ISSN 1870-4069

Ciencia de datos - fundamentos, algoritmos y aplicaciones: Una visión general

Leticia Gómez Rivera, Perfecto Malaquías Quintero Flores,
Sara Gutiérrez Carmona, Miguel Angel Bello Rivera

Tecnológico Nacional de México,
Campus Apizaco, Tlaxcala
México

{m21371205,perfecto.qf,d20371439,m21371204}@apizaco.tecnm.mx

Resumen. En la actualidad, se vive una época en la que los datos son producidos de manera masiva debido a los avances realizados en el tema de comunicaciones y gracias al monitoreo de los procesos de las empresas. Por este motivo, es necesario hablar sobre la ciencia de datos y sobre las ventajas que representa en el análisis de grandes conjuntos de datos para la extracción de información de calidad (smart data) empleados en la toma de decisiones que conduzcan a resultados exitosos. La finalidad de este trabajo es introducir dicho tema, dando en primer lugar, un panorama general del mismo. Para luego, realizar la revisión de conceptos fundamentales de la ciencia de datos. Posteriormente, se presenta una revisión de los algoritmos de aprendizaje máquina más comúnmente utilizados en el proceso de análisis de los grandes datos, específicamente en tareas de regresión, clasificación y agrupamiento. También se incluye la descripción de los procesos planteados por diferentes autores y las debilidades encontradas en ellos, para luego proponer la introducción de un nuevo paso en el proceso de ciencia de datos, que consiste en la definición de objetivos, el cual no se menciona en la literatura aunque es importante para alcanzar los resultados esperados. Finalmente, se describen ejemplos en los que se aplica la ciencia de datos en problemas del mundo real.

Palabras clave: Revisión literaria, ciencia de datos, anomalías de datos, proceso, algoritmos.

Data Science - Fundamentals, Algorithms and Applications: An Overview

Abstract. Nowadays, we live in an era in which data is massively produced due to the advances made in the field of communications and thanks to the monitoring of business processes. For this reason, it is necessary to talk about data science and the advantages it represents in the analysis of large data sets for the extraction of quality information (smart data) used in decision making that leads to successful results. The

purpose of this paper is to introduce this topic, giving first an overview of it. Then, a review of the fundamental concepts of data science is presented. Subsequently, a review of the most commonly used machine learning algorithms in the process of big data analysis, specifically in regression, classification and clustering tasks, is presented. It also includes a description of the processes proposed by different authors and the weaknesses found in them, and then proposes the introduction of a new step in the data science process, which consists in the definition of objectives, which is not mentioned in the literature although it is important to achieve the expected results. Finally, examples are described in which data science is applied to real-world problems.

Keywords: Literature review, data science, data anomalies, process, algorithms.

1. Introducción

De acuerdo con el sitio Go-Globe [1], en el año 2016, en tan solo 60 segundos se realizaban más de 2.3 millones de búsquedas en Google, se generaban más de 3.1 millones de likes, se realizaban más de 3 millones de publicaciones en Facebook y se enviaban más de 150 millones de e-mails, etc. Si se analiza esta información detalladamente, se puede notar que la cantidad de datos que se producen en la actualidad ha incrementado, debido al aislamiento que ha existido durante esta época de pandemia, donde la interacción con otras personas es llevada a cabo mediante herramientas digitales, las cuales almacenan información de las actividades cotidianas. De igual manera, gracias a la optimización de procesos, en las empresas se están empleando herramientas donde a través de sensores, cámaras y sistemas de control se registran los datos que están siendo obtenidos y que representan gran cantidad de información.

Al existir grandes volúmenes de datos, uno de los problemas principales que se presenta es cómo obtener información útil a través de ellos, ya que pueden existir datos irrelevantes que no aporten conocimiento, por lo que solo ocuparán espacio de almacenamiento de forma innecesaria.

Para abordar el reto que representa el contar con grandes cantidades de información se utiliza la ciencia de datos, la cual es un conjunto de métodos que son utilizados para extraer el valor de los datos y también es el descubrimiento del conocimiento, de esta forma el propósito de la ciencia de datos es encontrar estructuras significativas dentro de los datos para poder extraer información de calidad [2], empleando para esto algoritmos y tecnologías de aprendizaje automático, minería de datos, análisis predictivo y visualización de datos.

En [3] mencionan que la ciencia de datos es una teoría y metodología relacionada con la cadena de valor de los datos, cuya tarea principal es almacenar y procesar grandes cantidades de datos, los cuales son heterogéneos, como las imágenes, los videos, textos, etc., lo que los hace más complejos y con mayor nivel de incertidumbre.

La ciencia de datos tiene diversas aplicaciones en la vida cotidiana, como la conservación del medio ambiente a través del análisis de datos en redes sociales

[4], el estudio del rendimiento académico de estudiantes universitarios [5] y la búsqueda de métodos de aprendizaje a través de juegos que sean atractivos para los estudiantes, los cuales son utilizados en escuelas militares y de medicina [6], por mencionar algunas.

Para que la ciencia de datos pueda cumplir con su propósito de procesar los datos, utiliza a la estadística como una base que le permite reducir el ruido presente en ellos, filtrarlos y de esta manera, obtener la información necesaria para extraer conocimiento. Al respecto, en [7] se menciona que la ciencia de datos es una combinación entre la estadística y las ciencias de la computación, debido a que cuando trabajan en conjunto se puede realizar una mejor interpretación de los grandes volúmenes de datos. Por su parte, la estadística se encarga de realizar el análisis de información a través la reducción de dimensionalidad, inferencia, etc. y las ciencias de la computación almacenan la información, la filtran y la preparan para que pueda ser analizada.

A pesar de que los grandes volúmenes de datos proporcionan información relevante sobre la forma en la que se vive, presentan desventajas para comprender cuantitativamente los datos [8], ya que:

- Toman formas complejas.
- Las observaciones que se hacen de los datos no se realizan con un diseño experimental adecuado, por lo que existen sesgos y datos incompletos.
- Desafíos éticos, ya que se puede realizar una identificación personal a través de los datos.
- La transparencia algorítmica.

Contribuciones de este trabajo

El presente trabajo tiene cuatro propósitos principales, los cuales contemplan: El primer propósito es proporcionar una explicación de manera simple del campo de acción de la ciencia de datos e identificar las áreas donde está siendo aplicada, explicando las fuentes de donde provienen los datos y las bases en las que se sustentan tales aplicaciones. El segundo propósito consiste en realizar un análisis de cinco procesos de ciencia de datos, brindando una opinión desde la perspectiva de los autores de este artículo. El tercer propósito de esta revisión es plantear un nuevo paso dentro del proceso de ciencia de datos, el cual destaca que los objetivos de la investigación tienen un papel fundamental en el éxito de su implementación. Y por último, el cuarto propósito consiste en mostrar ejemplos prácticos que detallen claramente los conceptos presentes en el artículo, con la intención de que el lector los relacione con problemas y actividades cotidianas. Con el objetivo de cumplir estos propósitos, el artículo está organizado de la siguiente forma: En la sección dos se hace una introducción a los conceptos fundamentales de la ciencia de datos, la sección tres presenta un análisis sobre la clasificación de la ciencia de datos de acuerdo a los algoritmos de aprendizaje máquina aplicados, en la sección cuatro, se presenta una discusión referente al proceso de la ciencia de datos, la cual concluye con la propuesta de un proceso de ciencia de datos en el que se incluye una fase de definición de objetivos. Y

por último, en la sección cinco, se describe la revisión y análisis de cinco casos de estudio.

2. Conceptos fundamentales

2.1. Tipos de datos

Diariamente se generan grandes cantidades de información que provienen de diferentes fuentes, por ejemplo, el cuerpo humano, las instituciones bancarias, fábricas, oficinas, escuelas, entre otras. Al realizar la recolección de la información, los datos se presentan en diferentes formatos: imágenes, videos, textos, audios, etc., por lo cual, se requiere el uso de métodos estadísticos y de ciencia de datos para lograr extraer conclusiones que ayuden en la toma de decisiones. Uno de los puntos importantes a considerar en el momento de recolección y análisis de datos, es asegurar que no se violen los derechos sociales, profesionales y éticos de la sociedad [9].

A pesar de que existen diversas formas de almacenar los datos, en [10] mencionan que los dos tipos de datos más comunes son los numéricos y categóricos:

- **Datos numéricos**: Están compuestos principalmente por números, por lo que son datos de tipo cuantitativo, es decir, que pueden ser medidos. En este tipo de datos existen dos categorías:
  - Datos continuos: Tienen la característica de contemplar cualquier número de la recta numérica, por lo cual abarcan un conjunto de valores incontable. Algunos ejemplos de estos datos son: el peso de una persona, tiempo de espera en un banco, la temperatura corporal, velocidad de un automóvil, etc.
  - Datos de conteo: En este tipo de datos, solo se contemplan datos de tipo entero. Por ejemplo, el número de alumnos de una clase, número de mesas disponibles en un restaurante, la cantidad de ventanas en una casa, etc.
- **Datos categóricos**: Están constituidos por palabras, símbolos, frases, etc. Son de tipo cualitativo, es decir, datos que se refieren a las características pertenecientes a un objeto y por ese motivo pueden ser divididos en clases. Los datos categóricos pueden dividirse en dos categorías:
  - Datos ordenados u ordinales: La característica principal de este tipo de datos es que siguen un orden inherente. Ejemplo de este tipo de datos están presentes en la descripción de tallas de una prenda: 0-Talla pequeña, 1-Talla mediana, 2-Talla grande, 3-Talla extra grande.
  - Datos no ordenados o categóricos: No tienen un orden que seguir, por lo cual no se puede definir una categoría como anterior a otra. Ejemplos de este tipo de datos son: descripción del clima (lluvioso, templado, etc), la raza de una persona, un tipo de planta, etc.

Conjuntos de datos Cuando los datos son recolectados de manera periódica y con métricas a seguir se forman los **conjuntos de datos**. En estos conjuntos de datos, generalmente la información se agrupa, donde cada fila representa las instancias, ejemplos, registros, objetos, etc. y las columnas representan los atributos, propiedades, características, etc. de esas filas [11]. Un atributo representa una característica del objeto del cual se está obteniendo la información. Un ejemplo que se puede utilizar para describir esto es un auto. El auto representa el objeto o entidad, el conjunto de datos está conformado por los registros de autos que cumplen con cierta métrica, por ejemplo, autos seminuevos, y el color, marca, modelo, etc., representan los atributos de ese auto.

Análisis de datos Una de las partes más importantes en la ciencia de datos es el análisis de la información que está incluida en el conjunto de datos, la cual no se nota a simple vista, pero representa un punto crucial en el proceso de toma de decisiones. En [12] abordan dos enfoques para realizar este análisis, haciendo énfasis en la cantidad de variables:

1. Análisis univariante. En él, se van analizando uno por uno los atributos (variables) que contiene el conjunto de datos. Pueden ser de tipo numérico o categórico. Los atributos categóricos se pueden cambiar a numéricos mediante la codificación, mientras que los atributos numéricos se convierten a categóricos mediante la discretización.
2. Análisis bivariante. En este análisis se utilizan dos atributos, se determina si tienen asociación entre ellos y si es así se busca la fuerza de asociación. En caso contrario se abordan las diferencias existentes entre los dos atributos. Se puede realizar un análisis bivariante de tres maneras diferentes: Con dos atributos numéricos, con dos atributos categóricos y con un atributo numérico y uno categórico.

Por otro lado, en [22], dividen el análisis de la información en cuatro categorías, las cuales abordan el análisis desde una perspectiva en la cual, a través de los datos se pueden realizar predicciones o se puede determinar por qué ocurrieron sucesos específicos:

1. Análisis descriptivo: Este tipo de análisis ayuda a visualizar los acontecimientos que ocurrieron en el pasado con el objetivo de comprender los motivos de fracaso o éxito en ciertas situaciones, para que de esta manera los usuarios interpreten cómo es que esos acontecimientos pueden afectar los resultados futuros. Estos análisis describen los datos, los resumen y permiten que los usuarios puedan comprenderlos de manera simple.
2. Análisis de diagnóstico: La característica principal de este análisis es que permite a los usuarios entender lo que está sucediendo y por qué ocurrió, para que de esta manera se tomen decisiones que ayuden a mejorar lo ocurrido. Analiza los factores que ocasionan un resultado determinado.
3. Análisis predictivo: Este análisis proporciona información acerca de lo que podría ocurrir en el futuro. Su enfoque principal consiste en realizar

predicciones a partir de datos históricos que ayuden a determinar áreas de oportunidad y riesgos.

4. Análisis prescriptivo: No solo se trata de analizar los datos y predecir sucesos futuros, sino que ofrece sugerencias para extraer beneficios y aprovechar las predicciones. El análisis ayuda a optimizar el proceso de toma de decisiones, ya que anticipa qué va a ocurrir, cuándo y por qué razón. Es una guía para los usuarios, en la que infiere cómo le afectarán los hechos y sugiere la opción óptima.

2.2. Anomalías en la información

La información con la que trabaja la ciencia de datos presenta diferentes tipos de anomalías, esto ocurre debido a la naturaleza de los datos, o debido a los instrumentos con los cuales se hace la recolección de datos. Una anomalía es la información que no se encuentra dentro de los patrones normales de los datos.

Las anomalías representan información distinta a la habitual y en la actualidad proporcionan la ventaja de poder detectar sucesos inusuales, como la detección de invasiones en los sistemas y la detección de fraudes, además, ayudan en el análisis de la calidad de los datos, el escaneo de seguridad, el control de sistemas, etc. [13]. Es difícil establecer normas para la detección de anomalías, ya que, toman diferentes formas dependiendo de los tipos de datos que se estén analizando, pero es una tarea importante comprender el tipo de anomalía específica que se presenta en el conjunto de datos con el que se desea trabajar. En [14] se menciona que es importante identificar los tipos de anomalías y sus características en las áreas estadísticas, de ciencia de datos y de aprendizaje máquina, esto con el objetivo de comprenderlas y poder realizar una analítica adecuada de los datos, sin que exista información que altere los resultados obtenidos. Una de las razones fundamentales por la que es importante detectar anomalías en los conjuntos de datos es evitar que los algoritmos se ejecuten mal o que fallen por lo complejos que son los datos del mundo real.

Clasificar los tipos de anomalías es una tarea complicada debido a la variedad que existe por la naturaleza de los datos con los que se está trabajando, pero en [15] dividen a las anomalías en tres categorías:

-Anomalías colectivas: Este tipo de anomalías se forman debido a la combinación de muchos casos, por ejemplo, la secuencia de los datos que se presentan en los sistemas bancarios.

- Anomalías contextuales: Este tipo de anomalías no se distinguen sin la presencia de un contexto. Por ejemplo, un clima caluroso en época invernal mostraría una anomalía, pero si el clima se presenta sin la información de la temporada, se tomaría como un dato válido, sin presencia de anomalías.

- Anomalías puntuales: Son aquellas en donde una sola muestra dentro del conjunto de datos es diferente de las otras muestras.

Al contar con grandes cantidades de datos que se producen cada día y que necesitan ser analizadas para la toma de decisiones, surge el desafío de la detección de anomalías en estos grandes conjuntos de información. La alta dimensionalidad de los datos crea dificultad para la búsqueda de anomalías,

porque aumenta la cantidad de atributos presentes en la información y se necesitan más datos para generalizar estos atributos y detectar los valores atípicos, además, el ruido presente en estos datos afecta la efectividad con la que se puede abordar la búsqueda. Todo esto impacta directamente en las técnicas de detección de anomalías, ya que si se aumentan las dimensiones, se vuelve más complejo el conjunto de datos y aumentan los falsos positivos en la detección [16]. Se debe tener cuidado especial con el tema de las anomalías de los datos, porque gracias a su detección es posible realizar una mejor analítica y toma de decisiones. Si se detectan correctamente las anomalías es posible resolver diversos problemas cotidianos que abarcan desde la seguridad de las grandes empresas, hasta la seguridad en los procedimientos médicos a los cuales se someten tantas personas todos los días.

3. Clasificación de la ciencia de datos desde la perspectiva de algoritmos de aprendizaje automático utilizados

Dependiendo del tipo de datos con los que se trabaje, la ciencia de datos se puede clasificar en dos categorías que son: la ciencia de datos supervisada y no supervisada [2,28]:

- Ciencia de datos supervisada: En esta clasificación, como conjunto de entrenamiento, se utiliza un histórico de datos etiquetados y se busca obtener una función que es utilizada para clasificar datos no etiquetados. La predicción que se realiza tiene el objetivo de clasificar las variables de salida a partir del conjunto de variables de entrada en las que ya se conoce la categoría a la que pertenecen. Este tipo de ciencia de datos necesita suficientes registros etiquetados para que el modelo aprenda a partir de los datos.

- Ciencia de datos no supervisada: Este tipo de clasificación también es conocida como agrupación y se utiliza para definir categorías a partir de la asociación de los datos. El objetivo que persigue es descubrir patrones ocultos dentro de conjuntos de datos no etiquetados.

Por otra parte, en [29] dividen a la ciencia de datos en tres categorías, basadas en el tipo de aprendizaje al que pertenecen los algoritmos que emplea:

- Aprendizaje supervisado: Este tipo de aprendizaje es el que se utiliza con más frecuencia, en él, el programa sabe cuáles son las salidas que va a obtener. Utiliza variables independientes (atributos) para determinar la variable dependiente (clase).

- Aprendizaje no supervisado: Es menos utilizado que el aprendizaje supervisado. En este tipo de aprendizaje no se conocen las clases que existen dentro del conjunto de datos, por lo tanto, la información es agrupada dependiendo de las características que sean similares entre los registros. Este aprendizaje es menos preciso que el aprendizaje supervisado.

- Aprendizaje por refuerzo: En este aprendizaje, se realiza un proceso de entrenamiento de un modelo, el cual consiste en que la computadora interactúe con su entorno incierto y realice acciones. Después se le guía para obtener el resultado que se espera y se le proporcionan recompensas o penalizaciones

dependiendo de las decisiones tomadas. El objetivo de este tipo de aprendizaje es que las recompensas aumenten.

3.1. Algoritmos en la ciencia de datos

Antes de aplicar algún algoritmo de ciencia de datos es importante analizar el problema que se quiere resolver y a partir de ello, realizar un análisis de la naturaleza de los datos, su estructura, los tipos de datos presentes, la cantidad de registros que se tienen, si existen datos ausentes y en qué porcentaje, etc. Es posible implementar los algoritmos de ciencia de datos en cualquier lenguaje de programación, pero es mejor utilizar lenguajes especializados en esta área como lo son R, RapidMiner, Python, SAS Enterprise Miner, etc. [2].

Los algoritmos de la ciencia de datos pueden ser utilizados en diferentes tareas, que comprenden la **clasificación de datos, la regresión, el agrupamiento**, entre otras. Los algoritmos de ciencia de datos mencionados con mayor frecuencia en la literatura [17,18,19,20,21,22,23,24,25,26,27] son:

- **Regresión lineal**: Este tipo de método tiene el objetivo de describir la relación que existe entre una variable dependiente y una o más variables independientes. Las variables independientes y dependientes pueden diferenciarse mediante un supuesto que pertenece al análisis de regresión, en el cual se supone que las variables independientes son exógenas, es decir, que la variable independiente no puede afectarlas y que tampoco existen otras variables fuera del modelo que afecten a la variable dependiente ni a las variables independientes. Un ejemplo de regresión lineal es la predicción del número de ventas en una empresa (variable dependiente) a partir de factores como el costo de envío, el tiempo de entrega del producto y las formas de pago (variables independientes).
  En la Ecuación 1 obtenida de [25] se muestra cómo se calcula la relación lineal entre una variable independiente y una dependiente:

$$y = \beta_0 + \beta_1 x + \varepsilon, \tag{1}$$

  donde:
  - y es la variable dependiente.
  - x es la variable independiente
  - β_0 es la intersección en y o la constante.
  - β_1 es el coeficiente en x o la pendiente.
  - ε es el término de error (el cual refleja que la relación entre x y y no es exacta).
- **Regresión logística**: Este tipo de regresión se basa en la función sigmoidal o logística. Está basada en el principio de probabilidades, el cual se muestra en la ecuación 2 obtenida de [25] y que se define como la probabilidad de que ocurra un suceso (P) dividida por la probabilidad de que no ocurra:

$$probabilidades = \frac{P}{1-P}. \tag{2}$$

La regresión logística expresa el logaritmo natural de las probabilidades como una función lineal de una constante y $k-1$ variables independientes, lo cual se representa con las ecuaciones 3, 4, obtenidas de [25]:

$$\ln\left(\frac{P}{1-P}\right) = \beta_0 + \sum_{i=1}^{k-1} \beta_i x_i, \tag{3}$$

donde P es la probabilidad y β_i representa un cambio en las probabilidades logarítmicas, logrando un cambio en x.

La ecuación 3 puede ser reescrita en términos de las probabilidades P de la siguiente manera:

$$P = \frac{exp(\beta_0 + \sum \beta_i x_i)}{1 + exp(\beta_0 + \sum \beta_i x_i)}. \tag{4}$$

– **K-means**: Es un algoritmo de agrupamiento, por lo que utiliza datos no etiquetados, con el objetivo de encontrar conjuntos (grupos o clases) dentro de esos datos. En [17] mencionan cuatro pasos a seguir en este algoritmo: 1. Inicialización: Generar aleatoriamente los k centroides iniciales (centro de los clústeres). 2. Clasificación: Calcular las distancias de todos los puntos del conjunto a todos los centroides y asignar los datos al centroide más cercano. 3. Calcular los nuevos centroides de los datos. 4. Detener el algoritmo hasta que no se produzcan más cambios. Los pasos 2 y 3 se repiten mientras no se alcance este objetivo. En [12] se presenta la Ecuación 5 para minimizar la varianza total del grupo o la función de error cuadrático:

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} ||x_i^{(j)} - c_j||^2, \tag{5}$$

donde:
- J es la función objetivo
- k representa el número de clústeres
- n representa el número de casos
- $x_i^{(j)} - c_j$ es la función de distancia
- $x_i^{(j)}$ es el caso i
- c_j es el centroide para el clúster j

– **K-vecinos más cercanos**: Este tipo de algoritmo es eficaz tanto para la clasificación como para la regresión, pero es más utilizado en tareas de clasificación y predicción. El método trabaja con un conjunto de datos de entrenamiento etiquetado y un conjunto de datos de prueba no etiquetado. Los datos no etiquetados son clasificados en categorías, dependiendo de la cercanía que tengan con sus vecinos. En [21] definen los pasos a seguir para la implementación de este algoritmo: 1. Almacenar el conjunto de datos de entrenamiento. 2. Calcular la distancia Euclidiana con todos los puntos de datos de entrenamiento, para cada dato nuevo no etiquetado. 3.

Encontrar los k-vecinos más cercanos. 4. Asignar el punto no etiquetado a la clase que contenga la mayor cantidad de vecinos más cercanos. 5. Repetir el procedimiento hasta asignar cada punto no etiquetado a su clase correspondiente.

En la Ecuación 6 obtenida de [12] se muestra cómo se calcula la distancia Euclidiana, que es la más utilizada para determinar la etiqueta del dato a partir de los vecinos más cercanos:

$$\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}, \tag{6}$$

– **Naive Bayes:** Es un método basado en el Teorema de Bayes, el cual define una ecuación que describe la probabilidad de que ocurra un evento dada la probabilidad de eventos relacionados. La característica vuelve *ingenuo* a este algoritmo, es que asume que las variables del conjunto son independientes entre ellas, es decir, que la aparición de una variable no tiene nada que ver con las demás. Naive Bayes tiene tres variaciones para su clasificador:
  1. Bernoulli: Para conjuntos de datos binarios.
  2. Multinomial: Para conjuntos de datos discretos.
  3. Gaussiano: Para conjuntos de datos que se ajustan a una distribución normal.

  El Teorema de Bayes se muestra en la Ecuación 7 [27]:

$$P(L|caracteristicas) = \frac{P(caracteristicas|L) \times P(L)}{P(caracteristicas)}, \tag{7}$$

  donde:
  - $P(L)$ es la probabilidad de L antes de que se observen los datos.
  - $P(L|caracteristicas)$ es la probabilidad que se quiere calcular, la probabilidad de L dadas las $caracteristicas$.
  - $P(caracteristicas|L)$ es la probabilidad de las $caracteristicas$ dada una etiqueta L.
  - $P(caracteristicas)$ Es la probabilidad de las $caracteristicas$.

– **Árboles de decisión:** Este tipo de algoritmo se usa tanto para realizar clasificación como regresión. Los árboles de decisión están compuestos de nodos y ramas. Los nodos están conformados por las características de una categoría a clasificar y las ramas representan los valores que puede tomar el atributo. Este algoritmo pretende dividir el conjunto de datos en subconjuntos más pequeños.
  Si se utilizan los árboles de decisión para la clasificación es necesario calcular su *Entropía* y *Ganancia*. En [20], mencionan que la entropía se emplea para medir el grado de aletoriedad que existe en el conjunto de datos. El valor de la entropía se mide entre 0 y 1, siendo 0 el resultado esperado y 1 el

peor resultado que se puede obtener en el cálculo. Por otro lado, la ganancia de información es una métrica que informa de manera intuitiva sobre el conocimiento del valor de una variable aleatoria. Al contrario de la entropía, mientras mayor sea el valor de la ganancia de información, mejor.
Las Ecuaciones 8 y 9 fueron obtenidas de [12]. La Ecuación 8 se utiliza para calcular la entropía:

$$E(S) = \sum_{i=1}^{c} -p_i log_2 p_i, \tag{8}$$

donde:
- $E(S)$ representa la entropía de un atributo
- p_i representa la probabilidad del i-ésimo valor del atributo
Después, se divide el conjunto de datos en sus atributos y se calcula la entropía para cada uno.
Para obtener la ganancia de información se utiliza la Ecuación 9:

$$Ganancia(T, X) = Entropia(T) - Entropia(T, X), \tag{9}$$

donde:
- T es entropía antes de dividir el conjunto de datos.
- X es la entropía total de la división utilizando un atributo específico.

– **Máquinas de Soporte Vectorial:** Este tipo de método es ocupado para realizar clasificación lineal, no lineal, detección de valores atípicos y regresión. El objetivo de este algoritmo es encontrar un hiperplano con el que se separen los puntos de un vector en dos clases. Los puntos de datos cercanos al hiperplano son denominados vectores de soporte. Las principales limitaciones de este algoritmo son la velocidad y el tamaño de los datos, ya que no es adecuada en la clasificación de grandes conjuntos. Las ecuaciones 10 y 11 obtenidas de [23] tienen el objetivo de encontrar la mejor forma de separar el hiperplano:

$$wx - b = 0, \tag{10}$$

donde:
- w es un vector de valores reales
- x es un vector de características de entrada
- b es un número real
- wx representa $w^{(1)}x^{(1)} + w^{(2)}x^{(2)} + \cdots + w^{(D)}x^{(D)}$ y D representa el número de dimensiones del vector x

Para definir dos hiperplanos paralelos se hace uso de las siguientes ecuaciones:

$$\begin{aligned} wx - b &= 1, \\ wx - b &= -1. \end{aligned} \tag{11}$$

– **Bosques aleatorios:** Este tipo de algoritmo es utilizado para realizar predicciones y es adecuado para conjuntos de datos medianos y grandes. En

un bosque aleatorio se construyen muchos árboles de decisión individuales y después se promedian las predicciones realizadas por cada uno de ellos. Aunque los árboles de decisión son más fáciles de interpretar, los bosques aleatorios tienen mejores resultados realizando predicciones. Para controlar la profundidad de los árboles en un bosque aleatorio, es necesario definir desde la selección del modelo, el tamaño del subconjunto de variables predictoras [24].

– **DBSCAN:** Es uno de los algoritmos de agrupamiento más utilizados. Está enfocado en encontrar áreas de alta densidad en el espacio de distribución de los datos. Para medir la densidad es necesario tomar un punto del conjunto y encontrar los puntos más cercanos a él (formando una vecindad), haciendo uso de una métrica de distancia. Mientras más puntos se encuentren dentro de la vecindad, mayor densidad tendrá el clúster. Este algoritmo se emplea de forma recursiva, eligiendo un punto y verificando sus puntos vecinos para establecer la densidad.

  Para emplear el algoritmo DBSCAN es necesario tomar en cuenta dos valores [22]:
  - *epsilon*: Es un número positivo que sirve como métrica para medir la distancia máxima entre los dos puntos del clúster.
  - $MinPts$: Es un número natural que define un umbral mínimo para establecer un área de puntos como densa. Este parámetro es definido por el usuario.

4. Proceso de la ciencia de datos

En la ciencia de datos, se lleva a cabo un proceso para eliminar la información irrelevante en los conjuntos de datos, obtener la información importante de estos conjuntos y conocer la naturaleza de los registros con los que se está trabajando.

En [2] mencionan que el paso más importante para comenzar un proceso de ciencia de datos es la necesidad de analizar un problema, ya que sin una buena definición del problema no es posible aplicar la ciencia de datos. El proceso que proponen para llevar a cabo un proyecto de ciencia de datos consta de cinco pasos: 1. Obtener conocimiento previo del problema para conocerlo a profundidad, además de entender los datos relacionados a ese problema. 2. Preparar los datos, es decir, ajustar los datos de modo que se presenten en la forma requerida por los algoritmos de ciencia de datos (sin valores perdidos, con selección de características, sin anomalías en los datos, etc.). 3. Realizar la aplicación del modelo, es decir, representar los datos y las relaciones existentes entre ellos. 4. Integrar el modelo a su entorno de producción, aquí es donde se evalúa el modelo, el tiempo en el que responde a lo que se le solicita y el mantenimiento que requiere. 5. Obtener conocimiento acerca de los datos analizados, después de haber realizado la extracción de información no trivial a partir del conjunto de datos.

Por otro lado, en [30], señalan que la ciencia de la datos está relacionada con la gestión del conocimiento, por este motivo, proponen implementar los proyectos de ciencia de datos en el contexto del proceso del conocimiento, el cual está dividido en 5 pasos: 1. Establecer sistemas de información e infraestructura en donde se generen los datos. 2. Recopilar los datos de las fuentes establecidas, adquirir la información necesaria para llevar a cabo el proceso y realizar gestión del contenido. 3. Procesar, analizar y tratar la información/datos con el objetivo de reducir datos irrelevantes y extraer la información útil del conjunto. 4. Gestionar la información y compartir el conocimiento obtenido del análisis. 5. Usar la información y el conocimiento, aplicándolo en las áreas necesarias.

En el trabajo [31], proponen un entorno de trabajo para realizar el proceso de ciencia de datos, el cual consta de 8 pasos: 1. Especificar el problema que se necesita resolver, conociendo el dominio en el que se desarrolla el problema. 2. Descubrir los datos, es decir, buscar fuentes de datos ya existentes que estén relacionadas con el problema, antes de realizar una nueva recopilación de datos. 3. Establecer el cumplimiento de normas referentes al acceso, la difusión y la destrucción de los datos e introducir la información en plataformas de gestión. 4. Realizar la gestión de los datos donde se evalúa la calidad, la preparación y vinculación de la información. 5. Evaluar la idoneidad, este paso abarca desde tabulaciones y visualizaciones descriptivas hasta análisis complejos de los datos, y debe caracterizar el contenido informativo de los resultados. 6. Modelizar y realizar análisis estadísticos, lo cual es fundamental para obtener conclusiones sólidas obtenidas a partir de información incompleta. 7. Comunicar y difundir, es decir, compartir los datos, código que fue desarrollado, documentos referentes al trabajo y realizar presentaciones, conferencias, publicaciones, etc. 8. Realizar una revisión ética, proporcionando un conjunto de principios, en donde se incluyan consideraciones referentes a la vigilancia masiva, privacidad y soberanía de los datos.

Por otra parte, en el trabajo [32], proponen un método genérico a seguir en un proyecto de ciencia de datos, el cual fue obtenido después de analizar más de 150 actividades de ciencia de datos. El método consta de 10 pasos que son: 1. Identificar el problema o el fenómeno que requiere ser investigado y establecer cuál es el resultado que se espera. 2. Definir el problema utilizando el conocimiento del dominio, identificando factores críticos a analizar. 3. Formular la hipótesis a evaluar sobre los parámetros y modelos. 4. Diseñar el análisis desde el descubrimiento y adquisión de datos hasta el análisis e interpretación de resultados. 5. Garantizar la validez conceptual del diseño del análisis de datos. 6. Diseñar, probar y evaluar cada paso, seleccionando la clase de algoritmos pertinentes para la preparación de los datos y modelos, seleccionando y ajustando los algoritmos para satisfacer los requerimientos analíticos y garantizando la validez de la aplicación del análisis de datos. 7. Ejecutar una segmentación encauzada, asegurando que se cumplen los requerimientos. 8. Garantizar la validez de los resultados con respecto al problema investigado. 9. Interpretar los resultados con respecto a los modelos, métodos y requerimientos del análisis, y

evaluar los resultados. 10. Poner en funcionamiento y supervisar la segmentación encauzada y sus resultados.

A diferencia de los trabajos mencionados anteriormente, en [33] dividen el proceso de la ciencia de datos en nueve etapas, las cuales se presentan como una extensión del ciclo de vida de los datos. Las etapas son: 1. Realizar un diseño experimental, 2. Obtener o generar los datos y construir los modelos de datos, 3. Generar la hipótesis y explorar los datos, 4. Realizar una limpieza y organización de los datos, 5. Preparar los datos (valores perdidos y selección de características), 6. Realizar una estimación del modelo, 7. Llevar a cabo la simulación del algoritmo con los datos, 8. Visualizar los resultados obtenidos y por último, 9. Publicar el manuscrito del artefacto para que se pueda reusar y se puedan reproducir los resultados obtenidos durante el proceso.

Observando los pasos que se siguen en cada uno de los procesos de ciencia de datos que se han descrito, se pueden identificar algunos puntos débiles que ocasionarían que el proyecto no finalice con éxito, por ejemplo, en el trabajo [2] no se menciona un paso en el que se recolecten o generen los datos, a pesar de que el paso 1 contempla entender los datos involucrados en el problema. Otro de los puntos débiles encontrados en este proceso es que no incluye un paso donde se realice una publicación o un reporte de los hallazgos obtenidos, el cual es el producto final del proceso. Por otra parte, en el trabajo [30], a pesar de que el primer paso contempla establecer sistemas de información e infraestructura para la generación de datos, no se establece un paso anterior en donde se realice la observación del entorno para identificar un problema a resolver. Otro punto débil encontrado en el proceso es que, aunque en el paso 3 se hace referencia al análisis y tratamiento de los datos, no se refleja la implementación de algoritmos de ciencia de datos. En el trabajo [31], mencionan el uso de métodos estadísticos que ayuden a extraer información relevante a partir de los datos, pero en ninguno de sus pasos se establece la implementación de los algoritmos correspondientes al área de ciencia de datos. Otro de los puntos débiles identificados en este proceso, es que no se establece un paso en donde se haga una evaluación de los algoritmos implementados a lo largo del proceso, el cual es fundamental para medir el grado de precisión de los resultados obtenidos. Por otro lado, en el trabajo [32], el punto débil que se identificó, fue que no se establece un paso de publicación, el cual es importante para que los resultados obtenidos puedan ser validados y reproducidos por otros investigadores o personas interesadas en el tema. Por último, en el trabajo [33], el paso 1 menciona que se debe realizar un diseño experimental, pero al no contar con un paso anterior en donde se observe el entorno y se identifique el problema, no es posible la realización de dicho diseño.

Debido a los puntos débiles identificados en los cinco procesos de ciencia de datos analizados, es necesario establecer un proceso que contemple las etapas fundamentales para el desarrollo exitoso del proyecto, además, es necesario incluir un paso que no fue encontrado en los procesos anteriores y que representa la dirección a seguir en el proyecto de ciencia de datos que se necesita implementar: definir objetivos, ya que sin ellos, el desarrollo del proyecto de

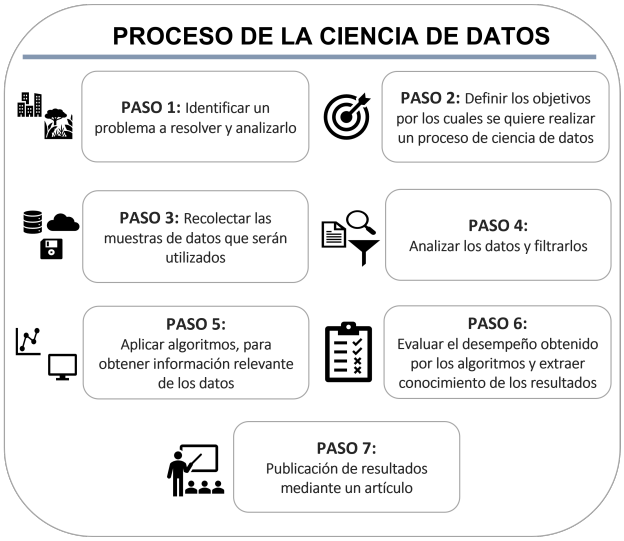

Fig. 1. Pasos del proceso a seguir en un proyecto de ciencia de datos. *Esta imagen es de autoría propia, en ella se contemplan los pasos más relevantes de los cinco procesos analizados, agregando la propuesta del Paso 2.

ciencia de datos no contempla una base sólida en la que apoyarse y sobre la cual dirigirse, lo que conlleva a que no se evalúe de manera adecuada el progreso del proyecto, ni los resultados obtenidos. Por ese motivo, en este artículo, se propone introducir al proceso de ciencia de datos una fase de definición de objetivos, y se enumeran los pasos fundamentales para desarrollar el proyecto, los cuales se obtuvieron de los cinco procesos analizados, esto se detalla en la Figura 1. El proceso consta de siete pasos que se describen a continuación:

- Paso 1. Identificar un problema a resolver y analizarlo, fundamentando el empleo de ciencia de datos en el proyecto: Este paso implica observar el entorno que nos rodea, ya sea en un entorno natural o social e identificar algún problema que esté conformado por grandes cantidades de datos, de las cuales sea necesario obtener información. Se analiza el problema, los atributos que lo conforman y se obtiene toda la información posible relacionada con él. En este paso es importante observar si el problema necesita resolverse mediante técnicas de ciencia de datos.
- Paso 2. Definir los objetivos, por los cuales, se requiere realizar un proceso empleando ciencia de datos: Con los objetivos se determina qué es lo que se quiere lograr al realizar el desarrollo del proyecto de ciencia de datos. Por ejemplo, si se realiza ciencia de datos sobre información de una empresa dedicada a las ventas, se puede establecer el objetivo de analizar los datos de sus productos en busca de defectos en su maquinaría o en el material que utiliza; o bien, se puede analizar qué productos son los que más se venden juntos para armar paquetes de venta.

- Paso 3. Recolectar las muestras de datos que serán utilizadas: Estas muestras pueden ser generadas o recolectadas del entorno en el cual haya sido identificado el problema, para esto, se necesita seguir una regla de recolección de atributos por cada registro.
- Paso 4. Analizar los datos y filtrarlos: En esta etapa, se emplean técnicas estadísticas para analizar los conjuntos de datos (obtener la media, mínimo y máximo de los datos, desviación estándar, etc.) que se obtuvieron en el paso anterior, también se filtran los datos, tomando en cuenta los datos ausentes en el conjunto.
- Paso 5. Aplicar algoritmos para obtener información relevante: Este paso consiste en aplicar algoritmos estadísticos o de aprendizaje máquina para lograr clasificar los datos o si es el caso, realizar la predicción de los resultados que serán obtenidos.
- Paso 6. Evaluar el desempeño obtenido por los algoritmos, extrayendo el conocimiento de los resultados: Se evalúa el grado de precisión con el que los algoritmos manejaron los conjuntos de datos para realizar las tareas asignadas y se analizan los resultados para obtener información relevante para la toma de decisiones.
- Paso 7. Publicación de resultados mediante un artículo: Este paso es importante para poder comunicar los resultados obtenidos a lo largo del proceso.

A pesar de que en esta sección se han descrito diferentes pasos en el proceso de la ciencia de datos, es importante identificar que el punto de partida es definir un problema que puede ser resuelto por medio de ciencia de datos, y al final, la meta es extraer información relevante sobre los datos analizados, los pasos intermedios pueden variar dependiendo de los tipos de datos que se quieran analizar y de las métricas para la obtención de información y de divulgación de resultados.

5. Ejemplos de aplicación de ciencia de datos

En esta sección, se presentan tres ejemplos de aplicaciones de la ciencia de datos en diversas áreas. Estas aplicaciones ayudan a obtener información útil e identifican aspectos del problema, que no hubieran sido detectados sin el análisis de los datos.

La ciencia de datos se puede aplicar en el área deportiva, realizando el monitoreo del estado en el que se encuentra un atleta. En [34] analizaron la actividad cerebral de atletas de tiro con arco, utilizando pruebas de electroencefalograma (EEG) y el algoritmo de Bosques Aleatorios. El propósito de realizar el monitoreo fue permitir a los entrenadores conocer el estado de los atletas antes de la competencia, ya que la carga psicológica, el control de los nervios, la toma de decisiones y la respuesta rápida son puntos importantes a considerar para que los atletas tengan éxito. A través de los datos extraídos de la prueba de EEC (como el índice de estado funcional cerebral, la entropía, los

neurotransmisores, etc.) clasificaron el estado competitivo de los atletas en cinco categorías: excelente, bueno, general, malo y muy malo. La precisión después de aplicar el algoritmo de Bosques Aleatorios fue del 89.74 %, la cual fue comparada con la precisión obtenida mediante modelos de Máquinas de Soporte Vectorial, que presentaron un rendimiento más bajo, con un 80.35 % de precisión.

Otro de los ejemplos en los que se puede aplicar la ciencia de datos es en [35], en este trabajo, los autores proponen un modelo que evalúa la calidad de enseñanza en un aula invertida. En este tipo de aulas, se utiliza un método de enseñanza en donde el estudiante es el pilar más importante, por lo cual, evalúa la calidad de la enseñanza que está recibiendo y también factores clave en su correcto aprendizaje. En este ejemplo, utilizaron el algoritmo de Máquinas de Soporte Vectorial enfocado en la regresión y seleccionaron 4000 grupos de datos para el entrenamiento del algoritmo y 500 grupos de datos para pruebas experimentales. Los datos utilizados como entrada del algoritmo son los indicadores de evaluación de expertos, docentes y estudiantes. Los indicadores contemplan una evaluación para la planificación docente, actitud docente, los dispositivos de enseñanza, entorno de enseñanza y la situación del aula. Después de que aplicaron el algoritmo de Máquinas de Soporte Vectorial analizaron cuatro aspectos: la transformación de los métodos de enseñanza, la abundancia de recursos didácticos, el cambio de iniciativa de aprendizaje de los estudiantes y la evaluación de calidad de la enseñanza en el aula invertida, obteniendo una precisión superior al 99.70 % y un error máximo de 0.04.

La ciencia de datos también puede emplearse en el área médica. En este ejemplo se describe cómo es que la ciencia de datos puede ayudar en el diagnóstico oportuno del cáncer oral. En [36] utilizan aprendizaje automático para identificar el cáncer oral por medio de imágenes, el cual es un método no invasivo y cómodo para los pacientes. Actualmente, se puede aplicar la ciencia de datos en información médica debido a que en estos años se han ido digitalizando los expedientes y análisis médicos. El cáncer oral es uno de los cánceres con tasas más altas de mortalidad, por lo cual su diagnóstico en etapas tempranas es importante, y para eso se necesita identificar las lesiones que se pueden transformar en malignas. En el ejemplo, la clasificación del cáncer se realizó con diferentes métodos de ciencia de datos: las Máquinas de Soporte Vectorial tuvieron una precisión de 82 % en la detección de mucosa normal y patológica. Por otra parte, los árboles de clasificación RepTree y J48Tree, tuvieron un 78.7 % de precisión en la detección de cáncer oral.

En el trabajo [37], se muestra un ejemplo en el cual se utilizan algoritmos de aprendizaje supervisado con el objetivo de predecir la presencia de COVID-19 en una persona. Este virus ha afectado a muchas personas a lo largo del mundo, ya que daña a los órganos del cuerpo debido a la inflamación generalizada que provoca. En este trabajo utilizan un conjunto de datos públicos que se encuentra en el sitio web Kagle, denominado *Síntomas y presencia de COVID-19*, el cual contiene 5434 registros con 20 características, entre las que se incluyen: fiebre, tos seca, dolor de cabeza, fatiga, hipertensión, etc. Los algoritmos de

Tabla 1. Valores de las métricas por cada algoritmo implementado.

Algoritmo	Precisión	Instancias bien clasificadas	Instancias mal clasificadas	Estadística Kappa	Error absoluto medio	Tiempo (s)
J48 DT	0.986	4144	60	0.972	0.024	0.03
RF	0.988	4154	50	0.976	0.023	0.18
SVM	0.988	4154	50	0.976	0.012	3.12
KNN	0.987	4149	55	0.973	0.022	0.01

Tabla 2. Resultados de Sensibilidad y Especificidad obtenidos por los algoritmos

Algoritmo	Sensibilidad	Especificidad
CNN	0.79	0.64
SVM	0.8	0.71
KNN	0.75	0.63
NBC	0.73	0.63

aprendizaje supervisado que utilizaron fueron: Árbol de decisión J48 (J48 DT), Bosques Aleatorios (RF), Máquinas de Soporte Vectorial (SVM) y K-Vecinos más Cercanos (KNN), implementados en la herramienta WEKA. En este ejemplo, utilizaron seis métricas para medir el rendimiento de los algoritmos: 1. Máxima precisión. 2. Mayor cantidad de instancias clasificadas correctamente. 3. Menor cantidad de instancias clasificadas incorrectamente. 4. Estadística Kappa (determina la confiabilidad de los resultados entre dos evaluaciones sobre la misma instancia). 5. Error Absoluto Medio más bajo y 6. Menor tiempo necesario para construir el modelo. Los resultados obtenidos para estas métricas se muestran en la Tabla 1, donde se puede visualizar que los algoritmos con mejor rendimiento al realizar la clasificación fueron Bosques Aleatorios y Máquinas de Soporte Vectorial, aunque el algoritmo con menor tiempo de implementación fue el de K-Vecinos más Cercanos.

La ciencia de datos también puede aplicarse en el desarrollo de herramientas que permitan el diagnóstico de la oncología mamaria. En [38], muestran un ejemplo en donde hacen uso de algoritmos de ciencia de datos para diagnosticar el cáncer de mama, a partir de un conjunto de imágenes que muestran la temperatura dentro de los tejidos y órganos. Los datos que se utilizaron para entrenar y probar los algoritmos fueron exámenes médicos reales de radiometría de microondas, en donde se incluyó la información de 302 pacientes, de los cuales 124 tenían un diagnóstico de cáncer. Las características de cada uno de esos registros incluían datos como: las temperaturas medidas, edad, temperatura ambiente, tamaño de los senos, diagnóstico, etc. Los algoritmos de ciencia de datos que fueron utilizados en este ejemplo contemplaban redes neuronales convolucionales (CNN), Máquinas de Soporte Vectorial (SVM), K-Vecinos más Cercanos (KNN) y el Clasificador Naive Bayes (NBC), los cuales mostraron el desempeño que se visualiza en la Tabla 2.

Como puede notarse, el algoritmo de Máquinas de Soporte Vectorial es el que obtuvo los mejores resultados en comparación con los otros tres algoritmos.

Incorporar algoritmos de ciencia de datos en la medicina es importante para que los médicos, en conjunto con los sistemas informáticos realicen mejores diagnósticos que contribuyan a mejorar la vida de las personas.

6. Conclusiones

En este artículo se mostraron algunos de los conceptos clave que involucra la ciencia de datos. Al ser un área tan amplia no se abordaron todas las tareas que se pueden realizar con ella, como la implementación de motores de recomendación o la detección de anomalías, etc., y tampoco los algoritmos empleados en estas tareas. Sin embargo, en este artículo se abordaron temas importantes que ayudan a comprender con mayor facilidad cómo es que la ciencia de datos y sus algoritmos están presentes en nuestro entorno real y cómo es que con acciones que son parte de nuestra rutina se generan grandes cantidades de información, que al ser analizadas adecuadamente pueden contribuir a mejorar la forma en que vivimos. En este artículo, se realizó un análisis de los conceptos clave que se manejan en el área de ciencia de datos, expresado desde el punto de vista de los autores de las fuentes bibliográficas estudiadas y de los autores del presente artículo.

Además, se dedicó una sección específica al proceso que se sigue en el desarrollo de proyectos de ciencia de datos, debido a que este proceso es primordial para obtener los resultados esperados al finalizar la implementación. Dentro de la misma sección, se tomaron algunos de los pasos de los procesos analizados y se estructuraron de tal manera que contemplaran las etapas elementales para el desarrollo exitoso del proyecto. También, se incluyó un paso que no fue identificado en los procesos analizados: Definir los objetivos. Este paso es fundamental en el desarrollo de un proyecto de ciencia de datos porque está enfocado en resolver un problema específico, pero puede ser abordado mediante diversas técnicas y algoritmos que producirían resultados diferentes. Por ejemplo, si uno de los objetivos es mejorar el tiempo en el que se obtienen resultados con un algoritmo, se elegirá el que sea más rápido, por el contrario, si el objetivo está enfocado en obtener un mayor porcentaje de rendimiento que se vea reflejado en las métricas de desempeño, entonces el método debe cambiar por el que produzca mejores resultados, por este motivo es importante definir desde el inicio el problema que se quiere resolver, pero también los objetivos que quieren ser alcanzados al finalizar el proyecto.

Por último, en este artículo se describieron algunas de las aplicaciones en las que se describen problemas analizados desde un enfoque de ciencia de datos, en los cuales se mencionan los datos utilizados, los algoritmos de aprendizaje máquina empleados y los resultados obtenidos. La ciencia de datos es una poderosa herramienta que sirve para mejorar aspectos de nuestro entorno que de otra manera tardarían mucho tiempo y utilizarían muchos recursos para ser resueltos, por esta razón, la investigación y desarrollo en esta área es importante

para enfrentar los problemas que se presentan en la actualidad y que involucran una gran cantidad de datos producidos de manera masiva.

Referencias

1. GO-GLOBE: Things That Happen Every 60 Seconds [Infographic]. https://www.go-globe.com/60-seconds/ (2022)
2. Kotu, V., Deshpande, B.: Data Science Concepts and Practice. 2nd edn. Morgan Kaufmann Publishers, U.S. (2019)
3. Xu, Z., Tang, N., Xu, C., Cheng, X.: Data science: connotation, methods, technologies, and development. Data Science and Management 1, 32–37, (2021)
4. Toivonen, T. et al.: Social media data for conservation science: A methodological overview. Biological Conservation 23, 298–315, (2019)
5. Padilla, V., Morales, S., Quintana, M., Flores, J., Herrera, O.: Reducción de la dimensión de registros de evaluaciones académicas aplicando el algoritmo K-means. Research in Computing Science 148(7), 515–526, (2019)
6. Alonso, C., Calvo, A., Freire M., Martínez, I., Fernández, B.: Applications of data science to game learning analytics data: A systematic literature review. Computers & Education 141, (2019)
7. Ceri, S.: On the role of statistics in the era of big data: A computer science perspective. Statistics and Probability Letters 136, 68–72, (2018)
8. Olhede, S., Wolfe, P.: The future of statistics and data science. Statistics and Probability Letters 136, 46–50, (2018)
9. Muqeeth, M., Kolhar, M., Al Ameen, A., Rahmath, M.: Data Science Techniques, Tools and Predictions. International Journal Of Recent Technology And Engineering (IJRTE), 8(6), 5661-5668, (2020)
10. Gutman, A., Goldmeier, J.: Becoming a Data Head, 1st edn. John Wiley & Sons, Inc., Indiana (2021)
11. Zaki, M., Meira, W.: Data Mining and Machine Learning, 2nd edn. Cambridge University Press, UK (2020)
12. Saedsayad.com: Data Mining Map. https://www.saedsayad.com/data_mining_map.htm. Last accessed 2 Feb 2022
13. Foorthuis, R.:On the nature and types of anomalies: a review of deviations in data. International Journal of Data Science and Analytics 12, 297–331, (2021)
14. Foorthuis, R.: A Typology of Data Anomalies. In IPMU, International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, pp. 26-38. Cádiz, Spain (2018)
15. Sangaiah A., Zhang, Z., Sheng, Q.: Computational intelligence for multimedia big data on the cloud with engineering applications, 1st edn. Academic Press, London (2018)
16. Thudumu, S., Branch, P., Jin, J., Singh, J.: A comprehensive survey of anomaly detection techniques for high dimensional big data. Journal of Big Data 7(42), (2020)
17. Ortega, J., Almanza, N., Vega, A., Pazos, R., Zavala, J., Martínez, A.: The K-Means Algorithm Evolution. 10.5772/intechopen.85447 (2019)
18. Berry, M., Mohamed, A., Yap, B.: Supervised and Unsupervised Learning for Data Science. Springer Nature, Switzerland (2020)
19. Kroese, D., Botev, Z., Taimre, T., Vaisman, R.: Data Science and Machine Learning: Mathematical and Statistical Methods. CRC Press, Boca Raton (2019)

20. Jijo, B., Mohsin, A.: Classification Based on Decision Tree Algorithm for Machine Learning. Journal of Applied Science and Technology Trends. 2. pp. 20-28, (2021)
21. Taunk, K., De, S., Verma, S., Swetapadma, A.: A Brief Review of Nearest Neighbor Algorithm for Learning and Classification. 2019 International Conference on Intelligent Computing and Control Systems (ICCS), pp. 1255-1260 (2019)
22. Balusamy, B., Abirami, N., Kadry, S., Gandomi, A.: Big Data: Concepts, Technology, and Architecture, 1st edn. John Wiley & Sons, Inc., USA (2021)
23. Burkov, A.: The Hundred-Page Machine Learning Book. Andriy Burkov (2019)
24. Schonlau, M., Zou, R.: The random forest algorithm for statistical learning. The Stata Journal, 20(1), 3–29, (2020)
25. Daniels, L., Minot, N.: An Introduction to Statistics and Data Analysis Using Stata. 1st edn. SAGE, California (2018)
26. Taulli, T.: Artificial Intelligence Basics. 1st. edn. Apress, USA (2019)
27. Voron, F.: Building Data Science Applications with FastAPI. 1st. edn. Packt Publishing Ltd., UK (2021)
28. Ghavami, P.: Big Data Management, 1st edn. De Gruyter, Berlin/Boston (2020)
29. Qamar, U., Summair M.: Data Science Concepts and Techniques with Applications. Springer Nature, Singapore (2020)
30. Chen, J., Ayala, B., Alsmadi, D., Wang, G.: Fundamentals of Data Science for Future Data Scientists. Analytics And Knowledge Management, 167–194 (2018)
31. Keller, S., Shipp, S., Schroeder, A., Korkmaz, G.: Doing Data Science: A Framework and Case Study. Harvard Data Science Review, 2(1), (2020)
32. Braschler, M., Stadelmann, T., Stockinger, K.: Applied Data Science. 1st edn. Springer, Switzerland (2019)
33. Stodden, V.: The data science life cycle: a disciplined approach to advancing data science as a science. Communications of the ACM 63(7), 58–66 (2020)
34. Li, X.: Athletes' State Monitoring under Data Mining and Random Forest. Journal of Sensors 2022, 1–11 (2022)
35. Fu, J., Li, J.: Teaching Quality Evaluation Model of a Flipped Classroom in Colleges and Universities Based on Support Vector Machine. Wireless Communications and Mobile Computing 2022, 1–12 (2022)
36. García-Pola, M., Pons-Fuster, E., Suárez-Fernández, C., Seoane-Romero, J., Romero-Méndez, A., López-Jornet, P.: Role of Artificial Intelligence in the Early Diagnosis of Oral Cancer. A Scoping Review. Cancers 13(18), 4600 (2021)
37. Villavicencio, C., Macrohon, J., Inbaraj, X., Jeng, J., Hsieh, J.: COVID-19 Prediction Applying Supervised Machine Learning Algorithms with Comparative Analysis Using WEKA. Algorithms, 14, 201 (2021)
38. Khoperskov, A., Polyakov, M.: Improving the Efficiency of Oncological Diagnosis of the Breast Based on the Combined Use of Simulation Modeling and Artificial Intelligence Algorithms. Algorithms, 15, 292 (2022)