceso de Ciencia de Datos general: análisis de datos, procesamiento de datos y análisis de resultados.
12) La metodología de Ciencia de Datos de IBM (FMDS) es de uso general para cualquier campo de investigación. Por otro lado, la metodología de Microsoft (TDSP) también es de uso general, pero está limitada al uso del paquete de programas implementados por Microsoft.
13) Las metodologías de gestión de proyectos con Ciencia de Datos están enfocadas en la administración del proyecto utilizando un enfoque de Ciencia de Datos. En todos los casos, las metodologías siempre plantean el trabajo multidisciplinario que en dichos proyectos interactúan.
14) Se identifica que las metodologías de Ciencia de Datos presentan dos proveedores directos, por un lado la comunidad científica y por otro la industria.
15) El uso de una metodología de Ciencia de Datos general no ha sido especificado en la literatura.
16) De manera complementaria, se consultó a ocho científicos de datos que actualmente trabajan en la industria. En ningún caso expresaron explícitamente el uso de una metodología específica de Ciencia de Datos, y reconocieron el uso de metodologías de minería de datos como la base para el proceso de datos.
17) Se identifica una tendencia a la implementación de la Ciencia de Datos para ayudar a la toma de decisiones en empresas, institutos o grupos de investigación.
18) El uso de la Ciencia de Datos es incorporado en dominios en los cuales hasta hace unos años se tomaban decisiones basadas en experiencias .

7.2 Trabajo futuro

Las metodologías de Ciencia de Datos seguirán teniendo suma relevancia en los próximos años y la generación de metodologías aumentará, por lo que se propone actualizar la presente revisión sistemática cada 2 años, con la finalidad de aumentar el número de metodologías identificadas. Se recomienda la ampliación de la búsqueda a otros idiomas con la finalidad de realizar búsquedas en otras bases de datos.

De igual forma, se propone realizar un análisis comparativo de las metodologías de Ciencia de Datos para el proceso de datos, metodologías para la gestión de proyectos, metodologías de investigación y de flujo de trabajo, con el objetivo de brindar información más detallada a los científicos de datos para poder seleccionar la metodología que mejor se adapte a sus necesidades de investigación.

Referencias

[1] J.W. Turkey, “The Future of Data Analysis”, *The Annals of Mathematical Statistics*, vol. 33, no. 1, Mar., pp. 1-67, 1962.

[2] L.S. Méndez, “Desarrollo de una aplicación de Ciencia de Datos”, Tesis de Maestría, Depto. CC, CENIDET, Cuernavaca, Mor., México, 2018.

[3] D. Donoho, “50 years of data science”, *Journal of Computational and Graphical Statistics*, vol. 26, no. 4, Aug., pp. 745-766, 2017.

[4] P. Naur, “Concise Survey of Computer Methods”, Sweden: Petrocelli, 1975.

[5] I.F. Gonzáleza, G. Urrútia y P. Alonso-Coello, “Revisiones sistemáticas y metaanálisis: bases conceptuales e interpretación”, *Revista española de cardiología*, vol. 64, no. 8, pp. 688-696, 2011.

[6] J. Mcguire, “What works in correctional intervention?”, in Offender Rehabilitation in Practice: Implementing and Evaluating Effective Programs, Toronto: *John Wiley & Sons Ltd.*, 2001, ch. 2, sec. 1, p.p. 25-43.

[7] B. Kitchenham y S. Charters, “Guidelines for Performing Systematic Literature Reviews in Software Engineering”, Technical Report, United Kindom, Keele University, vol. 1, pp. 1-65, 2007.

[8] A.V. Vega , “Revisión del estado del arte de los algoritmos K-means y sus Mejoras”, Tesis de Maestría, Depto. CC, CENIDET, Cuernavaca, Mor., México, 2017.

[9] L. Cao, “Data Science: A Comprehensive Overview”, *ACM Computing Surveys*, vol. 50, no. 43, pp. 1-42, 2017.

[10] I. Martínez, E. Viles Y I.G. Olaizola, “Data Science Methodologies: Current Challenges and Future Approaches”, *Big Data Research*, vol. 24, pp. 1-18, 2021.

[11] V. Moustaka, A. Vakali, L.G. Anthopoulos, “A Systematic Review for Smart City Data Analytics”, *ACM Computing Surveys*, vol. 51, no. 5, pp. 103-144, 2018.

[12] Y. Chen, H. Chen, A. Gorkhali, Y. Lu, Y. Ma y L. Li, “Big data analytics and big data science: a survey”, *Journal of management analytics.*, vol. 3, pp. 1-42, 2016.

[13]C.C Iberoamericano, “Manual Cochrane de Revisiones Sistemáticas de Intervenciones”, Barcelona, vol. 5.1.0, 2012.

[14] K.S. Khan, R. Kunz, J. Keijnen y G. Antes, “Five steps to conducting a systematic review”, *Journal of the royal society of medicine.*, vol. 96, pp. 118-121, 2003.

[15] Z. Halim y S. Khan, “A data science-based framework to categorize academic journals”, *Scientometrics*, vol. 119, no. 1, pp. 393-426, 2019.

[16] P. Brereton, B. A. Kitchenham, D. Budgen, MarkTurner y MohamedKhalil, “Lessons from applying the systematic literature review process within the software engineering domain”, *Journal of Systems and Software*, vol. 80, no. 4, pp. 571-583, 2007.

[17] I. Nunes, D. Jannach, “A systematic review and taxonomy of explanations in decision support and recommender systems”, *User Modeling and User-Adapted Interaction*, vol. 27, no. 3-5, pp. 393-444, 2017.

[18]F. Provost, T. Fawcett,“ Data Science for Business: What you need to know about data mining and data-analytic thinking”, USA: *O`Reilly Media,* 2013.

[19] J.W. Foreman, "Using Data Science to Transform Information into Insight”, Canada: *John Wiley & Sons*, 2014.

[20] C.O'Neil, R. Schutt, “Using Data Science to Transform Information into Insight”, USA: *O`Reilly Media*, 2013.

[21] Zumel, Nina, J. Mount, J. Porzak, "Practical data science with R”, USA: *Manning*, 2014.

[22] S. Gopi, "Python Data Science Cookbook", United Kingdom: *Packt Publishing Ldt*, 2015.

[23] Mueller, J. Paul, L. Massaron, "Python for data science for dummies", USA: *John Wiley & Sons*., 2015.

[24] P. Roger D., "Programming for data science", USA: *Leanpub,* 2015

[25]Christos KK, A. Syropoulos, "Steps in scala: an introduction to object-functional programming", USA: *Cambridge*, 2010.

[26] S. Steven S., "The Data Science Design Manual", USA: *Springer* , 2017.

[27] B. Kei D., “Reimaging Research Methodology as Data Science”, *Big Data Cognitive Computing.*, vol. 2, no. 1, pp. 1-13, 2018.

[28] J.F. Saltz, I. Shamshurin and K. Crowston,“Comparing Data Science Project Management Methodologies via a Controlled Experiment”, presented at the *50th Int. Conf. Syst. Sci.*, Hawaii, Jan. 04-07, 2017, pp. 1013 – 1022.

[29] T. Priebe and S. Markus, “Business Information Modeling: A Methodology for Data-Intensive Projects, Data Science and Big Data Governance”, presented at the *IEEE Int. Conf. Big Data*, Santa Clara, Calif., USA, Oct. 29 – Nov. 1, 2015, pp. 2056 – 2065.

[30] M. Khalaf, A.J. Hussain, D. Al- Jumeily, T. Baker, R. Keight, P. Lisboa, P. Fergus and A.S. Al Kafri,“A Data Science Methodology Based on Machine Learning Algorithms for Flood Severity Prediction”, *presented at the IEEE Cong. Evol. Comp. (CEC)*, Rio de Janeiro, Brasil, July 10-13, 2019, pp.1-8.

[31] F. Foroughi and P. Luksch, “Data science methodology for cibersecurity”, *Computers and Society*, Mar., pp.1 -14 , 2018.

[32] E. Graells G., V. Peña A. and L. Bravo, “Adoption-Driven Data Science for Transportation Planning: Methodology, Case Study, and Lessons Learned”, Sustainability, vol. 12, no. 15, Jul., pp. 1-14, 2020.

[33] J.S. Saltz, “Identifying the key drivers for teams to use a data science process methodology”, *presented at the 26th Europ. Conf. on Info. Syst.*, Porthmouth, Uk, Jun. 23-28, 2018, pp. 1-10.

[34] K. Gilbert, J. Izquierdo, M. Sanchez M., S.H. Hamilton, I. Rodriguez R. and G. Holmes, “Which method to use? An assessment of data mining methods in Environmental Data Science”, *Environmental Modelling Software.*, vol. 110, Oct., pp. 3-27, 2018.

[35] A.A. Mahabal, D. Crichton, S.G. Djorgovski, E. Law and J.S. Hughes, “From Sky to Earth: Data Science Methodology Transfer”, *Proceeding of the International Astronomical Union*, no. 325, Oct., pp. 17-26 ,2016.

[36] J.S. Saltz, R. Heckman and I. Shamshurin, “Exploring how different project management methodologies impact data science students”, presented at the *Euro. Conf. Inf. Systs*., Lisboa, no. 43, Jul., pp. 2939-2948, 2017 .

[37] S.N. Brohi, M. Marjani, I.A.T. Hashem, T.R. Pillai, S. Kaur and S. Amalathas, “A Data Science Methodology for Internet-of-Things*”,* presented at the *Int. Conf. for Emerg, Tech. in Comp.* London, UK., Aug. 19-20, 2019, pp. 178-186.

[38] P. F. Rangel, C. Aguilar, J. Cerrada M. and A. J. “Methodological Framework for Data Processing based on the Data Science Paradigm”, presented at the *XL Latin American Comp. Conf.*, Uruguay, pp. 1-12 , Nov. 15-19, 2014.

[39] C. O. Dumitru, G. Schwarz, F. Castel, J. Lorenzo and M. Datchu, “Artificial Intelligence Data Science Methodology for Earth Observation”, *presented at the Advanced Analytics and Artificial Int. Apps.*, Ali Soofastael, Ed., London, UK.: IntechOpen, 2019, ch. 6, pp. 1-23.

[40] D.K. Griffin, “The Big Three: A Methodology to Increase Data Science ROI by Answering the Questions Companies Care About”, *Journal of Cornell University: Computer Science in Machine Learning Journal*, pp. 1-43 ,Feb. 2020.

[41] G. George, E.C. Osinga, D. Lavie and B.A. Scott, “Big Data and Data Science Methods for Management Research”, *Academy of Management Journal,* vol. 59, no. 5, Aug., pp. 1493-1507, 2016.

[42] W. van der Aalst, “Processes Meet Big Data: Connecting Data Science with Process Science”, an issue IEEE TRANSACTIONS ON SERVICES COMPUTING, vol. 8, no. 6, issue: 6, pp. 810-819, Dec. 2015.

[43] E. Popova, T. M. Rodgers, X. Gong, A. Cecen, J.D. Madison and S.R. Kalidindi, “Process-Structure Linkages Using a Data Science Approach: Application to Simulated Additive Manufacturing Data”, *Integrating Materials and Manufacturing Innovation*, vol. 6, Mar., pp. 54-68, 2017.

[44] J.S. Saltz and I. Shamshurin, “Exploring the Process of Doing Data Science Via an Ethnographic Study of a Media Advertising Company”, *in 2015 IEEE Int. Conf. on Big Data*, Santa Clara, CA, 2015, pp. 2098-2105.

[45] E.Kristoffersen, O.O. Aremu, F. Blomsma, P. Mikalef and J. Li, “Exploring the Relationship Between Data Science and Circular Economy: An Enhanced CRISP-DM Process Model”, *presented ath the 2019 Conf. on e-Business, e-Services and e-Society: Digital Tf. for a Susble. Society in the 21st Century,* Trondheim, Norway, 2019, pp. 177-189.

[46] C.J. Costa and J. T. Aparicio, “POST-DS: A Methodology to Boost Data Science”, presented at the *2020 15th Iberian Conference on Information Systems and Technologies (CISTI),* Sevilla, Spain, 2020, pp. 1-6.

[47] F. Martínez-Plumed et al., “CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories,” in *IEEE Transactions on Knowledge and Data Engineering*, 2020, pp. 1-14.

[48] J. S. Saltz and I. Shamshurin, “Big data team process methodologies: A literature review and the identification of key factors for a project's success”, presented at the *2016 IEEE International Conference on Big Data (Big Data),* Washington, DC, 2016, pp. 2872-2879.

[49] B. Yu and R. Barter, “The Data Science Process: One Culture”, *an issue in Journal of the American Stat. Assoc*., vol. 115, issue: 530, Apr., pp. 672-674, 2020.

[50] M.T. Mullarkey, A. R. Hevner, T. G. Gill and K. Dutta, “Citizen Data Scientist: A Design Science Research Method for the Conduct of Data Science Projects”, presented at the *2019 International Conference on Design Science Research in Information Systems and Technology: Extending the Boundaries of Design Science Theory and Practice*, Worcester, MA, USA, 2019, pp. 191-205.

[51] N. Wickramage, “Quality Assurance for Data Science*”, in 2016 Future Technologies Conference*, San Francisco, USA, Dec. 6 – 7, 2016, pp. 307-309.

[52] K. K. Kurt, H. T. Atay, M. A. Çiçek, S. B. Karaca and S. Türkeli, "The Importance of Multidisciplinary in Data Science: Application of the Method in Health Sector to Telecommunication Sector", presented at the *2019 International Artificial Intelligence and Data Processing Symposium (IDAP)*, Malatya, Turkey, 2019, pp. 1-4.

[53] C. Li, M. Hains, J. Wallin and Q. Wu, "Applying Data Science Methods for Early Prediction of Undergraduate Student Retention", presented at the *2019 International Conference on Computational Science and Computational Intelligence (CSCI)*, Las Vegas, NV, USA, 2019, pp. 1337-1340,

[54] D.L. Delgado and R.P. Navarro, "Ciencia de datos y estudios globales: aportaciones y desafíos metodológicos", *in Colombia Int.*, no. 102, Abr. 2020, pp. 41-62.

[55] P. Cristaldo, E. Schab, C. Richard, R. Rivera, A. De Battista,S. Retamar and N. Herrera, "Adecuación de una Propuesta Metodológica de Enfoque "Híbrido" para la Gestión de Proyectos de Ciencia de Datos", presented at the *FRCU - GIBD : Grupo de Investigación en Bases de Datos - Comunicaciones a congresos*,Uruguay, Nov. 2018, pp. 1-12.

[56] N.S. Represa, "Elaboración e implementación de una propuesta metodológica para la evaluación y gestión de la calidad del aire mediante el enfoque de la ciencia de datos", Ph.D., dissertation, Univ. Valencia and Univ. De la Plata, 2020. Avaliable: https://riunet.upv.es/handle/10251/144645.

[57] J.P. Rave, J.C.C. Morales and F.G. Echavarría, "Metodología para explorar datos abiertos de accidentalidad vial usando Ciencia de Datos: Caso Medellín", *Revista Chilena de Ingeniería*, vol. 27, no. 3, 2019, pp. 495-509

[58] K. Allice, "*Top 11 Trusted (and Free) Search Engines for Scientific and Academic Research*", 2013. available on http://www.emergingedtech.com/2013/12/top–11–trusted–and–free–search–engines– for–scientific–and–academic–research/

[59] J. Brophyand and D. Bawden, "*Is Google enough? Comparison of an internet search engine with academic library resources*," *Aslib Proceedings*, vol. 57, no. 6, 2005, pp. 498-512.

[60] J. B. Rollins, "Foundational Methodology for Data Science," *IBM analytics*, 2015, pp. 1-6. Available on https://www.ibm.com/downloads/cas/6RZMKDN8

[61] G. Kumar, "Team Data Science Process Lifecycle," *Microsoft Azure*, 2020. Avaliable on https://docs.microsoft.com/en-us/azure/architecture/data-science-process/overview

[62] L. C. Molina, “Torturando los Dato Hasta que Confiesen”, *Departamento de Lenguajes y Sistemas Informáticos*, 2000.