UNIVERSIDAD DE CANTABRIA

FACULTAD DE CIENCIAS ECONÓMICAS Y EMPRESARIALES

GRADO EN ADMINISTRACIÓN Y DIRECCIÓN DE EMPRESAS

CURSO ACADÉMICO 2024-2025

TRABAJO FIN DE GRADO

Análisis de Datos y Big Data en la Gestión Empresarial
Un Enfoque desde la Ciencia de Datos

Data Analysis and Big Data in Business Management
A Data Science Approach

Autor

D. Jorge García Martínez

Director

Dr. D. Pedro Solana González

Santander, 18 de Junio de 2025

DECLARACIÓN RESPONSABLE

La persona que ha elaborado el TFG que se presenta es la única responsable de su contenido. La Universidad de Cantabria, así como quien ha ejercido su dirección, no son responsables del contenido último de este Trabajo.

En tal sentido, Don/Doña Jorge García Martínez se hace responsable:

1. De la AUTORÍA Y ORIGINALIDAD del trabajo que se presenta.
2. De que los DATOS y PUBLICACIONES en los que se basa la información contenida en el trabajo, o que han tenido una influencia relevante en el mismo, han sido citados en el texto y en la lista de referencias bibliográficas.

Asimismo, declara que el Trabajo Fin de Grado tiene una extensión de máximo 10.000 palabras, excluidas tablas, cuadros, gráficos, bibliografía y anexos.

Fdo.: GARCIA MARTINEZ JORGE - 72196443H

Firmado digitalmente por GARCIA MARTINEZ JORGE - 72196443H
Fecha: 2025.06.18 13:08:13 +02'00'

ÍNDICE

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

Análisis de Datos y Big Data en la Gestión Empresarial
Un Enfoque desde la Ciencia de Datos

Resumen

En la actualidad, las organizaciones almacenan cantidades masivas de datos que provienen de múltiples fuentes, provocando dificultades a la hora de tratar los datos de una manera eficiente. Es por esto por lo que surgen elementos como el Big Data y la ciencia de datos, facilitando este proceso y generando ventajas competitivas en las estrategias organizacionales a través de la optimización de los recursos y la precisión que suponen.

Este trabajo aborda la importancia de estas herramientas en la gestión empresarial, con un enfoque en la transformación de la toma de decisiones mediante técnicas que combinan aspectos estadísticos, matemáticos y computacionales.

Se analizan las posibles aplicaciones de estos conceptos en diferentes ámbitos, y para ello se aborda desde su definición hasta las oportunidades que presentan, además de cómo aplicarlos en un modelo práctico similar a una situación real.

Palabras clave: Big Data, ciencia de datos, toma de decisiones, optimización de los recursos, gestión empresarial.

Data Analysis and Big Data in Business Management
A Data Science Approach

Abstract:

Nowadays, organizations store massive amounts of data coming from multiple sources, causing difficulties when dealing with data in an efficient way. This is why elements such as Big Data and data science appear, easing this process and generating competitive advantages in organizational strategies through the optimization of resources and the precision they entail.

This paper addresses the importance of these tools in business management, with a focus on the transformation of decision making through techniques that combine statistical, mathematical, and computational aspects.

The possible applications of these concepts in different fields are analysed, from their definition to the opportunities they present, as well as how to apply them in a practical model like a real situation.

Keywords: Big Data, data science, decision making, optimisation of resources, business management.

1. MARCO GENERAL DEL TRABAJO

1.1 INTRODUCCIÓN

En la actualidad, el exponencial crecimiento de internet y la existencia masiva de fuentes de información han generado un aumento significativo en la complejidad del análisis y tratamiento de los datos. Esto sucede por el auge de las nuevas tecnologías, así como la digitalización de múltiples sectores, generando una sobreabundancia de información. Por ello, se ha convertido en un reto considerable la gestión y filtrado de esta, para así tener la capacidad de generar un valor real para aplicarlo en la toma de decisiones de la empresa.

La ciencia de datos es un conjunto de macroprocesos que contienen una serie de subprocesos complejos los cuales tratan de convertir los datos en patrones útiles para la toma de decisiones (Crespo, Alves y Soto, 2022), y surge como una disciplina clave en este aspecto, ya que se encarga de simplificar modelos complejos a partir del análisis y simulaciones, permitiendo contextualizar la información en escenarios reales. A través de la ciencia de datos, las empresas pueden mejorar su capacidad productiva, optimizando sus procesos y maximizando el aprovechamiento de sus datos, traduciéndose en una gran ventaja competitiva.

Por otro lado, el Big Data (cuya definición universal no ha sido establecida hasta el momento) trata de convertir esta gran cantidad de datos en información útil con el objetivo de facilitar y optimizar el proceso de decisión en el ámbito empresarial. No solo implica el almacenamiento de datos en infraestructuras digitales, sino también el desarrollo de nuevos métodos para el análisis de datos en tiempo real. Esto se logra a través del uso de programas y algoritmos computacionales especializados en el análisis de grandes volúmenes de datos sin filtrar.

Por ello, el Big Data y la ciencia de datos se complementan convirtiendo datos masivos en conocimiento útil. Se puede entender el Big Data como un fenómeno que simplifica la información para que la Ciencia de Datos lo convierta en conocimiento. (Pérez y Alvear, 2024). La sinergia de estos dos fenómenos permite mejorar el rendimiento de las organizaciones en múltiples campos, además de reducir los riesgos asociados.

En este trabajo se tratará en mayor profundidad la influencia de estos fenómenos en el sector empresarial tanto desde el ámbito teórico como desde el práctico, desarrollando un caso real para comprender de manera más clara la relevancia de estas herramientas en la era digital y su papel en la evolución de los modelos de negocio.

1.2 JUSTIFICACIÓN DEL TEMA

Durante el siglo XXI, el sector tecnológico ha sido uno de los más actualizados, permitiendo infinitas posibilidades gracias a nuevos métodos como nuevas fuentes de información, la implementación de internet en nuestras vidas y, más recientemente, las inteligencias artificiales. De hecho, la cantidad de datos anuales se haya multiplicado por 40, siendo la cifra de 2,8 *zettabits* en el año 2012 (Joyanes Aguilar, 2013), equivalente a casi 3 billones de *gigabytes,* mientras que para el año 2024, se estima que esta cantidad ha ascendido hasta los 120 *zettabits (*Taylor, 2024)

La gran cantidad de datos disponibles diariamente representa una ventaja para la organización, ya que están al alcance de nuestra mano. Sin embargo, no es tan sencillo obtener aquello que realmente se necesita. Cuando se manejan fuentes a gran escala, se puede generar un escenario en el que nuestra empresa sea incapaz de convertirlo en información útil para la correcta toma de decisiones.

Por otro lado, el impacto del Big Data no solo es vital en el ámbito económico, sino también en la transformación de los nuevos modelos de negocio (a partir del análisis de los datos, se pueden organizar los negocios para obtener ventajas competitivas), como por ejemplo la personalización total del servicio, la optimización predictiva o las plataformas impulsadas por datos (como Amazon o Uber).

Por ello, resulta fundamental el análisis de estos fenómenos tecnológicos, procediendo a explicarlos y situarlos en escenarios empresariales reales para su mejor comprensión.

1.3 OBJETIVOS DEL TRABAJO

1.3.1 Objetivo general

La finalidad principal de este trabajo es el análisis de las posibles aplicaciones de estos métodos de análisis de datos e información, además de explicar cómo pueden el Big Data y la ciencia de datos responder ante la inmensa cantidad de datos existente a través de procesos de simplificación y explicación de estos. Para ello, se aplicarán técnicas como *clustering,* regresión y clasificación sobre una base de datos real, la cual contiene las calificaciones de un conjunto de alumnos y determinadas variables que afectan de una forma u otra a las mismas, como las horas de estudio, las horas de sueño o la calidad de su alimentación, entre otras.

1.3.2 Objetivos específicos

- Objetivo E1: Aplicabilidad del Big Data e impacto en la empresa.
- Objetivo E2: Principales oportunidades y desafíos para las organizaciones.
- Objetivo E3: Procesos futuros en los que la ciencia de datos vaya a cobrar importancia.

2. MARCO TEÓRICO

2.1 FUNDAMENTOS

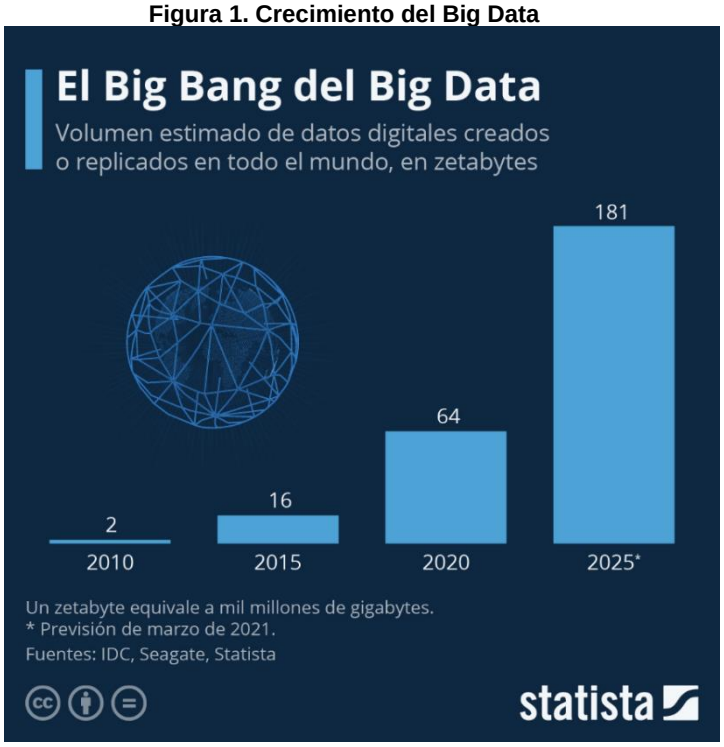
Al hablar de ciencia de datos nos referimos a macroprocesos que contienen una serie de subprocesos complejos, que buscan convertir los datos en patrones útiles para la toma de decisiones (Crespo, Alves y Soto, 2022). A su vez, la ciencia de datos abarca múltiples campos como son las matemáticas (mediante el lenguaje), la estadística (a través de las diferentes técnicas existentes) y la ciencia de la computación (herramientas de programación para manejar grandes volúmenes de datos), por lo que se puede concluir que es una disciplina híbrida.

El concepto de ciencia de datos debe su existencia a su necesidad, producida por la abundante cantidad de datos existente a raíz de las múltiples fuentes disponibles y a las formas de obtenerlos. Se generan datos en todo tipo de actividades; desde publicaciones en redes sociales hasta cámaras de seguridad, pasando por transacciones financieras o microchips de dispositivos electrónicos.

Como se señaló anteriormente, la cantidad de datos que se genera diariamente es muy abundante, habiéndose multiplicado por 40 en estos últimos 10 años. Son datos que provienen de diversas fuentes de información, pero además existen factores que han hecho que su crecimiento se haya acelerado:

- Expansión de Internet: Su accesibilidad y prestaciones han mejorado drásticamente estos últimos años y todo el mundo tiene la posibilidad de adentrarse en ellas y ser creador directo de información a través de diversos medios como Google, Facebook o Instagram, por ejemplo.
- *Internet of Things:* La conexión de objetos físicos a internet es una de las tendencias de estos últimos tiempos, mostrándose en múltiples formas diferentes como relojes inteligentes, altavoces, sensores, asistentes virtuales... e incluso las casas inteligentes. Estos nos facilitan nuestra vida, pero también son una gran fuente de datos.
- Crecimiento del e-commerce: Sobre todo en épocas como la pandemia del Covid-19, este fenómeno se desarrolló debido a la incapacidad de las personas de realizar compras físicas, sufriendo un gran crecimiento que se ha mantenido en el tiempo a través de grandes empresas como Amazon, AliExpress o Shein entre muchas otras.

Figura 1. Crecimiento del Big Data



Fuente: Statista

Es por esto por lo que estos dos conceptos (en este caso concreto el Big Data) se vuelven fundamentales en todo tipo de ámbitos, mostrándose como una herramienta esencial para la toma de decisiones. En primer lugar, la Ciencia de Datos es un campo de estudio que investiga cómo recoger, manejar y analizar datos para traducirlos en información útil (Berman, 2012), mientras que el Big Data trabaja con los datos para su mejor comprensión.

El Big Data se explica a través de sus cinco dimensiones: (Camargo-Vega *et al.*, 2015)

- Volumen: aumento masivo de datos registrado por la empresa genera una necesidad de filtrado para mejorar la toma de decisiones.
- Variedad: diferentes tipos de datos (estructurados, no estructurados y semiestructurados)
- Velocidad: rápida generación de datos en el entorno.
- Valor: los datos proporcionan información útil para la empresa, por lo que son un valioso activo.
- Veracidad: garantizar la calidad y fiabilidad de los datos para que sean precisos.

Relacionado con la variedad, se identifican los tres grandes grupos de datos, los cuales provienen de múltiples fuentes diferentes.

Los datos estructurados son aquellos con un formato o esquema fijo que, a su vez, poseen campos fijos. Son los datos de las hojas de cálculo, de las bases de datos relacionales y los archivos, fundamentalmente (Joyanes Aguilar, 2013). Por tanto, son datos bien definidos y especificados al detalle, además de con un orden establecido.

Otro tipo de datos son los no estructurados, que no pueden ser normalizados y tampoco poseen tipos definidos, ni siquiera están organizados bajo ningún patrón ni de manera relacional (Camargo-Vega *et al.*, 2015). Se puede entender que son datos sin orden, por lo que son más complejos a la hora de trabajar con ellos y necesitan de metadatos para su aplicación.

Por último, los datos semiestructurados son (a grandes rasgos) aquellos que no son ni datos en bruto, ni están tipificados de forma muy estricta como los datos estructurados (Abiteboul, 1996), pero si poseen etiquetas y otros marcadores que permiten separar los elementos dato (Joyanes Aguilar, 2013), como por ejemplo los correos electrónicos.

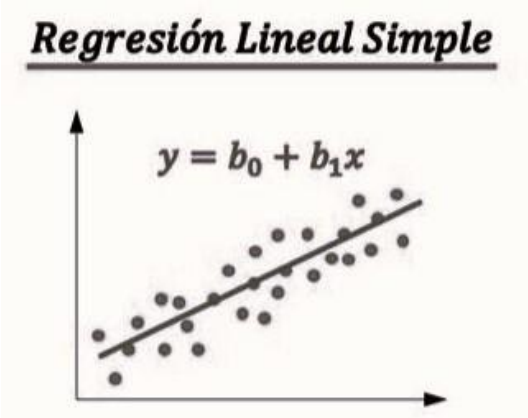
2.2 TÉCNICAS DE LA CIENCIA DE DATOS

La ciencia de datos utiliza diversas técnicas para analizar, interpretar y predecir patrones dentro de grandes volúmenes de datos. Dentro de los métodos de análisis, se puede diferenciar entre aprendizaje no automático (métodos de análisis convencionales, creados por el humano para resolver diferentes problemas sin el uso de algoritmos) y el aprendizaje automático (Machine Learning, que permite el análisis mediante el uso de algoritmos y modelos matemáticos, simplificando el proceso al no necesitar de trabajo humano). A su vez, el aprendizaje automático se divide en dos modalidades: El aprendizaje automático supervisado (predecir resultados conocidos a partir de datos etiquetados) y el no supervisado (descubre patrones en datos no etiquetados sin conocer las respuestas de antemano). En una tarea supervisada se busca capturar la relación entre algún input y algún output. Para ello, se poseen datos que contienen tanto la entrada como la salida, y se ingresa el resultado "correcto" para este conjunto de casos. En cambio, en una tarea no supervisada, la variable objetivo es desconocida, por lo que no es posible ajustar el modelo a salidas conocidas dadas las características de los inputs. (Lerena, 2019). Las principales técnicas aplicadas en la ciencia de datos, con sus definiciones y sus principales atributos, son las siguientes:

1. *Clustering*: Es una técnica de aprendizaje automático no supervisado, siendo un método estadístico multivariable para el agrupamiento automático de objetos, de forma que sujetos similares se ubiquen en el mismo grupo o "clúster", separando las muestras más diferentes en clústeres lo más lejanos posibles de las mismas. El algoritmo trata de particionar los objetos del modo optimo de acuerdo con alguna medida de validación, basada meramente en representaciones de los datos sin ningún conocimiento de membrecía a un grupo. (Lerena, 2019). A su vez, dentro del *clustering* existen diferentes algoritmos que permiten realizar diferentes técnicas:
    1.1. *K-Means*: Se selecciona el número de centroides (inician desde coordenadas aleatorias) y se asocia cada punto con el centroide más cercano. Tras esto, los centroides se reajustan hasta estabilizarse. Este agrupamiento se realiza minimizando la suma de distancias entre cada objeto y el centroide de su clúster. (Sáenz Villaverde, 2023).
    1.2. *DBSCAN (Density-Based Spatial Clustering of Applications with Noise*): Es un método para agrupar nubes de puntos que escalan de un gran número de puntos de datos con un tiempo de procesamiento relativamente corto. Realiza agrupaciones basadas en la densidad, pero no requiere un número predefinido de clústeres como en el método de *K-Means*. (Ming Ding *et al*., 2023)
    1.3. *Clústering* jerárquico: Similar al modelo de árboles de decisión, forma una jerarquía de clústeres anidados, permitiendo representar las similitudes entre los datos. Puede ser aglomerativo (comienza individualmente y se van agrupando en función de sus similitudes) o divisivos (parten del conjunto de elementos completos y se van separando los grupos que más diferentes sean entre ellos hasta quedarse con un número de clústeres que se considera óptimo). (Calvo, 2018)

2. Regresión: Esta técnica de aprendizaje supervisado se usa para analizar la relación existente entre dos o más variables, en las que una de ellas es dependiente al resto. Dicho de otra forma, permite conocer cómo afectan las variables independientes a la variable dependiente. Esto es aplicable en muchos ámbitos de la vida, como para la evaluación de riesgos, comprobar eficacias en cambios... así como en el mundo empresarial, donde se pueden analizar las diferentes ratios basándose en determinados cambios en las variables. Dentro del análisis de regresión, existen diferentes tipos aplicables a diversas situaciones.
    2.1. Lineal simple: Es el modelo básico, en el que se aplica un modelo matemático para explorar y cuantificar la relación existente entre la variable dependiente (Y) y una variable independiente (X). Esto se representa en una situación lineal con fines predictivos. (Espinoza, 2017)

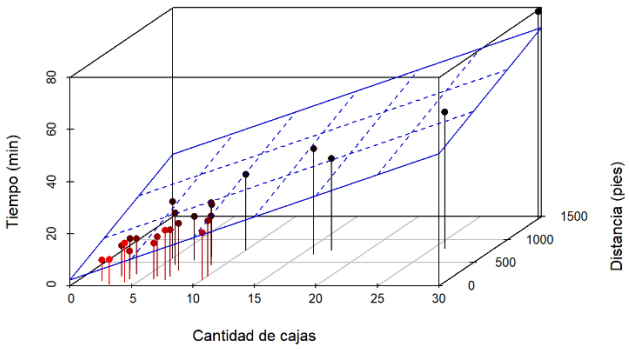
Figura 2. Regresión lineal simple



Fuente: Carpintero Ultralipoplus

    2.2. Lineal múltiple: A diferencia de la regresión lineal simple, en la que solo existe una variable independiente (X), en la versión de la regresión lineal múltiple hay más de una variable independiente.

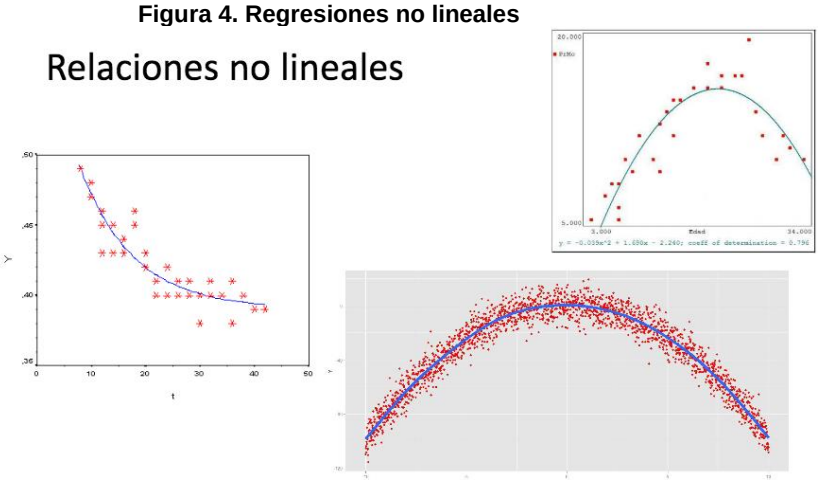
Figura 3. Regresión lineal múltiple



Fuente: github

2.3. No lineal: En este caso, la función resultante del análisis de regresión no es lineal, sino que adquiere formas diferentes como curvas. Esto se debe a que su ecuación es no lineal. Dentro de este tipo de análisis, existen diferentes variantes como la regresión polinómica (términos con potencias), exponenciales (términos con exponentes, donde la tasa de cambio de la variable dependiente es proporcional a su valor), logarítmicas...

Figura 4. Regresiones no lineales



Fuente: métodos numéricos merino

3. Clasificación: Es una técnica de aprendizaje supervisado en el que se trata de encontrar un modelo que asocie las características de entrada con un grupo. Dentro de este, se destacan los árboles de decisión (modelos en forma de árbol, donde se organizan los datos en una estructura jerárquica donde cada nodo representa una decisión sobre una variable, y cada rama muestra un resultado de esa decisión. Para su análisis, se utiliza el índice de Gini para la selección de las variables que mejor maximizan la homogeneidad de las particiones). (Zurita Herrera, 2024). Dentro de los árboles de decisión existen diferentes modelos, pero los más utilizados son:
    3.1. C4.5: Este método, desarrollado por Ross Quinlan, utiliza el concepto de entropía (mide el desorden) y se puede utilizar para datos discretos y concretos. (Syaraswati, Slamet y Winarno, 2017). Es un modelo mejorado del algoritmo ID3 (previamente creado por el mismo desarrollador), reduciendo drásticamente el sobreajuste “podando” el árbol.
    3.2. *CART (Classification and Regression Trees*): Es un modelo versátil, ya que también puede trabajar con regresiones. La principal diferencia respecto al modelo C4.5 es que produce árboles binarios. Además, este método permite trabajar con datos numéricos que son altamente sesgados. (Syaraswati, Slamet y Winarno, 2017)
    3.3. *CHAID (Chi-Square Automatic Interaction Detector*): Muy frecuente en procesos de segmentación de mercado, ya que por medio de la prueba Chi-cuadrado, permite generar árboles multiramificados para así dividir en segmentos los diferentes valores.

4. Redes neuronales: Son un subtipo de aprendizaje automático, basadas en el funcionamiento del cerebro humano y diseñados directamente para el reconocimiento de patrones y la resolución de problemas. Estas redes neuronales tienen tres capas diferentes en su estructura, siendo la primera de ellas la de entrada (esta capa recibe los datos), una o varias capas ocultas (procesan toda la información en función del peso y el umbral) y por último la capa de salida (esta capa tiene como función la generación de la predicción final). En función de su complejidad, existen diferentes tipos de redes neuronales, siendo estas las más importantes:
    4.1. Redes neuronales convolucionales: Aprovechan un perceptrón (unidad de red neuronal que realiza cálculos) multicapa que cuenta con varias capas convolucionales. Este tipo de redes usa frecuentemente principios de álgebra lineal para procesar patrones en imágenes y videos, por lo que es muy utilizado en el procesamiento de imágenes con Inteligencia Artificial, en reconocimiento facial o en procesamiento de lenguaje natural (NLP).
    4.2. Redes neuronales recurrentes: Este modelo guarda la salida generada por sus nodos procesadores y retroalimentándolos al algoritmo, mejorando así la capacidad de predicción y adquiriendo la capacidad de "aprender" de sus propios errores. Su uso más común es el relacionado con las aplicaciones de texto a voz y en predicciones de ventas y bolsa.
    4.3. Redes neuronales *Feedforward*: Este tipo de red neuronal hace que los datos viajen en una única dirección a través de varios nodos de procesamiento hasta llegar al nodo de salida. Están destinadas a procesar grandes cantidades de datos "ruidosos" y generar salidas "limpias". Las redes neuronales *feedforward* son la base del reconocimiento facial y el procesamiento del lenguaje natural. (Ortiz Domínguez, 2025)
5. Máquinas de soporte vectorial: Es un algoritmo automático supervisado creado para poder realizar trabajos de clasificación y/o regresión (de tal forma que se llama SVC para clasificación y SVR para regresión)(López, 2022). Se encarga de crear un hiperplano óptimo para lograr organizar las clases de datos de manera más eficaz. Esto lo logra mediante el uso de "*kernels*", que son funciones que permiten ilustrar los datos en un espacio de mayor dimensión para así capacitar su separación de una manera más sencilla. Los principales usos de esta técnica de análisis de datos son la detección de fraudes (historial de uso de tarjetas de créditos y transacciones fraudulentas), el análisis de imágenes (reconocimiento facial, detección de texto...) y para diagnósticos médicos (como la detección del cáncer o la clasificación de enfermedades)
6. Análisis de series temporales (ARIMA): Arima (*Autorregresive Integrated Moving Average*) es un modelo de medición estadística para predecir series temporales, es decir, para analizar posibles tendencias futuras. Esto es aplicable a las tendencias que puede tener una empresa en la cotización de sus acciones, por ejemplo.

2.3 PRINCIPALES OPORTUNIDADES Y DESAFÍOS

Son muchas las ventajas que proporcionan estos fenómenos, pudiéndose dividir en varias categorías, dependiendo de si están relacionadas con el proceso de toma de decisión, la optimización de los procesos, de personalización de servicios o la innovación en las actividades tecnológicas.

En primer lugar, la capacidad de aislar y filtrar la información que realmente se busca, para poder darle un enfoque mucho más directo y así aumentar la eficiencia redirigiendo los esfuerzos en la dirección correcta. Gracias a ello, no solo se logra una reducción significativa de los plazos de ejecución, sino que también se puede aumentar la competitividad y calidad, asegurándose de que las decisiones finales tomadas por la organización se han basado en una información clara y veraz.

Gracias a esta optimización de los procesos se logra una reducción de los costes operativos, aumentando nuestra rentabilidad y abriendo la posibilidad de invertir lo ahorrado en otras áreas, y la eficiencia operativa, además de mejorar considerablemente nuestra toma de decisiones. A su vez, permite generar predicciones conociendo los antecedentes, para así mitigar riesgos mediante la detección de fraudes y fallos de forma anticipada.

También son una poderosa arma competitiva, permitiendo el desarrollo de nuevos productos y modelos de negocio a partir de la innovación, además de una atención totalmente personalizada al cliente, facilitando la adaptación de la oferta a las necesidades específicas de cada uno de ellos y generando así una relación más estrecha y personal con los mismos, aumentando la fidelización.

En definitiva, si la organización es capaz de aprovechar al 100% estas ventajas, esta podrá dar un salto de calidad y eficiencia en su rendimiento, además de una reducción considerable en los costes y un aumento de la competitividad.

No son inconvenientes sino desafíos clave a los que se enfrenta la empresa en el proceso de implementación de estas nuevas tecnologías. La adopción de estas nuevas herramientas avanzadas no es sencilla, ya que no todas las organizaciones tienen la capacidad de hacerlo debido a la gran inversión inicial que supone la adición del software especializado, hardware y personal altamente cualificado necesario para aprovecharse de ello, además de los correspondientes gastos de mantenimiento y actualización que se requieren para garantizar el rendimiento óptimo a lo largo del tiempo. También se necesita de una infraestructura capaz de gestionar fuentes masivas de información que aportan datos constantemente, ya que la falta de esta infraestructura puede llegar a generar cuellos de botella en el procesamiento de toda esta información, llegando a perjudicar la capacidad de obtención de resultados de la empresa.

Debido a esto último, surge uno de los aspectos críticos a considerar por parte de las organizaciones, como son los riesgos en protección y seguridad de los datos obtenidos, los cuales pueden ser susceptibles de ser vulnerados mediante ciberataques o brechas del sistema y afectar directamente a la información que poseen de clientes, proveedores, cuentas bancarias... Para mitigar esto, es necesario que la empresa invierta en la implementación de protocolos de ciberseguridad los suficientemente potentes como para soportar estos posibles ataques.

Por ello, se puede concluir con que las organizaciones dispuestas a aplicar estos procesos en su actividad deberían adaptar un enfoque que combine la inversión en infraestructura tecnológica y la formación de sus empleados en este sector para así poder aprovechar al máximo el potencial del Big Data y la Ciencia de Datos.

3. ESTADO DEL ARTE Y REVISIÓN DE LA LITERATURA

El objetivo de este apartado es realizar una revisión de la literatura acerca de los dos fenómenos tratados en el trabajo, repasando diversos estudios y teorías de autores externos, además de sus diferentes resultados y conclusiones para así tener una visión más amplia.

3.1 CIENCIA DE DATOS

En la actualidad, el científico de datos es uno de los cargos cruciales en los aspectos económicos de las empresas, ya que deben predecir y tomar decisiones a partir de datos crudos para generar ventajas competitivas. En su artículo, (Davenport y Patil, 2012) lo consideran como uno de los trabajos más atractivos del siglo XXI, gracias a la importancia de traducir grandes cantidades de datos en soluciones prácticas, abriendo el camino para el éxito empresarial.

Algunos autores como (Dhar, 2013) se centran en el desarrollo de algoritmos avanzados y la importancia de construir sistemas inteligentes que puedan predecir determinados escenarios a partir de un conjunto de datos masivos, por medio de la automatización de estos procesos, para su posterior toma de decisiones. Asimismo, (Donoho, 2017) incide en la progresión de la ciencia de datos como una evolución de la estadística con un enfoque más amplio, señalando la importancia de la recopilación y visualización de los datos.

Por otro lado, autores como (Provost y Fawcett, 2013) indican que la ciencia de datos es un campo multidisciplinario que combina conocimientos en matemáticas, estadística, análisis de datos e informática, para generar valor mediante la extracción de información útil a partir de grandes volúmenes de datos estructurados y no estructurados. Con esto dejan claro que no es una simple forma de analizar datos, sino de obtener un beneficio a partir de ello, extrayendo lo realmente útil para su posterior aplicación en diferentes problemáticas. Por su parte, (Crespo, Alves y Soto, 2022) lo definen como macroprocesos que contienen una serie de subprocesos complejos, que buscan convertir los datos en patrones útiles para la toma de decisiones, pero resaltando que la forma en la que son generados los algoritmos es realmente influyente en los resultados. Por ejemplo, hacen referencia a que si los datos recogidos o los algoritmos discriminan en cierta medida (por ejemplo, temas de racismo), los pronósticos pueden contribuir a perpetuar tendencias de desigualdad en la sociedad.

Del mismo modo, (Cao, 2017) explica la importancia de no solo el proceso de la estadística y las técnicas de computación, sino de descubrir los patrones existentes en las distintas relaciones, para así poder conocer el contexto de los datos en su conjunto y traducir los resultados obtenidos en decisiones estratégicas adaptadas a los requerimientos de nuestra organización, por lo que entiende como vital la gestión final del resultado y no solo el proceso del análisis. Todo ello lo resume con la introducción del concepto de" pensamiento de ciencia de datos".

Desde un punto de vista más relacionado con la gestión organizacional, Skiena (2017) propone que además de los conocimientos en materia de matemáticas, estadística y ciencias de la computación, el gestor de datos debe poseer habilidades avanzadas en pensamiento crítico y razonamiento para poder tomar decisiones apoyándose en la información. Algo similar a la visión de (Martínez, Viles y Olaizola, 2021), que proponen una visión holística de la ciencia de datos en la que se debe dominar la gestión de

proyectos y de equipos, además del manejo de los datos para un correcto funcionamiento organizacional.

Para completar este apartado, se muestra una tabla en la que se recogen algunos estudios relacionados con el concepto que son realmente útiles para comprender la aplicabilidad real, señalando sus respectivos autores, el objetivo de estos y sus principales resultados.

Tabla 1. Ciencia de datos

AUTOR	**OBJETIVO DEL ESTUDIO**	**PRINCIPALES RESULTADOS**
(Kusuma, 2024)	Explorar la influencia de la ciencia de datos en las estrategias de marketing y cómo ha transformado los conocimientos tradicionales del consumidor.	Permite extraer información valiosa, detectar tendencias y predecir comportamientos, además de tomar decisiones basadas en datos precisos y en tiempo real para una mayor segmentación, compromiso y un mejor retorno de la inversión.
(Fu, 2024)	Analizar la aplicación de la ciencia de datos en la toma de decisiones de inversión y su mejora de la capacidad de previsión del mercado.	Con su capacidad y precisión, la ciencia de datos ha mejorado la eficacia de la toma de decisiones de inversión, además de su superioridad respecto a métodos tradicionales.
(Carmichael y Marron, 2018)	Desarrollar la relación entre las nuevas técnicas de análisis de datos y la estadística actual.	La principal diferencia es la ampliación que supone la ciencia de datos respecto a la estadística tradicional, incorporando sistemas computacionales y aportando una visión más amplia.
(Taylor, 2017)	Trata de proponer una idea de justicia y equidad en la forma en que las personas son visibilizadas a raíz de sus datos digitales para que exista cierta ética.	La ciencia de datos se ha basado en técnicas algorítmicas y no en la integración de principios, por lo que propone aportar visibilidad, antidiscriminación y desvinculación digital para lograr una sociedad más ética en materia de datos.
(Hedayetul, Shovon y Haque, 2012)	Estudiar la aplicabilidad de técnicas de ciencia de datos al entorno académico y analizar la influencia en el rendimiento de los estudiantes.	El uso de técnicas como *clustering* y árboles de decisión permitió predecir el GPA (*Grade Point Average*, las calificaciones de los alumnos), además de sugerir determinadas actividades personalizadas para mejorar su rendimiento escolar.

Fuente: Elaboración propia

3.2 BIG DATA

El concepto de Big Data ha crecido en gran medida en los últimos años en el sector empresarial, entre otros. A pesar de no tener una definición universal, algunos autores lo definen como cantidades masivas de datos que se acumulan con el tiempo que son difíciles de analizar y manejar utilizando herramientas comunes de gestión de bases de datos (Camargo-Vega *et al.*, 2015), mientras que para (Gartner, 2012), el Big Data son activos de información caracterizados por su alto volumen, velocidad y variedad, que demandan soluciones innovadoras y eficientes de procesado para la mejora del conocimiento y toma de decisiones en las organizaciones.

Por su parte, (Gandomi y Haider, 2015) indican la importancia de realizar tres tipos de procesos analíticos, como son análisis descriptivos, predictivos y prescriptivos, ya que sin ellos no es posible extraer el valor total a herramientas como el Big Data.

En un intento de analizar más a fondo esta herramienta, (Laney, 2001) explica la importancia de las tres 'V`, que son el volumen, la velocidad y la variedad. El volumen se refiere a la gran cantidad de datos generados y almacenados a los que las organizaciones deben hacer frente, que además proviene de múltiples fuentes. Por otro lado, la velocidad hace referencia a la fugacidad de generación de los datos, además de la capacidad de analizarlos que ello implica. Por último, la variedad se centra en los múltiples formatos en los que aparecen los datos, dificultando su gestión. Años más tarde, otros estudios como los realizados por (Geerts y O'Leary, 2022; Gandomi y Haider, 2015) van más allá, incorporando dos nuevos conceptos, dos nuevas V: La primera de ellas se refiere a la veracidad, la cual aporta certeza a los datos y así previniendo errores, mientras que la segunda V se refiere al valor, es decir, la capacidad de generar beneficios significativos a partir de estos datos.

La mayoría de los enfoques están orientados hacia el ámbito tecnológico, los cuales se basan en la implementación de estas nuevas herramientas y algoritmos para procesar grandes cantidades de datos. Por ejemplo, para (Ward y Barker, 2013) es un término que describe el almacenamiento y análisis de grandes conjuntos de datos grandes y/o complejos mediante una serie de técnicas que incluyen, entre otras, SQL y aprendizaje automático. Mientras que en otros estudios como el realizado por (McKinsey Global Institute, 2011) se hace hincapié en que estos datos a gran escala tienen un tamaño que está más allá de las capacidades de las herramientas del software de bases de datos típicas para capturar, almacenar, gestionar y analizar estos mismos.

Por otro lado, otro punto de vista del concepto va referido al enfoque organizacional, donde autores como (Chen *et al.*, 2014) indican que el Big Data hace referencia a los conjuntos de datos masivos que las herramientas de bases de datos no pueden capturar, almacenar, gestionar ni analizar y que contribuye a transformar la forma que tienen las organizaciones de gestionar la información para la toma de decisiones y obtener ventajas competitivas, resaltando la parte humana en el proceso de decisión. Asimismo, incidiendo en la parte humana, (Jenkins, 2013) señala que el Big Data no significa nada sin contexto e interpretación, el punto central de los datos es que nosotros (los seres humanos) podemos hacer algo con ellos, que solo nosotros podemos sacar conclusiones, hipótesis y averiguar porqué importan o no, es decir, que no podemos eliminar el factor humano.

Como en el apartado de la ciencia de datos, se aporta una tabla con varios estudios y sus principales resultados acerca del fenómeno del Big Data.

Tabla 2. Big Data

AUTOR	**OBJETIVO DEL ESTUDIO**	**PRINCIPALES RESULTADOS**
(Salgado Reyes, Guamba Gómez y Guerrero Flores, 2024)	Analizar la influencia de la tecnología de la información en la gestión empresarial, por medio de la adopción de métodos como el Big Data.	Estas tecnologías mejoran significativamente la eficiencia operativa y la toma de decisiones, a pesar de su compleja implementación. También permite digitalizar modelos de negocio y mejorar la experiencia de los clientes.
(Leon García, 2023)	Descubrir la relación entre la implementación del Big Data en las empresas y el desarrollo de la innovación.	Las capacidades del Big Data permiten que las organizaciones experimenten con mayor rapidez y precisión, acelerando el proceso de innovación.
(Cavlak y Cop, 2021)	Revelar la importancia del Big Data en el entorno del marketing digital.	El Big Data puede aportar a las empresas información de gran importancia acerca de sus clientes, para tomar decisiones personalizadas en base a ellos. Además, es una gran ventaja competitiva a la hora de mantener relaciones a largo plazo con los consumidores.
(Mittelstadt y Floridi, 2016)	Analizar el impacto ético que puede conllevar la automatización de los datos con herramientas como el Big Data.	A raíz de esta automatización, es posible que algunas tomas de decisiones basadas en estos datos afecten a los derechos de las personas, por lo que se proponen soluciones más éticas en la protección de la información.
(Daniel, 2015)	Explorar el uso y las oportunidades que ofrece el Big Data en el sector de la educación.	El Big Data permite analizar en tiempo real el progreso del estudiante para así intervenir de forma temprana, además de crear un aprendizaje personalizado para que su progreso sea más eficiente y sus resultados sean mejores.

Fuente: Elaboración propia

3.3 USOS GENERALES DEL BIG DATA Y CIENCIA DE DATOS

La ciencia de datos y el Big Data han permitido revolucionar diversos sectores a través de las diferentes técnicas que ofrecen, permitiendo manejar grandes volúmenes de datos para traducirlos en información realmente útil. Algunos de los campos más relevantes que abarca son, por ejemplo, el sector de la salud. Gracias al análisis de datos masivos, los científicos sanitarios pueden identificar patrones en enfermedades, para así predecir brotes o personalizar todo tipo de tratamientos. Son sistemas de gran ayuda en cuanto a detección temprana de enfermedades como el cáncer, suponiendo un gran avance respecto a los métodos previamente existentes. De hecho, actualmente se está desarrollando un programa llamado IPS (*International Patient Summary*), que tiene como meta identificar los datos clínicos más relevantes en casos de asistencia no programada o emergencias transfronterizas; aplicaciones usadas en el reconocimiento de lesiones dermatológicas que ayudan en el diagnóstico del cáncer de piel. (Nuño-Solinis *et al*., 2019)

También ha tenido un gran impacto en las nuevas formas de educar. Las instituciones educativas implementan estos sistemas para mejorar en los procesos de enseñanza y para ofrecer un aprendizaje más personalizado para el estudiante. Gracias a la captación y modelación de los datos, se abre la posibilidad de adentrarse en las áreas de mejora y a partir de ahí, diseñar diferentes estratégicas pedagógicas para potenciar las facultades del estudiante. Estos métodos tienen como objetivo la mejora de los procesos y de los resultados, la gestión masiva de programas, la experiencia de aprendizaje en tiempo real, la reducción de abandonos... Aunque, por otro lado, existen diferentes retos, siendo un proceso complejo de incrementar en las organizaciones, como la fragmentación de los datos o la inexactitud de la medición (la educación puede ser un campo más subjetivo que otros). (Filatro, 2024)

Por otro lado, ha permitido implementar nuevos sistemas de seguridad en las empresas. Las grandes organizaciones y gobiernos usan las funciones de la ciencia de datos en materia de ciberseguridad, para el análisis y detección de posibles amenazas, desde ataques cibernéticos hasta actividades terroristas. Otros métodos aplicables en este sector son las cámaras de vigilancia, las puertas con sellado virtual (a través de algoritmos de reconocimiento facial) e incluso el estudio de comportamientos inusuales entre grupos de personas. Estas funciones son vitales para la seguridad, permitiendo evitar grandes robos de información de millones de personas, como ya ha sucedido en ocasiones anteriores en ataques a la propia Policía Nacional. Se puede definir que la integración de la ciencia de datos en las operaciones de seguridad permite anticipar y prevenir los delitos cibernéticos. (Sanmorino, 2024).

El sector de la banca y las finanzas también se ha visto mejorado gracias a estas nuevas técnicas. Una gran variedad de posibilidades existe en este sector respecto al Big Data y la Ciencia de Datos, permitiendo a las instituciones financieras asegurar una experiencia personalizada al cliente, además de una gran seguridad en sus transacciones. Está permitiendo recopilar, analizar, limpiar y extraer información de los datos, además de sistemas de aplicaciones de calificación crediticia, aprobación de préstamos y pronósticos de tasas hipotecarias (Zheng *et al.*, 2024). Por otro lado, la incorporación de pagos digitales (locales e internacionales) desde los dispositivos móviles personales de cada uno, ofreciendo una mayor comodidad y fluidez.

Las empresas también han podido optimizar su logística y transporte en gran medida. Ha sido uno de los sectores más desarrollados gracias a las nuevas tecnologías, con el ejemplo de grandes empresas como Amazon o Uber, que han llevado la logística a otra dimensión. Estas tecnologías permiten a las empresas utilizar modelos predictivos para ofrecer a los clientes atención en tiempo real, optimizar las rutas de entrega y mejorar la eficiencia mediante datos sobre el clima o el tráfico, evitando retrasos. También son de gran utilidad en el transporte público, generando rutas mediante predicciones de afluencia de personas, para así reducir tiempos y ofrecer un servicio óptimo evitando las zonas de mayor congestión. Por tanto, se puede definir que es una gran herramienta en la planificación y que ofrece un seguimiento en tiempo real para que la interacción con el cliente sea más eficiente.

Otro de los sectores que ha crecido ha sido el del entretenimiento, además de que es uno de los que más se consumen a lo largo del día. Las grandes plataformas como Youtube, TikTok, Netflix o Spotify incorporan el análisis de datos para personalizar la experiencia a cada cliente, ofreciendo recomendaciones de contenido basado en el historial y los gustos de cada usuario. Esto se logra a partir del análisis de patrones de consumo, para predecir las tendencias y así conseguir una “fidelización” del consumidor. Esto se puede ver al navegar durante horas en aplicaciones como TikTok, en las que constantemente te ofrecen contenido llamativo y rápido, generando dopamina constante y así produciendo una gran adicción. También permiten evaluar nuevos proyectos teniendo en cuenta las posibles valoraciones de los clientes, como nuevas series o películas, teniendo en cuenta los posibles riesgos financieros.

Por último, la publicidad, que sigue el camino de la mejora del entretenimiento, ya que las empresas obtienen datos de consumidores basados en sus gustos o sus búsquedas recientes, para ofrecerte productos relacionados con ellos, como cuando navegas en Google y aparecen *banners* de algo que has buscado recientemente para llamar tu atención y tratar de venderte el producto.

3.4 APLICACIONES A LA GESTIÓN EMPRESARIAL

Siendo más concretos en la actividad empresarial, el uso de la analítica avanzada de datos permite a las organizaciones no solo mejorar procesos internos, sino transformar modelos de negocio (Lavalle *et al*., 2011). Áreas como la gestión de clientes, recursos y operaciones se ven beneficiadas por estas nuevas tecnologías, permitiendo que el rendimiento de la empresa mejore.

Con relación a la toma de decisiones, estas técnicas permiten visualizar información útil escondida tras grandes cantidades de datos, para así poder tomar decisiones vitales en una menos cantidad de tiempo y con una certeza mayor, además de incluso poder simular escenarios en los que la empresa puede encontrarse dependiendo del tipo de estrategia o decisión que elija.

La gestión de marketing es fundamental para la actividad empresarial, y a través de técnicas como árboles de decisión o regresiones, se tiene la capacidad de realizar predicciones basándose en las probabilidades teniendo en cuenta el comportamiento de los clientes. Es muy útil ya que, en materia de publicidad y marketing, se puede determinar antes de comenzar el proyecto si este tendrá éxito o no, además de seleccionar que factores priorizar. También se puede realizar una segmentación de clientes por medio del *clustering*, pudiendo diferenciar los grupos de clientes agrupándolos por características similares.

Por último, la optimización de los recursos permite obtener márgenes superiores, por lo que el uso de técnicas como redes neuronales (superan a los modelos tradicionales en la predicción de crisis financieras empresariales permite predecir rentabilidades y resultados para así saber qué alternativas tomar, además de poder detectar fraudes y errores para así reducir riesgos y asegurar beneficios. También es realmente útil para realizar predicciones de demanda, para elevar la eficiencia en el área de aprovisionamientos, evitando excesos que puedan causar unos mayores costes.

4. METODOLOGÍA

Este apartado describe la información aplicada en el caso práctico que voy a realizar a continuación. Este se basa en un análisis cuantitativo aplicado a una base de datos real, que trata de estudiar cómo afectan diferentes variables académicas, sociales y personales a los resultados académicos de un conjunto de alumnos del grado de Matemáticas, incidiendo en la calificación de un examen. Dentro de estas variables, se encuentran el género, la edad, las horas de estudio, las horas de redes sociales, las horas de sueño, las horas de Netflix, los estudiantes con empleo a horario partido, el porcentaje de asistencia, la calidad de la dieta, la frecuencia de ejercicio, el nivel de educación de los padres, la calidad del internet, la salud mental, la participación extracurricular y la calificación del examen, siendo esta última la variable dependiente.

El objetivo de este caso práctico es identificar patrones de comportamiento y resultados tras estas grandes cantidades de datos, con el fin de extraer conclusiones aplicables al entorno académico de los estudiantes. Mediante su análisis, se podrían tomar decisiones más precisas para mejorar el rendimiento académico de los alumnos a través de diferentes medidas como aprendizajes personalizados o seguimientos en tiempo real, para así corregir las desviaciones halladas.

La base de datos ha sido extraída de la plataforma Kaggel, la cual se encarga de publicar conjuntos de datos para su análisis. Esta en concreto se basa en la observación de alrededor de 1.000 estudiantes, con las 15 variables ya mencionadas. Por ello, se determina que todas son variables independientes, por lo que se busca analizar la incidencia de estas sobre la variable dependiente; la puntuación final obtenida por parte del alumno.

Se empleará el software R Studio para el análisis, introduciendo primeramente la base de datos, para después introducir los comandos relacionados con las técnicas seleccionadas y analizar los resultados.

Para su análisis, se utilizarán diferentes técnicas de análisis de datos como regresión (para determinar qué variables son las que más afectan al rendimiento académico), *clústering* (que permite agrupar a los alumnos basándose en patrones de resultados y comportamientos) y clasificación (un árbol de decisión en concreto, para clasificarlos según calificaciones).

5. DESARROLLO EMPÍRICO

Cabe señalar que en este apartado únicamente se tratarán los procedimientos de las técnicas de análisis, mientras que los resultados y su correspondiente explicación se incluirán en el siguiente capítulo.

En primer lugar, se debe introducir la base de datos en R Studio. Para ello, previamente se ha de haberlo descargado desde la página web de Kaggel, en un formato que permita ser importado a R Studio (en este caso, en formato csv). Además, se deben descargar los paquetes necesarios para operar, entre los que se encuentran "*readr*", "*broom*" y "*ggplot2*", entre otros.

```
> student_habits_performance <- read_csv("C:/Users/Jorge g/Desktop/student_habits_performance.csv")
```

Al introducir este comando, se ha agregado la base de datos al software, obteniendo un conjunto de datos llamado "*student_habits_performance*" con todas sus variables.

Figura 5. Datos prácticos

```
datos                          1000 obs. of 16 variables
    $ student_id                    : chr [1:1000] "S1000" "S10..
    $ age                           : num [1:1000] 23 20 21 23 ..
    $ gender                        : chr [1:1000] "Female" "Fe..
    $ study_hours_per_day           : num [1:1000] 0 6.9 1.4 1 ..
    $ social_media_hours            : num [1:1000] 1.2 2.8 3.1 ..
    $ netflix_hours                 : num [1:1000] 1.1 2.3 1.3 ..
    $ part_time_job                 : chr [1:1000] "No" "No" "N..
    $ attendance_percentage         : num [1:1000] 85 97.3 94.8..
    $ sleep_hours                   : num [1:1000] 8 4.6 8 9.2 ..
    $ diet_quality                  : chr [1:1000] "Fair" "Good..
    $ exercise_frequency            : num [1:1000] 6 6 1 4 3 1 ..
    $ parental_education_level      : chr [1:1000] "Master" "Hi..
    $ internet_quality              : chr [1:1000] "Average" "A..
    $ mental_health_rating          : num [1:1000] 8 8 1 1 1 4 ..
    $ extracurricular_participation: chr [1:1000] "Yes" "No" "..
    $ exam_score                    : num [1:1000] 56.2 100 34...
```

Fuente: Elaboración propia

5.1 Recta de regresión

Con ella, se podrá analizar la incidencia de las diferentes variables sobre la variable dependiente (la nota final del examen).

Para su aplicación, se le aporta un nombre cualquiera (en este caso, “modelo_reg) y posteriormente se vincula a la función lm para crear el modelo de regresión simple, incluyendo a variable dependiente y todas las independientes que se encuentran en la tabla.

```
modelo_reg <- lm(exam_score ~ study_hours_per_day + sleep_hours + social_media_hours +
                   netflix_hours + part_time_job + diet_quality + exercise_frequency +
                   attendance_percentage + internet_quality + mental_health_rating,
                 data = datos)
```

Tras crear la recta de regresión, se aplica el comando “*summary*” para que muestre todos los resultados obtenidos.

```
summary(modelo_reg)
```

Se obtienen los resultados de forma desordenada, por lo que mediante el comando “*tidy*”, sepide una tabla organizada y más comprensible, la cual se representa mediante el comando “*print*”.

```
tabla_coeficientes<-tidy(modelo_reg)
print(tabla_coeficientes)
```

5.2 *Clustering (k-means)*

Este método permite agrupar a los estudiantes en diferentes grupos en base a las variables independientes comunes entre ellos.

En primer lugar, se seleccionan las variables que se quieren analizar, las cuales deben ser numéricas con sentido entre ellas, ya que sino la agrupación no será lógica. En este caso, se han seleccionado variables como las horas de estudio, la frecuencia de ejercicio, la salud mental, las horas de sueño… Se crea el elemento "análisis_clúster", dándole el vector con todas las variables a agrupar.

```
#analisis cluster
analisis_cluster <- datos[,c("study_hours_per_day", "social_media_hours", "netflix_hours", "sleep_hours",
                             "attendance_percentage", "mental_health_rating", "exercise_frequency")]
```

Tras esto, se deben estandarizar o normalizar todos los datos, para que así las variables tengan una incidencia equitativa en el cálculo de las distancias de los *clústers*.

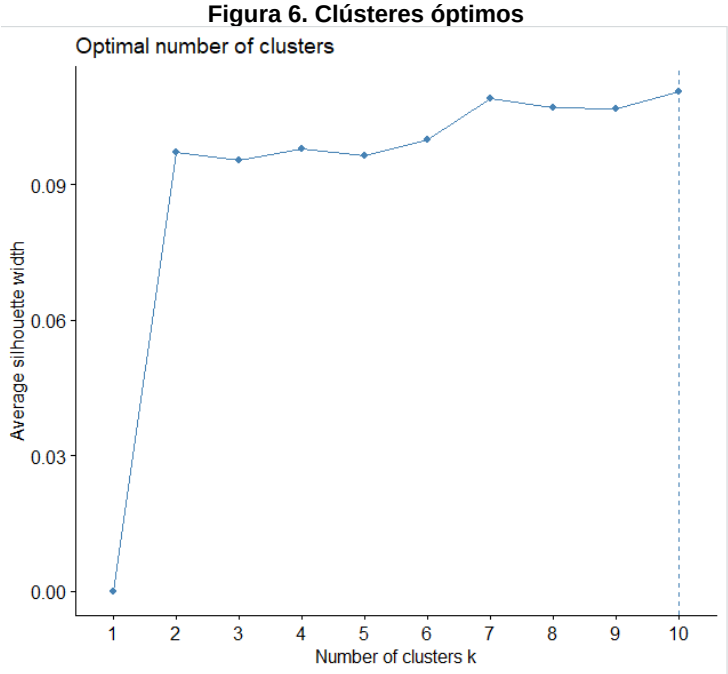
```
cluster_escalado<-scale(analisis_cluster)
```

Para averiguar el número de *clústers* que se necesitan, se introducen la siguiente función usando el método de la silueta, el cual determina el número óptimo indicando la maximización de la silueta.

```
fviz_nbclust(cluster_escalado, kmeans, method = "silhouette")
```

Al introducir este comando, se obtiene un gráfico en el que, tras analizarlo, se concreta que el número de *clústers* óptimo es 7, ya que es el punto donde la mejora deja de ser significativa además de ser un valor máximo. Se puede ver que quizás el x=10 sea ligeramente superior, pero no le precede ninguna mejora tan significativa.

Figura 6. Clústeres óptimos



Fuente: Elaboración propia

Una vez definido el número óptimo de *clústers* en nuestro análisis, se procede a aplicar el algoritmo *k-means* con la siguiente fórmula:

```
set.seed(100)
kmeans_rtdo<-kmeans(cluster_escalado,centers = 7, nstart = 25)
```

En ella se ubica la semilla número 100, que se utilizará siempre para este análisis, además de 7 "*centers*", que significa que el número óptimo de *clústers* son 7. En cuanto a "*nstart*", se introduce un número como 25 (por ejemplo) ya que significa que el algoritmo se ejecutará un total de 25 veces y seleccionando el mejor resultado, aumentando así la calidad del proceso. Ahora, solo queda analizar los resultados, para los que se introducen 3 comandos:

```
kmeans_rtdo$size
```

Este comando permite ver el tamaño de cada *clúster*, es decir, ver el número de alumnos que se encuentra dentro de cada uno de ellos.

```
kmeans_rtdo$centers
```

Por otro lado, este comando aporta los valores de cada *clúster* en cada variable, para sacar conclusiones del comportamiento de cada grupo de alumnos en relación con las mismas.

```
fviz_cluster(kmeans_rtdo, data = cluster_escalado, geom = "point", ellipse.type = "convex")
```

Por último, se obtiene una representación gráfica del análisis, para comprender la magnitud del algoritmo y observar la diferencia entre los múltiples grupos.

5.3 Clasificación (árbol de decisión)

En primer lugar, se crea una variable categórica, con la cual se va a determinar si una variable es alta o baja (en este caso, la puntuación del examen). Si es mayor que 75, será alta, mientras que si es menor será baja.

```
datos$objetivo<-ifelse(datos$exam_score >= 75, "Alto", "Bajo")
datos$objetivo<-as.factor(datos$objetivo)
```

Tras esto, se divide la base de datos en entrenamientos y en test real, con el objetivo de entrenar el modelo en una parte del estudio y evaluarlo en otra diferente.

```
set.seed(100)
division <- sample(1:nrow(datos),0.75*nrow(datos))
entrenamiento<-datos[division, ]
test <-datos[-division, ]
```

Ahora, se procede a crear el modelo del árbol de decisión con el siguiente comando, incluyendo todas las variables y basándose en la división del entrenamiento.

```
modelo <- rpart(objetivo ~ study_hours_per_day + sleep_hours + social_media_hours +
                  netflix_hours + attendance_percentage + mental_health_rating,
                data = entrenamiento, method = "class")
```

Para concluir, se podrá visualizar el árbol de decisión con el siguiente comando:

```
rpart.plot(modelo, type = 3, extra = 101)
```

Y para tener una visión más analítica, se realizará alguna predicción con los datos, las cuales van a permitir descubrir, por ejemplo, una matriz de confusión que permite evaluar el comportamiento del modelo realizado.

```
prediccion <- predict(modelo, test, type = "class")
table(test$objetivo, prediccion)
mean(prediccion == test$objetivo)
```

Tanto el resultado gráfico como el analítico se presentan en el apartado de resultados y discusión.

6. RESULTADOS Y DISCUSIÓN

6.1 Análisis de regresión

Tras ejecutar los comandos necesarios, se obtiene una tabla de resultados, donde se comenzará explicando el significado de cada columna:

En primer lugar, la columna "*estimate*" determina el efecto de la variable sobre la variable dependiente (en este caso el resultado en el examen), mientras que la segunda columna, llamada "*std.error*" representa el error del coeficiente. Por otro lado, "*statistic*" informa del valor t de la prueba, es decir, el efecto sobre la variable. Por último, el "*p.value*" representa el nivel de significancia estadística de la variable.

Una vez descritas las columnas, procedo a comentar los resultados.

Figura 7. Resultados de regresión

```
# A tibble: 13 × 5
   term                 estimate std.error statistic   p.value
   <chr>                   <dbl>     <dbl>     <dbl>     <dbl>
 1 (Intercept)             6.75      1.92       3.51  4.70e-  4
 2 study_hours_per_day     9.57      0.115     82.9   0
 3 sleep_hours             2.00      0.138     14.5   1.82e- 43
 4 social_media_hours     -2.60      0.144    -18.0   5.07e- 63
 5 netflix_hours          -2.27      0.157    -14.4   4.73e- 43
 6 part_time_jobYes        0.219     0.412      0.530 5.96e-  1
 7 diet_qualityGood       -0.685     0.377     -1.82  6.92e-  2
 8 diet_qualityPoor       -0.285     0.470     -0.605 5.45e-  1
 9 exercise_frequency      1.46      0.0835    17.4   1.65e- 59
10 attendance_percentage   0.143     0.0180     7.92  6.18e- 15
11 internet_qualityGood   -0.461     0.371     -1.24  2.14e-  1
12 internet_qualityPoor   -0.0637    0.500     -0.127 8.99e-  1
13 mental_health_rating    1.95      0.0595    32.8   3.27e-160
```

Fuente: Elaboración propia

Comenzando por las variables con un impacto positivo mayor sobre el rendimiento en el examen, destaco las horas de estudio al día, la salud mental y la frecuencia de ejercicio físico, ya que estas son variables con un coeficiente estimado alto y un "*p.value*" menor a 0,05, por lo que se determina que es estadísticamente significativo.

Sin embargo, hay variables como las horas de redes sociales y las horas de Netflix que impactan negativamente sobre la calificación. Además, a pesar de que es sorprendente, una buena dieta impacta de forma negativa aparentemente, aunque su nivel de significancia no es alto al ser mayor a 0,05.

El resto de las variables no son demasiado significativas o no influyen demasiado en el rendimiento académico, por lo que no son tan importantes como las anteriormente comentadas a la hora de tratar de corregirlas.

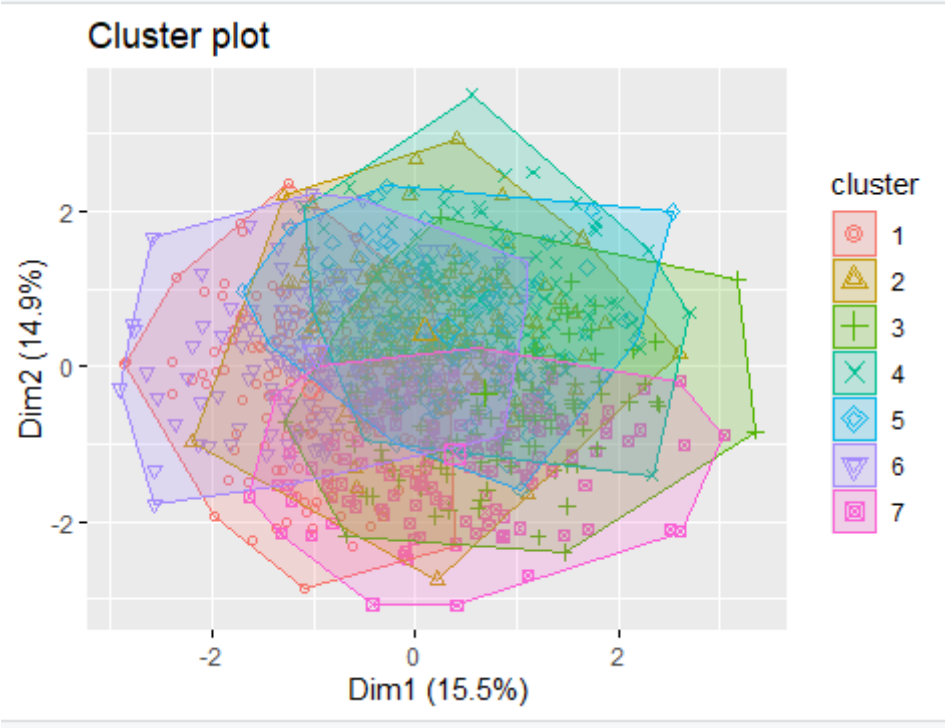
6.2 *Clúster*

Los principales resultados del análisis clúster se encuentran de 3 formas diferentes. La primera de ellas representa la cantidad de alumnos que se agrupan en cada *clúster* de los 7 sobre los que se ha trabajado. Como se puede observar, los alumnos varían entre los 124 y los 153, teniendo unas características similares entre los de su mismo grupo, pero diferentes respecto a los demás.

```
> kmeans_rtdo$size
[1] 146 141 153 124 140 147 149
```

A continuación, muestro la representación gráfica del análisis, el cual se explicará después más detenidamente:

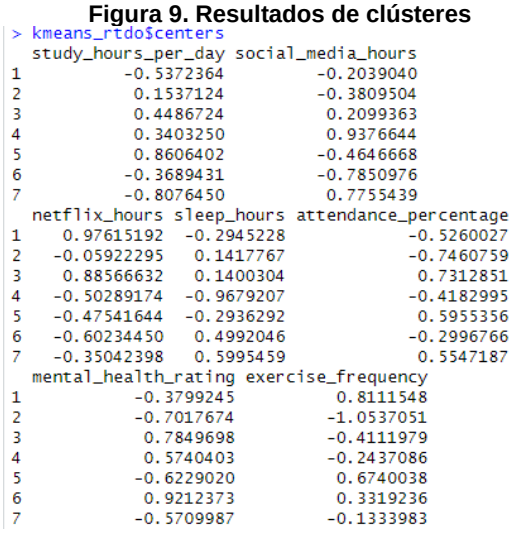
Figura 8. Gráfico de clústeres



Fuente: Elaboración propia

Se puede ver cómo se encuentran repartidos los alumnos dentro de los 7 clústers, diferenciados en colores. La mayoría de ellos se encuentran en ubicaciones centrales, pero hay algunos que se ubican muy alejados. Para su mejor comprensión, procedo con el informe analítico del mismo:

Figura 9. Resultados de clústeres



```
> kmeans_rtdo$centers
  study_hours_per_day social_media_hours
1          -0.5372364          -0.2039040
2           0.1537124          -0.3809504
3           0.4486724           0.2099363
4           0.3403250           0.9376644
5           0.8606402          -0.4646668
6          -0.3689431          -0.7850976
7          -0.8076450           0.7755439
  netflix_hours sleep_hours attendance_percentage
1    0.97615192  -0.2945228            -0.5260027
2   -0.05922295   0.1417767            -0.7460759
3    0.88566632   0.1400304             0.7312851
4   -0.50289174  -0.9679207            -0.4182995
5   -0.47541644  -0.2936292             0.5955356
6   -0.60234450   0.4992046            -0.2996766
7   -0.35042398   0.5995459             0.5547187
  mental_health_rating exercise_frequency
1           -0.3799245          0.8111548
2           -0.7017674         -1.0537051
3            0.7849698         -0.4111979
4            0.5740403         -0.2437086
5           -0.6229020          0.6740038
6            0.9212373          0.3319236
7           -0.5709987         -0.1333983
```

Fuente: Elaboración propia

En la imagen, se ven los valores que representa cada clúster respecto a las variables introducidas. Cada fila (de la fila 1 a la fila 7) de cada variable representa un *clúster* diferente, además de existir 7 variables.

La forma de leer y analizar los resultados es la siguiente: Los valores de cada fila representan al clúster respecto a esa variable, es decir; en la variable de horas de estudio al día, el clúster 7 representa un valor de -0,80 (valor negativo, indicando que se encuentra por debajo del promedio) mientras que el *clúster*, obtiene un valor de 0,86 (muy por encima del promedio). Como conclusión, se determina que:

En el clúster 1 se agrupan estudiantes con valores muy altos en horas de Netflix y muy bajos en horas de asistencia a clase y pocas horas de estudio, mientras que, en el 2, son estudiantes que no realizan prácticamente ejercicio y tampoco asisten a clase, además de que su salud mental está por debajo del promedio. Sin embargo, tampoco tienen valores muy altos en estudio ni en horas de sueño.

Por su parte, los alumnos del clúster 3 son los más responsables, teniendo valores muy altos en horas de estudio y asistencia, mientras que apenas usan Netflix.

El *clúster* 4 destaca por sus horas de *social media* y de salud mental, mientras que sus horas de sueño son muy pobres.

Por otro lado, el *clúster* 5 estudia y hace ejercicio, pero su salud mental no es la adecuada, mientras que el clúster 6 apenas usa Netflix ni *social media,* pero la salud mental es su fuerte.

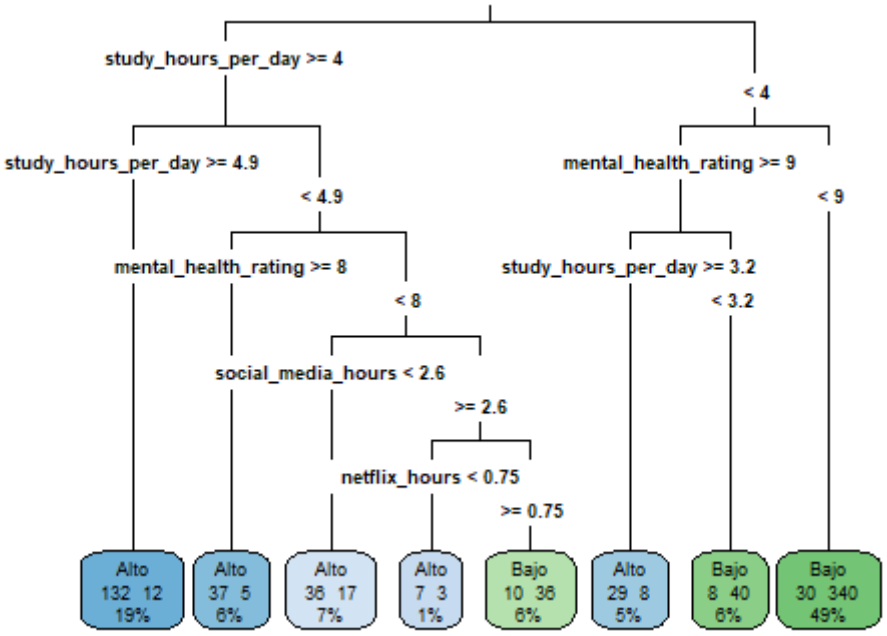
Por último, el clúster 7 es el grupo que menos estudia de todos, además de que su salud mental no es la adecuada. A pesar de ser el grupo que menos estudia, no usa apenas Netflix, mientras que las horas de social media si son altas respecto al promedio. Sus horas de sueño también destacan sobre el resto.

Se puede concluir que el análisis clúster es muy útil para agrupar a los estudiantes en grupos prototipo, segmentando el análisis y pudiendo observar patrones en su comportamiento comunes y directamente relacionados a su desempeño.

6.3 Clasificación (árbol de decisión)

El árbol queda representado de esta manera:

Figura 10. Árbol de decisión



Fuente: Elaboración propia

Se puede destacar que factores como una salud mental positiva refuerza el rendimiento, mientras que otros como el uso de redes sociales o de plataformas como Netflix producen lo contrario. Por otro lado, aunque es obvio, estudiar más horas al día es el factor principal en el rendimiento académico.

A través de los comandos introducidos para predecir el comportamiento del modelo, se obtienen los siguientes resultados:

```
> table(test$objetivo, prediccion)
      prediccion
       Alto Bajo
  Alto   75   14
  Bajo   22  139
```

Esto es una matriz de confusión. En las filas se representan los valores reales, mientras que, en las columnas, los valores predichos. Es decir, en este caso, existen 75 verdaderos positivos, 22 falsos positivos, 14 falsos negativos y 139 verdaderos negativos. En conclusión, el modelo analiza mejor los resultados bajos que los resultados altos.

```
> mean(prediccion == test$objetivo)
[1] 0.856
```

Por otro lado, este comando representa la exactitud del modelo, siendo esta de un 85,6%, un resultado bastante fiable.

7. CONCLUSIONES

A lo largo de este trabajo se ha analizado los conceptos de ciencia de datos y Big Data, explicando su crecimiento y cómo durante estos últimos años se han consolidado como algunas de las herramientas competitivas más potentes en el entorno empresarial, sobre todo en lo relacionado con la toma de decisiones.

Con relación a los objetivos propuestos inicialmente, se han analizado las posibles aplicaciones de estos fenómenos tanto en ámbitos generales (incluyendo sectores como la salud, las finanzas o la seguridad, entre otros) como más concretamente al ámbito empresarial (facilitando temas logísticos, de marketing y de optimización de recursos). Además, se han descrito las principales oportunidades y desafíos a los que la empresa, destacando la capacidad de extraer información aplicable a la empresa, a pesar de su costosa implementación.

Desde el punto de vista teórico, se han desarrollado los principales fundamentos y aplicaciones para la gestión empresarial, mediante su integración en los sistemas tecnológicos de las empresas, ya que a pesar de ser sistemas complejos de adaptar por temas de coste y de infraestructura, son inversiones que pueden llegar a situar a la empresa en un punto privilegiado respecto a la competencia, tanto en aspectos operativos como financieros, entre otros. No solo en el entorno puramente empresarial sino también en múltiples campos como la salud, seguridad, marketing y educación, donde permiten establecer métodos basados en las grandes cantidades de datos que se obtienen, para traducirlos en información simplificada que se pueda convertir en estrategias adecuadas, incluyendo asistencias personalizadas y en tiempo real para tener un mayor control sobre todas las áreas de la gestión.

Por otro lado, en el apartado práctico se han aplicado diferentes técnicas de análisis de datos como son las regresiones, los árboles de decisión y el clustering para trabajar con bases de datos reales, como si de una empresa se tratase, para sacar conclusiones acerca del rendimiento académico de un conjunto de alumnos y determinar qué variables externas son las que más les afectan, y así poder corregir de una forma temprana y correcta las desviaciones negativas que se hallen. De esta forma, queda reflejado cómo a través de estas técnicas, se puede extraer información útil y con un valor estratégico alto de grandes volúmenes de datos, siendo uno de los principales objetivos de este trabajo. Sin duda, la realización del apartado práctico ha sido lo más enriquecedor del trabajo, ya que ha permitido comprender el alcance que aportan estas herramientas, además de incitar a la imaginación de la aplicabilidad de estas a gran escala.

En definitiva, los conceptos de Big Data y ciencia de datos no solo representan una oportunidad en el ámbito tecnológico, sino también una nueva forma de gestionar modelos de negocio, otorgando la capacidad de traducir todos los datos que las organizaciones poseen en información útil que permita ser más eficiente y flexible a la hora de adaptarse a los rápidos cambios en el entorno, situando el rendimiento de las empresas que apliquen estas herramientas correctamente en un nivel superior en todos los ámbitos.

8. LIMITACIONES Y LINEAS FUTURAS DEL TRABAJO

A pesar de haber podido realizar el trabajo con éxito, se han hallado algunas limitaciones a la hora de la búsqueda de la información que han provocado que este sea más laborioso.

En primer lugar, la cantidad y calidad de las fuentes de información acerca de la temática tratada no es demasiado numerosa aun, existiendo una gran cantidad de trabajos en otras lenguas y muy pocos en castellano, restringiendo la riqueza del análisis. Además, la mayoría de los estudios encontrados no estaban disponibles de forma pública, siendo necesario su alquiler o compra. Esto no solo provoca que la temática sea más compleja de abordar, sino que es más difícil aún contrastar las fuentes disponibles.

Por otro lado, otra de las limitaciones halladas se relaciona con el uso de software como R Studio, ya que no se han encontrado demasiadas formas de aprender a usarlo de forma correcta.

Esto se puede deber a que es un tema relativamente nuevo, en el que aun falta mucho desarrollo (tanto teórico como técnico), y por eso creo que en los próximos años será posible realizar análisis con una profundidad y una calidad superior, gracias a las nuevas técnicas y estudios que se puedan aplicar.

En relación con lo anterior, las posibles líneas futuras podrían tratar de análisis más extensos de los conceptos, así como casos prácticos de aplicabilidad a sectores determinados e incluso con algoritmos más complejos.

A modo de conclusión, tanto este trabajo como los demás estudios existentes son únicamente el comienzo de la existencia de estos términos, pero lo realmente grande está por venir, con nuevas informaciones y técnicas que abrirán un mundo de posibilidades.

REFERENCIAS

Abiteboul, S. (1996) 'Querying semi-structured data', *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 1186, pp. 1–18. Disponible en: https://doi.org/10.1007/3-540-62222-5_33.

Bukaita, W. et al. (2025) 'Principles of Data Science'. Disponible en: https://www.researchgate.net/publication/387796782_Principles_of_Data_Science (Accedido: 17 Febrero 2025).

Calvo, D. (2018). 'Clúster Jerárquicos: Estrategia aglomerativa vs divisiva'. Disponible en: https://www.diegocalvo.es/cluster-jerarquicos-estrategia-aglomerativa-vs-divisiva/ (Accedido: 31 Marzo 2025).

Camargo-Vega, J.J. *et al.* (2015) 'Conociendo Big Data', *Revista Facultad de Ingeniería*, 24(38), pp. 63–77. Disponible en: http://www.scielo.org.co/scielo.php?script=sci_arttext&pid=S0121-11292015000100006&lng=en&nrm=iso&tlng=es (Accedido: 22 Enero 2025).

Cao, L. (2017) 'Data science: A comprehensive overview', *ACM Comput. Surv*, 50(43). Disponible en: https://doi.org/10.1145/3076253.

Carmichael, I. y Marron, J.S. (2018) 'Data Science vs. Statistics: Two Cultures?', *Japanese Journal of Statistics and Data Science* [Preprint]. Disponible en: https://doi.org/10.1007/s42081-018-0009-3.

Cavlak, N. y Cop, R. (2021) 'The role of Big Data in digital marketing', *Advanced Digital Marketing Strategies in a Data-Driven Era*, pp. 16–33. Disponible en: https://doi.org/10.4018/978-1-7998-8003-5.CH002.

Chen, Min *et al.* (2014) 'Big Data: A Survey', *Mobile Netw Appl*, 19, pp. 171–209. Disponible en: https://doi.org/10.1007/s11036-013-0489-0.

Crespo, F., Alves, T. y Soto, M. (2022) 'Ciencia de Datos, Inteligencia Artificial, y sus impactos sobre la sociedad', *Observatorio Económico*, (169), pp. 9–11. Disponible en: https://doi.org/10.11565/OE.VI169.474.

Daniel, B. (2015) 'Big Data and analytics in higher education: Opportunities and challenges', *British Journal of Educational Technology*, 46(5), pp. 904–920. Disponible en: https://doi.org/10.1111/BJET.12230.

Davenport, T.H. y Patil, D.J. (2012) *'HBR.ORG Spotlight on Big Data'*.

Dhar, V. (2013) 'Data science and prediction'. Disponible en: https://sci-hub.st/10.1145/2500499 (Accedido: 7 Mayo 2025).

Donoho, D. (2017) '50 Years of Data Science', *Journal of Computational and Graphical Statistics*, 26(4), pp. 745–766. Disponible en: https://doi.org/10.1080/10618600.2017.1384734;CTYPE:STRING:JOURNAL.

Espinoza. M (2017). 'Análisis de correlación y regresión lineal en función de estudiantes matriculados en la Universidad Ecotec'. Disponible en: https://www.researchgate.net/publication/349107216_ANALISIS_DE_CORRELACION_Y_REGRESION_LINEAL_EN_FUNCION_DE_ESTUDIANTES_MATRICULADOS_DE_LA_UNIVERSIDAD_ECOTEC (Accedido: 31 Marzo 2025).

Filatro, A. (2024) 'Ciencia de Datos en Educación - capítulo 1'. Disponible en: https://www.researchgate.net/publication/377233705_Ciencia_de_Datos_en_Educacion_-_capitulo_1 (Accedido: 28 Marzo 2025).

Fu, Y. (2024) 'Application of Data Science in Investment Decision and Its Influence on Market Forecast', *Frontiers in Business, Economics and Management*, 17(1), pp. 317–319. Disponible en: https://doi.org/10.54097/9MAC5Q70.

Gandomi, A. y Haider, M. (2015) 'Beyond the hype: Big data concepts, methods, and analytics', *International Journal of Information Management*, 35(2), pp. 137–144. Disponible en: https://doi.org/10.1016/J.IJINFOMGT.2014.10.007.

Gartner (2012*). 'Definition of Big Data - IT Glossary.'* Disponible en: https://www.gartner.com/en/information-technology/glossary/big-data (Accedido: 16 Mayo 2025).

Geerts, G.L. y O'Leary, D.E. (2022) 'V-Matrix: A wave theory of value creation for big data', *International Journal of Accounting Information Systems*, 47, p. 100575. Disponible en: https://doi.org/10.1016/J.ACCINF.2022.100575.

Hedayetul, M., Shovon, I. y Haque, M. (2012) 'An Approach of Improving Student's Academic Performance by using K-means clustering algorithm and Decision tree', *IJACSA) International Journal of Advanced Computer Science and Applications*, 3(8). Disponible en: www.ijacsa.thesai.org (Accedido: 15 Mayo 2025).

Jenkins, T. (2013) 'The End Of Judgment? Big-Data Analytics Still Needs People.' Disponible en: https://www.forbes.com/sites/netapp/2013/04/24/big-data-human/

Joyanes Aguilar, L. (2013) *Big Data Análisis de grandes volúmenes de datos en organizaciones E-Books & Papers for Statisticians*.

Kusuma, J. (2024) 'Data Science in Marketing: How Analytics are Reshaping Consumer Insights', *Advances: Jurnal Ekonomi & Bisnis*, 2(2). Disponible en: https://doi.org/10.60079/AJEB.V2I2.234.

Laney, D. (2001). '3D Data Management: Controlling Data Volume, Velocity and Variety. Application Delivery Strategies'.

Leon Garcia, O. (2023) 'Impacto de las capacidades de big data en la innovación empresarial'. Disponible en: https://www.researchgate.net/publication/370743218_Impacto_de_las_capacidades_de_big_data_en_la_innovacion_empresarial (Accedido: 21 Enero 2025).

Lerena, O. (2019) 'Métodos y aplicaciones de la ciencia de datos para las políticas de CTI: redes sociales, minería de textos y clustering'. Disponible en: https://www.researchgate.net/publication/334668096_Metodos_y_aplicaciones_de_la_ciencia_de_datos_para_las_politicas_de_CTI_redes_sociales_mineria_de_textos_y_clustering (Accedido: 29 Marzo 2025).

López, E. (2022) 'Eficiencia del truco de Kernel en la predicción de una máquina de soporte vectorial'. Disponible en: https://www.researchgate.net/publication/374584475_Eficiencia_del_truco_de_Kernel_en_la_prediccion_de_una_maquina_de_soporte_vectorial (Accedido: 3 Abril 2025).

Martínez, I., Viles, E. y Olaizola, I.G. (2021) 'Data Science Methodologies: Current Challenges and Future Approaches'.

McKinsey Global Institute (2011). ‘Big data: The next frontier for innovation, competition, and productivity.’ Disponible en: www.mckinsey.com/mgi. (Accedido: 17 Mayo 2025).

Ming Ding et al. (2023). Improved and optimal DBSCAN for Embedded Applications Using High-Resolution Automative Radar. Disponible en: https://www.researchgate.net/publication/346096076_Improved_and_Optimal_DBSCAN_for_Embedded_Applications_Using_High-Resolution_Automotive_Radar

Mittelstadt, B.D. y Floridi, L. (2016) ‘The Ethics of Big Data: Current and Foreseeable Issues in Biomedical Contexts’, *Law, Governance and Technology Series*, 29, pp. 445–480. Disponible en: https://doi.org/10.1007/978-3-319-33525-4_19.

Nuño-Solinis, R. (2019) ‘Ciencia de Datos y Big Data en Salud’. Disponible en: https://www.researchgate.net/publication/337170811_Ciencia_de_Datos_y_Big_Data_en_Salud (Accedido: 28 Marzo 2025).

Ortiz Domínguez, M. (2025) ‘Redes neuronales artificiales’, *Ingenio y Conciencia Boletín Científico de la Escuela Superior Ciudad Sahagún*, 12(23), pp. 38–44. Disponible en: https://doi.org/10.29057/ESCS.V12I23.14132.

Pérez, N y Alvear, A. (2024) ‘CIENCIA DE DATOS PARA LA INNOVACIÓN’. Disponible en: https://www.researchgate.net/publication/380450351_CIENCIA_DE_DATOS_PARA_LA_INNOVACION (Accedido: 20 Enero 2025).

Provost, F. y Fawcett, T. (2013) ‘Introduction: Data-Analytic Thinking’, *Data Science for Business: What You Need to Know About Data Mining and Data-Analytic Thinking.* Edited by M. Loukides and M. Blanchette, pp. 1–18. Disponible en: https://www.researchgate.net/publication/256438799_Data_Science_for_Business (Accedido: 6 Mayo 2025).

Salgado Reyes, N., Guamba Gómez, A. y Guerrero Flores, R. (2024) ‘El impacto de la tecnología de la información en la gestión empresarial’, *Nexus Research Journal*, 3(2), pp. 17–34. Disponible en: https://doi.org/10.62943/NRJ.V3N2.2024.101.

Sanmorino, A. (2024) ‘The Role of Data Science in Enhancing Web Security’. Disponible en: https://www.researchgate.net/publication/386090049_The_Role_of_Data_Science_in_Enhancing_Web_Security (Accedido: 28 Marzo 2025).

Skiena, S.S. (2017) ‘The Data Science Design Manual’. Disponible en: https://doi.org/10.1007/978-3-319-55444-0.

Syaraswati, R.A., Slamet, I. y Winarno, B. (2017) ‘Classication of Status of the Region on Java Island using C4.5, CHAID, and CART Methods’, *Journal of Physics: Conference Series*, 855(1). Disponible en: https://doi.org/10.1088/1742-6596/855/1/012053.

Taylor, L. (2017) ‘What is data justice? The case for connecting digital rights and freedoms globally,’ *Big Data and Society*, 4(2) Disponible en: https://doi.org/10.1177/2053951717736335.

Taylor, P. (2024*). ‘Data growth worldwide 2010-2028’.* Disponible en: https://www.statista.com/statistics/871513/worldwide-data-created/ (Accedido: 21 Enero 2025).

Ward, J.S. y Barker, A. (2013) ‘Undefined By Data: A Survey of Big Data Definitions’. Disponible en: http://bigdatawg.nist.gov/home.php. (Accedido: 16 Mayo 2025).

Zheng, X. *et al.* (2024) 'Data Science in Finance: Challenges and Opportunities', *AI (Switzerland)*, 5(1), pp. 55–71 Disponible en: https://doi.org/10.3390/AI5010004.

Zurita, J.A, (2024) 'Evaluación de la precisión en el pronóstico de la inflación en Bolivia: Random forest y árboles de decisión vs Arima', *Revista Compás Empresarial*, 15(39), pp. 52–80. Disponible en: https://doi.org/10.52428/20758960.V15I39.1227.