Una propuesta de formación en estadística y ciencia de datos para profesores de matemáticas

Johan Santiago Cárdenas Román

Keanu Narnovarick Guerrero Castro

Facultad de Ciencia y Tecnología

Departamento de Matemáticas

Universidad Pedagógica Nacional

Bogotá, Colombia

2025

Una propuesta de formación en estadística y ciencia de datos para profesores de matemáticas

Johan Santiago Cárdenas Román

Cód. 2020240011

Keanu Narnovarick Guerrero Castro

Cod. 2020210026

Trabajo de grado presentado como requisito parcial para optar al título de:

Licenciado en Matemáticas

Director:

Prof. César Rendón Mayorga, Mg.

Departamento de Matemáticas

Universidad Pedagógica Nacional

Facultad de Ciencia y Tecnología

Bogotá, Colombia

2025

Agradecimientos

Quiero expresar mi agradecimiento a Dios por permitirme llegar hasta aquí, por darme salud, claridad y fortaleza para culminar este trabajo.

A mi familia, por ser mi soporte permanente y por creer en mí incluso cuando yo dudaba. Su apoyo emocional, económico y espiritual fue esencial a lo largo de este proceso.

A mis amigos, quienes han sido compañeros fieles en esta trayectoria llena de altos y bajos, siempre dispuestos a escuchar, acompañar y motivar.

A la Universidad Pedagógica Nacional, por abrirme sus puertas y brindarme una formación integral, crítica y humana. A este lugar le debo grandes aprendizajes y experiencias inolvidables.

A los docentes del Departamento de Matemáticas, quienes con su compromiso y pasión por la enseñanza marcaron significativamente mi camino académico. En especial, mi gratitud para los profesores Óscar Molina, Alberto Suárez y Benjamín Sarmiento, quienes, con su dedicación, claridad y esa "milla extra" lograron que las matemáticas se convirtieran en una experiencia fascinante y significativa.

Gracias a todos los que, de una u otra manera, hicieron parte de este proceso y dejaron huella en mi formación personal y profesional.

Santiago Cárdenas

Dedicatoria

A mi novia Mafe, quien ha sido un pilar fundamental en mi vida. Gracias por tu apoyo, tu paciencia, tu compañía en los momentos difíciles y por enseñarme un amor tan especial, de esos que crecen cada día un infinito más.

Dedico también este logro a Mattie, Lila e Iris, quienes han estado conmigo en los momentos más duros de la carrera, brindándome compañía, calma y alegría. Y, de manera muy especial, a mi maestro de la educación, el profesor César Rendón, quien ha sido guía, inspiración y acompañante en cada momento importante de mi proceso formativo. Gracias por mostrarme la belleza de la estadística y la Ciencia de Datos, y por impulsarme siempre a ir un paso más allá.

Santiago Cárdenas

A Johana mi madre, por su apoyo incondicional y su fe firme en cada paso que di; a Cresencia mi abuela, por la serenidad y la paz que siempre supo transmitirme; y a Diego mi tío, por su cuidado constante y su presencia paternal. Ellos, con su amor y fortaleza, fueron el sostén que me permitió avanzar incluso en los días más exigentes. A Arena, mi compañera fiel, que con su compañía silenciosa iluminó las noches largas de trabajo y me recordó que nunca estuve solo en este camino. Y agradezco también al amor que encontré durante esta carrera, amor que apañó mis esfuerzos, celebró mis logros y resignificó esta etapa de mi vida.

A todos ellos, por haberse convertido en parte esencial de este logro, dedico este esfuerzo.

Keanu Guerrero

Resumen

Se presenta una propuesta formación en estadística y ciencia de datos dirigido a los estudiantes de la Licenciatura en Matemáticas de la Universidad Pedagógica Nacional a través del diseño de una cartilla. Esta busca fortalecer competencias estadísticas y de los futuros profesores, promoviendo una comprensión crítica y contextualizada del análisis de datos. La cartilla se fundamenta en la necesidad de superar los enfoques tradicionales de enseñanza centrados en la memorización de procedimientos y algoritmos, para avanzar hacia una educación estadística y de ciencia de datos que fomente el razonamiento, la interpretación crítica y la toma de decisiones informadas. Además, reconoce que la formación actual debe incluir no solo los conceptos básicos de la estadística descriptiva, sino también el análisis de grandes volúmenes de datos mediante software especializado. Desde esta perspectiva, esta propuesta integra herramientas de ciencia de datos que permiten modelar, visualizar y procesar información compleja, acercando a los estudiantes a prácticas contemporáneas de análisis basadas en datos reales y en la automatización computacional.

El diseño de la cartilla se estructura a partir del ciclo de datos sugerido por Lee y Delaney (2022), integrando fases de problematización, plan, análisis y conclusiones. Así mismo, se incorporan herramientas tecnológicas como R, Excel y la IA que permiten realizar análisis exploratorios y descriptivos de datos reales.

El resultado del trabajo de grado incluye el material didáctico (cartilla) que orienta la búsqueda de fomentar el pensamiento estadístico en un contexto educativo a través de la integración de la ciencia de datos.

Palabras clave: estadística, ciencia de datos, formación docente, educación estadística, pensamiento estadístico.

Abstract

This undergraduate thesis presents a proposal for designing a statistics and data science workbook aimed at students in the Mathematics Education Program at the Universidad Pedagógica Nacional. The workbook seeks to strengthen the statistical and technological competencies of future teachers, promoting a critical and contextualized understanding of data analysis. It is grounded in the need to move beyond traditional teaching approaches centered on the memorization of procedures and algorithms, advancing instead toward a form of statistical and data science education that fosters reasoning, critical interpretation, and informed decision-making. Furthermore, it recognizes that current teacher preparation must include not only basic concepts of descriptive statistics, but also the analysis of large datasets using specialized software. From this perspective, the proposal integrates data science tools that enable the modeling, visualization, and processing of complex information, bringing students closer to contemporary data-based analytical practices and computational automation.

The design of the workbook is structured around the data cycle suggested by Lee and Delaney (2022), incorporating phases of problem formulation, planning, analysis, and conclusions. Likewise, it integrates technological tools such as R, Excel, and AI to facilitate exploratory and descriptive analyses of real datasets.

The outcome of this thesis includes the didactic material (workbook) that guides its aims to foster statistical thinking in an educational context through the integration of data science.

Keywords: *statistics, data science, teacher training, statistics education, statistical thinking.*

Tabla de contenido

Capítulo 1. Aspectos Generales

1.1. Introducción

En la actualidad, la estadística y la ciencia de datos[1] ocupan un lugar central en la comprensión de fenómenos sociales, educativos, económicos y tecnológicos. En un mundo profundamente influenciado por la información, las decisiones se apoyan cada vez más en el análisis de datos, la modelación y el razonamiento basado en evidencia. En este contexto, la formación inicial de profesores de matemáticas enfrenta el reto de ir más allá de la enseñanza tradicional centrada en procedimientos, para incorporar herramientas, enfoques y modos de pensamiento propios de la ciencia de datos y de la investigación estadística contemporánea.

El presente trabajo de grado surge como respuesta a esta necesidad. Su propósito es el diseño de un material, tipo cartilla, de estadística y ciencia de datos dirigido a los estudiantes de la Licenciatura en Matemáticas de la Universidad Pedagógica Nacional. La propuesta del material se concibe como una guía formativa que integra fundamentos teóricos, actividades prácticas con datos reales, reflexión pedagógica y uso de herramientas tecnológicas como RStudio, Excel y aplicaciones de inteligencia artificial. La cartilla se organiza a través de modelos reconocidos para la investigación en educación estadística como el ciclo de datos (Lee y Delaney, 2022), que permite organizar y comprender el trabajo estadístico como un proceso sistemático de planteamiento de preguntas, análisis, validación e interpretación.

[1] En el Capítulo 2 se provee de una definición de ciencia de datos. Por ahora, basta con mencionar que se trata de un área que integra saberes matemáticos-estadísticos, computacionales y contextuales en aras del análisis de grandes volúmenes de datos.

A lo largo del documento se desarrolla un análisis teórico y metodológico que sustenta el diseño de la cartilla. En el Capítulo 1 (Marco teórico) se presenta una revisión de los fundamentos históricos y conceptuales de la ciencia de datos, resaltando sus vínculos con la estadística, la informática y la matemática, así como sus principales definiciones, clasificaciones y debates disciplinares. Además, se exploran los modos de pensamiento esenciales para su comprensión: los pensamientos computacional, estadístico y matemático. Este capítulo incluye también la Tabla PHO (Procesos-Habilidades-Objetos): instrumento que sintetiza los modos de pensamiento computacional, estadístico y matemático para interpretar problemas de manera integrada dentro del ciclo de datos y apoyar decisiones basadas en evidencia. Un instrumento diseñado para articular procesos, habilidades y objetos matemáticos, que orienta la integración entre teoría, práctica y tecnología dentro de la cartilla.

En el Capítulo 2 se desarrolla en profundidad las fases del ciclo de investigación estadística y del ciclo de datos contemporáneo, destacando cómo estos marcos orientan la formulación de problemas, la recolección y depuración de la información, la exploración inicial, la modelación, la validación y la comunicación de resultados. La articulación de ambos ciclos permite comprender el análisis de datos como un progreso integral; este capítulo se constituye como un componente central del Marco Teórico, ya que provee los fundamentos conceptuales que sustentan la propuesta pedagógica de la cartilla y guía la manera en que se abordan las actividades de análisis, interpretación y toma de decisiones basadas en datos.

En el mismo Capítulo 2 se encuentra el Marco Estadístico, que presenta los conceptos y procedimientos esenciales para el análisis de datos en contextos educativos, haciendo énfasis en la regresión logística, las métricas de evaluación y las técnicas de validación de modelos. Estos elementos proporcionan la base analítica que permite comprender, justificar e interpretar los

procesos de modelación empleados en la cartilla, además de evidenciar su aplicabilidad en situaciones escolares reales.

Posteriormente, en el Capítulo 3 (Aspectos metodológicos), se describe el proceso de construcción de la propuesta, organizado en cuatro etapas: revisión teórica, organización de la cartilla según el ciclo de datos, diseño de la cartilla didáctica y análisis de las observaciones derivadas del proceso. Cada etapa del trabajo aporta elementos teóricos, metodológicos y didácticos que, al integrarse, dan forma a un enfoque pedagógico actualizado y coherente con las necesidades actuales de la educación estadística y la ciencia de datos. Esta articulación asegura que la propuesta sea sólida y viable para su implementación en procesos de formación docente.

Por su parte, el Capítulo 4 (Diseño del Material) presenta el proceso que condujo al diseño de la cartilla propuesta. En este capítulo se describen las decisiones tomadas respecto a la selección de contenidos, la organización de los temas y la formulación de actividades y tareas dirigidas al lector. La cartilla se presenta íntegramente en el apéndice del documento.

Finalmente, el Capítulo 5 (Conclusiones), sintetiza los aportes de la propuesta al fortalecimiento de competencias estadísticas, tecnológicas y pedagógicas de los futuros profesores de matemáticas, destacando la importancia de articular teoría, práctica y herramientas digitales en la enseñanza actual de la estadística.

En conjunto este trabajo busca aportar a la transformación de la educación estadística en la formación inicial del profesor, procurando la incorporación de la ciencia de datos como un campo emergente que fortalece el pensamiento crítico, el análisis riguroso y la toma de decisiones fundamentales en evidencia.

1.2. Antecedentes

En los últimos años, la enseñanza de la estadística ha adquirido una creciente relevancia en la formación de los profesores, especialmente ante los desafíos que plantea la sociedad del conocimiento y el vertiginoso desarrollo de la ciencia de datos. Actualmente, la educación demanda que los educadores no solo dominen procedimientos matemáticos, sino que sean capaces de interpretar, analizar y comunicar información de manera crítica y fundamentada. En esta línea, un antecedente clave es la publicación de las GAISE II [*Guidelines for Assessment and Instruction in Statistics Education*] (2020), documento curricular de uso extendido en la educación estadística y que incorpora explícitamente la ciencia de datos como componente formativo desde los niveles escolares. Asimismo, el libro *Guide to Teaching Data Science* (Hazzan y Mike, 2023) aporta perspectivas relevantes sobre los modos de pensamientos necesarios en la educación basada en datos y presenta un recorrido por los principales hitos que demarcan el desarrollo de la ciencia de datos en tanto disciplina y, en particular, su llegada al sistema escolar en diferentes lugares del mundo.

Por otra parte, es reconocido que autores como Batanero (2009) y Franklin et al. (2005) resaltan que conceptos fundamentales como la variabilidad, la incertidumbre, la probabilidad y la interpretación de datos constituyen el núcleo del pensamiento estadístico. Estos conceptos son esenciales para que los profesores puedan promover en sus estudiantes procesos de razonamiento basados en evidencia y una comprensión crítica de la información, elementos indispensables para la educación del siglo XXI.

No obstante, investigaciones recientes, como la de Chávez et al. (2021), evidencian que la enseñanza estadística aún se encuentra centrada en procedimientos mecánicos y algoritmos predefinidos, relegando la comprensión conceptual y el razonamiento crítico. Este enfoque

limitado dificulta el desarrollo de competencias para analizar información de manera autónoma y para enfrentar problemas complejos de la vida cotidiana y de la sociedad del conocimiento. Estos hallazgos subrayan la importancia de repensar la didáctica de la estadística, promoviendo estrategias que integren el uso de tecnologías digitales, la interpretación de datos en contextos significativos y la aplicación de áreas emergentes, como la ciencia de datos, en tanto herramienta para resolver problemas reales.

En el contexto colombiano, el libro «Introducción a la Ciencia de Datos en R. Un enfoque práctico» (Pérez Castillo, 2020) de la Universidad Distrital Francisco José de Caldas, constituye un referente relevante de propuesta de formación en ciencia de datos. Aunque el texto no delimita explícitamente su público objetivo, su orientación aplicada y su carácter institucional permiten inferir que está pensado principalmente para estudiantes universitarios y profesionales en proceso de formación en ciencia de datos y estadística.

Esta obra propone un enfoque de aprendizaje activo, centrado en la práctica y la experimentación con datos reales mediante el uso de RStudio. La utilización de R permite a los profesores y estudiantes explorar conjuntos de «datos auténticos», en alusión a aquellos provenientes de situaciones reales, que conservan la complejidad, variabilidad y posibles inconsistencias propias del mundo real. El uso de RStudio permite a profesores y estudiantes realizar análisis descriptivos, crear visualizaciones gráficas y aplicar modelos estadísticos, facilitando así la comprensión de conceptos abstractos mediante la experiencia directa con los datos. Por ejemplo, mediante ejercicios de limpieza de datos, generación de gráficos y análisis de tendencias, los estudiantes pueden observar cómo la variabilidad y los patrones emergen de datos reales, fortaleciendo su capacidad de razonamiento estadístico y su competencia en la interpretación de información. Este enfoque evidencia la creciente necesidad de integrar

herramientas digitales en la enseñanza de la estadística, de manera que los futuros profesores puedan formar estudiantes capaces de interpretar datos en contextos reales y diversos, fortaleciendo así la alfabetización estadística desde etapas tempranas.

En este sentido, diversas experiencias pedagógicas reportadas en la literatura han demostrado que la integración de RStudio y la ciencia de datos en el aula puede transformar la enseñanza de la estadística. Estudios como los de Çetinkaya-Rundel y Ellison (2016) y Baumer et al. (2014) muestran que el trabajo con proyectos basados en datos reales como encuestas, bases abiertas o registros institucionales fomenta que los estudiantes formulen preguntas, exploren patrones y generen conclusiones fundamentadas. Asimismo, se ha evidenciado que el uso de visualizaciones interactivas y técnicas de modelación en R facilita la comprensión de conceptos centrales como variabilidad, dispersión y asociación, al vincular la teoría estadística con situaciones auténticas (Çetinkaya-Rundel y Hardin, 2020). De igual forma, prácticas que incorporan simulación, predicción y experimentos computacionales fortalecen el pensamiento estadístico y computacional, preparando a los estudiantes para enfrentar problemas reales en entornos complejos basados en datos (Hardin et al., 2015).

En consecuencia, los aportes de estos autores y referentes muestran la urgencia de diseñar estrategias didácticas innovadoras que no solo enseñen técnicas estadísticas, sino que promuevan una alfabetización estadística crítica, fundamentada en la interpretación, el análisis y la comunicación de información. La articulación entre estadística, ciencia de datos y herramientas digitales no solo amplía las competencias de los profesores, sino que también prepara a los estudiantes para enfrentar un mundo en el que la información y los datos constituyen recursos estratégicos para la toma de decisiones informadas y el desarrollo de pensamiento crítico. Así, la enseñanza de la estadística se convierte en un eje central para la formación integral, en la que la

comprensión de la variabilidad, la incertidumbre y los patrones de datos se traduce en habilidades útiles para la vida académica, profesional y social, potenciando la capacidad de los estudiantes para actuar con autonomía, juicio crítico y sentido ético en contextos complejos y basados en datos.

1.3. Justificación

En los últimos años, la formación inicial de profesores de matemáticas ha enfrentado un desafío ampliamente señalado en la literatura: la insuficiente preparación en estadística, manejo de datos y herramientas computacionales modernas. Diversos estudios (*v. g.*, Batanero y Díaz, 2011; Franklin et al., 2020; Gould, 2017) coinciden en que muchos programas de formación docente continúan centrados en enfoques tradicionales, basados en la memorización de procedimientos y el cálculo manual, lo que dificulta el desarrollo de un pensamiento estadístico profundo y contextualizado. A esto se suma la creciente presencia de la ciencia de datos en la educación escolar, reflejada en documentos internacionales como GAISE II (2020), que recomiendan explícitamente introducir procesos de análisis de datos, visualización y razonamiento computacional desde edades tempranas. Esta situación ha generado una brecha entre las demandas actuales de la enseñanza estadística y la preparación que reciben los futuros profesores.

En el contexto colombiano, esta problemática también ha sido identificada en investigaciones recientes y en los lineamientos curriculares del MEN, que reconocen la necesidad de incorporar tecnologías, análisis de datos y actividades basadas en contextos reales dentro de la formación docente. Sin embargo, los programas de formación inicial aún presentan limitaciones en la integración sistemática de herramientas como RStudio, la exploración de datos reales y el desarrollo de competencias para el análisis crítico de información. Existe, por tanto,

una necesidad académica y formativa de diseñar propuestas curriculares que articulen estadística, tecnología y didáctica, permitiendo a los futuros docentes experimentar procesos auténticos de indagación y análisis de datos.

En este escenario, el diseño de un material de estudio de estadística y ciencia de datos responde directamente a esta problemática identificada y compartida por la comunidad académica. Este trabajo no se justifica por los resultados que eventualmente pueda generar la cartilla, sino porque aporta a un campo en el que múltiples investigadores han señalado vacíos: la integración de la ciencia de datos en la formación inicial docente y la necesidad de fortalecer las habilidades de análisis, interpretación y comunicación de información basada en datos. La propuesta se alinea con discusiones contemporáneas sobre el rol del pensamiento estadístico, computacional y crítico en la educación matemática, y contribuye a la construcción de rutas formativas que preparen a los futuros profesores para los retos educativos del siglo XXI.

De este modo, la importancia del proyecto radica en que aborda una necesidad reconocida internacionalmente, ofrece una respuesta situada al contexto colombiano y se vincula con debates actuales en educación estadística. Además, abre un espacio para seguir investigando y desarrollando propuestas que integren estadística, ciencia de datos y tecnología en la formación docente, un campo que continúa en consolidación y que requiere aportes teóricos y curriculares.

1.4. Objetivos

1.4.1. Objetivo General

Diseñar un material de estudio sobre estadística y ciencia de datos, dirigido a los estudiantes de la Licenciatura en Matemáticas de la Universidad Pedagógica Nacional, que

articule fundamentos conceptuales y procedimentales de estas áreas, con el propósito de proveer una herramienta para apoyar su formación profesional inicial.

1.4.2. Objetivos específicos

1. Analizar los principales fundamentos teóricos de la ciencia de datos con el propósito de identificar los principios y enfoques que fundamenten el diseño de una cartilla didáctica centrada en la enseñanza de la estadística y la ciencia de datos.
2. Sistematizar la información referente a técnicas estadísticas como el análisis descriptivo, el análisis de componentes principales (ACP) y la regresión logística, con el fin de construir un marco conceptual que oriente el desarrollo de la cartilla.
3. Examinar en qué consiste el ciclo de datos para establecer criterios de organización y secuenciación de las actividades de la cartilla.
4. Configurar una primera versión del material didáctico (cartilla), fundamentada en la revisión teórica, que articule contenidos, tareas y guías didácticas para apoyar la compresión de conceptos básicos de ciencia de datos y el análisis estadístico.

Capítulo 2. Marco Teórico

El presente capítulo desarrolla los fundamentos conceptuales que sustentan el diseño de una cartilla de estadística y ciencia de datos. Su propósito es establecer un marco de referencia sólido que permita comprender cómo se articula la educación estadística, la ciencia de datos [2]y los enfoques contemporáneos de análisis de información dentro de la formación inicial de profesores de matemáticas. Para ello, el capítulo se organiza en varias secciones que abordan, de manera progresiva, los conceptos esenciales que fundamentan la propuesta del trabajo.

En la primera sección se presenta la ciencia de datos, examinada desde sus orígenes tanto en la estadística como en la informática. Se ofrecen definiciones ampliamente reconocidas en la literatura, se presentan distintas clasificaciones del campo y se discute su consolidación como disciplina emergente. A partir de estas revisiones se establecen los elementos centrales que permiten comprender su relación con los procesos de enseñanza y aprendizaje.

Posteriormente, se aborda la ciencia de datos como disciplina aplicada a la educación, destacando las implicaciones que tiene para la formación docente y la alfabetización basada en datos. Esta discusión permite situar el papel de la ciencia de datos en contextos educativos contemporáneos y justificar su pertinencia dentro de los procesos formativos de profesores.

En una sección siguiente se introducen los modos de pensamiento asociados al análisis de datos computacional, estadístico y matemático, considerados como procesos cognitivos fundamentales para la comprensión, exploración e interpretación de información. Asimismo, se

[2] A lo largo del documento, **CD** (Mayúsculas) se designa la ciencia de datos como disciplina interdisciplinaria, mientras que **cd** hace referencia al ciclo de datos.

presenta la Tabla de pensamiento holístico orientado a datos (PHO) como integración conceptual de estos modos, lo que permite articularlos dentro del ciclo de datos y fortalecer la interpretación global de fenómenos.

Finalmente, se describe el ciclo de datos contemporáneo, que ofrece un marco de referencia para organizar actividades de indagación, análisis e interpretación dentro del diseño de la cartilla.

En conjunto, estas secciones proporcionan los referentes teóricos necesarios para comprender las decisiones pedagógicas y metodológicas asumidas en el trabajo, así como los fundamentos conceptuales que sustentan la propuesta formativa.

2.1 Ciencia de datos

La ciencia de datos surge de una integración interdisciplinaria entre las matemáticas, la estadística y la informática. Aunque sus técnicas se aplican ampliamente en ámbitos como los negocios, la biología o la educación, estos no forman parte de su constitución disciplinar, sino que representan campos de aplicación. Como la plantea Dhar (2013), La ciencia de datos (CD en adelante) combina principios fundamentales de estadística, ciencias de la computación y matemáticas para extraer conocimiento útil a partir de grandes volúmenes de datos. Así, a primera vista, nada en la CD parecería ser nuevo, los métodos y herramientas que utiliza derivan, en su mayoría, de disciplinas que históricamente han trabajado con datos, como la estadística, la informática, la minería de datos o la bioinformática. Sin embargo, se describe brevemente el desarrollo de la CD, desde su origen hasta la actualidad con el fin de identificar por qué es un área distinta de las citadas y emergente en los últimos tiempos.

2.1.1 Los orígenes de la CD en la estadística

Donoho (2017) describe el papel de la estadística en la creación de la CD. El autor retoma al artículo de Tukey de 1962 «El futuro análisis de datos», en el que se presenta una visión prospectiva y amplia del campo de la estadística, cuyo enfoque propuesto es el análisis de datos en el lugar de la inferencia estadística, postura consonante con diversos sectores de la comunidad estadística que promovía un cambio de enfoque el cual permitiera superar la rigidez del modelo matemático como eje central del análisis.

En ese contexto, surge la noción de *aprendizaje basado en datos*, entendida como una estrategia metodológica en la cual el conocimiento, los patrones y las estructuras emergen directamente de los datos, sin imponer un «supuesto» modelo teórico estricto. A diferencia de la inferencia estadística clásica, que parte de supuestos específicos sobre distribuciones o relaciones funcionales, el aprendizaje basado en datos favorece la exploración, la visualización, el descubrimiento de patrones y la adaptabilidad a datos complejos.

Esta combinación entre estadística e informática, según Donoho (2017, p. 747) «implica la recopilación, gestión, procedimientos, análisis, visualización e interpretación de grandes cantidades de datos heterogéneos asociados con una diversa gama de aplicaciones científicas, disciplinarias e interdisciplinarias». Esté enfoque amplio y complejo, reconoce el autor, puede resultar desconcertante en un primer momento para los estadísticos, ya que representa un cambio sustancial respecto a los enfoques tradicionales de la disciplina. No obstante, en las intenciones metodológicas de estas perspectivas, como el aprendizaje basado en datos, se reconocen ya coincidencias con las pretensiones que más adelante abanderaría la CD.

Al buscar más información sobre la expresión «ciencia de datos», se encuentra la siguiente definición en el código de ciencia de datos de la *Data Science Association* (2013): «científico de datos se refiere a un profesional que utiliza métodos científicos para extraer y crear significado a partir de datos sin procesar». Si bien esta definición se centra en el perfil profesional, ofrece una aproximación indirecta a los fundamentales de la disciplina, ya que el conjunto de tareas que desempeña este profesional refleja los elementos centrales de la CD. Es decir, con base en la definición, se puede afirmar que la CD busca la utilización de métodos científicos para la extracción y creación de significados a partir de datos.

Desde la perspectiva estadística, la anterior definición de científico de datos se completa con la formulación clásica de estadística como «la ciencia que se encarga de recopilar, organizar, analizar e interpretar datos para la toma de decisiones» (Spiegel y Stephens, 2009). Para un estadístico, esta definición permite establecer puntos de contacto con el rol del científico de datos, especialmente en lo que respecta a la extracción de conocimiento a partir de datos cuantitativos. Sin embargo, la definición clásica puede resultar limitada frente al enfoque contemporáneo de la CD, dado que históricamente la estadística se ha centrado, en gran medida en la inferencia a partir de muestras pequeñas, bajo condiciones controladas y con supuestos teóricos bien definidos a causa de las limitaciones tecnológicas de épocas anteriores.

En las últimas décadas, la relación entre la estadística y la CD ha sido objeto de intensos debates dentro de la comunidad profesional. Contrario a ciertas percepciones simplificadas, los estadísticos no tratan todos los datos por igual: el tamaño, la estructura y la calidad de la información condicionan profundamente las técnicas aplicables y la validez de las conclusiones. Por su parte, la CD se distingue por una apertura metodológica más amplia, la integración

imprescindible de tecnologías digitales, el trabajo con datos masivos y la necesidad de adaptar los análisis a contextos heterogéneos y multidisciplinarios.

Este panorama ha generado tensiones conceptuales y profesionales. La estadística se encuentra en un momento particular, pues muchas de las actividades que han sido históricamente centrales en su desarrollo, análisis, visualización, inferencia y construcción de modelos; etiquetado como componentes de la CD, lo que ha llevado a cuestionamientos sobre su identidad y su rol actual. Como resultado, diversas organizaciones y figuras influyentes del campo han expresado públicamente sus posturas, las cuales reflejan la complejidad del debate:

- ¿No somos ciencia de datos? Columna de la presidenta de la ASA, Marie Davidian, en AmStat News, julio de 2013
- Un gran debate: ¿Es la ciencia de datos simplemente un "*rebranding*" de la estadística? Martin Goodson, coorganizador de la reunión de la *Royal Statistical Society* el 19 de mayo de 2015, sobre la relación entre la estadística y la ciencia de datos, en publicaciones en internet que promocionan dicho evento.
- ¿Por qué necesitamos la ciencia de datos si hemos tenido estadística durante siglos? Irving Wladawsky-Berger *Wall Street Journal*, informe del CIO, 2 de mayo de 2014
- La ciencia de datos es estadística. Para Karl Broman (2013) de la Universidad de Wisconsin asume que «cuando los físicos hacen matemáticas, no dicen que estás haciendo ciencia numérica. Están haciendo matemáticas. Si analizas datos, estás haciendo estadística. Puedes llamarlo ciencia de datos, informática, analítica o lo que sea, pero sigue siendo estadística. Puede que no te guste lo que hacen algunos estadísticos. Puede que

sientas que no comparten tus valores. Puede que te avergüencen. Pero eso no debería llevarnos a abandonar el término estadística».

- Para Gelman (2013), se plantea una postura provocadora: asume que la estadística sería la parte menos importante dentro de la CD. Esta afirmación genera controversia, ya que la estadística aporta aspectos esenciales como el diseño de experimentos, el análisis de muestras y la inferencia, que son fundamentales incluso en contextos de datos masivos.

Jeff Wu utilizó por primera vez el término CD en una conferencia ante la Academia China de Ciencias en Pekín como nombre alternativo para la estadística en 1985 (Wu al, 2021). Doce años después, en 1997, impartió una conferencia titulada «¿Estadística = Ciencia de Datos?» con motivo de su nombramiento como Cátedra H. C. Carver en la Universidad de Michigan (Wu, 1997). En su conferencia, Wu presentó su visión sobre las futuras direcciones de la estadística, incluyendo el manejo de datos grandes y complejos, el uso de redes neuronales y métodos de minería de datos, y la representación y exploración del conocimiento mediante algoritmos computacionales. Wu también sugirió un nuevo currículo de estadística más equilibrado, con mayor énfasis en la recopilación de datos, una base científica y matemática para el modelado, la computación para sistemas grandes y complejos, y un componente interdisciplinario que, para los estudiantes de pregrado, implicaba cursar una especialización en ciencias cognitivas, mientras que para los estudiantes de posgrado implicaba cursar entre el 30 % y el 50 % del currículo fuera del departamento de estadística.

Donoho (2017) propuso un cambio, introduciendo la idea del ciclo de CD, con el argumento de que los ciclos de datos propios de la estadística debían ampliarse y reorientarse. En particular, en primer lugar, propone una primera configuración fundacional de lo que más tarde se

consolidaría como CD. No se trata aún de un ciclo cerrado, sino de una estructura embrionaria que integra tres elementos fundamentales:

- Un contenido intelectual específico, centrado en la extracción de conocimiento a partir de datos.
- Una organización comprensible, que permite articular métodos, herramientas y procesos de análisis.
- Una confianza explícita en la prueba empírica, es decir, en la validación por medio de la experiencia como criterio último de verdad.

Aunque Donoho (2017) no formula explícitamente un ciclo en el sentido literal del término, la interacción entre estos tres componentes permite vislumbrar una dinámica propia de producción de conocimiento basada en datos: una práctica orientada al aprendizaje inductivo, empíricamente validado y metodológicamente estructurado. Según estas ideas, las matemáticas no se calificarían como una ciencia empírica, ya que su estándar máximo de validez se basa en la consistencia lógica y la demostrabilidad formal. En contraste, el análisis de datos supera estos tres criterios (contenido intelectual, organización comprensible y validación por experiencia), lo que permite considerarlo como una ciencia. No se define por un objeto de estudio específico, sino por un problema transversal: ¿cómo extraer conocimiento a partir de datos?

En este sentido, Cleveland (2001), planteó seis enfoques de actividad dentro de la CD, junto con una distribución sugerida en porcentajes del esfuerzo que debía dedicarse a cada uno. Estas actividades son:

- Investigaciones multidisciplinarias (25%)
- Modelos y métodos para datos (20%)

- Computación con datos (15%)
- Pedagogía (15%)
- Evaluación de herramientas (5%)
- Teoría (20%)

Esta propuesta buscaba establecer a la CD como una disciplina con identidad propia, más allá de la estadística, integrando competencias de múltiples áreas y asignándoles un valor estratégico dentro de la práctica profesional.

A partir de diversas discusiones sobre qué debe entenderse por CD, Donoho (2017) propone en su artículo *"50 Years of Data Science"* una definición estructurada de la disciplina. Su contribución consiste en organizarla en seis divisiones fundamentales, que describen tanto los campos de trabajo como las actividades operativas involucradas en un ciclo completo de análisis de datos. Estas divisiones no constituyen un listado arbitrario; Donoho las presenta como componentes interrelacionados que permiten entender qué hace un científico de datos y cómo se articula su labor en la práctica.

Las seis divisiones propuestas por Donoho son:

1. Recopilación, preparación y exploración de datos: Obtención, limpieza, integración y análisis preliminar para comprender la estructura y calidad de la información.

2. Representación y transformación de los datos: Codificación, normalización, reformulación y estructuración de los datos para hacerlos analizables.

3. Computación con datos: Uso de algoritmos, programación, infraestructura computacional y herramientas digitales para procesar información a gran escala.

4. Modelos de datos: Construcción, ajuste, validación e interpretación de modelos estadísticos, predictivos o explicativos.

5. Visualización y presentación de datos: Diseño de gráficos, narrativas y recursos que permitan comunicar resultados de manera clara y eficaz.

6. Ciencia sobre la ciencia de datos: Reflexión meta científica sobre métodos, prácticas, reproducibilidad y estándares profesionales.

Donoho plantea que estas seis divisiones conforman un primer ciclo operativo de la CD, útil para describir tanto su alcance disciplinar como su naturaleza multidimensional. Su propuesta busca mostrar que la CD no es únicamente estadística, sino una actividad que integra métodos computacionales, principios estadísticos, prácticas de ingeniería y procesos de comunicación, todo enmarcado en un flujo de trabajo coherente. Está propuesta busca integrar enfoques previos como los de Cleveland (2001) y Chambers (2008), quienes ya habían señalado la necesidad de abordar el análisis de datos desde una perspectiva interdisciplinaria. La innovación de Donoho (2017) radica en presentar un ciclo completo que estructura de manera explícita los distintos niveles de trabajo dentro de la ciencia de datos, desde la recolección inicial hasta la reflexión sobre la propia práctica científica. A continuación, se comenta las seis divisiones propuestas por el autor:

I. Recopilación, preparación y exploración de datos

Esta etapa incluye la adquisición de datos a partir de diversas fuentes, desde los métodos clásicos del diseño experimental hasta tecnologías modernas como sensores GPS, redes sociales, y monitores de búsqueda (Google Trends, NGram Viewer), el volumen y variedad de datos disponibles ha aumentado exponencialmente.

La preparación implica tareas críticas como la limpieza, transformación y organización de datos. Estas acciones permiten abordar problemas comunes como valores faltantes, formatos inconsistentes o variables irrelevantes. En términos actuales, se habla de «limpieza y manipulación de datos». La exploración de datos, influenciada por el enfoque de Análisis Exploratorio de Datos (AED) propuesto por John Tukey (1977) busca comprender patrones, anomalías o relaciones significativas antes de aplicar modelos formales.

II. Representación y transformación de los datos

Los datos se presentan en múltiples formatos, desde simples archivos de texto hasta base de datos distribuidas (SQL, NoSQL, flujos en tiempo real). El científico de datos debe adaptar estructuras y aplicar transformaciones que revelen información útil.

También se utilizan representaciones matemáticas específicas: por ejemplo, la transformada de Fourier para señales acústicas o *wavelets*[3] para imágenes y sensores. Estas herramientas permiten condensar y organizar la información para su análisis posterior.

[3] Las *wavelets* son funciones matemáticas utilizadas para descomponer señales o datos en distintos niveles de resolución, permitiendo analizar simultáneamente detalles locales y tendencias globales.

Se señala aquí que todo científico de datos debería conocer y utilizar varios lenguajes para el análisis y el procesamiento de datos. Estos pueden incluir lenguajes populares como R y Python, pero también lenguajes específicos para transformar y manipular texto, así como para gestionar procesos computacionales complejos. No es sorprendente participar en proyectos ambiciosos que utilizan media docena de lenguajes en conjunto.

Más allá del conocimiento básico de los lenguajes, los científicos de datos necesitan mantenerse al día con los nuevos lenguajes para utilizarlos eficientemente y comprender los problemas más profundos asociados con la eficiencia computacional. La computación en clústeres y en la nube, y la capacidad de ejecutar cantidades masivas de trabajos en dichos clústeres, se ha convertido en un componente sumamente poderoso del panorama computacional moderno. Para aprovechar esta oportunidad, los científicos de datos desarrollan flujos de trabajo que sistematizan las fases de recolección, procesamiento, modelación y comunicación de resultados.

III. Visualización y presentación de datos

La visualización de datos, en un extremo, se solapa con los gráficos elementales del AED (*v. g.*, histogramas, diagramas de dispersión, gráficos de series temporales), pero en la práctica moderna puede llevar a extremos mucho más elaborados. Los científicos de datos suelen dedicar mucho tiempo a decorar gráficos simples con colores o símbolos adicionales para incorporar un nuevo factor importante, y a menudo cristalizan su comprensión de un conjunto de datos mediante el desarrollo de un nuevo gráfico que lo codifica. Los científicos de datos también crean paneles para supervisar los canales de procesamiento de datos que acceden a datos en *streaming* o ampliamente distribuidos.

IV. Modelos de datos

En la práctica, los científicos de datos combinan herramientas y enfoques provenientes de ambas culturas de modelado. La primera corresponde al modelado generativo (*Random Forest),* que consiste en proponer un modelo estocástico capaz de explicar cómo podrían haberse generado los datos y, a partir de él, derivar procedimientos para inferir las propiedades del mecanismo subyacente. Este enfoque, en términos generales, coincide con la tradición de la estadística académica y sus desarrollos posteriores. La segunda es el modelado predictivo, en el que se construyen métodos que predicen con precisión sobre un universo de datos determinado, es decir, un conjunto de datos concreto muy específico. Esto coincide, a grandes rasgos, con el aprendizaje automático moderno y sus derivados industriales.

V. Ciencia sobre ciencia de datos

Los científicos de datos hacen ciencia sobre la CD cuando identifican flujos de trabajo de análisis/procesamientos comunes, por ejemplo, utilizando datos sobre su frecuencia de ocurrencia en algún ámbito académico o empresarial; cuando miden la efectividad de los flujos de trabajo estándar en términos del tiempo humano, los recursos informáticos, la validez del análisis u otras métricas de rendimiento, y cuando descubren fenómenos emergentes en el análisis de datos, por ejemplo, nuevos patrones que surgen en los flujos de trabajo de análisis de datos o artefactos perturbadores en los resultados de análisis publicados.

El alcance aquí también incluye el trabajo fundamental para posibilitar dicha ciencia en el futuro, como la codificación de la documentación de análisis y conclusiones individuales en un formato digital estándar para la posterior recopilación de metaanálisis.

A medida que el análisis de datos y el modelado predictivo se convierten en una empresa global cada vez más distribuida, la «ciencia sobre la ciencia de datos» adquirirá una importancia cada vez mayor.

La propuesta de Cleveland (2001), fue visionaria al anticipar la evolución de la CD como un campo autónomo, aunque enraizado en la estadística. Al identificar estas seis áreas claves desde la computación y la teoría hasta la pedagogía y la evaluación de herramientas, estableció un marco que no solo amplía el alcance técnico de la estadística, sino que también reconoce la naturaleza interdisciplinaria del análisis de datos en contextos reales.

En las dos décadas siguientes, es evidente que la visión de Cleveland se adelantó a una transformación que ahora se materializa plenamente: la CD ha emergido como una disciplina autónoma, con identidad propia, pero que mantiene la estadística como un componente esencial. El enfoque estadístico aporta el rigor necesario para extraer inferencias válidas, controlar la incertidumbre y generar modelos que no solo describen los datos, si no que permitan comprender y predecir fenómenos complejos.

2.1.2 Los orígenes de la CD en la informática

En 1996, Peter Naur propuso al editor de *Communications of the ACM*[4] una nueva terminología y un nuevo nombre para la informática, que enfatiza la centralidad de los datos en la computación (Naur, 1966). El objetivo de Naur era definir la datalogía (*datalogy*), como «la ciencia de la naturaleza y el uso de los datos» pues según Naur, la datalogía, podría ser un sustituto

[4] *Communications of the ACM* es la revista oficial de la Association for Computing Machinery (ACM), la mayor organización mundial de profesionales en computación. Publica artículos de investigación, revisiones y discusiones sobre avances en informática, ciencia de datos, software y tecnologías emergentes, siendo una fuente reconocida y altamente citada en la comunidad científica.

adecuado de la informática. Con ello otras ramas del conocimiento empezaron a abordar el análisis de datos desde nuevas perspectivas.

Una de ellas fue la minería de datos, una subdisciplina que surgió inicialmente en el cruce entre la estadística aplicada, la gestión de la información y la informática. Aunque sus fundamentos prácticos ya eran utilizados por estadísticos y analistas de datos, fue en la década de 1990 cuando esta metodología fue formalmente adoptada por la comunidad informática (Fayyad et al., 1996). No obstante, esta adopción no estuvo exenta de controversias: muchos estadísticos mostraron escepticismo frente a la minería de datos, al no considerarla una estrategia de investigación científicamente aceptada.

Pese a estas críticas, el enfoque ganó rápidamente notoriedad. Lovell (1993) destaca cómo el concepto se popularizó en círculos informáticos, dando lugar al desarrollo del denominado extraer conocimiento en bases de datos (*Knowledge Discovery in Data bases*, KDD). Este término fue acuñado por Gregory, Piatetsky y Shapiro en 1989 , en el que más tarde sería reconocido como el primer taller especializado en KDD. A diferencia de otros enfoques más centrados en métodos o algoritmos, el KDD hacía énfasis en que el conocimiento y no simplemente los datos debía ser el producto final de un proceso de descubrimiento riguroso. A partir de entonces, los términos «minería de datos» y «descubrimiento de conocimiento» comenzaron a utilizarse de forma casi intercambiable (Piatetsky-Shapiro, 1990), ampliando el horizonte práctico y conceptual de lo que hoy se entiende por CD.

Para el año 2000, unos diez años después de ese primer taller de KDD, Piatetsky-Shapiro (2000, p 59 -61) predijo tres ideas bastantes importantes:

1. A medida que aumenta la potencia computacional y las capacidades de almacenamiento de datos, la minería de datos y el KDD se convertirán en importantes métodos de investigación.
2. La minería de datos y las herramientas de KDD se integrarán gradualmente mejor en las bases de datos comerciales.
3. Surgirán nuevas aplicaciones de minería de datos y KDD en el comercio electrónico y el descubrimiento de fármacos.

Hoy en día la minería de datos se utiliza en diversos ámbitos de aplicación con diversos fines, García-Flores (2022), asegura que la minería de datos se emplea en distintas actividades como el mercadeo, inversiones financieras, detección de fraudes, entre otras. En estas aplicaciones la minería de datos se utiliza para extraer datos significativos de conjuntos de datos mediante técnicas estadísticas, de aprendizaje automátic, matemáticas e inteligencia artificial, (Al-Hashedi y Magalingam, 2021). Las técnicas de minería de datos también pueden utilizar la exploración de datos para integrar grandes volúmenes de información no estructurada, identificar correlaciones significativas y extraer patrones útiles (Su y Wu, 2021).

Esta capacidad de transformar datos en conocimiento accionable permitió que la minería de datos se consolidará como una fase esencial dentro del proceso de descubrimiento de conocimiento en bases de datos (KDD), integrándose así en la evolución histórica de la CD. Aunque sus raíces se encuentran en la informática y la estadística aplicada, su desarrollo en la década de 1990, especialmente tras el auge del KDD, contribuyó de forma determinante a consolidar el paradigma moderno de la CD, al reunir métodos computacionales, algoritmos estadísticos y aplicaciones prácticas de análisis automatizado.

Paralelamente al auge de la minería de datos y el KDD, otras ramas de la investigación en informática se enfrentaron a los retos de recopilar y gestionar grandes cantidades de datos, que no podían recopilarse, almacenarse ni analizarse mediante técnicas de almacenamiento convencionales (Cox y Ellsworth, 1997), aseguran que la CD depende en gran medida de la informática y de los dispositivos informáticos para recopilar y almacenar datos, analizarlos, presentar análisis y conclusiones y desarrollar sistemas basados en análisis y resultados.

La CD, tal como se concibe hoy, es el resultado de un proceso histórico interdisciplinario que integró aportes de la estadística, la informática y el análisis de datos aplicado. Desde los primeros aportes conceptuales de Tukey y Peter Huber, pasando por las propuestas estructuradas de Cleveland y Donoho, hasta la consolidación del KDD y la minería de datos en la década de 1990, la CD ha evolucionado desde una práctica empírica hasta convertirse en un campo formal de conocimiento.

El auge de la minería de datos y el descubrimiento de conocimiento en bases de datos no solo impulsó la automatización del análisis, sino que transformó la forma en que las organizaciones perciben el valor estratégico de los datos. Esto abrió paso a nuevas metodologías capaces de extraer patrones, correlaciones y conocimiento a partir de volúmenes masivos de información, aun cuando estos no presentan relaciones explícitas entre sí. En este contexto, la informática desempeñó un papel crucial al proporcionar las herramientas y capacidades técnicas necesarias para el almacenamiento, procesamiento, visualización y comunicación de resultados.

En síntesis, comprender los orígenes, debates y marcos conceptuales que han configurado la CD permite situarla dentro de un campo en constante transformación. Con este panorama, en la sección siguiente se presentan diversas definiciones que permiten precisar su alcance

disciplinar y establecer un punto de referencia conceptual para el marco teórico del presente trabajo.

2.1.3 Definiciones de CD

La diversidad de enfoques, prácticas y orígenes que confluyen en la CD ha dado lugar a múltiples definiciones del campo, cada un enfatizado aspecto distintos según su tradición disciplinar o propósito aplicado. Presentar este conjunto de definiciones permite identificar los elementos comunes y las variaciones conceptuales que existen en la literatura, así como delimitar el sentido en el que la CD será entendida en este trabajo. En particular, estas definiciones ofrecen un punto de referencia para comprender el tipo de actividades, competencias y procesos que la caracterizan, lo cual resulta fundamental para sustentar la propuesta formativa desarrollada más adelante. A partir de este marco, a continuación, se presentan algunas de las definiciones más influyentes en el ámbito académico reciente.

En la actualidad no hay una definición única de CD, pues según Wickham et al. ([2016] 2023) la CD «te permite convertir datos en bruto en comprensión, conocimiento e ideas.». De manera similar, Leek y Peng (2020) sostienen que la CD es «el proceso de formular una pregunta cuantitativa que pueda responderse con datos, recolectar y limpiar los datos, analizarlos y comunicar la respuesta a la pregunta a una audiencia relevante.». Por su parte, Baumer et al. (2021) consideran la CD una «ciencia de extraer información significativa de los datos.» y Timbers et al. (2022) la definen como «el proceso de generar conocimiento a partir de datos mediante procesos reproducibles y auditables.». Para Cassel y Topi (2015) la CD es un proceso que incluye todos los aspectos de recopilación, limpieza, organización, análisis, interpretación y visualización de hechos representados por los datos sin procesar.

Por otra parte, Craiu (2019) sostiene que la falta de certeza sobre qué es la CD es importante porque «¿quién puede realmente decir qué hace que alguien sea poeta o científico*?*». El autor continúa diciendo que un científico de datos es «alguien con formación en investigación de datos que se adhiere a enfoques o principios en la implementación de métodos estadísticos y tiene habilidades eficientes de computación» (*Data Science Association*, 2013).

Francis Edgeworth, el economista y estadístico del siglo XIX, consideraba la estadística como la ciencia «de aquellos medios que se presentan por fenómenos sociales», (Edgeworth, 1885). En cualquier caso, una característica de esta definición es que no trata los datos como *terra nullius*, o tierra de nadie. Los estadísticos tienden a ver los datos como el resultado de algún proceso que nunca se puede conocer, pero que se trata de usar para llegar a comprender. Muchos estadísticos se preocupan profundamente por los datos y la medición; sin embargo, hay muchas investigaciones en estadística donde ese tipo de cosas apenas aparecen, pertenecen a otro ámbito. Pero ese nunca es realmente el caso.

Gran parte de la discusión sobre la CD se centra en el término *ciencia*, pero como señalan Wickham et al. ([2016] 2023), también es fundamental atender al término *datos*. Esta perspectiva destaca que muchos científicos de datos son generalistas interesados en una amplia variedad de problemas, y que lo que une estos intereses es la necesidad de recolectar, limpiar y preparar datos desordenados. Con frecuencia, son los detalles específicos de estos datos los que requieren más tiempo, los que se actualizan con mayor rapidez y los que demandan la atención más cuidadosa durante el proceso analítico.

Aunque se han propuesto numerosas definiciones de CD, aún no existe un consenso sobre una única formulación aceptada. Esta dificultad se debe a la naturaleza multifacética del campo,

que puede entenderse simultáneamente como una ciencia, un método de investigación, una disciplina, un flujo de trabajo o incluso como una profesión. En este contexto, el propósito de esta sección no es establecer una definición definitiva que todavía no existe, sino mostrar la variedad de perspectivas presentes en la literatura, identificar los elementos comunes que emergen entre ellas y delimitar el sentido específico en el que la CD será entendida en este trabajo. De este modo, el listado de definiciones constituye un insumo conceptual para precisar el enfoque adoptado y orientar la construcción del marco teórico.

2.2 Pensamiento en CD

Si bien muchas disciplinas integran múltiples enfoques teóricos y metodológicos la CD se distingue por articular de manera explícita modos de pensamiento estadístico, computacional y matemático en torno a problemas centrados en datos, más allá de la aplicación aislada de técnicas o algoritmos. Esta sección explora tres enfoques clave que sustentan el pensamiento en CD: el pensamiento computacional, asociado a la informática y la capacidad de diseñar soluciones automatizadas; el pensamiento estadístico, vinculado con la inferencia, la variabilidad y la toma de decisiones basada en datos; y el pensamiento matemático, que proporcionan el lenguaje formal y la estructura lógica necesaria para modelar fenómenos complejos. A continuación, se comentan brevemente cada uno de estos.

2.2.1 Pensamiento computacional

El pensamiento computacional es reconocido hoy en día como una habilidad cognitiva esencial en la sociedad del conocimiento, con una aplicabilidad cada vez más amplia en múltiples disciplinas. Introducido originalmente por Seymour Papert en 1990, y más tarde redefinido y popularizado por Jeannette Wing en 2006, este enfoque subraya la relevancia de las

ideas fundamentales de la informática no solo en el ámbito técnico, sino en la resolución de problemas cotidianos y en la toma de decisiones informadas en diversos contextos (Boholano, 2017; Harper, 2018). Así, el pensamiento computacional ha pasado de ser una competencia especializada a consolidarse como una habilidad transversal clave en la sociedad digital contemporánea.

A lo largo del tiempo, distintos autores han ofrecido definiciones que enriquecen la comprensión del pensamiento computacional. Günbatar (2019), por ejemplo, lo describe como un proceso de resolución de problemas que implica la capacidad de diseñar soluciones implementables por una persona, una computadora o una combinación de ambas. Este enfoque destaca no sólo la dimensión técnica, sino también la flexibilidad del pensamiento computacional como herramienta intelectual.

A diferencia de otras habilidades, no se enfoca exclusivamente en la enseñanza de contenidos disciplinares específicos, sino que promueve la adquisición de conocimientos amplios, multidisciplinarios y transferibles. Se trata de una forma de pensamiento que puede aplicarse en diversos campos, desde las ciencias hasta las humanidades, pasando por la ingeniería, la economía, la medicina y, por supuesto, la CD.

Una de las grandes fortalezas del pensamiento computacional es su aplicabilidad más allá de la informática. Se ha demostrado que su incorporación en áreas como la educación, la ciencia, el arte o la gestión empresarial fomenta enfoques más estructurados, colaborativos e innovadores para la resolución de problemas. Además, este tipo de pensamiento contribuye al desarrollo de habilidades blandas esenciales en el siglo XXI, como el trabajo en equipo, la organización del tiempo, la comunicación efectiva y la planificación estratégica.

2.2.2 Pensamiento estadístico

El pensamiento estadístico constituye una de las competencias esenciales dentro del campo de la CD. Su consolidación como enfoque educativo y metodológico se remonta a los trabajos de W. Edwards Deming en 1986, quien enfatizó la importancia de comprender la variabilidad en los procesos de calidad y producción. Posteriormente, David Moore (1990) profundizó en su aplicación pedagógica, promoviendo su incorporación en la enseñanza de la estadística, y destacando su papel como herramienta cognitiva que va más allá de la simple aplicación de técnicas.

El pensamiento estadístico se define como la capacidad para comprender la esencia, estructura y variabilidad inherentes a los datos del mundo real. Según Ben-Zvi y Garfield (2004), este tipo de pensamiento implica no sólo entender el *por qué* y el *cómo* de las investigaciones estadísticas, sino también captar y aplicar los grandes conceptos que las sustentan. En este sentido, no basta con conocer fórmulas o procedimientos; es fundamental desarrollar una visión crítica y contextualizada de los datos, los métodos y los resultados.

Este enfoque se fundamenta en una serie de principios clave:

- **Reconocimiento de la variabilidad.** Los datos del mundo real presentan siempre algún grado de variación. Esta puede deberse a errores de medición, factores externos, sesgos o simplemente a la naturaleza aleatoria de los fenómenos. Reconocer y aceptar esta variabilidad es el primer paso hacia un análisis riguroso y significativo (Wild y Pfannkuch, 1999).
- **Aplicación de métodos adecuados.** El pensamiento estadístico exige saber cuándo y cómo aplicar técnicas específicas para el análisis de datos, considerando sus

supuestos, limitaciones y objetivos. Esta competencia es clave para evitar errores de interpretación y para tomar decisiones fundamentadas.

- **Comprensión del muestreo.** La inferencia estadística se basa en el análisis de muestras. Entender cómo se seleccionan, cuál es su representatividad y qué conclusiones pueden extraerse de ellas es fundamental para realizar investigaciones válidas (Moore, 1990; Ben-Zvi y Garfield, 2004).
- **Gestión de modelos estadísticos.** Más allá del uso mecánico de modelos, el pensamiento estadístico promueve una comprensión profunda de sus fundamentos, utilidad y limitaciones. Se trata de saber interpretar los modelos, ajustarlos al contexto y evaluar su pertinencia.
- **Contextualización.** Los datos no existen en el vacío. Todo análisis debe considerar el contexto en el que se generan, recolectan y analizan los datos, para que las conclusiones tengan sentido y relevancia (Cobb y Moore, 1997).
- **Proceso integral.** El pensamiento estadístico abarca todas las etapas de la investigación, desde la formulación de preguntas hasta la interpretación y comunicación de resultados. Esto incluye la recolección de datos, el análisis, la visualización y la reflexión crítica sobre los hallazgos (Franklin et al., 2005).
- **Evaluación crítica.** Finalmente, una característica esencial de este pensamiento es la capacidad de criticar y evaluar continuamente los resultados obtenidos, considerando la validez, la confiabilidad y la robustez de las conclusiones.

En el ámbito del aprendizaje automático y la CD, el pensamiento estadístico cobra una relevancia aún mayor. Permite comprender la naturaleza de los datos de entrenamiento, identificar posibles sesgos, evaluar el rendimiento de los modelos y garantizar que las

conclusiones extraídas sean sólidas y reproducibles. La falta de pensamiento estadístico puede conducir a errores graves, como el sobreajuste (*overfitting*), la mala interpretación de correlaciones espurias o la generalización indebida de patrones aprendidos.

2.2.3 Pensamiento matemático

El pensamiento matemático, por su parte, es otra piedra angular en la formación en CD. Su importancia radica en la capacidad para razonar abstractamente, modelar fenómenos complejos, y aplicar principios formales en contextos variables. A diferencia del pensamiento computacional o estadístico, el pensamiento matemático se centra en el uso de estructuras lógicas, definiciones rigurosas y relaciones entre conceptos que permiten construir soluciones generalizables y teóricamente fundamentadas.

El pensamiento matemático no solo permite desarrollar modelos más sólidos y eficientes, sino que también proporciona el marco teórico para validar su funcionamiento. Sin esta base, los algoritmos de aprendizaje automático corren el riesgo de convertirse en cajas negras sin justificación ni interpretación.

Además, la habilidad para abstraer, generalizar, establecer relaciones lógicas y construir argumentos matemáticos es fundamental para enfrentar problemas nuevos y no rutinarios, lo cual es una constante en el entorno dinámico de la CD. Como señala Devlin (2000), las matemáticas no son solo un conjunto de técnicas, sino una forma de pensar y estructurar la realidad.

2.2.4 Tabla de Procesos, Habilidades y Objetos (PHO)

La Tabla PHO (Procesos – Habilidades – Objetos Matemáticos) surge como un instrumento diseñado para articular de forma estructurada los procesos matemáticos fundamentales y las herramientas tecnológicas necesarias para la enseñanza de la estadística y la CD en el contexto de la propuesta planteada en este trabajo. La intención es brindar una guía que

facilite tanto la planificación docente como el desarrollo progresivo de competencias por parte de los estudiantes.

La creación de esta tabla responde a la necesidad de trascender el enfoque tradicional de enseñanza estadística, centrado en la repetición de algoritmos o la construcción de gráficos aislados, hacia una propuesta que promueva el pensamiento estadístico, el trabajo por proyectos, la toma de decisiones informada y el uso consciente de la tecnología. Además, busca visibilizar la riqueza didáctica que implica el cruce entre los procesos matemáticos, los objetos propios del pensamiento estadístico y las herramientas actuales de la CD.

Asimismo, la viabilidad de integrar la Tabla PHO al diseño del material que se quiere diseñar radica en que responde a marcos curriculares reconocidos como los del NCTM (2000) y los lineamientos de educación estadística en formación docente. Es flexible y adaptable a diferentes contextos y niveles de profundidad, permitiendo ajustar actividades según el avance de los estudiantes.

Por otra parte, se considera que esta Tabla PHO puede facilitar la planificación por fases de un espacio de formación, ya que cada parte del ciclo de datos puede alinearse con una combinación específica de procesos, habilidades, objetos y herramientas. Así, conecta lo conceptual con lo práctico, permitiendo que el diseño del espacio electivo virtual sea coherente tanto con las necesidades del futuro profesor de matemáticas que participa del curso como con el carácter aplicado de la CD.

Finalmente, la Tabla PHO contribuye a explicitar conexiones con otras disciplinas ya que permite abordar problemas reales que conectan con otras áreas del conocimiento y con contextos socialmente relevantes.

2.3 Ciclo de Datos

En la Sección 2.1, dedicada a la historia de la CD se ha mostrado cómo esta disciplina ha surgido como una intersección entre la estadística, la informática y el conocimiento del dominio. Su desarrollo ha sido complejo, no lineal y, en muchos casos, acompañado de controversias. Desde sus orígenes, la CD ha estado atravesada por debates acerca de su definición, sus límites, sus enfoques metodológicos y su relación con otras áreas del conocimiento.

A medida que el cd fue consolidándose en el ámbito académico y profesional, diversos autores y especialistas contribuyeron a su construcción conceptual por medio de conferencias, artículos científicos, mesas de discusión, foros interdisciplinarios e incluso contenidos divulgativos como memes, blogs o pódcast. Este conjunto de espacios académicos y divulgativos han permitido visibilizar su impacto y enriquecer la discusión sobre su papel en la investigación, la educación y la industria.

En consecuencia, el abordaje de grandes volúmenes de datos ha requerido el uso de metodologías propias de la minería de datos y de técnicas basadas en redes neuronales, capaces de identificar patrones no evidentes dentro de la información masiva. Estas aproximaciones no solo exigen una alta capacidad computacional, sino que también introducen un cambio de paradigma en el análisis, en el que se privilegia un enfoque empírico guiado por los datos, en lugar de depender exclusivamente de modelos teóricos preestablecidos. Este enfoque requiere, además, una comprensión mecanicista, es decir, el entendimiento de los procesos y relaciones causales que subyacen a los patrones observados en los datos, con el fin de interpretar los resultados de manera coherente dentro del contexto del dominio de aplicación. Bajo esta perspectiva, la CD no se limita a identificar correlaciones, sino que busca construir conocimiento

significativo que oriente la toma de decisiones fundamentadas, proceso que se operacionaliza mediante el cd.

En coherencia con lo anterior, Berman et al. (2016) propone el siguiente cd (Figura 7).

Figura 1. *El ciclo de vida (The data life cycle) (Berman et al., 2016)*

El informe de 2016 del Grupo de Trabajo de Ciencia de Datos del Comité Asesor de Ciencias de la Computación e Ingeniería de la Información de la NSF presentó una perspectiva más amplia del flujo de trabajo de la ciencia de datos, conocido como el ciclo de vida de los datos (Berman et al., 2016). El esquema (ver figura 7), amplía la perspectiva del ciclo de investigación estadística al considerar las etapas operativas y éticas del manejo de datos. El modelo comprende cinco fases: adquisición, limpieza, uso o reutilización, publicación y preservación o destrucción de los datos. A diferencia del ciclo PPDAC, que enfatiza el planteamiento del problema, el análisis y las conclusiones, este enfoque incorpora explícitamente la gestión, difusión y gobernanza de los datos, así como un marco transversal de ética, políticas regulatorias contexto del dominio. De este modo, el ciclo de datos constituye un paradigma

sociotécnico que articula prácticas estadísticas, computacionales y organizacionales de la CD contemporánea.

En la educación sobre educación en CD se reconoce que no existe un único modelo de cd adecuado para todos los contextos formativos, sino que su selección debe responder a los objetivos pedagógicos y al nivel de los estudiantes. En consecuencia, distintos autores han propuesto enfoques diferenciados del flujo de trabajo en CD, enfatizando ciertas fases según el perfil del aprendiz, el contexto educativo y la finalidad del curso. Esta diversidad de modelos plantea la necesidad de reflexionar sobre qué ciclo de datos resulta más pertinente en cada escenario formativo, lo cual se discute a continuación mediante algunas consideraciones orientadoras:

• Los estudiantes de posgrado y los investigadores se beneficiarán al aprender un ciclo de vida que enfatice las fases del flujo de trabajo de CD orientadas a la investigación (p. ej., recopilación y exploración de datos).

• Los estudiantes que no realizan investigaciones (p. ej., estudiantes de secundaria) deberían familiarizarse con uno de los modelos más simples.

• Consideraciones similares deben orientar la selección del flujo de trabajo de Ciencia de Datos en cursos de aprendizaje automático.

Sin embargo, el énfasis tradicional en algoritmos y técnicas de modelación puede generar una comprensión parcial del proceso, en la que las fases de adquisición, limpieza y exploración de datos son percibidas como secundarias o instrumentales. Esta reducción del ciclo de datos a la fase de modelación restringe la formación del pensamiento analítico y crítico, al invisibilizar las

decisiones metodológicas que anteceden y condicionan la construcción del conocimiento basado en datos. Con esta problemática Pfister (2015) propone el siguiente ciclo de datos (ver Figura 8).

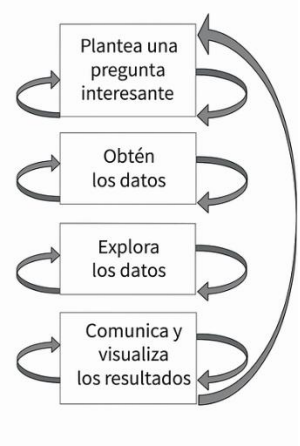


Figura 8. *The agile data science workflow (based on Pfister et al., 2015)*

Según Mike and Hazzan (2022a, 2022b)) El ciclo de datos en la CD se define como el conjunto de etapas iterativas y complementarias que permiten transformar datos brutos en conocimiento mediante un proceso de investigación sistemático. Este ciclo no sigue una secuencia lineal, sino que se caracteriza por una retroalimentación constante entre sus fases, lo que posibilita la mejora continua en la comprensión del fenómeno estudiado.

Actualmente el ciclo de vida de los datos proporciona un marco de alto nivel para representar las etapas de los datos a lo largo de su ciclo de vida (Demestichas y Daskalakis, 2020). Adaptando y ampliando el concepto de "flujos de información" de Schimmelpfennig (2016), el ciclo de vida de los datos agrícolas descrito aquí se simplifica en un proceso de tres etapas (Figura 9):

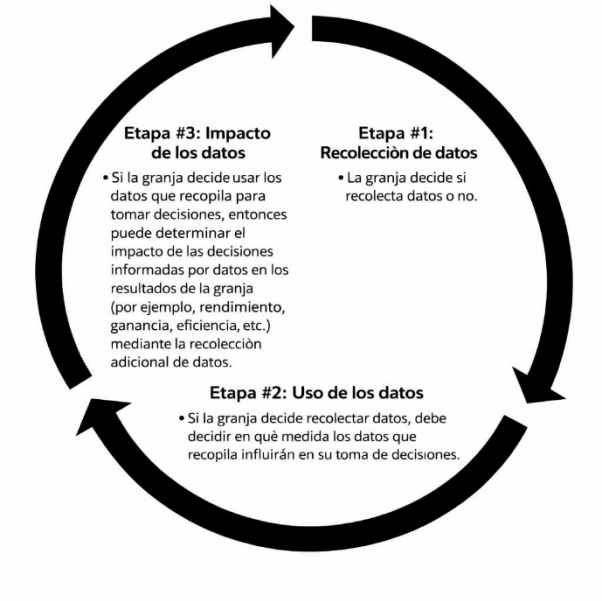


Figura 9. Flujos de información de Schimmelpfennig

1. **Recopilación de datos.** Las explotaciones agrícolas deciden si recopilar datos o no.

2.**Uso de los datos**. Las explotaciones agrícolas que recopilan datos deciden en qué medida estos influirán en su toma de decisiones.

3. **Impacto de los datos. L**as explotaciones agrícolas cuyos datos influyen en sus decisiones evalúan el impacto de las decisiones basadas en datos en la producción, la eficiencia, la rentabilidad, etc. de la explotación.

Este ejemplo permite visualizar de manera más completa cómo los datos dejan de ser menos registros o cifras aisladas para convertirse en herramientas estratégicas que apoyan la gestión agrícola. A través de su recopilación sistemática, los agricultores pueden obtener información relevante sobre el estado de sus cultivos, los recursos disponibles y las condiciones del entorno. Cuando estos datos se utilizan de forma efectiva en la toma de decisiones, permiten optimizar procesos como el riego, la fertilización, el control de plagas y la planificación de la cosecha. Finalmente, al evaluar el impacto de las decisiones basadas en datos, los productores

pueden identificar mejoras en la eficiencia, la productividad y la rentabilidad de la explotación, cerrando así un ciclo que transforma la información en conocimiento accionable y contribuye al desarrollo de una agricultura más sostenible y competitiva.

La CD no solo es una disciplina técnica, sino también un enfoque epistémico que propone una nueva forma de pensar los fenómenos y resolver problemas mediante el uso de datos. Uno de los pilares conceptuales más importantes en esta área es el ciclo de datos, una representación estructurada del proceso que va desde la recolección de datos hasta la toma de decisiones informadas y la generación de impacto real. Esta idea ha sido defendida por diversos autores clave en el desarrollo de la CD como Donoho (2017), Jeff Wu, Peter Norvig, y más recientemente por Thompson et al. (2021) en el campo de la agricultura de precisión.

En este sentido, el ciclo de datos no es simplemente una secuencia lineal, sino una dinámica iterativa que permite que los sistemas de análisis aprendan y evolucionen en función de los resultados obtenidos. Basado en modelos como el "*Farm Data Lifecycle*" desarrollado por Thompson et al. (2021), el ciclo se compone de tres grandes etapas:

1. **Recolección de datos**. Esta etapa implica la decisión consciente de capturar datos relevantes del entorno, ya sea mediante sensores, software, encuestas u otros mecanismos. La calidad y frecuencia de los datos recolectados condicionan todo el proceso posterior. Aunque la agricultura de precisión constituye un ejemplo paradigmático de esta fase debido a su alta instrumentación y dependencia de datos en tiempo real, la recolección de datos es una etapa transversal en múltiples dominios, incluyendo la salud, la educación, las ciencias sociales y la industria. En todos estos contextos, las decisiones sobre qué datos capturar, con qué frecuencia y mediante qué

instrumentos condicionan la validez de los análisis posteriores y la calidad de las inferencias realizadas.

2. **Uso de datos**. Aquí se define el grado en que los datos recopilados se utilizan para influir en las decisiones. No basta con almacenar datos; se requiere transformarlos en conocimiento accionable. Esta etapa resalta la necesidad de habilidades técnicas (análisis estadístico, visualización, modelado) y cognitivas (pensamiento crítico, contextualización) que permiten interpretar los datos y establecer relaciones causales o predictivas. Según el estudio citado, muchas decisiones clave como el manejo de nutrientes o la tasa de siembra en el ámbito agrícola se ven altamente influenciadas por estos análisis.

3. **Impacto de los datos**. Una vez que los datos se han utilizado para tomar decisiones, se debe evaluar su impacto real en los resultados (por ejemplo, en eficiencia, rendimiento, costos). Esta evaluación retroalimenta el ciclo, promoviendo mejoras continuas. En la práctica, esto no solo implica observar métricas cuantitativas, sino también adoptar un enfoque reflexivo sobre la validez de los modelos empleados y la calidad de las inferencias realizadas.

En el marco de la fase de impacto de los datos, es posible afirmar que el ciclo de datos constituye una herramienta metodológica poderosa y adaptable, capaz de guiar procesos de investigación, análisis y toma de decisiones en una amplia gama de contextos. Su estructura modular y cíclica permite su aplicación tanto en ámbitos educativos como industriales, desde proyectos de análisis social hasta estudios de rendimiento agrícola, así como en el desarrollo de algoritmos de aprendizaje automático y sistemas de recomendación en entornos digitales. Esta

versatilidad evidencia que el impacto de los datos trasciende un dominio específico y se configura como un eje transversal en la producción contemporánea de conocimiento.

En este contexto surge el ciclo de datos propuesto por Lee y Delaney (Figura 10), el cual se presenta como una evolución de modelos previos como el PPDAC. A continuación, se describen sus fases, con especial énfasis en la fase de impacto de los datos, relevante para la formación del futuro profesor.

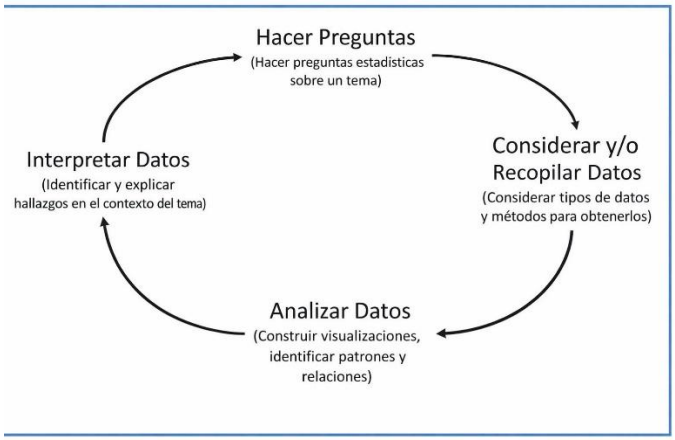


Figura 10. *Flujos de información*

Este nuevo enfoque reinterpreta el proceso de investigación con datos desde las exigencias contemporáneas de la ciencia de datos, donde la gestión de grandes volúmenes de información como se puede deducir en la Figura 10, donde se presenta el análisis mediado por herramientas computacionales y la visualización efectiva se vuelven componentes indispensables. Así, el ciclo de datos amplía la mirada tradicional al integrar, de manera equilibrada, el pensamiento estadístico, computacional y matemático, acompañado de consideraciones éticas sobre el uso responsable de la información.

Dado su carácter actualizado, flexible y formativo, este ciclo será el marco conceptual que guiará el desarrollo de la cartilla, sirviendo como hilo conductor para las actividades y orientaciones pedagógicas que se presentarán en adelante.

Análisis de partes del ciclo de datos

La fase de análisis de datos ocupa un lugar central en el flujo de trabajo, permitiendo la interpretación de la información obtenida. Aunque tradicionalmente se utiliza el modelado a través de métodos de aprendizaje automático, esta etapa puede realizarse también mediante técnicas alternativas como pruebas estadísticas o análisis cualitativo, con el objetivo de extraer información valiosa y validar las hipótesis de investigación (Mike y Hazzan, 2022).

Para Lee y Delaney (2022), este ciclo organiza de forma estructurada las etapas esenciales que todo proceso basado en datos debe seguir, permitiendo un enfoque sistemático, replicable y reflexivo. Si bien su forma circular refleja la naturaleza iterativa del análisis de datos, cada uno de sus pasos cumple una función específica en la generación de conocimiento a partir de datos. A continuación, se describen los cuatro componentes generales del ciclo de datos propuesto por los autores.

1. *Ask Questions* (Plantear preguntas)

Todo proceso de análisis de datos comienza con una pregunta estadística significativa, la cual debe estar formulada en función de un contexto específico. Esta etapa es análoga al paso "Problema" del ciclo PPDAC (Wild y Pfannkuch, 1999), donde se identifican los objetivos de la investigación. En ciencia de datos, esta etapa es crucial, ya que define el marco conceptual de todo el análisis. De acuerdo con Grolemund y Wickham (2019), una buena pregunta no solo guía el análisis, sino que también determina qué datos deben recolectarse, cómo deben analizarse y qué resultados serán considerados relevantes.

2. *Consider and/or Collect Data* (Considerar o recolectar datos)

Una vez planteada la pregunta, es necesario decidir qué tipo de datos se requieren y cómo se van a obtener. Esta fase incluye tanto la recolección directa como el uso de datos

preexistentes. Equivale al paso de “Planificación” en el ciclo PPDAC. En el entorno actual de ciencia de datos, esto implica elegir entre fuentes estructuradas o no estructuradas, definir variables clave, asegurar la calidad de los datos y garantizar su trazabilidad. Además, se toman decisiones éticas importantes, como el consentimiento informado o el respeto a la privacidad. Como señalan Baumer et al. (2017), esta etapa también implica una reflexión sobre el contexto sociotécnico del origen de los datos.

3. *Analyze Data* (Analizar datos)

Durante esta etapa, se realiza el análisis estadístico propiamente dicho, utilizando técnicas que van desde la estadística descriptiva hasta el aprendizaje automático. Aquí también se generan visualizaciones que ayudan a identificar patrones, relaciones y anomalías. Aunque en el modelo PPDAC esta parte se asocia a “Análisis”, en el ciclo de datos moderno se reconoce el papel protagónico de la visualización, un aspecto que en el enfoque PPDAC está presente pero no centralizado. La representación visual de los datos permite no solo comprender, sino también comunicar de manera efectiva los hallazgos a distintos públicos, como resaltan Cairo (2016) y Few (2009).

4. *Interpret Data* (Interpretar datos)

El análisis no está completo sin una adecuada interpretación. En esta etapa se busca dar sentido a los resultados dentro del contexto del problema planteado. Aquí se integran conocimientos del dominio para validar o refutar hipótesis, reconocer limitaciones del estudio, y generar conclusiones informadas. En el modelo PPDAC, este paso corresponde a “Conclusiones”. En ciencia de datos, además, esta interpretación puede desembocar en la generación de nuevas preguntas, cerrando así el ciclo de forma natural y fomentando la mejora continua del proceso analítico.

En suma, el ciclo de datos de Lee y Delaney no se presenta como una ruptura con modelos anteriores como el PPDAC, sino como una evolución adaptada al contexto contemporáneo de la ciencia de datos, caracterizada por grandes volúmenes de datos, técnicas computacionales avanzadas y una creciente necesidad de visualización efectiva. A través de este enfoque, se promueve una enseñanza integral que abarca tanto el pensamiento estadístico como el computacional y el matemático, integrando habilidades técnicas, cognitivas y éticas necesarias para la práctica responsable y efectiva en ciencia de datos.

Una de las mayores fortalezas del ciclo de datos es su flexibilidad conceptual. Tal como se observa en los estudios de Donoho (2017) y Thompson et al. (2021), el ciclo no impone una única forma de proceder, sino que proporciona un marco adaptable que puede ser ajustado según el tipo de datos, el dominio del problema y los objetivos específicos del proyecto. Esto permite a los equipos interdisciplinarios conformados por científicos de datos, estadísticos, ingenieros, diseñadores y expertos del dominio colaborar de forma efectiva y coherente, respetando la lógica del proceso de análisis sin sacrificar creatividad o especialización.

Además, al seguir un enfoque cíclico e iterativo, se reconoce que el análisis de datos no es un proceso lineal ni definitivo, sino un diálogo continuo entre las preguntas que nos hacemos, los datos que recolectamos, los métodos que aplicamos y las interpretaciones que obtenemos. En este sentido, el ciclo de datos también incorpora elementos del método científico, permitiendo la generación de hipótesis, su validación empírica y la retroalimentación de conocimiento hacia nuevas preguntas. Esto lo convierte en una herramienta idónea no solo para resolver problemas prácticos, sino también para fomentar el pensamiento crítico y la formación científica en estudiantes y profesionales.

Por otra parte, asumir que «cualquier ciclo de datos puede funcionar con el proyecto adecuado» implica reconocer que la eficacia del modelo no reside únicamente en su diseño estructural, sino también en su implementación contextualizada. Es decir, la utilidad del ciclo de datos dependerá de su adecuada integración con los recursos disponibles (software, hardware, conocimiento experto), la calidad de los datos, la claridad de los objetivos y, sobre todo, del compromiso analítico y ético de quienes lo ejecutan. No todos los contextos requerirán las mismas herramientas, ni todas las preguntas podrán responderse con el mismo nivel de profundidad, pero el ciclo de datos brinda un andamiaje versátil sobre el cual construir soluciones pertinentes.

En este trabajo se adopta el ciclo de datos de Lee y Delaney como marco metodológico, al considerarlo una evolución del ciclo PPDAC adaptada a las demandas contemporáneas de la ciencia de datos, caracterizadas por la disponibilidad masiva de datos, el uso de métodos computacionales y la necesidad de comunicación efectiva de resultados (Lee & Delaney, 2022). Su flexibilidad conceptual permite ajustarlo a distintos contextos educativos y de investigación, promoviendo la integración del pensamiento estadístico, computacional y matemático, así como el desarrollo de competencias técnicas y éticas (Donoho, 2017; Thompson et al., 2021)

Desde una perspectiva educativa, este ciclo proporciona un marco estructurado para diseñar experiencias de aprendizaje centradas en la resolución de problemas reales mediante datos, lo cual resulta coherente con enfoques actuales de alfabetización en datos y educación en ciencia de datos. Al incorporar fases explícitas de adquisición, exploración, análisis, interpretación y comunicación de datos, el modelo facilita la articulación entre teoría y práctica,

permitiendo que los estudiantes transiten de la formulación de preguntas a la toma de decisiones fundamentadas.

Asimismo, el ciclo de datos responde a una concepción epistemológica contemporánea del conocimiento científico como proceso iterativo y reflexivo, en el que las preguntas, los datos y los modelos se retroalimentan continuamente. Esta visión coincide con el pensamiento estadístico propuesto por Wild y Pfannkuch (1999), pero amplía su alcance al integrar componentes computacionales y de gestión de datos propios de la ciencia de datos moderna. En este sentido, el ciclo no solo orienta el análisis, sino que estructura la construcción de conocimiento en contextos complejos y dinámicos.

Desde el punto de vista metodológico, el modelo de Lee y Delaney permite sistematizar prácticas propias de la ciencia de datos, tales como la limpieza, transformación, modelación y visualización de datos, integrándolas en un flujo coherente y replicable. Esto resulta especialmente relevante en contextos educativos, donde se busca que los estudiantes comprendan no solo los resultados del análisis, sino también los procesos subyacentes que los generan. El ciclo, por tanto, funciona como un andamiaje conceptual para la enseñanza de proyectos de análisis de datos con sentido social, educativo o científico.

Finalmente, la adopción de este ciclo se justifica por su potencial para articularse con la Tabla PHO y con el diseño de la cartilla didáctica propuesta, en la medida en que permite vincular procesos matemáticos, habilidades cognitivas y objetos matemáticos dentro de un marco integrador. De este modo, el ciclo de datos no se limita a ser una referencia teórica, sino que se constituye en un eje organizador del material didáctico, orientando tanto la selección de

contenidos como la formulación de actividades y tareas para los futuros profesores de matemáticas.

2.4 Marco estadístico

En la cartilla diseñada se presentan dos técnicas estadísticos fundamentales para el análisis y la comprensión de los datos: la regresión logística y el análisis de componentes principales (ACP). La selección de estas técnicas responde al propósito de incluir un ejemplo representativo de aprendizaje supervisado (regresión logística) y otro de aprendizaje no supervisado (ACP), permitiendo así ilustrar dos formas complementarias de abordar problemas en ciencia de datos. Con ello, los estudiantes pueden reconocer cómo se modelan relaciones cuando existen variables objetivo-definidas y, al mismo tiempo, cómo es posible explorar la estructura interna de los datos cuando no se cuenta con una categoría previamente establecida. Asimismo, se incorporan procedimientos básicos de validación de modelos, esenciales para analizar la variabilidad y garantizar interpretaciones confiables.

2.4.1 Regresión logística

La regresión logística es un modelo estadístico utilizado para predecir la probabilidad de ocurrencia de un evento binario (éxito o fracaso) en función de un conjunto de variables independientes $X_1, X_2, X_3, \dots X_k$ que pueden ser cuantitativas o cualitativas.

A diferencia de la regresión lineal, la variable dependiente Y es categórica y toma valores 0 o 1.

Sea Y_i una variable binaria para el individuo i, y $X_i = (X_{i1}, X_{i2}, X_{i3}, \dots, X_{ik})$ el vector de predictores. La probabilidad condicional de éxito se expresa como:

$$P(X_i) = \pi_i = \frac{e^{\beta_0+\beta_1 X_{i1}+\beta_2 X_{i2}+\beta_3 X_{i3}+\cdots+\beta_k X_{ik}}}{1+e^{\beta_0+\beta_1 X_{i1}+\beta_2 X_{i2}+\beta_3 X_{i3}+\cdots+\beta_k X_{ik}}}$$

En donde π_i representa la probabilidad de éxito para la observación i.

El modelo puede reescribirse en términos del logit, definido como el logaritmo de la razón de probabilidades:

$$log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \cdots + \beta_k X_{ik}$$

Este modelo se estima mediante el método de máxima verosimilitud, que busca los parámetros $\hat{\beta}$ que maximizan la siguiente función:

$$L(\beta) = \prod_{i=1}^{n} \pi_i^{Y_i}(1-\pi_i)^{1-Y_i}$$

El logaritmo de la función de verosimilitud (*log-likelihood*) se expresa como:

$$l(\beta) = \sum_{i=1}^{n}[Y_i\ log\ (\pi_i)\ + (1-Y_i) log\ (1-\pi_i)]$$

La estimación se realiza iterativamente usando métodos numéricos como Newton - Raphson o *Iteratively Reweighted Least Squares* (IRLS) Hosmer et al. (2013).

Cada coeficiente β_j se interpreta en términos del odds ratio (OR) que expresa cuánto cambian las probabilidades relativas de que ocurra un evento cuando una variable explicativa aumenta en una unidad. Un OR mayor que 1 indica que el evento se vuelve más probable; un OR menor que 1 señala que es menos probable; y un OR igual a 1 implica que la variable no modifica la probabilidad del evento.:

$$OR_j = e^{\beta_j}$$

Lo cual indica el cambio multiplicativo en la razón de probabilidades por cada unidad de incremento en X_j, manteniendo las demás variables constantes.

2.4.2 Análisis de componentes principales

El análisis de componentes principales (ACP) es una técnica descriptiva multivariada cuyo objetivo es reducir la dimensionalidad de un conjunto de variables correlacionadas sin perder información relevante. Esta reducción facilita la interpretación de los datos y permite visualizar las relaciones entre observaciones y variables.

Sea una matriz de datos X de dimensión $n \times p$, con n observaciones y p variables estandarizadas ($\underline{X}_j = 0, s_j = 1$). La matriz de covarianza muestral se define como:

$$S = \frac{1}{n-1} X'X$$

El ACP busca un conjunto de vectores $a_1, a_2, \dots, a_p$ (auto vectores) que satisfacen:

$$Sa_j = \lambda_j a_j$$

Donde λ_j Son los autovalores asociados, los cuales representan la varianza explicada por cada componente principal.

Los componentes principales se obtienen como combinaciones lineales:

$$Z_j = a'_j X$$

Y cumplen que $Var(Z_j) = \lambda_j$ y que Z_j y Z_k son incorrelacionadas para cualquier $j \neq k$. El porcentaje de varianza total explicada por el conjunto de primeras m componentes se calcula como:

$$V_m = \frac{\sum_{j=1}^{m} \lambda_j}{\sum_{j=1}^{p} \lambda_j} \times 100\%$$

El ACP puede interpretarse como una rotación del sistema de coordenadas en el espacio p-dimensional de las variables, de modo que los nuevos ejes(componentes) capturen la máxima varianza de los datos.

En análisis exploratorios, el ACP facilita la creación de *biplots* que muestran simultáneamente la posición de las observaciones y la correlación entre variables, lo cual resulta útil para visualizar agrupamientos o patrones (Jolliffe y Cadima, 2016).

2.4.3 Validación de modelos

La validación de modelos busca garantizar la capacidad predictiva y generalización de un modelo estadístico, evitando el sobreajuste (*overfitting*). Este proceso evalúa qué tan bien se comporta el modelo con datos no utilizados en el ajuste (James et al., 2021). Comúnmente se divide el conjunto de datos en dos partes:

- Entrenamiento (*training set*): entre el $70\ y\ 80\%$ de los datos, usados para ajustar el modelo.
- Prueba (*test set*): entre el $30 - y\ 20\%$ de datos restante, respectivamente, usados para evaluar el desempeño del modelo construido con el conjunto de entrenamiento.

El error de predicción medio (EPM) se estima como:

$$EPM = \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} (y_i - \hat{y}_i)^2$$

En modelos de clasificación, este error puede sustituirse por la proporción de clasificación incorrectas.

2.4.4 Validación cruzada

La validación cruzada *k-fold* divide los datos en k subconjuntos de igual tamaño. El modelo se entrena en $k - 1$ subconjuntos y se valida en el restante.

El error promedio se calcula como:

$$CV(k) = \frac{1}{k} \sum_{i=1}^{k} E_i$$

Donde E_i representa el error del modelo en el i-ésima partición del conjunto de datos, dentro del procedimiento de validación cruzada.

Este método proporciona una estimación más estable del error de generalización que una simple partición entrenamiento – prueba.

2.4.5 Métricas de desempeño

En modelos logísticos o de clasificación se emplean métricas adicionales:

$$Exactitud = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Sensibilidad\ (Recall) = \frac{TP}{TP + FN}, Precisión = \frac{TP}{TP + FP}$$

$$F1 = 2 \times \frac{Precisión\ \times Recall}{Precisión + Recall}$$

Donde TP, TN, FP, FN representan los verdaderos positivos, verdaderos negativos, falsos positivos y falsos negativos respectivamente. Para evaluar qué tan bien predice el modelo, se comparan los resultados predichos con los reales, y de ahí surgen las siguientes categorías:

Verdadero positivo (TP): el modelo predijo que un estudiante tendría un puntaje alto y, efectivamente, su puntaje real fue alto.

Verdadero negativo (TN): el modelo predijo que un estudiante tendría un puntaje bajo y realmente lo obtuvo.

Falso positivo (FP): el modelo predijo que un estudiante tendría un puntaje alto, pero en realidad fue bajo.

Falso negativo (FN): el modelo predijo que el estudiante tendría un puntaje bajo, pero su resultado real fue alto.

Otra métrica ampliamente utilizada es el área bajo la curva ROC (AUC), que mide la capacidad del modelo para distinguir entre clases. Un AUC cercano a 1 indica un modelo con alta discriminación; un valor de 0.5 sugiere un desempeño aleatorio.

La métrica F1 se define como la media armónica entre Precisión y Recall, según la fórmula:

$$F1 = 2 \times \frac{Precisión \times Recall}{Precisión + Recall}$$

La precisión indica qué proporción de las predicciones positivas realizadas por un modelo son correctas, mientras que el Recall mide qué proporción de los casos positivos reales fueron correctamente identificados. El $F1$ combina estas dos medidas, penalizando los desbalances; por lo tanto, un valor alto de $F1$ solo se alcanza cuando precisión y Recall son altas simultáneamente. Esta métrica es especialmente útil en conjuntos de datos desbalanceados, ya que proporciona una evaluación más equilibrada del desempeño del modelo que considerar la precisión o el Recall por separado.

Capítulo 3. Aspectos Metodológicos

Este trabajo toma elementos de la estrategia investigativa del experimento de enseñanza, orientada al diseño y análisis de secuencias didácticas que promuevan aprendizajes significativos en contextos auténticos de aula (Steffe y Thompson, 2000). El experimento de enseñanza no se concibe como una intervención puntual, sino como un proceso iterativo, reflexivo y colaborativo que combina la planificación, la implementación y el análisis interpretativo de la experiencia.

El punto de partida metodológico es una conjetura de aprendizaje, según la cual la organización de las actividades en torno al ciclo de datos formulado por Lee y Delaney (2022) permite que los estudiantes comprendan la estadística como una práctica social significativa, desarrollen habilidades de análisis y formulen problemas basados en datos reales.

El ciclo de datos, entendido como una secuencia iterativa de fases que incluyen la formulación de preguntas, la recolección de datos, el análisis, la interpretación y la comunicación de resultados, constituye un eje estructurante del presente trabajo. Su importancia radica en que permite organizar de manera sistemática el proceso de trabajo con datos, desde la problematización inicial de un fenómeno hasta la toma de decisiones fundamentadas y la difusión del conocimiento generado. En el contexto educativo, este ciclo ofrece un marco pedagógico que orienta a los estudiantes en el desarrollo progresivo del pensamiento estadístico, computacional y matemático, al tiempo que articula habilidades de indagación, modelación, razonamiento y comunicación. Por esta razón, se considera fundamental explicitar cada una de sus fases, de modo que la cartilla didáctica no solo presente herramientas técnicas, sino que guíe a los futuros profesores en la comprensión integral del proceso de investigación basada en datos

3.1. Software R

El software R, en su entorno RStudio, fue empleado como herramienta metodológica central por su potencia en análisis estadístico, visualización de datos y modelación. R se utilizó en las fases de recolección, depuración y análisis, aplicando técnicas no solamente descriptivas sino también técnicas como la regresión logística, el ACP y la validación de los modelos, descritas en detalle en el Marco Estadístico.

La elección de R frente a otros programas responde a su carácter libre y de código abierto, su amplia comunidad académica y su versatilidad para el análisis de datos educativos.
Como señalan James et al (2021), R constituye una herramienta que permite unir teoría y práctica estadística en un mismo espacio, facilitando la comprensión profunda del ciclo analítico.

Por otra parte, a continuación, se describen los procedimientos aplicados en cada fase del análisis, con el propósito de mostrar cómo estas técnicas se integraron operativamente dentro del ciclo de datos trabajado en la cartilla.

3.2. Etapas para el desarrollo del trabajo

Etapa 1: revisión sobre CD

La primera etapa consistió en una revisión bibliográfica sobre los fundamentos conceptuales, metodológicos y didácticos de la Ciencia de Datos. Se analizaron las propuestas de Donoho (2017), Ridgway (2016) y Batanero (2009), quienes destacan la necesidad de integrar la estadística, la computación y la modelación como ejes del razonamiento moderno basado en datos. Esta revisión permitió establecer los principios teóricos de la cartilla y orientar la selección de contenidos, enfatizando el valor del pensamiento estadístico y el uso crítico de la información.

Etapa 2: propuesta de organización de la cartilla según el ciclo de datos

En esta fase se diseñó la estructura general de la cartilla tomando como eje organizador el ciclo de datos propuesto por Lee y Delaney (2022), que incluye las etapas de formular preguntas, recolectar datos, analizar e interpretar resultados y comunicar hallazgos.

Cada fase se tradujo en propósitos, contenidos y tareas orientadas al trabajo con datos reales y al uso de herramientas digitales, especialmente el software R. Esta organización permitió estructurar una secuencia progresiva que guía al lector a través del ciclo analítico de manera articulada.

Etapa 3: diseño de la cartilla

La tercera etapa se centró en la elaboración de la cartilla. El documento articula explicaciones conceptuales, ejemplos con datos reales y guías de trabajo práctico en RStudio, permitiendo a los lectores recorrer las etapas del ciclo de datos mediante actividades de recolección, limpieza, visualización y análisis estadístico.

Cada sección combina teoría, código y orientaciones de uso, con el fin de ofrecer un recurso claro y aplicable para el aprendizaje autónomo del análisis de datos en contextos educativos.

Etapa 4: ajustes pendientes y proyección de implementación

La última etapa estuvo orientada a la revisión y ajuste final del material. Sin embargo, dado que la cartilla no fue implementada en un contexto real, quedaron aspectos por desarrollar. Entre ellos se encuentran la validación de las actividades con estudiantes, la experimentación con los conjuntos de datos propuestos y la incorporación de ejemplos adicionales derivados de la práctica. Asimismo, persistieron ideas por ampliar, como el diseño de rúbricas específicas para evaluar el trabajo con datos y la inclusión de casos más robustos de modelación.

En conjunto, estos elementos constituyen líneas de trabajo futuras para fortalecer la versión final de la cartilla y su posible aplicación en escenarios educativos.

3.3. **Uso de la Tabla PHO**

La Tabla PHO funcionará como un marco de referencia y planificación para el desarrollo del material (cartilla) y de los proyectos investigativos que se implementarán durante el mismo. Su uso será doble:

1. **Diseño curricular**:

Permitirá definir claramente qué se espera que hagan los estudiantes, con qué contenidos matemáticos y mediante qué medios tecnológicos. Se usará para alinear objetivos de aprendizaje con prácticas pedagógicas pertinentes y herramientas actuales del análisis de datos.

2. **Desarrollo de proyectos:**

La Tabla PHO [Procesos – Habilidades - Objetos] guiará a los estudiantes en la selección y uso de herramientas estadísticas y tecnológicas adecuadas para sus investigaciones. Adicionalmente, permitirá mapear los procesos cognitivos y didácticos involucrados en cada paso del ciclo de datos, promoviendo una enseñanza reflexiva. Finalmente, ofrecerá una estructura clara para diseñar informes, visualizar datos, validar modelos y comunicar resultados de manera efectiva.

La Tabla PHO (Tabla 1) se propone como un instrumento de articulación entre procesos matemáticos, habilidades cognitivas y objetos matemáticos vinculados al trabajo con datos. Su estructura busca evidenciar la relación entre las acciones que realizan los estudiantes al resolver problemas basados en datos, las competencias que se espera desarrollar y los conceptos matemáticos involucrados. De este modo, la tabla orienta el diseño de la cartilla al establecer correspondencias explícitas entre el cd, los procesos matemáticos y las tareas propuestas,

permitiendo una integración coherente entre teoría, práctica y tecnología en el aula. Se prevé que la Tabla PHO aporte:

- Una visión amplia e integrada del trabajo estadístico en el aula, desde lo teórico hasta lo aplicado.
- Una herramienta concreta para implementar en un espacio electivo virtual basado en el ciclo de investigación estadística y el ciclo de datos.
- Un apoyo sistemático para construir proyectos de análisis de datos con sentido social, en los que los estudiantes puedan vivenciar cada etapa del trabajo estadístico y desarrollar competencias clave para su futuro como profesores de matemáticas.

Procesos	Habilidades	Objetos matemáticos
La formulación, tratamiento y resolución de problemas	Identificar variables relevantes en una situación real y formular preguntas estadísticas	Variable categórica y numérica
	Resolver problemas usando análisis de tendencias de datos históricos	Medidas de tendencia central y análisis temporal
	Formular preguntas de investigación a partir de fenómenos sociales o educativos observables	Variables estadísticas relevantes, tipos de datos
	Identificar errores en la recolección y limpieza de datos, proponiendo soluciones	Datos faltantes, valores atípicos
	Formular preguntas investigables a partir de fenómenos reales o escolares	Preguntas estadísticas, tipos de variables
	Elegir técnicas de muestreo adecuadas para responder a preguntas específicas	Muestreo aleatorio, estratificado, sistemático
	Traducir preguntas del contexto al lenguaje estadístico	Variables, población, muestra, parámetros
	Resolver problemas relacionados con la variabilidad y la incertidumbre en los datos	Distribuciones, errores muestrales
	Determinar el tamaño de muestra necesario para cierta confiabilidad	Cálculo del tamaño de la muestra usando fórmulas estadísticas
La modelación	Construir diagramas para representar la distribución de datos (manual o digitalmente)	Diagramas de barras, histogramas, diagramas de caja
	Utilizar diagramas de dispersión para visualizar relaciones entre variables	Gráficos bivariados, matriz de correlación
	Crear modelos lineales o no lineales para interpretar relaciones entre variables	Regresión lineal/múltiple, relaciones funcionales
	Simular escenarios futuros con base en datos recolectados	Ajuste de curvas, proyecciones estadísticas
	Ajustar modelos lineales simples a datos reales	Regresión lineal

	Evaluar la pertinencia del ajuste de un modelo	Coeficiente de determinación (R^2), residuales
	Modelar relaciones no lineales simples en fenómenos reales	Regresión cuadrática, exponencial, logarítmica
	Usar modelos de predicción en contextos educativos o sociales	Árboles de decisión simples
La comunicación	Relacionar el concepto de variabilidad con fenómenos sociales (ej. desigualdad, cambio climático)	Desviación estándar, varianza, rango intercuartílico
	Conectar el análisis de datos con otras disciplinas: economía, biología, etc.	Variables estadísticas de fenómenos reales
	Presentar oralmente un análisis de datos usando visualizaciones como apoyo	Datos tabulados, gráficas comparativas
	Explicar en un foro colaborativo las decisiones tomadas durante el análisis	Ciclo de Datos
	Elaborar informes interpretando hallazgos y su relevancia en contextos sociales o escolares	Informes con narrativas basadas en datos
	Producir infografías para comunicar visualmente los resultados del análisis	Visualización compuesta (gráficos + texto + íconos)
	Grabar presentaciones o pódcast explicando el proyecto y sus implicaciones	Argumentación, divulgación de datos
	Diseñar presentaciones orales con visualización de datos	Tableros de control, s
	Redactar informes argumentando con base en datos	Argumentos estadísticos, inferencias
	Explicar a otros la diferencia entre correlación y causalidad con ejemplos	Correlación, causalidad, variables de confusión
	Elaborar infografías para divulgar resultados de investigaciones escolares	Gráficos, mapas de datos

El razonamiento	Evaluar críticamente la calidad de los datos usados en una investigación	Muestreo, confiabilidad, sesgos
	Justificar la elección de medidas de centralización según la forma de la distribución	Media, mediana, moda
	Evaluar si una correlación implica causalidad en un conjunto de datos	Correlación lineal, regresión
	Comparar distintos modelos y justificar la elección del más apropiado	Error cuadrático medio, validación cruzada
	Contrastar hipótesis formuladas con base en evidencia estadística	Pruebas de hipótesis, p-valor, intervalos de confianza
	Justificar la elección de un método de análisis según el tipo de datos	Tipos de variables, tipo de análisis
	Identificar supuestos en modelos usados en ciencia de datos	Normalidad, independencia, homocedasticidad
	Inferir propiedades de una población a partir de una muestra	Intervalos de confianza, test de hipótesis
	Cuestionar resultados estadísticos difundidos en medios	Sesgo, mala representación, errores de interpretación
	Evaluar si un gráfico representa adecuadamente la información	Gráficos manipulados, ejes truncados, escalas engañosas
La formulación, comparación y ejercitación de procedimientos	Diseñar y automatizar rutinas para limpiar, organizar y analizar grandes volúmenes de datos	Algoritmos de limpieza, pipelines
	Crear rutinas de análisis para diferentes tipos de variables y estructuras de datos	Funciones personalizadas, scripts
	Comparar métodos de visualización para determinar cuál comunica mejor los datos	Tipos de gráficos, percepción visual

	Diseñar procedimientos para la transformación de variables cualitativas a cuantitativas	Codificación
	Comparar diferentes estrategias de imputación de datos faltantes y evaluar su impacto	Medias, moda, interpolación, regresión
	Automatizar tareas repetitivas de análisis de datos	Funciones, bucles, scripts reutilizables
	Contrastar el uso de medidas robustas vs tradicionales (media vs mediana, desviación típica vs rango intercuartílico)	Medidas de tendencia y dispersión
	Establecer criterios para seleccionar tipos de gráficos en función del público objetivo	Gráficos de cajas, dispersión con densidad
	Comparar el uso de representaciones tabulares frente a visuales para comunicar resultados	Tablas de frecuencia, s, histogramas interactivos
	Proponer rutinas para la validación de resultados en contextos escolares o sociales	Repetición del análisis con nuevas muestras, validación cruzada

Tabla 1. Tabla PHO

Para ilustrar la estructura y utilidad de la tabla presentada, a continuación, se analiza brevemente el funcionamiento de dos de sus filas. Estas muestran cómo se articulan los procesos matemáticos, las habilidades que se esperan desarrollar y los objetos matemáticos involucrados dentro de la cartilla y la cartilla.

Proceso	**Habilidad**	**Objeto Matemático**
La formulación, tratamiento y resolución de problemas	Identificar variables relevantes en una situación real y formular preguntas estadísticas	Variable categórica y numérica

Esta fila muestra cómo la cartilla orienta a los estudiantes hacia la problematización inicial del ciclo PPDAC (Problema, Plan, Datos, Análisis, Conclusiones), propuesto por Wild y Pfannkuch, (1999) como un modelo para guiar investigaciones estadísticas. Este ciclo organiza el proceso de trabajo con datos en cinco fases: la formulación del problema, la planificación del estudio, la recolección de datos, el análisis de la información y la elaboración de conclusiones. El modelo PPDAC permite articular los procesos de indagación, modelación y comunicación en contextos educativos, proporcionando un marco sistemático para el desarrollo del pensamiento estadístico y la toma de decisiones basadas en evidencia. En este sentido, el futuro profesor debe reconocer los elementos esenciales de una situación real susceptible de ser investigada mediante datos. La habilidad se centra en desarrollar la capacidad de formular preguntas investigables a partir de la identificación de variables relevantes.

Por otra parte, es posible ejemplificar cómo se relacionan los procesos, las habilidades y los objetos matemáticos en la enseñanza de la estadística y la ciencia de datos. En la Tabla 1 se muestra, por ejemplo, el proceso de modelación, que requiere la habilidad de crear modelos lineales o no lineales para interpretar relaciones entre variables, y tiene como objetos matemáticos asociados la regresión lineal o múltiple y otras relaciones funcionales. Esta representación permite visualizar de manera estructurada cómo cada proceso del pensamiento matemático se conecta con las habilidades que los estudiantes deben desarrollar y los conceptos o herramientas matemáticas que deben manejar.

Proceso	Habilidad	Objeto matemático
La modelación	Crear modelos lineales o no lineales para interpretar relaciones entre variables.	Regresión lineal/múltiple, relaciones funcionales

En este caso, la fila pertenece al proceso de modelación, fase fundamental del ciclo de análisis de datos en la cual se construyen representaciones matemáticas para interpretar fenómenos. La habilidad seleccionada evidencia el tránsito hacia prácticas propias de la ciencia de datos, donde el estudiante formula modelos que permiten explicar o predecir comportamientos a partir de datos reales. Los objetos matemáticos involucrados, la regresión lineal y múltiple y otras relaciones funcionales apuntan a la formalización de modelos utilizados contemporáneamente, reforzando la intención de la cartilla de superar enfoques basados únicamente en manuales de cálculo para incorporar herramientas analíticas y tecnológicas actuales.

Capítulo 4. Diseño del Material

Este capítulo describe el proceso de diseño, estructuración y elaboración de la cartilla sobre estadística y CD, producto central del presente trabajo de grado. Se exponen su propósito formativo, los criterios utilizados para seleccionar los temas, las decisiones estéticas y técnicas tomadas durante su construcción y la manera en que se articula con el ciclo de datos adoptado previamente en el marco metodológico. Este capítulo constituye el desarrollo del proyecto, mostrando de manera explícita cómo se materializó la propuesta planteada.

4.1. Público objetivo y propósito formativo

La cartilla está dirigida principalmente (pero no exclusivamente) a estudiantes de la Licenciatura en Matemáticas que hayan cursado estadística básica, sin restringirse a un semestre específico. Está pensada para servir como un recurso autoguiado que permita a los futuros docentes familiarizarse con el análisis de datos desde una perspectiva introductoria, situada y accesible. Las competencias esperadas se centran en:

- Comprender y utilizar lenguaje estadístico elemental.
- Realizar análisis exploratorios con datos reales.
- Utilizar software especializado para manipular, visualizar y modelar datos.
- Formular e interpretar resultados dentro de un ciclo de investigación basado en datos.

Estas competencias buscan preparar a los futuros profesores para un contexto educativo donde la alfabetización y la comunicación basada en datos son cada vez más necesarias.

4.2. Criterios para la selección de los temas de la cartilla

La cartilla desarrolla dos conjuntos de datos principales: una base de datos de Pokémon y los resultados de las pruebas Saber 11 (ICFES) del periodo 2020-2. La selección de estos conjuntos de datos responde a criterios pedagógicos y de pertinencia.

En primer lugar, la base de datos de Pokémon es sencilla, accesible, se espera que, para los estudiantes, lo que facilita la introducción a conceptos estadísticos básicos y al manejo inicial del software R. Al tratarse de un conjunto de datos limpio, pequeño y conocido, permite practicar tareas esenciales como importar información, depurar variables, explorar distribuciones y graficarlas sin dificultad técnica.

Por otro lado, el análisis de los resultados de las pruebas Saber 11 2020-2 constituye un tema directamente relevante para la formación docente. Trabajar con los datos del periodo posterior a la pandemia (2020-2) permite un acercamiento a la compresión de fenómenos reales que afectan la educación colombiana, analizar brechas, formular preguntas sobre rendimiento y caracterización estudiantil, y desarrollar modelos predictivos e interpretativos. Así, los futuros profesores se aproximan a un tipo de información que es fundamental en la toma de decisiones educativas.

4.3. Estructura general de la cartilla

La presente cartilla ha sido diseñada para acompañar a estudiantes de licenciatura en matemáticas en el desarrollo de habilidades fundamentales para el análisis y la comunicación de datos. Su propósito es ofrecer una ruta formativa progresiva, que inicia con conceptos básicos y evoluciona hacia ejercicios aplicados con datos reales, empleando el lenguaje estadístico R y el entorno de trabajo RStudio. A lo largo del texto, el lector encontrará actividades que promueven

el pensamiento crítico[5], la interpretación de resultados y la toma de decisiones informadas a partir de evidencia.

La cartilla está organizada en módulos secuenciales. Cada sección construye sobre la anterior, permitiendo un avance natural dentro del ciclo de análisis de datos: obtener – limpiar – explorar – modelar – interpretar – comunicar. El recorrido inicia con nociones esenciales del análisis de datos, continúa con el aprendizaje del software RStudio, avanza hacia prácticas aplicadas con bases de datos (Pokémon y Saber 2020- 2), y culmina con una reflexión final que busca integrar aprendizajes.

Cada módulo está diseñado para fomentar el pensamiento crítico y la toma de decisiones basadas en evidencia, integrando teoría, práctica y reflexión pedagógica. Con ello, se espera que los futuros educadores desarrollen competencias que les permitan incorporar el análisis de datos en su ejercicio profesional, comprender fenómenos educativos mediante información cuantitativa y orientar a sus estudiantes en el uso responsable e informado de los datos en el aula y en la vida cotidiana. A continuación, se presenta el contenido detallado de la cartilla, organizado por módulos y temáticas, para guiar el proceso de formación paso a paso.

Introducción

1. ¿Qué es el análisis de datos?
2. ¿Por qué es indispensable que los futuros profesores de matemáticas analicen y comuniquen datos?

[5] El término pensamiento crítico se entiende como la capacidad de analizar, evaluar y reflexionar sobre información o resultados de manera rigurosa, identificando supuestos, sesgos y posibles limitaciones, para tomar decisiones fundamentadas. Esta definición se inspira en enfoques de enseñanza de la estadística y la ciencia de datos (Facione, 2015).

Módulo 1. Fundamentos para el análisis de datos

1. Fundamentos del lenguaje estadístico
2. Introducción a R y RStudio
3. Instalación del software
    - Instalación de R
    - Instalación de RStudio
4. Explicación de la interfaz: de RStudio
    - Instalación de paquetes

Módulo 2. "Atrápalos con R": primeros pasos con datos

1. Atrápalos con R: primeros análisis de datos
2. Selección de la muestra
3. Carga de la muestra en RStudio

Módulo 3. Investigación aplicada: SABER 11 2020-2

1. Análisis de resultados SABER 11 2020-2
    - Investigación basada en el ciclo de datos
2. Preguntas de investigación y análisis de datos en RStudio
3. Selección y descarga de la muestra
4. Material: Diccionario catálogos abiertos

Módulo 4. Modelos estadísticos y análisis

1. Regresión logística (simple y múltiple)
2. Interpretación de resultados estadísticos
    - Rangos y p-valores

- Interpretación del coeficiente de correlación
- Interpretación de p-valores en correlación

3. Diagnóstico del modelo
    - Rangos y criterios formales del VIF
4. Explicación detallada de los resultados obtenidos
5. Síntesis interpretativa
    - Factores que explican el alto rendimiento en la prueba SABER 11 2020-2
6. Cierre del ciclo de datos en regresión
    - Interpretación final y conclusiones

La organización de esta cartilla busca facilitar un recorrido formativo claro, progresivo y aplicado, permitiendo que el lector transite desde conceptos básicos del análisis de datos hasta la interpretación y comunicación de resultados estadísticos con herramientas reales.

4.4. Decisiones estéticas y de diseño

La cartilla fue elaborada en Word. Los títulos se diseñan con tipografías de tamaño adecuado y llamativas que facilitan la navegación entre secciones y la identificación de categorías de contenido.

No se siguió un criterio estético formal más allá de privilegiar la claridad visual y el atractivo para los estudiantes. Las imágenes fueron creadas por los autores, adaptadas o generadas a partir de recursos propios del trabajo y ubicadas estratégicamente para acompañar explicaciones y ejemplos.

Las secciones teóricas, las tareas y los ejemplos computacionales están claramente diferenciados mediante cuadros, encabezados y variaciones cromáticas, lo cual contribuye a la organización interna del contenido.

4.5. Tareas y actividades integradas

Las actividades distribuidas en la cartilla cumplen varias funciones:

- Profundizar la comprensión conceptual mediante preguntas de investigación.
- Desarrollar habilidades técnicas a través de la ejecución de código en R.
- Promover la reflexión sobre resultados y sobre el rol del análisis de datos en contextos educativos reales.

Entre las tareas se incluyen desafíos de exploración, ejercicios comparativos, pequeñas simulaciones y análisis guiados. Cada actividad está alineada con alguna fase del ciclo de datos, asegurando coherencia entre contenido, propósito y metodología.

4.6. Articulación con el ciclo de datos.

La estructura de la cartilla incorpora específicamente el ciclo de datos y lo desarrolla en dos proyectos distintos (Pokémon y Saber 11 2020-2). En cada uno se recorre las etapas de plantear preguntas, obtener o preparar los datos, analizarlos y finalmente interpretar los resultados. De este modo, el ciclo no se presenta solo como un marco teórico, sino como una ruta práctica de trabajo que guía los procedimientos del lector.

Cada conjunto de actividades muestra de manera independiente cómo se aplica el ciclo de datos, permitiendo evidenciar que este modelo es adaptable a distintos tipos de información y objetivos analíticos.

Capítulo 5. Conclusiones

En este caso, se presenta cómo los objetivos específicos guiaron el proceso de revisión conceptual, desarrollo analítico y construcción de la cartilla, articulando los fundamentos teóricos con las prácticas de análisis de datos propuestas en el documento.

El primer objetivo específico se cumplió mediante la revisión bibliográfica de autores como Batanero, Franklin y Lee y Delaney, cuyos aportes permitieron comprender los principios del pensamiento estadístico, la importancia de la variabilidad y el papel de la CD en la educación contemporánea. Esta fundamentación quedó reflejada en el marco teórico y en la estructura conceptual de la cartilla.

El segundo objetivo específico se alcanzó a través de la ejecución de diversos códigos en RStudio, que permitieron practicar la importación, limpieza y transformación de datos, la creación de visualizaciones y la implementación de modelos básicos como el ACP o la regresión logística. También se ejercitaron procesos de estadística descriptiva en Excel, lo cual permitió afianzar la comprensión de conceptos fundamentales. Estas evidencias quedaron documentadas en capturas, scripts y explicaciones distribuidas en los anexos y a lo largo del desarrollo del trabajo, las cuales muestran de manera secuencial el proceso analítico llevado a cabo.

En relación con el tercer objetivo específico, este se cumplió con la elaboración de la cartilla que organiza las actividades en correspondencia con las fases del ciclo de datos. La cartilla ofrece ejercicios, explicaciones conceptuales y actividades prácticas en RStudio, además de integrar orientaciones pedagógicas pensadas para docentes en formación que buscan comprender cómo trabajar con datos reales en el aula. De esta manera, el documento final se consolida como un recurso que articula teoría, análisis y práctica.

Respecto al cumplimiento del objetivo general, puede afirmarse que este se alcanzó satisfactoriamente, en tanto se logró diseñar una cartilla estructurada alrededor del ciclo de datos que permite comprender y practicar procesos esenciales del análisis estadístico en contextos educativos. El uso de datos reales, la incorporación de herramientas computacionales y la integración entre fundamentos teóricos y procedimientos aplicados evidencian que el proyecto cumplió con la meta de fortalecer la formación en CD desde una perspectiva educativa.

No obstante, es importante señalar que algunos aspectos quedaron pendientes debido a los límites de tiempo y alcance del proyecto. No se trabajaron otros tipos de datos como textos, datos geoespaciales o imágenes, ni se exploraron softwares adicionales relevantes en el campo de la CD, como Python, Power BI o Tableau. Tampoco se avanzó en técnicas analíticas más complejas, como modelos predictivos avanzados, *clustering* especializado o árboles de decisión. Asimismo, varios elementos de la cartilla quedaron abiertos para su futura mejora, especialmente aquellos relacionados con el diseño de una versión digital o interactiva y la incorporación de casos más robustos de modelación. Estos aspectos constituyen líneas de desarrollo para versiones posteriores o para un eventual proyecto de investigación.

Este trabajo aportó elementos significativos tanto académicos como profesionales. Permitió adquirir una comprensión sólida de la historia y evolución de la Ciencia de Datos, así como apropiarse de conceptos centrales relacionados con variabilidad, incertidumbre y pensamiento crítico. La introducción estructurada al uso de RStudio posibilitó superar la barrera inicial frente a la programación y comprender la lógica del trabajo con código. Junto con ello, el empleo de Excel contribuyó al fortalecimiento de habilidades de estadística descriptiva y organización de datos.

En términos tecnológicos, la experiencia favoreció el aprendizaje de procedimientos esenciales como la manipulación de datos, la creación de visualizaciones, la depuración de bases y la construcción de modelos estadísticos básicos. Estas competencias digitales resultan valiosas tanto para la enseñanza como para la investigación. Desde una perspectiva profesional, el proceso permitió producir una guía que puede ser utilizada en futuras prácticas docentes y fortaleció habilidades que hacen parte del perfil contemporáneo del educador, como el pensamiento computacional, el análisis de información, la interpretación crítica de datos y la comunicación de resultados. También se obtuvieron beneficios personales, como el uso más técnico de herramientas de edición, la mejora de la redacción académica y el reconocimiento del potencial de aprender a programar desde cero.

El desarrollo del proyecto no estuvo exento de desafíos. La curva de aprendizaje inicial de RStudio representó un reto importante, especialmente en la comprensión del código. Las limitaciones de tiempo dificultaron la profundización en técnicas avanzadas o la exploración de otros softwares, y la búsqueda de bases de datos educativas depuradas también supuso un obstáculo. A esto se añadió la necesidad de ajustar el lenguaje académico a los estándares del trabajo de grado y de organizar adecuadamente la cartilla para que resultara clara y pertinente para docentes en formación. Sin embargo, estas dificultades se abordaron mediante tutorías, lectura constante y práctica continua, lo que permitió avanzar de manera significativa en los objetivos propuestos y consolidar un producto coherente y útil para la formación docente en CD.

El desarrollo de esta cartilla deja sentadas bases sólidas para avanzar en nuevas propuestas dentro de la educación estadística y la CD, y abre un conjunto de posibilidades que no pudieron ser abordadas en el marco de este trabajo. Entre estas proyecciones, una de las más relevantes es la necesidad de aplicar la cartilla con estudiantes reales, con el fin de evaluar su pertinencia, su

claridad y su impacto en situaciones auténticas de aula. Esta implementación permitiría recoger evidencias empíricas sobre cómo los futuros docentes comprenden y aplican el ciclo de datos, así como identificar ajustes necesarios en las actividades o en la secuencia propuesta.

Asimismo, queda abierta la posibilidad de incorporar otros tipos de datos que ampliarán la riqueza analítica del material, tales como datos de movilidad, redes sociales, clima, educación o texto. La inclusión de estos formatos permitiría diversificar los problemas estadísticos y computacionales que pueden abordarse, acercando a los usuarios de la cartilla a escenarios más próximos a los que se encuentran en la práctica profesional de la CD.

Otra proyección importante es el desarrollo de una aplicación o plataforma digital que acompañe la cartilla y facilite la interacción con los datos, el acceso a ejemplos guiados y la visualización dinámica de procedimientos estadísticos. Esta herramienta podría contribuir a modernizar el recurso y hacerlo más accesible para distintos perfiles de estudiantes. De igual manera, se vislumbra la posibilidad de extender la propuesta hacia áreas más avanzadas, como *machine learning*, minería de datos o analítica educativa, campos que resultan especialmente relevantes en los debates contemporáneos sobre formación docente y toma de decisiones informadas.

También se considera valiosa la integración de visualizaciones interactivas mediante entornos como Shiny, Python o Power BI, que permitirían enriquecer la experiencia del usuario y fomentar una comprensión más profunda de los datos y sus patrones. Finalmente, queda abierta la perspectiva de convertir la cartilla en un recurso educativo abierto para programas de licenciatura y de emprender investigaciones que examinen el impacto del ciclo de datos en la formación docente, con el fin de fortalecer el papel de la estadística y de la CD en la educación actual. Estas

líneas futuras representan oportunidades para ampliar, profundizar y consolidar el potencial formativo del trabajo aquí realizado.

Capítulo 6. Referencias

Al-Hashedi & Magalingam (2021): Al-Hashedi, S. A., & Magalingam, P. (2021). A review of data mining techniques in social media. *International Journal of Advanced Computer Science and Applications, 12*(1), 1–9. https://doi.org/10.14569/IJACSA.2021.0120101

Batanero, C. (2009). *Retos para la formación estadística de los profesores*. Actas do II Encontró de Probabilidades de Estadística na Escola, 52–71.

Batanero, C. (2009, November). *Statistics education at non-university levels: Current opportunities and challenges*. Paper presented at the IX Congreso Galego de Estatística e Investigación de Operacións, Ourense, Spain. Retrieved from https://www.sgapeio.es/descargas/congresos_SGAPEIO/ourense_2009/resumenes/Batanero-sgapeio2009.pdf

Batanero, C., & Díaz, C. (2011). *Estadística con proyectos*. Universidad de Granada.

Baumer, B., Kaptein, R., & Hortsmeyer, M. (2014). Data science in the statistics curriculum: Preparing students to "think with data." *The American Statistician, 68*(4), 265–270. https://doi.org/10.1080/00031305.2014.972984

Baumer, B. S., Çetinkaya-Rundel, M., Bray, A., Loi, L., & Horton, N. J. (2021). Modern data science for statistics. *Journal of Statistics and Data Science Education, 29*(1), 1–16. https://doi.org/10.1080/10691898.2020.1848485

Baumer, B. S., Kaplan, D. T., & Horton, N. J. (2017). Modern data science for statistics. *Journal of Computational and Graphical Statistics, 26*(4), 859–882. https://doi.org/10.1080/10618600.2017.1330205

Ben-Zvi, D., & Garfield, J. (Eds.). (2004). *The challenge of developing statistical literacy, reasoning and thinking*. Dordrecht: Kluwer Academic Publishers. https://doi.org/10.1007/1-4020-2278-6

Berman, F., Fox, G., Hey, T., & Hey, J. (2016). *Data lifecycle: Managing data from creation to preservation*. In *Science in the cloud: The next generation of scientific computing* (pp. 25–42). Morgan & Claypool Publishers. https://doi.org/10.2200/S00620ED1V01Y201602AIM033

Boholano, H. B. (2017). Smart social networking: 21st century teaching and learning skills. *Research in Pedagogy, 7*(1), 21–29. https://doi.org/10.17810/2015.45

Broman, K. W. (2013, July). *Data science is statistics*. *Amstat News*. American Statistical Association. https://www.amstat.org/asa/files/pdfs/AmstatNews-July13.pdf

Cassel, L., & Topi, H. (2015). Data science education: A survey of challenges and opportunities. *ACM Inroads, 6*(2), 58–61. https://doi.org/10.1145/2764917

Cairo, A. (2016). *The truthful art: Data, charts, and maps for communication*. New Riders.

Cobb, G. W., & Moore, D. S. (1997). Mathematics, statistics, and teaching. *American Mathematical Monthly, 104*(9), 801–823. https://doi.org/10.1080/00029890.1997.11990637

Cox & Ellsworth (1997): Cox, M., & Ellsworth, D. (1997). Application-controlled demand paging for out-of-core visualization. *Proceedings of the 8th IEEE Visualization Conference*, 235–244. https://doi.org/10.1109/VISUAL.1997.663888

Chambers, J. M. (2008). Software for data analysis: Programming with R. Springer. https://doi.org/10.1007/978-0-387-75936-4

Chávez, J., Hernández, A., & Rodríguez, M. (2021). Enseñanza de la estadística en educación básica: enfoques tradicionales y desafíos para el desarrollo del pensamiento estadístico. *Revista Latinoamericana de Educación Matemática, 14*(2), 45–63.

Cleveland, W. S. (2001). Data science: An action plan for expanding the technical areas of the field of statistics. International Statistical Review, 69(1), 21–26. https://doi.org/10.1111/j.1751-5823.2001.tb00477.x

Craiu, R. V. (2019). Statistics and data science: A Canadian perspective. *Canadian Journal of Statistics, 47*(4), 563–580. https://doi.org/10.1002/cjs.11525

Data Science Association. (2013). *Data Science Code of Professional Conduct.* Recuperado de https://dev.datascienceassn.org/code-conduct

Davidian, M. (2013, July). *Aren't we data science? Amstat News*, 438. American Statistical Association.

Devlin, K. (2000). *The math gene: How mathematical thinking evolved and why numbers are like gossip*. New York: Basic Books.

Demestichas, P., & Daskalakis, E. (2020). 5G on the horizon: Key challenges for data science. *IEEE Communications Magazine, 58*(3), 62–67. https://doi.org/10.1109/MCOM.001.1900490

Donoho, D. (2017). *50 years of data science. Journal of Computational and Graphical Statistics, 26*(4), 745–766. https://doi.org/10.1080/10618600.2017.1384734

Dhar, V. (2013). *Data science and prediction*. Communications of the ACM, 56(12), 64–73.

Fayyad, U. M., Piatetsky-Shapiro, G., & Smyth, P. (1996). *From data mining to knowledge discovery in databases*. AI Magazine, 17(3), 37–54. https://doi.org/10.1609/aimag.v17i3.1230

Facione, P. A. (2015). *Critical thinking: What it is and why it counts* (rev. ed.). Insight Assessment.

Few, S. (2009). *Now you see it: Simple visualization techniques for quantitative analysis*. Analytics Press.

Edgeworth, F. Y. (1885). Methods of statistics. *Journal of the Royal Statistical Society, 48*(2), 181–217. https://doi.org/10.2307/2979443

Franklin, C., Kader, G., Mewborn, D., Moreno, J., Peck, R., Perry, M., & Scheaffer, R. (2005). *A curriculum framework for K–12 statistics education: GAISE report*. American Statistical Association.

Franklin, C., Bargagliotti, A., Case, C., Kader, G., Scheaffer, R., & Spangler, D. (2020). *Guidelines for assessment and instruction in statistics education (GAISE) report II: A framework for statistics and data science education*. American Statistical Association.

Günbatar, M. S., Bakirci, H., & Karalar, H. (2019). STEM teaching intention and computational thinking skills of pre-service teachers. *Education and Information Technologies, 24*(2), 1615–1629. https://doi.org/10.1007/s10639-018-9849-5

García-Flores, D. E., Quezada-Lozada, T. C., & Quezada-Lozada, G. A. (2022). Prospectiva de políticas de acceso y permanencia a la educación superior de grupos vulnerables. *Saberes Andantes*, *4*(Especial), 170-186.

Gelman, A. E. (2013, 14 de noviembre). *Statistics is the least important part of data science*. Statistical Modeling, Causal Inference, and Social Science.

Gould, R. (2017). Data literacy is statistical literacy. *Statistics Education Research Journal*, 16(1), 22–25.

Grolemund, G., & Wickham, H. (2019). *R for data science: Import, tidy, transform, visualize, and model data*. Sebastopol, CA: O'Reilly Media. https://r4ds.had.co.nz

Hardin, J., Hoerl, R., Horton, N., Nolan, D., Baumer, B., Hall-Holt, O., Murrell, P., Peng, R., Roback, P., Temple Lang, D., & Ward, M. (2015). Data science in statistics and statistics in data science: Teaching the next generation. *The American Statistician, 69*(2), 138–148. https://doi.org/10.1080/00031305.2015.1032432

Harper, S. R. (2018). *Race matters in college*. Sterling, VA: Stylus Publishing.

Hazzan, O., & Mike, K. (2023). *Guide to teaching data science*. Springer.

Hazzan, O., & Mike, K. (2022a). *Teaching data science: Foundations and approaches*. Springer.

Hazzan, O., & Mike, K. (2022b). *Practical applications in data science education*. Springe

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (3rd ed.). Wiley.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning with applications in R* (2nd ed.). Springer.

Jolliffe, I. T., & Cadima, J. (2016). *Principal component analysis: A review and recent developments. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374*(2065), 1–16.

Leek & Peng (2020**):** Leek, J. T., & Peng, R. D. (2020). *The elements of data analytic style*. Johns Hopkins University. https://leanpub.com/datastyle

Lee, V. R., & Delaney, V. (2022). Identifying the content, lesson structure, and data use within pre-collegiate data science curricula. *Journal of Science Education and Technology, 31*(1), 81–98. https://doi.org/10.1007/s10956-021-09932-1

Lovell, M. C. (1993*). Data mining*. Review of Economics and Statistics, 75(1), 1–12. https://doi.org/10.2307/2109630

Moore, D. S. (1990). *Uncertainty*. In L. A. Steen (Ed.), *On the shoulders of giants: New approaches to numeracy* (pp. 95–137). Washington, DC: National Academy Press.

Naur, P. (1966). *The science of datalogy.* Communications of the ACM, 9(7), 485.

National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics*. Reston, VA: NCTM.

Piatetsky-Shapiro, & vG. Frawley (Eds*.),* (1989). *Knowledge discovery in databases*: An

Piatetsky-Shapiro, G. (1990). *What is data mining and KDD?* SIGKDD Explorations, 1(1), 1–5.

Pfister, H. (2015). & Blitzstein, J. *CS109: Data science course materials — Data science workflow*. Harvard University.

Pérez Castillo, J. N. (2020). *Introducción a la Ciencia de Datos en R. Un enfoque práctico*. Universidad Distrital Francisco José de Caldas.

Ridgway, J. (2016). *Implications of the data revolution for statistics education*. *International Statistical Review, 84*(3), 528–549.

Royal Statistical Society. (2015, May 19). *What is data science?* [Meeting and panel discussion]. Royal Statistical Society, London, UK.

Çetinkaya-Rundel, M., & Ellison, S. R. (2016). A fresh start for statistics: R, RStudio, and tidyverse in introductory courses. *Journal of Statistics Education, 24*(2), 1–12. https://doi.org/10.1080/10691898.2016.1190199

Çetinkaya-Rundel, M., & Hardin, J. (2020). Computing in the statistics curriculum: Data science topics and pedagogical approaches. *Journal of Statistics Education, 28*(2), 1–19. https://doi.org/10.1080/10691898.2020.1804497

Steffe, L. P., & Thompson, P. W. (2000). *Teaching experiment methodology: Underlying principles and essential elements.* In A. E. Kelly & R. A. Lesh (Eds.), Handbook of research design in mathematics and science education (pp. 267–306). Lawrence Erlbaum Associates.

Schimmelpfennig, D. (2016). *Farm profits and adoption of precision agriculture* (Economic Research Report No. 217). United States Department of Agriculture, Economic Research Service. https://www.ers.usda.gov/publications/pub-details/?pubid=80326

Spiegel, M. R., & Stephens, M. A. (2009). Statistics: The art and science of learning from data. *International Statistical Review, 77*(2), 147–163. https://doi.org/10.1111/j.1751-5823.2009.00089.x

Su, J., & Wu, J. (2021). Data science education in China: Current status and future directions. *Journal of Statistics Education, 29*(2), 1–15. https://doi.org/10.1080/10691898.2021.1917372

Timbers, T., Campbell, T., & Lee, M. (2022). *Data science: A first introduction*. Victoria, BC: University of British Columbia. https://ubc-dsci.github.io/introduction-to-datascience

Tukey, J. W. (1977). *Exploratory data analysis*. Addison-Wesley.

Thompson, K., McFarland, D. A., Khanna, S., Domingue, B. W., & Pardos, Z. A. (2021). Education data science: Past, present, future. *AERA Open, 7*(1), 1–12. https://doi.org/10.1177/23328584211052055

Wild, C. J., y Pfannkuch, M. (1999). *Statistical thinking in empirical enquiry*. *International Statistical Review, 67*(3), 223–265.

Wu, S., Wang, H., Huang, D., Zhu, X., Pan, R., Zhou, J., Gao, Y., Ma, Y., Zhu, Y., Qi, H., Li, X., Cai, L., & Hu, Q. (2021). Data science and artificial intelligence: A statistical

perspective. *National Science Review, 8*(11), nwab123. https://doi.org/10.1093/nsr/nwab123

Wu, C. F. J. (1997, August). *Statistics = Data Science?* Inaugural lecture for the H. C. Carver Chair in Statistics, University of Michigan, Ann Arbor, MI.

Wickham et al. ([2016] 2023): Wickham, H., Çetinkaya-Rundel, M., & Grolemund, G. (2016/2023). *R for data science: Import, tidy, transform, visualize, and model data*. Sebastopol, CA: O'Reilly Media. https://r4ds.had.co.nz

UNIVERSIDAD PEDAGÓGICA NACIONAL

DEPARTAMENTO DE MATEMÁTICAS

CIENCIA DE DATOS EN ACCIÓN: NAVEGANDO EL ANÁLISIS CON R

JOHAN SANTIAGO CÁRDENAS ROMÁN

KEANU NARNOVARICK GUERRERO CASTRO

Ciencia de datos en acción: navegando el análisis con R

Autores:

Johan Santiago Cárdenas Román

Keanu Narnovarick Guerrero Castro

Universidad Pedagógica Nacional

Departamento de Matemáticas

Bogotá, Colombia

2025

Título: Ciencia de datos en acción: navegando el análisis con R

Autores: Johan Santiago Cárdenas Román, Keanu Narnovarick Guerrero Castro

Programa: Licenciatura en Matemáticas

Institución: Universidad Pedagógica Nacional (UPN)

Bogotá, Colombia

2025

Revisión y asesoría:

César Guillermo Rendón Mayorga

Universidad Pedagógica Nacional

Esta cartilla fue desarrollada como trabajo de grado para optar al título de licenciado en matemáticas.

Contacto de los autores:

jscardenasr@upn.edu.co

knguerreroc@upn.edu.co

Diseño y diagramación:

Los autores

Tabla de contenido

Índice de Figuras

Índice de Tablas

Capítulo 1. Introducción a la Ciencia de Datos

1.1. Introducción

En un mundo cada vez más orientado por los datos, la capacidad de analizarlos, interpretarlos y generar conocimiento a partir de estos se ha convertido en una competencia fundamental en múltiples disciplinas. En ese contexto surge la ciencia de datos, como un campo que integra estadística, computación y conocimiento del contexto para transformar datos en decisiones informadas y soluciones innovadoras.

Esta cartilla tiene como propósito brindar al lector algunas herramientas conceptuales y prácticas básicas sobre ciencia de datos. A lo largo de los capítulos, se abordan estrategias para trabajar con diferentes tipos de datos, estructurados y no estructurados, y se promueve el desarrollo de habilidades que permitan extraer conocimiento a partir del análisis de datos, tales como:

- Limpieza y preparación de datos (manejo de valores faltantes, recodificación y organización de información).
- Visualización e interpretación descriptiva para identificar patrones, distribuciones y relaciones entre variables.
- Aplicación de técnicas específicas que se estudian en esta cartilla, como la regresión logística, el análisis discriminante lineal y el análisis de componentes principales.

- Lectura crítica y argumentada de resultados estadísticos, incluyendo interpretación de coeficientes, p-valores, medidas de ajuste y varianza explicada.
- Uso de herramientas computacionales, especialmente R y RStudio, para ejecutar procedimientos de análisis de datos.

De manera más particular, esta cartilla surge como una posibilidad para desarrollar competencias básicas en ciencia de datos en profesores de matemáticas en formación profesional inicial, aunque no se pretende excluir a otros públicos que puedan estar interesados en su lectura.

Además, se proponen diversas tareas numeradas, distribuidas estratégicamente en los capítulos. Estas tareas no constituyen ejercicios rutinarios, sino actividades orientadas a promover la comprensión conceptual, la apropiación progresiva de herramientas computacionales y el desarrollo de habilidades analíticas propias de la ciencia de datos. Cada tarea cumple una función específica dentro del proceso formativo: algunas buscan reforzar conceptos fundamentales, otras favorecen la exploración autónoma de datos, y varias están diseñadas para articular las fases del ciclo de la ciencia de datos (preguntar, recolectar, analizar, interpretar y comunicar). De este modo, las tareas operan como momentos de aplicación, reflexión y consolidación del aprendizaje, permitiendo al lector transitar de la comprensión teórica a la práctica fundamentada.

El objetivo es que el lector comprenda cómo la ciencia de datos puede apoyar la toma de decisiones estratégicas y operativas dentro de

organizaciones y proyectos en distintos campos de conocimientos. Además, se destaca el uso de entornos como RStudio, Excel y herramientas de inteligencia artificial, los cuales constituyen plataformas versátiles y poderosas para llevar a cabo análisis de datos, desarrollar aplicaciones y explorar oportunidades de innovación basadas en evidencia.

Al finalizar el estudio de esta cartilla, se espera que el lector esté en capacidad de:

- Desarrollar habilidades fundamentales para el análisis de datos, tales como la limpieza, estructuración, visualización e interpretación básica de conjuntos de datos.
- Reconocer y aplicar dos técnicas específicas de ciencia de datos que se estudian en esta cartilla: la regresión logística y las técnicas multivariadas (como el análisis discriminante lineal y el análisis de componentes principales ACP).
- Evaluar oportunidades de progreso que surgen del uso de la ciencia de datos en la toma de decisiones en proyectos académicos.

Para aprovechar de manera óptima esta cartilla, se recomienda que el lector cuente con conocimientos básicos en probabilidad y estadística (medidas de tendencia central, medidas de dispersión, distribución normal).

Si bien aquí se repasan algunos conceptos esenciales necesarios para el análisis de datos, esta revisión debe entenderse como un refuerzo puntual, no como una exposición completa de la materia. Una

formación previa en estos temas facilita la comprensión de métodos aplicados, permitiendo enfocarse en su uso práctico dentro del ciclo de la ciencia de datos. Al respecto se sugiere revisar textos de estadística como Freund, J. E. (2014). *Estadística matemática con aplicaciones*., o de probabilidad como Ross, S. M. (2010). *Introducción a la probabilidad y estadística para ingeniería y ciencias*.

Finalmente, esta cartilla está organizada en capítulos que avanzan desde conceptos fundamentales hasta análisis aplicados. El Capítulo 1 introduce los principios básicos del análisis de datos y su relevancia en la formación docente. El Capítulo 2 presenta los fundamentos de R y RStudio. El Capítulo 3 desarrolla un primer proyecto de análisis descriptivo basado en datos de Pokémon como ejercicio introductorio. El Capítulo 4 expone el proceso de análisis de una base de datos real del ICFES 2020-2. El Capítulo 5 aborda la regresión logística y la selección de variables. Siguiendo las directrices de APA 7, las capturas de pantalla que cumplen un propósito estrictamente procedimental no se numeran como figuras, por lo cual no todas las imágenes contenidas en este documento cuentan con un número de figura asociado.

1.2. ¿Qué es el análisis de datos?

Antes de abordar el uso de herramientas como R y RStudio en los capítulos siguientes, es necesario comprender qué implica el análisis de datos, qué objetivos persigue y cuál es su papel dentro de procesos de investigación y toma de decisiones.

En este sentido, el capítulo inicia con una definición clara del análisis de datos y de las etapas básicas que lo componen: inspección, limpieza, transformación, descripción y modelado de datos para que el lector tenga un panorama conceptual sólido antes de pasar a procedimientos técnicos. Con esta base, es posible establecer un puente coherente hacia los capítulos posteriores, donde se introduce el uso de R como herramienta para aplicar estos procesos en situaciones reales.

El análisis de datos es el proceso de inspeccionar, limpiar, transformar, describir y modelar datos con el objetivo de descubrir información útil, obtener conclusiones y apoyar la toma de decisiones. Puede entenderse como un puente entre los números y la comprensión del fenómeno que representan. Una forma intuitiva de pensarlo es compararlo con una investigación detectivesca: los datos son las pistas y el análisis es el método con el que se organiza la información para revelar la historia que contienen.

Aunque este capítulo se enfoca en el análisis de datos, es pertinente situarlo dentro del panorama más amplio de la ciencia de datos, disciplina que integra estadística, programación, visualización, aprendizaje automático y comunicación de resultados. La ciencia de datos no es un campo nuevo, pero ha experimentado una explosión en las últimas décadas. Históricamente, las estadísticas descriptivas y las pruebas de hipótesis eran el núcleo de la estadística. Con la llegada de la era digital, la capacidad de recolectar volúmenes masivos de datos a bajo costo, junto con el avance de la computación, ha transformado el

análisis de datos en una disciplina mucho más amplia, abarcando técnicas de aprendizaje automático (*machine learning*) o inteligencia artificial (IA). Hoy se considera un ciclo iterativo, en el que la comprensión de un conjunto de datos a menudo conduce a nuevas preguntas y a un análisis más profundo.

La importancia del análisis de datos hoy en día es inmensurable:

- En Negocios y Finanzas: permite entender el comportamiento del cliente, optimizar precios, predecir tendencias de mercado y gestionar riesgos financieros.
- En Salud: facilita la investigación de enfermedades, el desarrollo de nuevos tratamientos, la personalización de la medicina y la optimización de la gestión hospitalaria.
- En Ciencias e Ingeniería: impulsa descubrimientos en física de partículas, genética, climatología y diseño de materiales, al dar sentido a experimentos y simulaciones masivas.
- En Gobierno y Políticas Públicas: ayuda a entender patrones demográficos, evaluar la efectividad de programas sociales y optimizar la asignación de recursos.
- En Deportes y Entretenimiento: permite el análisis de rendimiento de atletas, la personalización de contenidos y la predicción de preferencias de la audiencia.

Capítulo 2. Fundamentos de R y RStudio

2.1. Fundamentos del lenguaje estadístico de RStudio

Este capítulo tiene como propósito introducir al lector en el uso del entorno estadístico R, una herramienta clave en el trabajo actual con datos.

R es un lenguaje de programación de alto nivel y un entorno computacional ampliamente utilizado en estadística, ciencia de datos y visualización de datos. Su popularidad se debe a múltiples factores: es software libre y de código abierto, cuenta con una vasta colección de paquetes especializados y permite ejecutar análisis en múltiples plataformas (Windows, macOS, Linux). Además, ofrece una enorme flexibilidad para implementar modelos estadísticos clásicos y modernos, incluyendo pruebas de hipótesis, análisis multivariado, modelos lineales y no lineales, series de tiempo, minería de datos, entre otros.

A lo largo de este capítulo, se guiará al lector por los elementos básicos del uso de R en el, desde la importación y manipulación inicial de datos, pasando por su exploración gráfica, hasta su preparación para análisis posteriores. Este recorrido básico permitirá al lector comprender cómo el uso de herramientas como R se convierte en una pieza estratégica dentro de cualquier flujo moderno de trabajo en ciencia de datos.

2.2. Instalación de R Studio

El uso de R en el análisis estadístico y en la ciencia de datos requiere de un entorno que facilite la interacción con el lenguaje y optimice el proceso de programación. En este sentido, RStudio se ha posicionado como el entorno de desarrollo integrado (IDE) más utilizado para R, gracias a su interfaz intuitiva, la integración de herramientas de análisis y visualización, así como la posibilidad de gestionar proyectos y paquetes de forma organizada. En lo que sigue, se comenta cómo instalar R y RStudio en un computador[1].

2.2.1 ¿Cómo instalar R?

1. Buscar «RStudio Desktop» en un buscador (p. ej. Google). Se obtiene un resultado igual o similar a la de la Figura 1

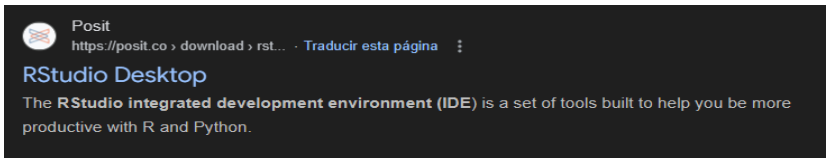


Figura 1. Sitio Web RStudio

2. Buscar la sección *Install R* en la página y presionar «*Download and install R*» (Figura 2)

[1] Las indicaciones que se presentan han sido desarrolladas para un equipo con sistema operativo Windows 11. Sin embargo, tanto R como RStudio tienen versiones para MacOS y Linux, cuya instalación es similar.

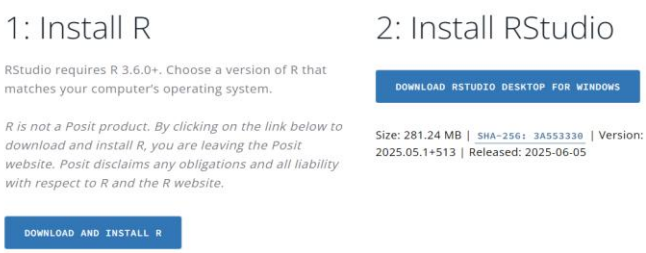


Figura 2. Instalación de R

3. Seleccionar «*Download R for Windows*» (Figura 3)

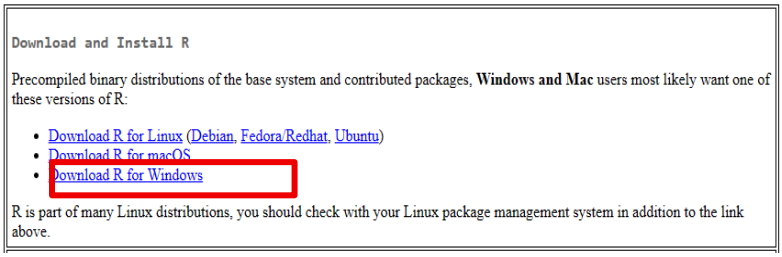


Figura 3. Download R For Windows

4. Seleccionar «*install R for the first time*» (Figura 4):

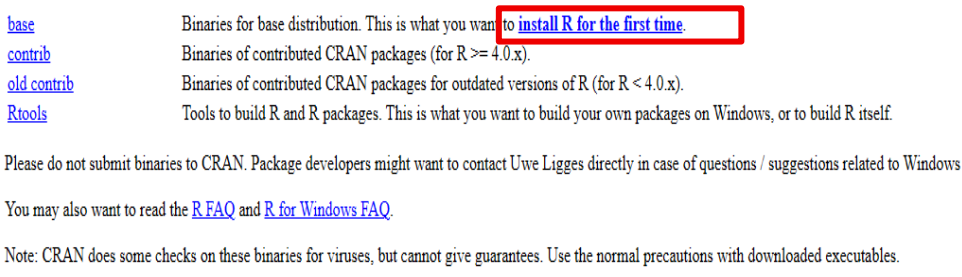


Figura 4. Instalación

5. Presionar «*Download R-4.5.1 for Windows*» (Figura 5):

Figura 5. Carpeta

6. Abrir el archivo descargado, que se encontrará en la carpeta de descargas; a continuación, seleccionar el idioma de su preferencia (Figura 6).

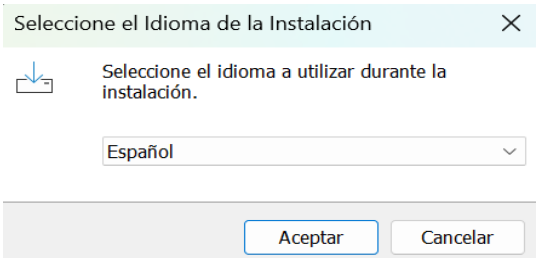


Figura 6. Selección del idioma

7. Presionar siguiente hasta finalizar el proceso

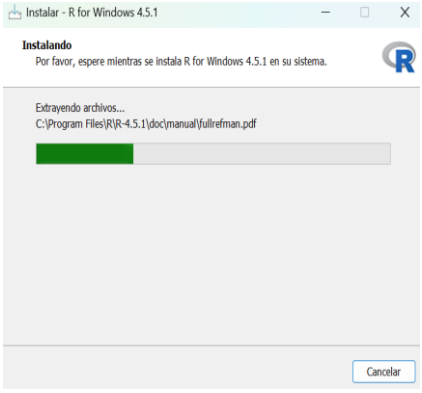


Figura 7. Finalización del proceso

8. Finalmente, en el escritorio deberán aparecer el ícono de R

Nota importante: Para instalar RStudio, una vez completados los pasos de instalación de R, el usuario debe regresar a la misma página principal en la que se encuentra la opción *Install R*, tal como se mostró en el paso 2. En ese mismo apartado debe seleccionarse ahora la opción “Install RStudio”, la cual redirige a la página oficial de descarga de *RStudio Desktop*. Allí se elige la versión gratuita y compatible con el sistema operativo correspondiente, se descarga el archivo ejecutable y se sigue el asistente de instalación aceptando las opciones predeterminadas.

2.3. Explicación de la interfaz de RStudio

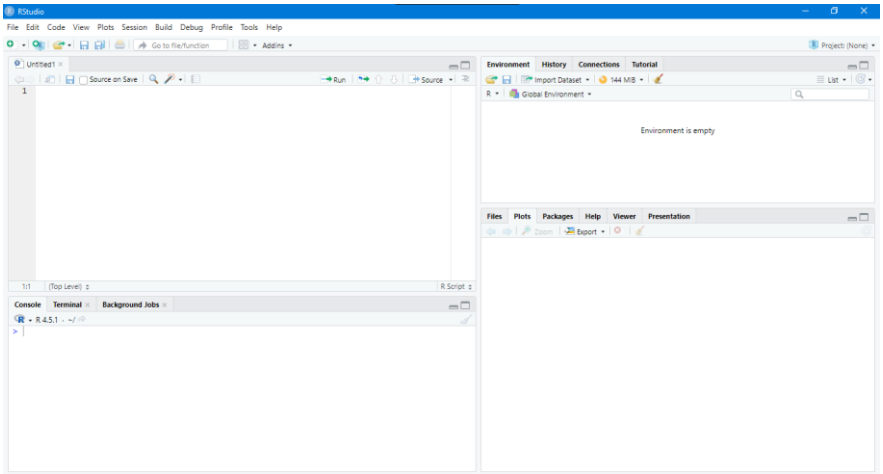
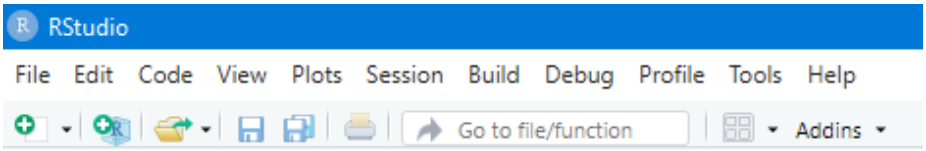


Figura 8. Interfaz de RStudio

La interfaz que se muestra en la Figura 8 corresponde al entorno que aparece al abrir RStudio. Está organizada en la barra de herramientas y cuatro paneles principales, cada uno con funciones específicas que se comentan enseguida.

1. Barra de menús y herramientas (parte superior)



- **Barra de menús**: contiene opciones como *File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, Help*.
- **Barra de herramientas**: con íconos para abrir archivos, guardar, ejecutar código, crear proyectos, entre otros. Sirve para manejar archivos, ejecutar scripts, depurar código y configurar la sesión.

2. Panel superior izquierdo – Editor de scripts

- Aquí se escriben las instrucciones a través de código de R.

- Tiene botones como: ***Run*** 

(ejecutar), ***Source***  (ejecutar todo elscript), **Guardar**, **Comentar**, etc. Es el espacio principal para programar y redactar código reproducible.
- Icono de  Guardar el documento.

3. Panel inferior izquierdo – *Consola, Terminal* y *Jobs* (ver Figura 9)

- ***Consola***: aquí se observan los resultados de la ejecución del código.
- ***Terminal (script y barras de herramientas)*** : permite usar la línea de comandos del sistema operativo sin salir de RStudio.
- ***Background Job (Archivos, Gráficas, paquetes, ayuda visor)***: muestra tareas que se ejecutan en segundo plano.

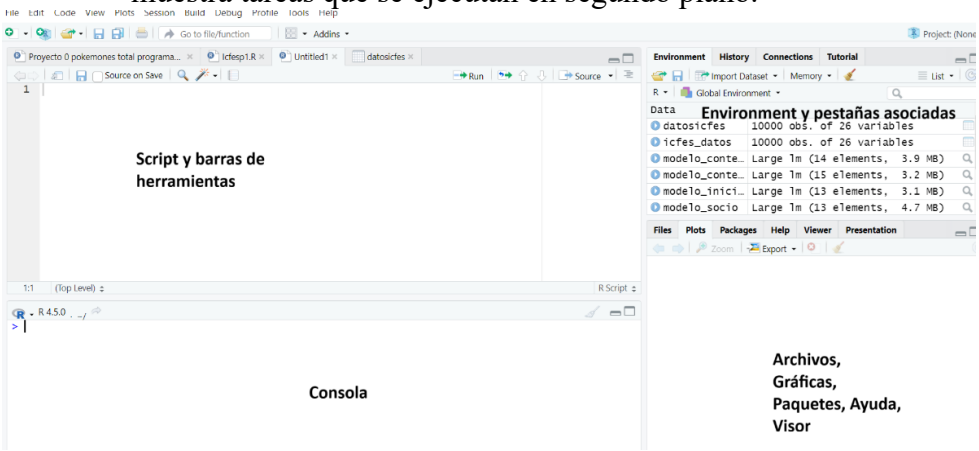


Figura 9. Descripción interfaz RStudio

4. Panel superior derecho – *Environment* y pestañas asociadas (ver Figura 9)

- ***Environment***: lista de objetos creados en la sesión (dataframes, vectores, variables, funciones).
- ***History***: historial de comandos ejecutados.
- ***Connections***: conecta R con bases de datos externas.
- ***Tutorial***: acceso a tutoriales interactivos.

5. Panel inferior derecho – Archivos, Gráficas, Paquetes, Ayuda, Visor (ver Figura 9)

- ***Files***: explorador de archivos en el directorio de trabajo.
- ***Plots***: muestra las gráficas generadas.
- ***Packages***: lista y administración de paquetes instalados.
- ***Help****:* documentación de funciones y paquetes.
- ***Viewer***: muestra reportes HTML, aplicaciones Shiny o contenido web generado en R.
- ***Presentación***: pestaña para presentaciones creadas con R Markdown.

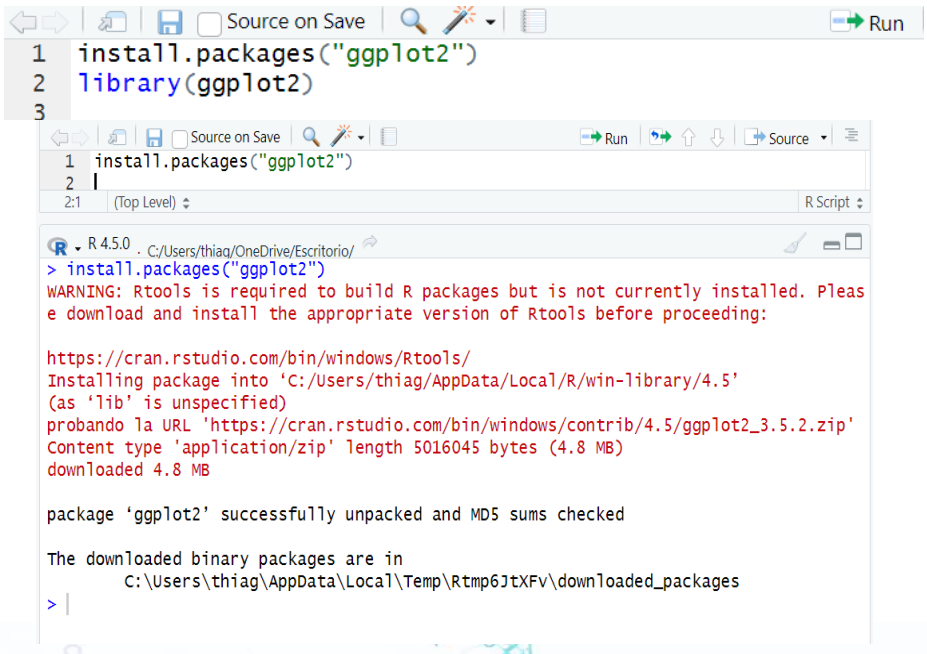
2.4. Instalación de paquetes

La instalación de paquetes es un paso fundamental en el uso de R, ya que permite al usuario adaptar el entorno de trabajo a las necesidades de su proyecto y aprovechar herramientas desarrolladas por la comunidad de programadores de R. En esta sección se presentan de manera sencilla los pasos necesarios para instalar y

cargar paquetes. Para empezar a utilizar las funciones del paquete, es necesario cargarlo en la sesión de trabajo de la siguiente manera:

1. En el espacio para la escritura de código o *script*, escribir el comando install.packages(*nombre del paquete*) y presionar Control + Enter o seleccionar la opción de "*run*". Por ejemplo, install.packages(ggplot2), para instalar el paquete ggplot2 que permite la construcción de gráficas estadísticas visualmente llamativas en R. Una vez se ejecuta esta instrucción, en la parte inferior izquierda (consola) aparece el nombre del paquete instalado y características generales (Figura 9).

Para empezar a utilizar las funciones del paquete, es necesario cargarlo en la sesión del trabajo con la función library, de la siguiente manera:



Nota importante: la instalación de un paquete debe hacerse solo una vez. Sin embargo, cada vez que se quiera usar en una nueva sesión de RStudio, se debe cargar con la función library().

Capítulo 3. Primeros Análisis: Caso Pokémon

3.1. Atrápalos con R: primeros análisis de datos

En esta sección se va a emplear una base de datos de la franquicia de videojuegos Pokémon, para realizar algunos análisis de datos básicos. Aunque no se corresponde con una base de datos de un contexto real, se considera que funciona bien como un ejercicio introductorio, previo a analizar bases de datos reales, las cuales usualmente son más complejas de manejar.

Pokémon es una franquicia de videojuegos nacida en 1996 y que consiste, básicamente, en capturar criaturas ficticias y avanzar en una historia. Los pokemones se utilizan para hacer combates con otros jugadores. Por lo tanto, estos pokemones tienen estadísticas asociadas a sus puntos de vida. Las estadísticas clave de un Pokémon incluyen puntos de salud, ataque, defensa, ataque especial, defensa especial y velocidad. Estas estadísticas, a menudo abreviadas como HP, ATTACK, DEFENSE, SP. ATK, SP. DEF y SPEED respectivamente, son variables numéricas que definen la capacidad de un Pokémon para infligir daño, resistir ataques y moverse rápidamente en combate. La importancia de estas estadísticas varía según el estilo de juego y las estrategias individuales de cada entrenador. Algunos pueden priorizar el aumento del ataque para derrotar rápidamente a los oponentes, mientras que otros pueden optar por aumentar la defensa para resistir

más golpes. La elección de Pokémon y la optimización de sus estadísticas pueden marcar la diferencia entre la victoria y la derrota en combate.

En esta actividad se empleará un archivo en formato Excel que contiene una tabla con las estadísticas numéricas de distintos Pokémon. Esta base de datos incluye variables como HP, Attack, Defense, Special Attack, Special Defense y Speed, todas ellas registradas como valores cuantitativos que permiten realizar análisis descriptivos y comparativos. Aunque se trata de un conjunto de datos ficticio proveniente del universo Pokémon resulta especialmente útil como primer acercamiento, pues su estructura es clara, homogénea y fácil de manipular en RStudio. Además, introduce al lector en procesos fundamentales como importar datos, revisar su estructura, seleccionar columnas y generar visualizaciones básicas, antes de avanzar hacia el trabajo con bases de datos reales más grandes y complejas.

3.2. Selección de la muestra

Para estudiar las estadísticas de los Pokemones, se ha diseñado un proceso de selección de muestra mediante muestreo aleatorio simple en Excel, asignando un número (pseudo)aleatorio a cada individuo y ordenando de mayor a menor para luego tomar los n primeros, siendo n el tamaño deseado para la muestra; este procedimiento garantiza la aleatoriedad del muestreo. El muestreo se realiza en Excel y no directamente en R, ya que se tiene como

propósito que el lector se familiarice primero con una herramienta de uso común y accesible, antes de introducir la sintaxis de R para generar números aleatorios. Excel permite visualizar de manera inmediata el procedimiento, lo que facilita la comprensión conceptual del muestreo; posteriormente, cuando se trabaja con R en capítulos más avanzados, se retoma este proceso empleando funciones propias del lenguaje para que el lector pueda comparar ambos enfoques. Para ello, se tomó una muestra aleatoria de 151 Pokémon de una población de 800 pokemones[2], utilizando como fuente de datos la información publicada en Pokémon Database.

Nota: En la página web de Pokemon Database no se encuentran las columnas de Generation y Legendary. Además, los nombres cambian entre la página y la base de datos, sin embargo, siguen directamente relacionados.

[2] La base de datos empleada se basa en información factual de Pokémon (*The Pokémon Company*). Se utiliza exclusivamente para fines educativos y no incluye material visual o textual protegido por derechos de autor.

Observación	Nombre	Total	HP	Attack	Defense	Sp. Atk	Sp. Def	Speed	Generation	Legendary
720	HoopaHoopa	600	80	110	60	150	130	70	6	False
713	Avalugg	514	95	117	184	44	46	28	6	False
707	Klefki	470	57	80	91	80	87	75	6	False
707	Klefki	470	57	80	91	80	87	75	6	False
705	Sliggoo	452	68	75	53	83	113	60	6	False
704	Goomy	300	45	50	35	55	75	40	6	False
681	AegislashBlac	520	60	150	50	150	50	60	6	False
681	AegislashBlac	520	60	150	50	150	50	60	6	False
674	Pancham	348	67	82	62	46	48	43	6	False
674	Pancham	348	67	82	62	46	48	43	6	False
668	Pyroar	507	86	68	72	109	66	106	6	False
657	Frogadier	405	54	63	52	83	56	97	6	False
656	Froakie	314	41	56	40	62	44	71	6	False
653	Fennekin	307	40	45	40	62	60	60	6	False
647	KeldeoOrdina	580	91	72	90	129	90	108	5	False
641	TornadusInca	580	79	115	70	125	80	111	5	False
629	Vullaby	370	70	55	75	45	65	60	5	False
626	Bouffalant	490	95	110	95	40	95	55	5	False
625	Bisharp	490	65	125	100	60	70	70	5	False
625	Bisharp	490	65	125	100	60	70	70	5	False
618	Stunfisk	471	109	66	84	81	99	32	5	False
612	Haxorus	540	76	147	90	60	70	97	5	False
611	Fraxure	410	66	117	70	40	50	67	5	False
610	Axew	320	46	87	60	30	40	57	5	False
603	Eelektrik	405	65	85	70	75	70	40	5	False
602	Tynamo	275	35	55	40	45	40	60	5	False

Figura 10. Base de datos Pokémon

En la Figura 10 se muestra un fragmento de la base de datos Pokémon organizada en forma de tabla. Cada fila corresponde a un Pokémon específico y cada columna representa una variable asociada a sus características. Las columnas (variables) visibles son:

- **Observación:** número identificador consecutivo de cada Pokémon en la base de datos.
- **Nombre:** el nombre del Pokémon (por ejemplo, *HoopaHoopa*, *AegislashBla*, *Pancham*).
- **Total:** suma de todas las estadísticas base del Pokémon.
- **HP (puntos de salud):** mide la resistencia del Pokémon en combate.
- **Attack (ataque físico):** poder de los movimientos físicos.
- **Defense (defensa física):** capacidad de resistir ataques físicos.

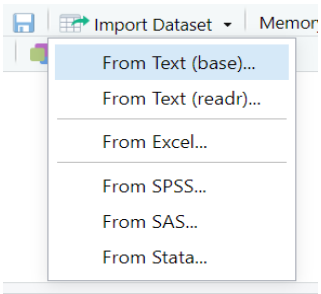
- **Sp. Atk (ataque especial):** poder de los movimientos especiales.
- **Sp. Def (defensa especial):** capacidad de resistir ataques especiales.
- **Speed (velocidad):** determina qué Pokémon ataca primero en combate.
- **Generation:** generación a la que pertenece el Pokémon (por ejemplo, 5 o 6).
- **Legendary (Legendario):** variable booleana (True/False) que indica si el Pokémon es legendario o no.

Nota: si no se conoce el procedimiento para realizar un muestreo aleatorio simple en Excel, nos podemos dirigir a una IA generativa (p. ej. Copilot, Chat GPT). En su defecto, el video Muestreo Aleatorio Simple en Excel explica el paso por paso de cómo realizar un muestreo aleatorio simple en Excel. O en el video (Muestreo Aleatorio Simple en R) se indica cómo podría realizarlo directamente desde RStudio. Los siguientes códigos QR conducen a los vídeos en cuestión.

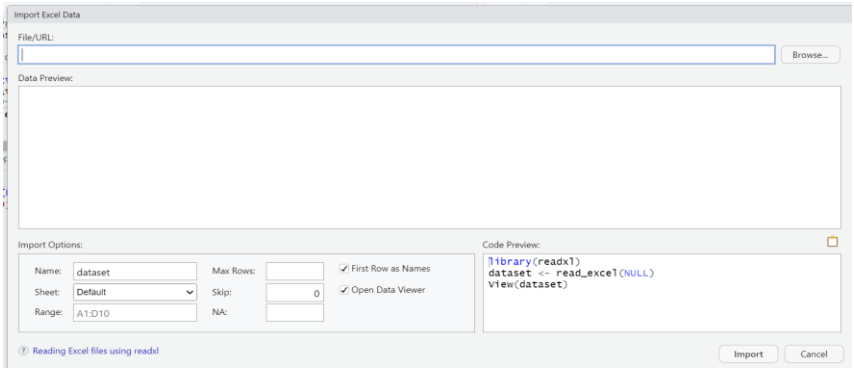
Muestreo aleatorio simple en Excel	*Muestreo aleatorio simple en R*

3.3. Importar a R Studio la muestra

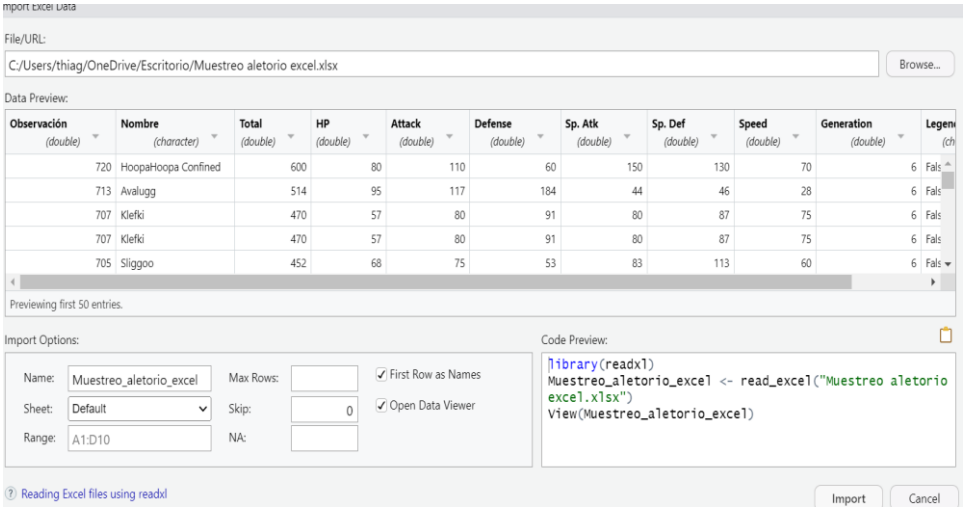
Una vez terminado el muestreo aleatorio simple podemos cargar el archivo, en este caso en la interfaz de RStudio seleccionamos "*import Dataset*" y nos aparecerán los siguientes indicadores:



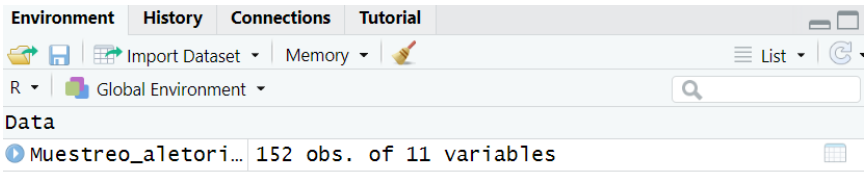
Seleccionamos la opción *From Excel* y aparecerá el siguiente recuadro:



Seleccionamos la opción de *Browse* y cargamos el archivo:



Al importarlo nos aparecerán los datos que seleccionamos y clicamos en *Import*. Automáticamente quedan cargados como se muestra a continuación:



Finalmente, se utiliza el siguiente código:

```
install.packages("psych")
```

El paquete psych incluye funciones tales como: cálculo de medias, medianas, varianzas, desviaciones estándar, correlaciones, etc.

Es un paquete especializado en análisis psicométrico, estadística descriptiva avanzada y métodos de análisis multivariado.

Organización de los datos

En este ejercicio se propone utilizar los datos de Pokémon[3] como un recurso didáctico para la familiarización con el lenguaje de programación R y con las primeras nociones de análisis de datos. Aunque los pokemones cuentan con múltiples estadísticas como defensa, velocidad o puntos de vida, en este caso el análisis se enfocará exclusivamente en la variable *Attack* (Ataque). Esta delimitación permite simplificar el proceso inicial, facilitar la interpretación de resultados y sentar las bases para familiarizarse mejor con RStudio.

Escribimos el siguiente código:

```
attach(Muestreo_aletorio_excel)

datos<-data.frame(HP,Attack,Defense,`Sp. Atk`,`Sp. Def`,Speed)
        View(Muestreo_aletorio_excel)
```

***Explicación*:**

attach() sirve para acceder directamente a las variables de una base de datos sin tener que escribir el nombre del objeto cada vez.

[3] La base de datos de Pokémon a trabajar se llama “Muestreo_aletorio_excel”.

- datos<-data.frame() crea un objeto llamado "datos" que será un data frame (tabla de datos en R). Dentro de este data.frame se incluyen las variables de (HP, Attack, Defense, `Sp. Atk`, `Sp. Def`, Speed), que se utilizan en nuestro caso.

Esto es necesario porque queremos organizar las estadísticas clave en una sola tabla sobre la cual podamos hacer análisis y comparaciones.

Tarea 1	Formula una hipótesis basada en las estadísticas disponibles. Por ejemplo: *Los Pokémon con ataques especiales presentan mayor velocidad que aquellos con ataques no especiales".* Procura establecer una justificación para tu hipótesis.

1. Resumen general de los datos

El primer paso para analizar las estadísticas de los Pokémon consiste en obtener un panorama general de la información. Con un solo comando en R podemos acceder a valores representativos como el mínimo, el máximo, la media, la mediana y los cuartiles de cada variable numérica. La función summary proporciona un resumen de algunas medidas numéricas de las variables que hacen parte de la data frame datos, como se muestra en la Figura 12.

summary(datos)

```
> summary(datos)
       HP              Attack          Defense          Sp..Atk
 Min.   :  1.00   Min.   : 20.00   Min.   : 20.00   Min.   : 20.00
 1st Qu.: 50.00   1st Qu.: 55.00   1st Qu.: 50.00   1st Qu.: 45.00
 Median : 60.50   Median : 72.00   Median : 60.00   Median : 60.00
 Mean   : 65.89   Mean   : 76.57   Mean   : 68.89   Mean   : 69.05
 3rd Qu.: 75.25   3rd Qu.: 94.25   3rd Qu.: 85.00   3rd Qu.: 90.00
 Max.   :170.00   Max.   :180.00   Max.   :200.00   Max.   :180.00
    Sp..Def          Speed
 Min.   : 20.00   Min.   : 15.00
 1st Qu.: 50.00   1st Qu.: 43.00
 Median : 65.00   Median : 60.00
 Mean   : 65.32   Mean   : 63.53
 3rd Qu.: 80.00   3rd Qu.: 80.00
 Max.   :130.00   Max.   :160.00
>
```

Figura 11. Summary datos Pokémon

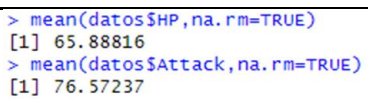
Al observar la información de la variable *Attack*, esta tiene una mediana de 72 y se encuentra más cercana al valor mínimo que el valor máximo, se puede deducir que la mayoría de pokemones tiende a tener el ataque entre los valores del primer y tercer cuartil. Los cuartiles ayudan a identificar cómo se agrupan los Pokémon en diferentes niveles de rendimiento. Sin embargo, hay que recordar que el primer y segundo cuartil corresponde al 50% de los datos.

3.3.1 Medidas de tendencia central

Las medidas de tendencia central permiten identificar un valor representativo de la distribución de los datos. En el caso de los pokemones, conocer la media y la mediana del ataque, defensa o

velocidad nos ayuda a comprender cuál es el rendimiento «típico» en combate y a analizar si la distribución de los valores es equilibrada o está sesgada. En el código utilizado, gracias a la función attach() es posible acceder directamente a las variables del conjunto de datos escribiendo únicamente el nombre de cada una, por ejemplo, simplemente «**HP**» en lugar de «**datos$HP**»[4], lo que facilita la escritura y lectura del código. Si no se tiene en cuenta no afectará para la ejecución del código.

Cuando se calculan estadísticas descriptivas en R, como la media o la mediana, es posible que algunas variables contengan valores faltantes, conocidos como **NA.** En estos casos, funciones como **mean()** o **median()**, pueden devolver un error o un resultado no deseado si no se les indica como manejar esos valores. Para resolverlo, estás funciones incluyen el argumento na.rm=TRUE, que instruye a R a eliminar los valores faltantes antes de hacer el cálculo.

mean(datos$Attack,na.rm=TRUE)
 *Figura 12. Mean (HP)*

[4] En R, el símbolo $ se utiliza para acceder a una columna específica dentro de un data frame. Funciona como un operador que permite llamar una variable por su nombre dentro de un conjunto de datos.

```
median(HP)
median(datos$Attack, na.rm = TRUE)
```

```
> median(datos$HP, na.rm = TRUE)
[1] 60.5
> median(datos$Attack, na.rm = TRUE)
[1] 72
```

Figura 13. Median

Se observa en la Figura 12 que la media de la variable *Attack* es 76.57237, esto indica que si, por ejemplo, todos los pokemones tuvieran el mismo poder físico para el ataque, sería precisamente alrededor de 76.6. Así mismo, considerando que la media va de 20 a 180 puntos (Figura 12), una media alrededor de 76 implica que el 'poder ofensivo típico' de un Pokémon no es ni particularmente débil ni extraordinariamente fuerte. Por su parte, la mediana es 72 (Figura 13), lo que indica que el ataque de la mitad de los pokemones están por debajo y el de la otra mitad por encima de este valor.

Tarea 2	1. Elige un Pokémon de tu base de datos (puede ser uno al azar o tu favorito). 2. Extrae tres estadísticas clave del Pokémon elegido: Ataque, Defensa y Velocidad. 3. Usando las funciones respectivas de R, calcula y compara el valor de cada variable para el Pokémon elegido, con la media y la mediana del conjunto de datos.

	4. Responde las siguientes preguntas: - ¿En cuántas unidades el ataque del Pokémon difiere de la media? ¿Es esta diferencia estadística o prácticamente relevante? - ¿Qué indica la comparación entre media y mediana sobre la distribución del ataque? - ¿La combinación de defensa y velocidad posiciona al Pokémon dentro del cuartil superior, medio o inferior de la población? Justifique.

3.3.2 Medidas de dispersión

Para los valores de las variables de los pokemones, la desviación estándar, la varianza, el rango y el rango intercuartil permiten evaluar si las estadísticas están concentradas alrededor de la media o si hay una gran variabilidad entre diferentes criaturas. A continuación, en la Figura 14, se muestran los cálculos de la desviación estándar, la varianza, el rango y el rango intercuartílico para el data frame *datos*:

```
sd(datos$Attack, na.rm = TRUE) # desviación estándar
var(datos$Attack, na.rm = TRUE)     # varianza
```

```
range(datos$Attack, na.rm = TRUE)  # mínimo y máximo
IQR(datos$Attack, na.rm = TRUE)    # rango intercuartil
```

```
> sd(datos$Attack, na.rm = TRUE) # desviación estándar
[1] 31.10665
> var(datos$Attack, na.rm = TRUE)     # varianza
[1] 967.6239
> range(datos$Attack, na.rm = TRUE)  # mínimo y máximo
[1]  20 180
> IQR(datos$Attack, na.rm = TRUE)     # rango intercuartil
[1] 39.25
>
```

Figura 14. Medidas de dispersión

Como se observa en la Figura 14, la desviación estándar es 31.11, lo que significa que, en promedio, los valores se alejan 31.11 unidades con respecto de la media. La varianza, cuyo valor es aproximadamente 967, representa el promedio de los cuadrados de esas desviaciones. Dado que la varianza no tiene una cota superior ni inferior, no es correcto calificarla directamente como “alta” o “baja”, ya que su magnitud depende de la escala de los datos. Sin embargo, puede llegar a ser comparada en el caso de que la variable tenga distintos momentos como temporales o procedentes de diferentes grupos o poblaciones.

Adicionalmente, conviene recordar que la varianza tiene el inconveniente de quedar expresada en las unidades de medida elevadas al cuadrado, lo que en la mayoría de las ocasiones dificulta su interpretabilidad.

Tarea 3	1. Calcula las medidas de dispersión para todas las variables base "(HP, At Attack, Defense, Sp. Atk, Sp. Def, Speed y Total)." 2. Interpreta los resultados obtenidos en el numeral anterior respondiendo: • ¿Cuál variable presenta menor dispersión en la población? • ¿Cuál tiene mayor desviación estándar? • ¿Qué característica muestra mayores diferencias entre los Pokémon?

Para interpretar adecuadamente la dispersión, es preferible comparar la desviación estándar con la media del conjunto de datos, con la dispersión de otras variables o con el rango posible de la variable analizada. Una alternativa útil es emplear el coeficiente de variación (CV), que se obtiene dividiendo la desviación estándar entre la media. Este indicador expresa la variabilidad de forma relativa y

permite determinar si la dispersión es grande o pequeña en relación con el valor promedio del ataque de los Pokémon.

3.3.3 Cuantiles

Los cuantiles dividen una distribución de datos en un número dado de partes iguales. Por ejemplo, los cuartiles y percentiles dividen la distribución en cuatro o cien partes iguales, respectivamente. En el contexto de los Pokémon, los cuartiles permiten identificar, por ejemplo, el valor de ataque que separa al 25% de los Pokémon más débiles del resto o al 75% de los más fuertes.

En el código de R, la función c() se utiliza para agrupar varios valores en un solo vector, como en c(25, 50, 75), lo que permite calcular simultáneamente diferentes percentiles o cuartiles dentro de funciones como quantile(), como se indica en la Figura 15.

```
#Cuartiles y percentiles
quantile(datos$Attack, probs = c(0.25, 0.5, 0.75),
na.rm = TRUE)
```

```
> quantile(datos$Attack, probs = c(0.25, 0.5, 0.75), na.rm = TRUE)
  25%   50%   75%
55.00 72.00 94.25
>
```

Figura 15. Cuartiles y Percentiles

El primer cuartil (Q1) de la variable *Attack* es 55, lo que significa que el 25% de los Pokémon tienen un valor de ataque igual o

inferior a este. El segundo cuartil (Q2) es 72, es igual a la mediana y divide a la mitad el conjunto de datos. El tercer cuartil (Q3) es igual a 94.25, indicando que el 75% de los Pokémon tienen un valor de ataque igual o inferior a este.

Tarea 4	1. En el ejemplo se puede ver que se usa la variable *Attack*, calcula en R los cuartiles para las demás variables. 2. Resuelve las siguientes preguntas, respectivamente: ● Según la categoría de tipo de Pokémon (por ejemplo: Fuego, Agua, Planta, Eléctrico, etc.), ¿qué tipo de Pokémon presenta el cuartil 1 (Q1) más alto en la estadística analizada (p. ej., Attack, Defense o Speed ● ¿Qué categoría de tipo de Pokémon presenta el valor de Q3 más alto en la estadística analizada (por ejemplo, Attack, Defense o Speed)? ¿Qué indica esto en comparación con los demás tipos de Pokémon?

3.4. Gráficos estadísticos básicos

La visualización de datos facilita la comprensión de patrones y distribuciones. Histogramas, diagramas de caja y gráficos de barras permiten observar la forma de la distribución de las estadísticas de los

Pokémon, la presencia de valores atípicos y la proporción de categorías. A continuación, se comentan algunos aspectos generales de tres distintos tipos de gráficos estadísticos convencionales. En las figuras 16 y 17, se muestra un ejemplo de cada una, así como el código para construirlas.

- **Histograma**

La forma de la distribución se representa mediante histogramas, especialmente útiles cuando se trabaja con variables continuas, ya que permiten observar cómo se distribuyen los datos a lo largo de un rango de valores. A través de estos gráficos se puede analizar la tendencia central, la dispersión, la simetría o sesgo (hacia la izquierda o hacia la derecha) y la curtosis.

EL histograma de la variable **Attack** muestra una distribución amplia y asimétrica hacia la derecha, donde la mayoría de los Pokémon se concentran en valores entre 40 y 120 puntos de ataque, reflejando un poder ofensivo moderado en la mayor parte de la muestra; sin embargo, se observan algunos casos extremos con ataques muy altos, superiores a 130 y hasta 180, que corresponden a especies destacadas como Dragonite, Tyranitar o Haxorus, mientras que en el extremo inferior aparecen Pokémon básicos con ataques con puntajes bajos (30–50), lo que evidencia una estructura en la que predominan los valores intermedios pero con una cola alargada hacia la derecha por la presencia de individuos excepcionalmente fuertes.

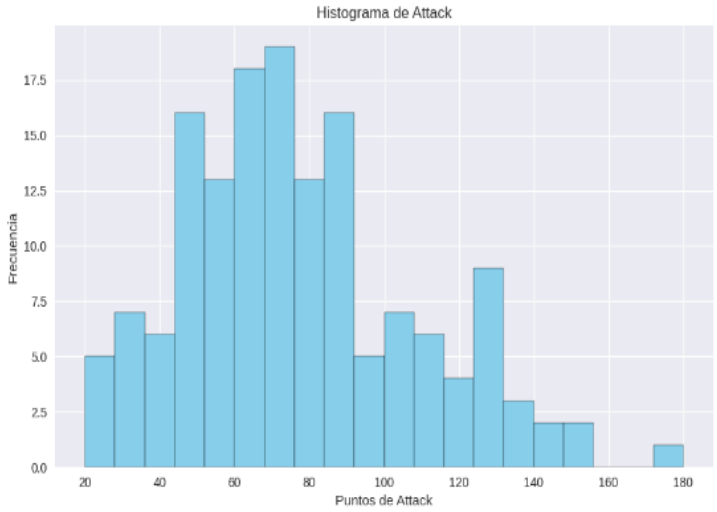


Figura 16. Histograma Attack

- **Diagrama de caja y bigotes (*boxplot*)**: un diagrama de caja y bigotes permite observar la ubicación de la mediana, los cuartiles, el rango y posibles valores atípicos del conjunto de datos.

- **Diagrama de barras:** se utiliza usualmente para variables categóricas, comparando frecuencias o proporciones de cada categoría de manera visual.

A continuación, se presenta los códigos para la representación de los datos:

```
hist(datos$Attack)                    # histograma
```

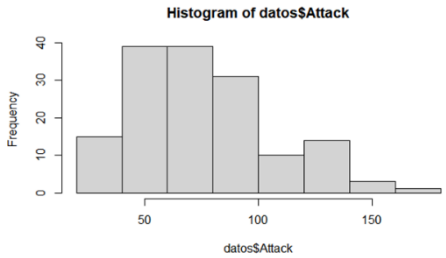


Figura 17. Histograma Pokémon

```
boxplot(datos$Attack)                 # diagrama de caja
```

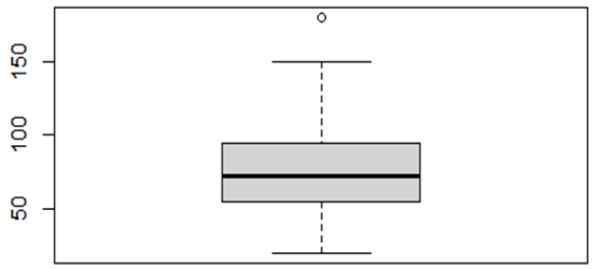
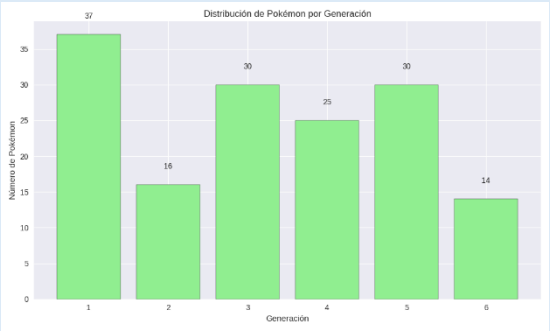


Figura 18. Diagrama de cajas Pokémon

3.5. Cierre del capítulo

Tarea 5	Realiza el diagrama de barras con respecto a la variable de *Generation*. Escribe al menos tres conclusiones de dicha variable a partir del diagrama graficado.  *Figura 19. Diagrama de barras Pokémon por Generación*

El análisis de las estadísticas de los Pokémon permitió introducir de manera práctica y didáctica el uso de RStudio, comprendiendo cómo importar bases de datos, organizarlas y aplicar criterios de selección para conformar muestras de estudio. Este ejercicio inicial no solo facilitó el acercamiento al entorno de programación, sino que también mostró la utilidad de la estadística descriptiva en contextos fáciles de manejar, sin las complejidades que comportan los datos de bases asociadas a realidades sociales, económicas, biológicas, entre otras. Así, damos la bienvenida al

siguiente proyecto: “Análisis de resultados ICFES 2020-2: Una investigación basada en el ciclo de datos”, en el cual sí se empleará una base de datos asociada a un contexto real y se hará una introducción a herramientas más propias de la ciencia de datos.

ANÁLISIS DE DATOS ICFES

INVESTIGACIÓN EDUCATIVA

Capítulo 4. Análisis ICFES 2020-2: una investigación basada en el ciclo de datos

A diferencia del proyecto inicial, en esta ocasión se trabajará con una base de datos de carácter más extenso y complejo: los resultados de las pruebas Saber 11 del año 2020, segundo periodo.

La elección de esta base de datos responde al interés particular que despierta en el ámbito educativo, dado que no solo contiene los puntajes obtenidos en cada una de las áreas evaluadas y el puntaje global, sino que además incluye información complementaria de gran relevancia: las respuestas a cuestionarios de contexto sobre condiciones socioeconómicas y tecnológicas en los hogares (acceso a internet, disponibilidad de computador, consola de videojuegos, entre otros), así como variables relacionadas con las instituciones educativas (ubicación rural o urbana, oficial o privada, carácter distrital, códigos y nombres de colegio, entre otros). Esta riqueza de variables convierte a la base en un insumo altamente pertinente para realizar un análisis integral.

El foco de interés se centra especialmente en los resultados de los puntajes de la prueba, por lo que las preguntas de indagación

estarán orientadas a comprender cómo se distribuyen dichos resultados en función de factores asociados tanto a los estudiantes como a las instituciones. En este sentido, resulta valioso explorar posibles relaciones entre las condiciones de estudio y los desempeños obtenidos, lo que abre la posibilidad de generar reflexiones pedagógicas y sociales de gran importancia.

El eje central de este proyecto radica en el análisis de bases de datos extensas y complejas, en las cuales se hace indispensable el uso de la programación para su manejo, limpieza y análisis. Este ejercicio no solo permite afianzar habilidades técnicas en R, sino también consolidar la capacidad de abordar de manera autónoma cualquier base de datos amplia, a partir de preguntas de interés personal, académico o social.

Asimismo, la elección de la base del ICFES 2020-2 no es arbitraria: corresponde al periodo inmediatamente posterior a la pandemia de COVID-19, un contexto que afectó profundamente los procesos de enseñanza y aprendizaje. Para los futuros profesores de matemáticas, resulta especialmente relevante comprender cómo se comportaron los resultados en este escenario, ya que permite reconocer el impacto que tuvo la educación remota, la brecha digital y las desigualdades socioeconómicas en el rendimiento académico de los estudiantes.

Antes de seleccionar la muestra, es importante contextualizar la base de datos utilizada. La base del ICFES 2020-2 corresponde a

los resultados oficiales de la prueba Saber 11 aplicados en el segundo semestre de 2020, disponibles públicamente en el portal de Datos Abiertos del ICFES[5]. Este archivo contiene información individual de cada estudiante aproximadamente 557.000 registros organizada en formato CSV (Comma-Separated Values, es decir, valores separados por comas), con variables de tipo demográfico, socioeconómico y académico (puntajes por área y puntaje global). Este formato de texto plano organiza la información en filas (registros) y columnas (variables), lo que facilita su importación y procesamiento en entornos como R y RStudio.

Para la propuesta de este proyecto se tomó una muestra aleatoria simple de 10 000 registros. La decisión de trabajar con una muestra obedece a la necesidad de asegurar un análisis eficiente y manejable en términos computacionales, especialmente al trabajar en R. Este tamaño de muestra permite capturar la diversidad de la población original sin comprometer la representatividad de los resultados y garantiza que los procedimientos estadísticos y de programación puedan ejecutarse de manera fluida en cualquier computador actual.

Luego de haber seleccionado la muestra y revisado las variables de la base de datos del ICFES 2020-2, que incluyen información de puntajes en las distintas áreas evaluadas, el puntaje

[5] https://www.icfes.gov.co/investigaciones/data-icfes/ ; sitio web de Datos abiertos del Icfes.

global, así como factores asociados al contexto socioeconómico, tecnológico e institucional, se opta por centrar las preguntas de indagación con base en los cuatro tipos clave de análisis de datos planteados por Davenport y Harris (2007):

- El análisis descriptivo pregunta: "¿Qué pasó?"
- El análisis predictivo pregunta: "¿Qué podría pasar en el futuro?"
- El análisis prescriptivo pregunta: "¿Qué se debe hacer a continuación?"
- El análisis diagnóstico pregunta: "¿Por qué sucedió esto?"

De los cuatro tipos de análisis mencionados se seleccionan dos enfoques específicos, de acuerdo con los métodos que se van a exponer en este capítulo. Estos métodos son la regresión logística y el análisis en componentes principales.

Para la regresión logística se adopta un análisis predictivo, ya que el modelo se centra en estimar la probabilidad de que un estudiante alcance un alto rendimiento académico a partir de variables contextuales como el acceso a internet, la tenencia de computador o la naturaleza del colegio. En particular, se plantea la pregunta de análisis: *¿Qué factores del contexto socioeconómico y escolar permiten predecir la probabilidad de obtener un alto rendimiento en las pruebas ICFES?*

En el caso del análisis de componentes principales (ACP), se realiza un análisis descriptivo, orientado a explorar las relaciones y patrones entre los puntajes de las distintas áreas evaluadas, representándolos en un mapa vectorial que permite observar cómo se agrupan y correlacionan las dimensiones del desempeño académico. Para ello, la pregunta que moviliza el estudio es: *¿Qué relaciones existen entre los puntajes por área y cómo se agrupan las competencias evaluadas en función de su aporte al rendimiento global?*

4.1. Selección de la muestra

Dado que el propósito de esta cartilla es introducir al lector en el análisis estadístico, se decidió no trabajar directamente con la base original sin depurar. En su lugar, se emplea una versión previamente organizada, en la cual se ajustaron inconsistencias y se estandarizaron algunas variables. Por ejemplo, las respuestas categóricas “Sí” y “No” fueron recodificadas como 1 y 0, con el fin de facilitar su uso en modelos estadísticos.

Una vez preparada esta base depurada, se seleccionó una muestra aleatoria simple de 10 000 registros, un tamaño adecuado para garantizar un análisis eficiente y manejable en R sin perder representatividad respecto de la población total. En las secciones siguientes se describen los pasos para descargar la base completa y el proceso general seguido para su preparación.

4.2. Pasos para descargar la muestra

A continuación, se presenta un procedimiento para localizar y descargar cualquier base de datos publicada en la página Data Icfes.

1) Entrar a la página oficial de DataIcfes

Ingrese a: https://www.icfes.gov.co/investigaciones/data-icfes/. En esa página encontrará el acceso al nuevo portal Data Icfes, el inventario de bases y el calendario de publicaciones como se muestra en las Figuras 20 y 21.

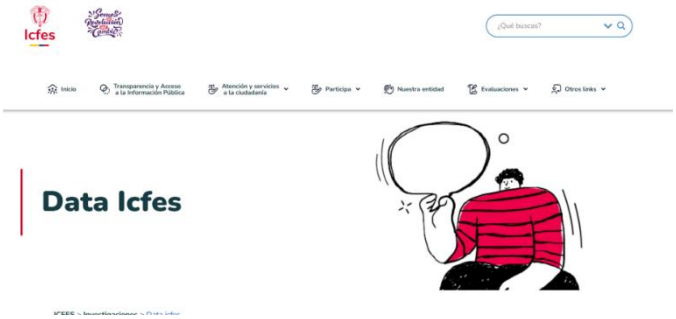


Figura 20. Página Icfes

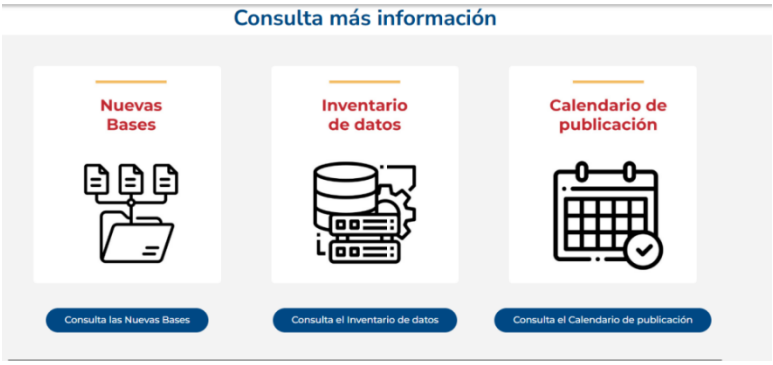


Figura 21. Pestañas de ingreso

2) Registrarse (si aplica)

En la parte superior aparece el enlace “Regístrate al Data Icfes”, como se observa en la Figura 22 (redirige a un formulario para crear credenciales). Si la base que desea requiere credenciales, complete el registro y espere las instrucciones/credenciales.

Figura 22. Registro Icfes

3) Localizar la base concreta (navegador)

Use los enlaces “Consulta las Nuevas Bases” o “Consulta el Inventario de datos” para ver qué conjuntos están disponibles (Saber 11, Saber Pro, cruces, años, niveles de observación, etc.). El inventario muestra qué años y qué nivel de observación incluye cada base. ICFES

Figura 23. Bases y Documentación

Busque la base Saber 11 2020-2 en el inventario o en la lista de nuevas bases.

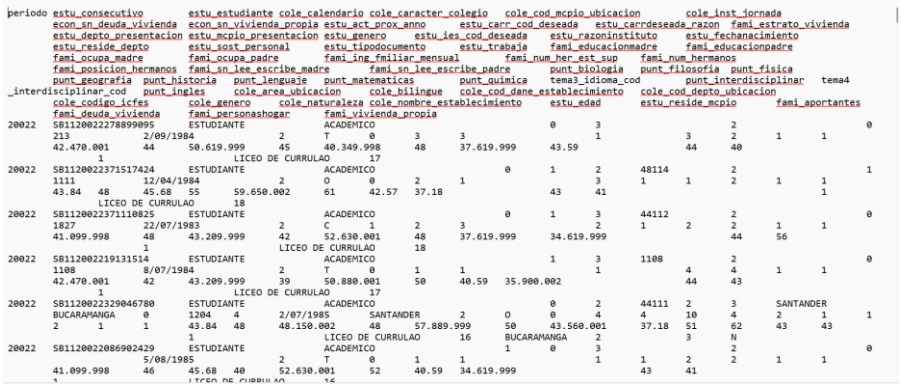
4) Descargar el archivo (navegador)

En el enlace de la base normalmente encontrará un archivo comprimido (.zip) o un archivo plano (.txt). Haga clic en el enlace de descarga y guarde el archivo localmente (Botón derecho → “Guardar enlace como...” si necesita obtener la URL directa)

Nota: Si la descarga está protegida por credenciales, use las credenciales proporcionadas luego del registro.

5) Descomprimir / comprobar formato

Descomprima el archivo.zip que fue descargado con el explorador de archivos. Data Icfes entrega muchos archivos en formato texto plano (.txt) con delimitadores (por ejemplo |, ¬, , o ;) y Codificación Unicode (UTF-8). Los datos se muestran de la siguiente manera:



periodo estu_consecutivo estu_estudiante cole_calendario cole_caracter_colegio cole_cod_mcpio_ubicacion cole_inst_jornada
econ_sn_deuda_vivienda econ_sn_vivienda_propia estu_act_prox_anno estu_carr_cod_deseada estu_carrdeseada_razon fami_estrato_vivienda
estu_depto_presentacion estu_mcpio_presentacion estu_genero estu_ies_cod_deseada estu_razoninstituto estu_fechanacimiento
estu_reside_depto estu_sost_personal estu_tipodocumento estu_trabaja fami_educacionmadre fami_educacionpadre
fami_ocupa_madre fami_ocupa_padre fami_ing_fmiliar_mensual fami_num_her_est_sup fami_num_hermanos
fami_posicion_hermanos fami_sn_lee_escribe_madre fami_sn_lee_escribe_padre punt_biologia punt_filosofia punt_fisica
punt_geografia punt_historia punt_lenguaje punt_matematicas punt_quimica tema3_idioma_cod punt_interdisciplinar tema4
_interdisciplinar_cod punt_ingles cole_area_ubicacion cole_bilingue cole_cod_dane_establecimiento cole_cod_depto_ubicacion
cole_codigo_icfes cole_genero cole_naturaleza cole_nombre_establecimiento estu_edad estu_reside_mcpio fami_aportantes
fami_deuda_vivienda fami_personashogar fami_vivienda_propia

20022 SB11200222788990 95 ESTUDIANTE ACADEMICO 0 3 2 0
213 2/09/1984 2 T 0 3 3 1 3 2 1 1
42.470.001 44 50.619.999 45 40.349.998 48 37.619.999 43.59 44 40
1 LICEO DE CURRULAO 17
20022 SB11200223715174 24 ESTUDIANTE ACADEMICO 0 1 2 48114 2 1
1111 12/04/1984 2 O 0 2 1 3 1 1 2 1 1
43.84 48 45.68 55 59.650.002 61 42.57 37.18 43 41 1
LICEO DE CURRULAO 18
20022 SB11200223711108 25 ESTUDIANTE ACADEMICO 0 1 3 44112 2 0
1827 22/07/1983 2 C 1 2 3 2 1 2 2 1 1
41.099.998 48 43.209.999 42 52.630.001 48 37.619.999 34.619.999 44 56
1 LICEO DE CURRULAO 18
20022 SB11200221913151 4 ESTUDIANTE ACADEMICO 1 3 1108 2 0
1108 8/07/1984 2 T 0 1 1 1 4 4 1 1
42.470.001 42 43.209.999 39 50.880.001 50 40.59 35.900.002 44 43
1 LICEO DE CURRULAO 17
20022 SB11200223290467 80 ESTUDIANTE ACADEMICO 0 2 44111 2 3 SANTANDER
BUCARAMANGA 0 1204 4 2/07/1985 SANTANDER 2 0 0 4 4 10 4 2 1 1
2 1 1 43.84 48 48.150.002 48 57.889.999 50 43.560.001 37.18 51 62 43 43
1 LICEO DE CURRULAO 16 BUCARAMANGA 2 3 N
20022 SB11200220869024 29 ESTUDIANTE ACADEMICO 1 0 3 2 0
5/08/1985 2 T 0 1 1 1 1 2 2 1 1
41.099.998 46 45.68 40 52.630.001 52 40.59 34.619.999 43 41
1 LICEO DE CURRULAO 16

Figura 24. Base de datos periodo 2020-2

Para que los datos sean más fáciles de tratar, se discriminan los datos de tal forma que se continúan en el paso 6.

6) Abrir con Excel

Si quiere abrir en Excel (pasos recomendados por el sitio web ICFES):

1. Abra un libro nuevo en Excel.
2. Datos → Obtener datos externos → Desde un archivo de texto.

3. Seleccione el .txt descomprimido.
4. En el asistente marque Delimitados y elija Unicode (UTF-8) como origen del archivo.
1. Siguiente, marque otro e ingrese el símbolo del delimitador.
2. Finalizar y Aceptar.

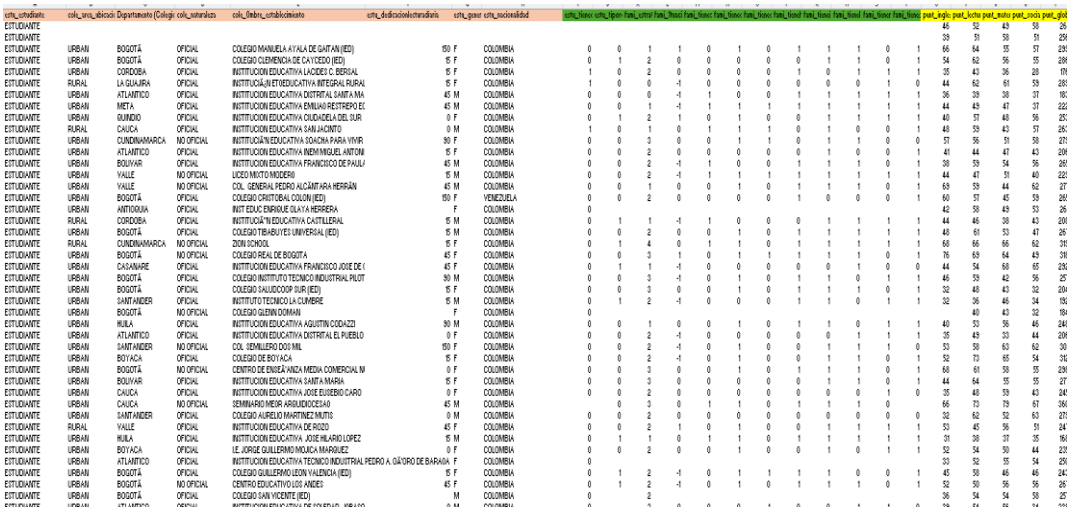


Figura 25. Base de Datos periodo 2020-2

Solo se trabajará con las siguientes variables (las demás se pueden eliminar para optimizar los tiempos de cálculo y procesamiento en el computador): Estu_estudiante .

1. cole_area_ubicacion
2. Departamento (Colegio)
3. cole_naturaleza
4. cole_0mbre_establecimiento
5. estu_nacionalidad

6. estu_tieneetnia
7. estu_tiporemuneracion
8. fami_estratovivienda
9. fami_1tuacioneco0mica
10. fami_tieneautomovil
11. fami_tienecomputador
12. fami_tieneconsolavideojuegos
13. fami_tienehor0microogas
14. fami_tieneinternet
15. fami_tienelavadora
16. fami_tienemotocicleta
17. fami_tieneserviciotv
18. punt_ingles
19. punt_lectura_critica
20. punt_matematicas
21. punt_sociales_ciudadanas
22. punt_global

7. Una vez concluido el proceso de selección de variables, se realiza la recodificación de aquellas que presentaban respuestas categóricas del tipo “Sí/No”. Para tal fin, se asigna el valor 1 a las respuestas afirmativas (“Sí”) y 0 a las respuestas negativas (“No”).

Esta transformación se efectuó con el propósito de facilitar el tratamiento estadístico de los datos, dado que, en R, numerosos métodos de análisis, como la regresión logística o los modelos lineales generalizados, requieren que las variables dicotómicas estén escritas

en términos de ceros y unos. Además, la codificación binaria permite interpretar de manera más clara los resultados, en tanto el valor 1 representa la presencia de una característica o condición, y el valor 0 su ausencia.

Por otra parte, para las variables numéricas, los valores faltantes o ausentes fueron reemplazados por la mediana, mientras que en las variables categóricas se utilizó la moda. Cabe señalar que estos valores imputados no corresponden a observaciones reales, sino que fueron calculados estadísticamente para mantener la coherencia de los datos. Este procedimiento se fundamenta en los lineamientos propuestos por Allison (2001) y Little y Rubin (2019), quienes destacan que la imputación simple es una alternativa válida para el tratamiento de datos faltantes cuando se busca preservar la estructura y distribución de la muestra sin alterar significativamente los resultados inferenciales.

4.3. Material Diccionario catálogos abiertos.

El presente material se apoya en el Diccionario del Examen Saber 11°, elaborado por el ICFES, el cual constituye una herramienta técnica fundamental para comprender la estructura y el significado de las bases de datos oficiales. Este documento detalla las variables académicas y contextuales recolectadas en las pruebas de Estado, describe sus categorías de respuesta y establece las codificaciones empleadas.

Contar con un diccionario de variables resulta esencial no solo para garantizar la correcta interpretación de la información, sino también para desarrollar procesos de análisis de datos transparentes, reproducibles y pedagógicamente pertinentes. En el marco de esta cartilla, el diccionario se convierte en un recurso clave que permite a los futuros profesores de matemáticas familiarizarse con los datos reales del sistema educativo colombiano, comprender su naturaleza cuantitativa y cualitativa, y utilizarlos como insumo en la formación estadística y crítica de sus estudiantes.

Nota: El Diccionario de Variables del Examen Saber 11° se encuentra disponible en el portal oficial de DataIcfes, dentro de la sección de documentación técnica que acompaña a las bases de datos abiertas del ICFES. Allí se puede acceder tanto al diccionario actualizado, que describe de manera detallada cada variable y sus categorías de respuesta, como al inventario histórico de las pruebas aplicadas. Este recurso es de libre consulta y puede descargarse desde la página web del ICFES en el apartado de Investigaciones → DataIcfes

Capítulo 5. Análisis estadístico

5.1. Regresión logística

La regresión logística es una técnica estadística utilizada especialmente cuando se desea predecir o explicar variables categóricas a partir de un conjunto de factores asociados. En el contexto de los datos del ICFES, esta técnica permite analizar, por ejemplo, la probabilidad de que un estudiante alcance un determinado nivel de desempeño o rendimiento académico (variable dependiente), considerando variables explicativas como el puntaje en áreas específicas, las condiciones socioeconómicas, el acceso a recursos educativos o las características del entorno escolar (variables independientes).

La regresión logística es adecuada cuando la variable dependiente es dicotómica (por ejemplo, “alto rendimiento” frente a “bajo rendimiento”). El modelo estima la probabilidad de ocurrencia de un evento (por ejemplo, obtener un puntaje global superior a cierto umbral), a partir de la combinación lineal de predictores mediante una función logística.

A continuación, se presenta el código de RStudio para el desarrollo de la regresión logística. **Nota:** No olvidar que es necesario importar o cargar la base de datos, como se ha explicado en la sección 3.1 de “Atrapalos con R”.

Paso 1: instalación y cargas de paquetes.

```
install.packages(c("readxl", "dplyr", "ggplot2", "car",
"MASS", "psych",
                   "ResourceSelection", "pROC", "caret",
"glmnet"))
```

1.1 Carga de librería.

```
library(readxl)
library(dplyr)
library(ggplot2)
library(car)
library(MASS)
library(psych)
library(ResourceSelection)
library(pROC)
library(caret)
library(glmnet)
library(FactoMineR)
library(factoextra)
```

Paso 2: transformación de las variables

Antes de realizar cualquier análisis estadístico, es necesario transformar algunas columnas de la base de datos que aparecen como texto (caracteres) en variables categóricas o factores.

Esto es importante porque la base del ICFES contiene múltiples variables que representan categorías, por ejemplo:

- género (Hombre / Mujer),
- tipo de colegio (Calendario A / Calendario B),
- zona (Urbano / Rural),
- respuestas como “Sí” / “No”,
- niveles socioeconómicos,

En el archivo original estas variables se presentan como cadenas de texto. Sin embargo, R no interpreta automáticamente que estos textos representan categorías, y para algunos análisis (tablas, gráficos, regresiones, modelos multivariados) es indispensable que estén declarados como factores.

```
datos<- data.frame(icfes_datos)

attach(datos)
```

Esta línea de código analiza la base de datos:

```
The following objects are masked from datos (pos = 3):

    ...1, cole_0mbre_establecimiento, cole_area_ubicacion,
    cole_naturaleza, Departamento..Colegio., estu_estudiante,
    estu_nacionalidad, estu_tieneetnia, estu_tiporemuneracion,
    fami_1tuacionecoOmica, fami_estratovivienda, fami_tieneautomovil,
    fami_tienecomputador, fami_tieneconsolavideojuegos,
    fami_tienehorOmicroogas, fami_tieneinternet, fami_tienelavadora,
    fami_tienemotocicleta, fami_tieneserviciotv, punt_global,
    punt_ingles, punt_lectura_critica, punt_matematicas,
    punt_sociales_ciudadanas
```

Figura 26. Analis de la base de datos

Antes de continuar con los análisis estadísticos, es fundamental revisar la estructura de los datos importados en R. La instrucción str(datos) permite visualizar el tipo de cada variable (numérica, carácter, lógica, etc.) y su contenido básico. Al aplicar este comando se observa que varias columnas aparecen como caracteres, aunque en realidad corresponden a categorías, como nacionalidad del estudiante, tipo de establecimiento, ubicación del colegio o posesión de bienes.

```
# estructura
str(datos)
datos$estu_nacionalidad<-as.factor(datos$estu_nacionalidad)
datos$estu_estudiante<-as.factor(datos$estu_estudiante)
datos$cole_area_ubicacion<-as.factor(datos$cole_area_ubicacion)
datos$Departamento..Colegio.<-as.factor(datos$Departamento..Colegio.)
datos$cole_naturaleza<-as.factor(datos$cole_naturaleza)
datos$cole_0mbre_establecimiento<-
as.factor(datos$cole_0mbre_establecimiento)
```

```
datos$estu_tiporemuneracion<-as.factor(datos$estu_tiporemuneracion)
datos$fami_tieneautomovil<-as.factor(datos$fami_tieneautomovil)
datos$fami_tienecomputador<-as.factor(datos$fami_tienecomputador)
str(datos)
```

```
> # estructura
> str(datos)
'data.frame':   10000 obs. of  24 variables:
 $ ...1                        : num  73859 246601 177653 157186 127765 ...
 $ estu_estudiante             : chr  "ESTUDIANTE" "ESTUDIANTE" "ESTUDIANTE" "ESTUDIANTE" "ESTUDIA
NTE" ...
 $ cole_area_ubicacion         : chr  "URBAN" "URBAN" "URBAN" "URBAN" ...
 $ Departamento..Colegio.      : chr  "CAUCA" "VALLE" "META" "LA GUAJIRA" ...
 $ cole_naturaleza             : chr  "OFICIAL" "OFICIAL" "OFICIAL" "OFICIAL" ...
 $ cole_0mbre_establecimiento  : chr  "I.E. NUCLEO TÃ‰CNICO AGROPECUARIO" "INSTITUCIO
N EDUCATIVA  JORGE ELIECER GAITAN" "INSTITUCION EDUCATIVA NUEVA ESPERANZA" "INSTITUCI
ON MANUEL ANTONIO DAVILA" ...
 $ estu_nacionalidad           : chr  "COLOMBIA" "COLOMBIA" "COLOMBIA" "COLOMBIA" ...
 $ estu_tieneetnia             : num  1 0 0 0 0 0 0 0 1 0 ...
 $ estu_tiporemuneracion       : num  1 0 0 1 0 0 0 0 0 0 ...
 $ fami_estratovivienda        : num  2 2 1 4 1 2 3 0 3 3 ...
 $ fami_1tuacioneco0mica       : num  1 0 0 1 0 1 1 0 0 0 ...
 $ fami_tieneautomovil         : num  0 0 0 0 0 1 1 0 0 1 ...
 $ fami_tienecomputador        : num  1 1 0 1 1 1 1 0 0 1 ...
 $ fami_tieneconsolavideojuegos: num  0 0 0 0 1 1 1 0 0 0 ...
 $ fami_tienehor0microogas     : num  1 1 0 0 0 0 1 0 0 0 ...
 $ fami_tieneinternet          : num  1 1 0 1 1 1 1 0 1 1 ...
 $ fami_tienelavadora          : num  1 0 0 1 1 1 1 0 0 1 ...
 $ fami_tienemotocicleta       : num  1 1 0 1 1 1 0 0 0 0 ...
 $ fami_tieneserviciotv        : num  1 1 1 0 0 1 1 0 0 1 ...
 $ punt_ingles                 : num  39 43 40 41 73 35 52 35 47 58 ...
 $ punt_lectura_critica        : num  48 48 49 53 48 62 58 40 55 60 ...
 $ punt_matematicas            : num  56 44 58 53 66 55 61 31 60 68 ...
 $ punt_sociales_ciudadanas    : num  52 26 39 38 47 51 63 52 54 59 ...
 $ punt_global                 : num  235 195 236 242 273 271 298 202 285 299 ...
>
```

Figura 27. Estructura de las variables

Nótese que, después de ejecutar las líneas de código anteriores, las variables que originalmente aparecían como texto ahora se encuentran correctamente convertidas en factores. Esto puede verificarse al volver a ejecutar str(datos), donde se observa que las columnas transformadas muestran el tipo “Factor” junto con el número de niveles que las componen. Esta conversión es fundamental para que R reconozca estas columnas como variables categóricas y las procese adecuadamente en los análisis posteriores.

```
'data.frame':   10000 obs. of  24 variables:
 $ ...1                          : num  73859 246601 177653 157186 127765 ...
 $ estu_estudiante               : Factor w/ 1 level "ESTUDIANTE": 1 1 1 1 1 1 1 1 1 1
...
 $ cole_area_ubicacion           : Factor w/ 2 levels "RURAL","URBAN": 2 2 2 2 2 2 2 2
2 2 ...
 $ Departamento..Colegio.        : Factor w/ 33 levels "ORTE SANTANDER",..: 12 31 22 20
1 7 8 16 31 6 ...
 $ cole_naturaleza               : Factor w/ 2 levels "NO OFICIAL","OFICIAL": 2 2 2 2 2
2 1 2 2 1 ...
 $ cole_Ombre_establecimiento    : Factor w/ 4740 levels "ORMAL SUPERIOR",..: 1709 2724
3889 4371 286 2633 1138 3651 2 129 ...
 $ estu_nacionalidad             : Factor w/ 6 levels "ARGENTINA","COLOMBIA",..: 2 2 2
2 2 2 2 2 2 2 ...
 $ estu_tieneetnia               : num  1 0 0 0 0 0 0 0 1 0 ...
 $ estu_tiporemuneracion         : Factor w/ 2 levels "0","1": 2 1 1 2 1 1 1 1 1 1 ...
 $ fami_estratovivienda          : num  2 2 1 4 1 2 3 0 3 3 ...
 $ fami_ltuacionecoOmica         : num  1 0 0 1 0 1 1 0 0 0 ...
 $ fami_tieneautomovil           : Factor w/ 2 levels "0","1": 1 1 1 1 1 2 2 1 1 2 ...
 $ fami_tienecomputador          : Factor w/ 2 levels "0","1": 2 2 1 2 2 2 2 1 1 2 ...
 $ fami_tieneconsolavideojuegos: num  0 0 0 0 1 1 1 0 0 0 ...
 $ fami_tienehorOmicroogas       : num  1 1 0 0 0 0 1 0 0 0 ...
 $ fami_tieneinternet            : num  1 1 0 1 1 1 1 0 1 1 ...
 $ fami_tienelavadora            : num  1 0 0 1 1 1 1 0 0 1 ...
 $ fami_tienemotocicleta         : num  1 1 0 1 1 1 0 0 0 0 ...
 $ fami_tieneserviciotv          : num  1 1 1 0 0 1 1 0 0 1 ...
 $ punt_ingles                   : num  39 43 40 41 73 35 52 35 47 58 ...
 $ punt_lectura_critica          : num  48 48 49 53 48 62 58 40 55 60 ...
 $ punt_matematicas              : num  56 44 58 53 66 55 61 31 60 68 ...
 $ punt_sociales_ciudadanas      : num  52 26 39 38 47 51 63 52 54 59 ...
 $ punt_global                   : num  235 195 236 242 273 271 298 202 285 299 ...
```

Figura 28. Cambio de factor de las variables

Paso 3: carga de datos y selección de muestra 70% y 30%

Para evaluar la capacidad de generalización del modelo, la muestra se divide en un conjunto de entrenamiento (70%), utilizado para ajustar y construir el modelo a partir de los datos disponibles, y un conjunto de prueba (30%), destinado a evaluar su desempeño sobre observaciones no utilizadas durante el proceso de entrenamiento. Esta separación evita evaluar el modelo con los mismos datos utilizados para ajustarlo, garantizando una valoración más objetiva de su desempeño. A continuación, el código para hacer la separación en los subconjuntos de datos.

```
# Número total de observaciones
n <- nrow(datos)
```

```
# Índices aleatorios para el 70%
set.seed(123)  # Para reproducibilidad
indices_train <- sample(seq_len(n), size = 0.7 * n)

# Crear subconjuntos
datosicfes<-datos[indices_train, ]
test <- datos[-indices_train, ] # 30%

# Comprobar tamaños
cat("Tamaño del conjunto de entrenamiento:", nrow(datosicfes), "\n")
cat("Tamaño del conjunto de prueba:", nrow(test),"\n")
```

```
> # Número total de observaciones
> n <- nrow(datos)
> # Índices aleatorios para el 70%
> set.seed(123)  # Para reproducibilidad
> indices_train <- sample(seq_len(n), size = 0.7 * n)
> # Crear subconjuntos
> datosicfes<-datos[indices_train, ]
> test <- datos[-indices_train, ] #Pendiente 30%
> # Comprobar tamaños
> cat("Tamaño del conjunto de entrenamiento:", nrow(datosicfes), "\n")
Tamaño del conjunto de entrenamiento: 7000
> cat("Tamaño del conjunto de prueba:", nrow(test),"\n")
Tamaño del conjunto de prueba: 3000
```

En el código anterior, el comando set.seed(123) permite que los números aleatorios generados por la función sample, sean siempre los mismos. Así, el número específico que va en la función set.seed no es relevante; lo importante es que se garantiza que el

procedimiento pueda replicarse con los mismos resultados siempre.

Tarea 6	Modifica la partición de 70% y 30% por una división 80% para el conjunto de entrenamiento y 20% para el de prueba. Ejecuta nuevamente el código y responde las siguientes preguntas: - ¿Cómo cambió el tamaño de cada conjunto? - ¿Qué ventajas y desventajas podría tener usar más datos para entrenar y menos para evaluar? - ¿Se mantiene la reproducibilidad cuando cambias el tamaño de la muestra? Explica por qué.

Paso 4: preparación de las variables

Antes de avanzar con los análisis, es necesario preparar algunas variables que serán utilizadas en los modelos estadísticos. En primer lugar, se seleccionan las columnas categóricas que requieren un tratamiento específico; para ello se crea un objeto que almacena únicamente las variables de este tipo. Además, se construye una nueva variable denominada alto_rendimiento, la cual clasifica a los estudiantes según su puntaje global en la prueba Saber 11. Esta variable es fundamental para los análisis de regresión posteriores, ya que permite distinguir entre estudiantes que alcanzaron un desempeño alto y quienes no.

vars_categoricas<-datosicfes[,c(2,3,5,6,7,9)]

colnames(datosicfes)

datosicfes$alto_rendimiento <- ifelse(datosicfes$punt_global >= 250, 1, 0)

```
> vars_categoricas<-datosicfes[,c(2,3,5,6,7,9)]
> #Las variables nombradas son estu_nacionalidad", "estu_estudiante", "cole_area_ubic
ion",
> #"cole_naturaleza", "cole_0mbre_establecimiento",
> #"estu_tiporemuneracion"
> colnames(datosicfes)
 [1] "...1"                          "estu_estudiante"
 [3] "cole_area_ubicacion"           "Departamento..Colegio."
 [5] "cole_naturaleza"               "cole_0mbre_establecimiento"
 [7] "estu_nacionalidad"             "estu_tieneetnia"
 [9] "estu_tiporemuneracion"         "fami_estratovivienda"
[11] "fami_1tuacioneco0mica"         "fami_tieneautomovil"
[13] "fami_tienecomputador"          "fami_tieneconsolavideojuegos"
[15] "fami_tienehor0microogas"       "fami_tieneinternet"
[17] "fami_tienelavadora"            "fami_tienemotocicleta"
[19] "fami_tieneserviciotv"          "punt_ingles"
[21] "punt_lectura_critica"          "punt_matematicas"
[23] "punt_sociales_ciudadanas"      "punt_global"
> # Crear variable binaria de rendimiento
> datosicfes$alto_rendimiento <- ifelse(datosicfes$punt_global >= 250, 1, 0)
```

Figura 29. Preparación de las variables

Nota: Las variables nombradas (2,3,5,6,7,9) son las variables “estu_nacionalidad”, “estu_estudiante”, “cole_area_ubicacion” ,“cole_naturaleza", "cole_0mbre_establecimiento", y “estu_tiporemuneracion"

Tarea 7	Modifica la construcción de la variable alto_rendimiento definiendo un nuevo umbral según tu propio criterio. Por ejemplo, podrías usar rangos como: 0 a 300 → bajo rendimiento 301 a 500 → alto rendimiento Luego responde: - ¿Cómo cambia la proporción de estudiantes clasificados como alto rendimiento con tu nuevo umbral? - ¿El nuevo criterio es más exigente o flexible que el anterior (250)? Explica. - ¿Cómo crees que afectará esta decisión al modelo de regresión logística más adelante?

La variable *alto_rendimiento* se construye a partir del puntaje global del ICFES usando un umbral de 250 puntos. Este valor se toma como referencia porque corresponde a la media teórica de la prueba, cuyo rango total va de 0 a 500 puntos. Así, 250 representa el punto medio de la escala y funciona como un criterio razonable para diferenciar desempeños: los estudiantes con un puntaje mayor a 250 se clasifican como 1 (alto rendimiento), mientras que aquellos con puntajes menores o iguales a 250 se clasifican como 0 (bajo

rendimiento). Esta distinción permite construir una variable binaria interpretativa y consistente con la estructura de la prueba.

Paso 5: preparación para método LASSO

Una vez organizada la base de datos, conviene seleccionar el subconjunto de variables independientes que harán parte del modelo de regresión logística. En estadística, no siempre se prefiere utilizar todas las variables disponibles para la construcción de un modelo, pues ello puede llevar a problemas de sobreajuste, dificultad en la interpretación contextual de los resultados o de demasiada carga computacional. Así, lo que se busca es un número óptimo de variables que permitan la configuración de un modelo suficientemente preciso sin sobrecargarlo. Uno de los métodos que existen para la selección de variables es el LASSO (*Least Absolute Shrinkage and Selection Operator*).

El método LASSO se utiliza porque permite obtener un modelo más simple, interpretativo y con mejor capacidad predictiva. Su principal ventaja es que realiza selección automática de variables: penaliza los coeficientes y reduce a cero aquellos predictores que no aportan información relevante. Esto evita incluir variables redundantes o poco útiles.

Además, LASSO es especialmente útil cuando existe un número elevado de variables o cuando algunas están correlacionadas, ya que previene el sobreajuste y mejora la generalización del modelo a nuevos datos. Finalmente, al eliminar predictores innecesarios,

produce un modelo más fácil de explicar y aplicar en contextos educativos. A continuación, el código en R para su implementación.

```
library(glmnet)

# Detectar columnas con un solo valor
colnames(datosicfes)[sapply(datosicfes, function(x) length(unique(x)) == 1)]
datosicfes_clean <- datosicfes[ , sapply(datosicfes, function(x)
length(unique(x)) > 1)]

# Preparar los datos (solo variables numéricas o dummy)
x <- model.matrix(alto_rendimiento ~ . - 1, data = datosicfes_clean)
#Cambio de naturaleza de la variable(para categorica)
y<-as.factor(datosicfes$alto_rendimiento)

# Ajuste del modelo LASSO
modelo_lasso <- cv.glmnet(x, y, family = "binomial", alpha = 1)
```

Nota: Este código demora en ejecutar debido a que analiza todos los modelos posibles.

```
# Coeficientes seleccionados
coef(modelo_lasso, s="lambda.min")
```

Tarea 8	1. Identifica cuáles variables categóricas fueron convertidas en *dummies*[6] dentro de la matriz, para ello ejecuta "colnames(x)" 2. Consulta qué representa exactamente esta variable *dummy* en el método LASSO

Paso 6: modelo logístico a partir de LASSO

Una vez preparadas las variables categóricas y numéricas, se procede a ajustar un modelo de regresión logística con el fin de identificar qué factores explican con mayor fuerza la probabilidad de que un estudiante alcance un alto rendimiento en el puntaje global del ICFES. La regresión logística es adecuada para este análisis porque la variable dependiente alto_rendimiento es binaria (1 = alto rendimiento, 0 = no alto rendimiento).

En este paso, primero se estima un modelo más amplio que incluye varias variables predictoras relacionadas con condiciones familiares y puntajes por áreas. Posteriormente, se ajusta un segundo modelo con los predictores que resultaron significativos, lo que permite obtener una versión más parsimoniosa y estable del modelo final. A continuación, se calculan las probabilidades predichas para

[6] Una variable *dummy* es aquella que solo toma los valores **0** y **1** para indicar la pertenencia o no de una observación a una categoría específica. Por ejemplo, si la variable "tipo de colegio" tiene las categorías *Oficial* y *Privado*, una variable dummy podría tomar el valor 1 si el colegio es privado y 0 si es oficial. Cuando existe más de una categoría, se generan varias variables dummy, una por cada categoría, tomando el valor 1 cuando la observación pertenece a dicha categoría y 0 en caso contrario.

cada estudiante y el valor del logit, que representa la razón logarítmica entre la probabilidad de éxito y la probabilidad de no éxito.

```
modelo_lasso <- glm (alto_rendimiento ~ fami_tieneautomovil + fami_tienehor0microogas + fami_tienecomputador + estu_tiporemuneracion +
                     fami_tieneserviciotv + punt_lectura_critica + punt_sociales_ciudadanas,
                    data = datosicfes, family = binomial)
summary(modelo_lasso)
modelo_resultante <- glm (alto_rendimiento ~ fami_tieneautomovil + fami_tienecomputador +
                          punt_lectura_critica + punt_sociales_ciudadanas,
                         data = datosicfes, family = binomial)
summary(modelo_resultante)
datosicfes$prob <- predict(modelo_resultante, type = "response")
# Calcular el logit (log (p / (1 - p)))
datosicfes$logit <- log (datosicfes$prob / (1 - datosicfes$prob))
```

Nota: El logit (log (p / (1 - p))), corresponde al logaritmo de la razón entre la probabilidad de éxito y la probabilidad de no éxito. Esta transformación es necesaria porque las probabilidades siempre toman valores entre 0 y 1, mientras que el logit puede asumir cualquier valor en la recta real, lo que facilita el ajuste del modelo y la interpretación

de los coeficientes. En este proyecto, el logit se calcula para cada estudiante a partir de la probabilidad estimada por el modelo, permitiendo analizar cómo cambian las posibilidades de alcanzar un alto rendimiento a medida que varían las variables explicativas.

Paso 7: rangos, p-valores e interpretación (argumentación)

Una vez que el método LASSO identifica las variables más relevantes, estas se emplean para ajustar un modelo de regresión logística tradicional. En esta etapa, el modelo se estima completamente y se obtienen los coeficientes asociados a cada predictor, junto con sus p-valores. Estos p-valores permiten determinar si el efecto de cada variable es estadísticamente significativo en la explicación de la probabilidad de alcanzar un alto rendimiento. A partir de los coeficientes también se calcula el *odds ratio*[7], una medida que indica cómo cambian las probabilidades relativas del evento cuando un predictor aumenta en una unidad: valores mayores que 1 sugieren un incremento en la probabilidad del alto rendimiento, mientras que valores menores que 1 indican una disminución.

En el análisis de regresión, cada coeficiente estimado se acompaña de un p-valor que permite evaluar su significancia estadística. Dicho p-valor se deriva de una prueba de hipótesis en la que la hipótesis nula establece que el coeficiente es igual a cero

[7] El *odds ratio* es una medida que se utiliza en la regresión logística para interpretar el efecto de un predictor

($H_0: \beta = 0$), es decir, que la variable no ejerce influencia sobre la probabilidad de alto rendimiento académico. La hipótesis alternativa ($H_1: \beta \neq 0$) plantea que el coeficiente difiere de cero y, por tanto, la variable sí contribuye al modelo. De esta manera, los p-valores asociados a cada predictor indican si el efecto observado es estadísticamente significativo: valores inferiores a 0.05 permiten rechazar la hipótesis nula y confirmar la relevancia de la variable, mientras que valores superiores sugieren que su aporte no es significativo y puede considerarse prescindible en el modelo final.

Elemento	Interpretación
Coeficiente > 0	Aumenta la probabilidad de alto rendimiento.
Coeficiente < 0	Disminuye la probabilidad de alto rendimiento.
p-valor < 0.05	Efecto estadísticamente significativo.
p-valor ≥ 0.05	No significativo (puede eliminarse del modelo final).

Tabla 1. p-valores

El uso de estos criterios se justifica porque, después de la etapa de penalización y selección del LASSO, el modelo reducido debe verificarse estadísticamente para asegurar que las variables retenidas no solo fueron seleccionadas por el algoritmo, sino que también

presentan efectos significativos en una regresión logística estándar. Esto fortalece la validez del modelo final.

En el ejemplo analizado, al utilizar un nivel de significancia de $\alpha = 0.05$, se observa que las variables: fami_tieneautomovil, fami_tienecomputador, lectura_crítica, y sociales_y_ciudadanas presentan p-valores menores a 0.05, lo cual indica que su efecto sobre la probabilidad de obtener alto rendimiento es estadísticamente significativo. En consecuencia, estas variables se mantienen en el modelo final como predictores importantes.

Tarea 9	1. Modifica el modelo agregando una variable que NO quedó seleccionada por LASSO 2. Compara los p-valores y coeficientes de esta nueva variable con los del modelo final. - ¿La variable añadida es significativa? - ¿Su coeficiente cambia notablemente la interpretación del modelo?

Paso 8: representaciones gráficas y coeficientes de correlación

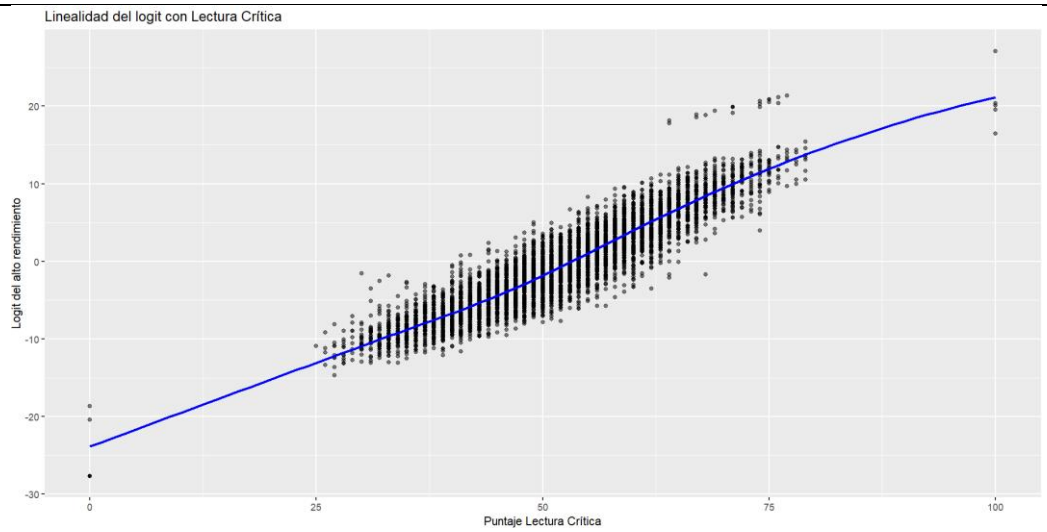
Aunque la regresión logística no exige linealidad entre las variables independientes X y la variable dependiente Y, sí requiere una relación aproximadamente lineal entre las variables continuas y el

logit de la probabilidad (Hoster, D et (2013). El uso de coeficientes de correlación ayuda a detectar si existe una asociación razonablemente estable entre las variables. A continuación, el código para obtener los coeficientes de correlación lineal entre cada variable independiente y la dependiente.

Nota: Aunque la variable dependiente original del modelo es *alto_rendimiento*, en los análisis de correlación que siguen se utiliza el logit como variable de referencia. Esto se debe a que el logit representa la transformación continua de la probabilidad estimada por el modelo logístico, lo cual permite evaluar su asociación con otras variables numéricas mediante métodos de correlación.

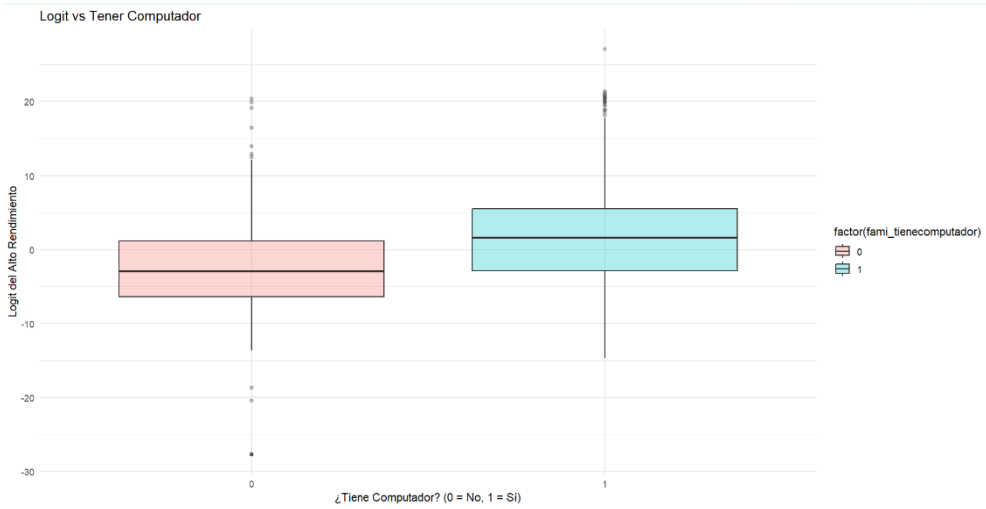
```
library(ggplot2)

# Lista de variables continuas

colnames(datosicfes)

vars_continuas<-datosicfes[,c(21,23,24)]

#Puntaje Sociales y ciudadanas

cor.test(datosicfes$logit,vars_continuas$punt_sociales_ciudadanas,method="spearman")

cor.test(datosicfes$logit,vars_continuas$punt_sociales_ciudadanas,method="pearson")

plot(datosicfes$logit,vars_continuas$punt_sociales_ciudadanas)
```

```
##abline(lm())-- Recta de regresión

#Puntaje Lectura critica

cor.test(datosicfes$logit,vars_continuas$punt_lectura_critica,
method="spearman")

cor.test(datosicfes$logit,vars_continuas$punt_lectura_critica, method="pearson")

plot(datosicfes$logit,vars_continuas$punt_lectura_critica)

ggplot(datosicfes, aes(x = punt_lectura_critica, y = logit)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "loess", se = FALSE, color = "blue") +
  labs(title = "Linealidad del logit con Lectura Crítica",
       x = "Puntaje Lectura Crítica",
       y = "Logit del alto rendimiento")
```




```
# Correlaciones con variables discretas

# Automóvil

cor.test(datosicfes$logit,datosicfes$fami_tieneautomovil,method="spearman")

# Computador

cor.test(datosicfes$logit,datosicfes$fami_tienecomputador,method="spearman")

# Boxplot: Relación entre tener computador y logit

ggplot(datosicfes, aes(x = factor(fami_tienecomputador), y = logit, fill = factor(fami_tienecomputador))) +

  geom_boxplot(alpha = 0.3) +

  labs(title = "Logit vs Tener Computador",

       x = "¿Tiene Computador? (0 = No, 1 = Sí)",
```

```
    y = "Logit del Alto Rendimiento") +
  theme_minimal()
```

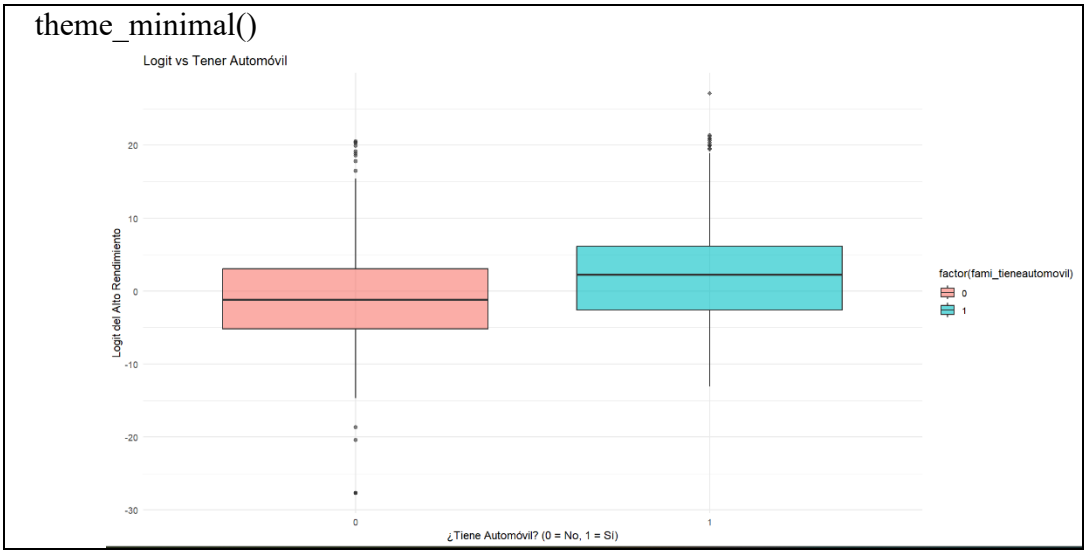



```
#
```

Boxplot: Relación entre tener automóvil y logit

```
ggplot(datosicfes, aes(x = factor(fami_tieneautomovil), y = logit, fill = factor(fami_tieneautomovil))) +
  geom_boxplot(alpha = 0.6) +
  labs(title = "Logit vs Tener Automóvil",
       x = "¿Tiene Automóvil? (0 = No, 1 = Sí)",
       y = "Logit del Alto Rendimiento") +
```

theme_minimal()



El coeficiente de correlación lineal de Pearson (representado como ρ o r para el caso poblacional o el muestral, respectivamente) permite cuantificar la fuerza y dirección de la relación lineal entre dos variables. Su valor oscila entre -1 y 1; los extremos indican relaciones perfectas y un valor cercano o igual a 0 indica ausencia de relación lineal. Para interpretar su magnitud se utiliza la siguiente clasificación:

- $0.00 - 0.19$**: Muy débil**
  La relación entre las variables es casi inexistente. Los cambios en una variable apenas se asocian con cambios en la otra.
- $0.20 - 0.39$**: Débil**
  Existe una relación, pero es limitada. Las variables se mueven juntas de manera poco consistente.

- $0.40 - 0.59$**: Moderado**
  La relación es clara y visible. Las variables tienen una asociación apreciable, aunque no determinante.

- $0.60 - 0.79$**: Fuerte**
  Hay una relación sólida. Las variaciones en una variable están fuertemente asociadas con cambios en la otra.

- $0.80 - 1.00$**: Muy fuerte**
  La relación es extremadamente alta. Las variables prácticamente se comportan de manera coordinada.

Nota. Esta escala se aplica al valor absoluto del coeficiente. El signo del coeficiente indica la dirección de la relación: positiva (cuando ambas variables aumentan juntas) o negativa (cuando una aumenta mientras la otra disminuye).

Esta clasificación permite evaluar si cada predictor tiene el potencial de aportar información relevante al modelo logístico o al proceso de selección de variables mediante LASSO. Es decir, se buscan las correlaciones más altas o significativas para el modelo.

Por otra parte, el p-valor indica si la correlación observada es estadísticamente significativa, es decir, si es muy poco probable que esa relación se deba únicamente al azar. Su interpretación es la siguiente:

- $p < \alpha$:
  La correlación es significativa. Existe evidencia estadística

suficiente para afirmar que las dos variables están relacionadas.

- $p > \alpha$:
  No hay evidencia suficiente para afirmar que exista una relación real entre las variables.

En este caso particular, dado que $\alpha = 0.05$:

Tarea 10	1. Selecciona una variable continua diferente a las evaluadas en el ejemplo. 2. Crea un gráfico alternativo. Puedes elegir uno de los siguientes: - *Código del gráfico de violín* ggplot(datosicfes, aes(x = 1, y = punt_matematicas)) + geom_violin(fill = "lightblue", alpha = 0.4) + labs(title = "Distribución de Puntaje de Matemáticas", x = "", y = "Puntaje Matemáticas") - *Código del gráfico de densidad*

	ggplot(datosicfes, aes(x = punt_matematicas)) + geom_density(alpha = 0.6, fill = "lightblue") + labs(title = "Gráfico de Densidad del Puntaje de Matemáticas", x = "Puntaje Matemáticas", y = "Densidad") + theme_minimal() Responde las siguientes preguntas: - ¿El gráfico que elegiste muestra patrones interesantes (asimetrías, grupos, valores atípicos)? - ¿Crees que esta variable podría ayudar al modelo logístico o no aporta mucho?

- Esto indica que todas las correlaciones evaluadas son altamente significativas desde el punto de vista estadístico.
- Como consecuencia, se puede concluir que todas las variables incluidas presentan algún grado de asociación con la variable dependiente, lo cual justifica su posterior evaluación dentro del modelo.

Paso 9: ausencia de multicolinealidad

En regresión logística es importante que las variables explicativas no estén demasiado correlacionadas entre sí, ya que esto:

- Distorsiona los coeficientes.
- Aumenta la varianza.
- Hace inestables las estimaciones.
- Puede inflar los errores estándar.
- Puede producir signos o magnitudes incoherentes.

Este problema se conoce como multicolinealidad.

Para detectarlo, una herramienta posible es el *Variance Inflation Factor* (VIF), que mide cuánto aumentan las varianzas de los coeficientes debido a correlación entre predictores. A continuación, el código para su implementación:

```
library(car)

vif(modelo_resultante)
```

```
> vif(modelo_resultante)
     fami_tieneautomovil     fami_tienecomputador     punt_lectura_critica
                1.065283                 1.065546                 1.009868
punt_sociales_ciudadanas
                1.009597
```

El VIF es un indicador que permite evaluar el nivel de multicolinealidad entre los predictores de un modelo. Valores altos de VIF indican que una variable está altamente correlacionada con otras, lo cual puede distorsionar las estimaciones del modelo y afectar su interpretación. Los rangos comúnmente aceptados son:

- **VIF = 1 → Sin correlación**
  La variable no presenta relación lineal con las demás. Es la condición ideal.
- **VIF entre 1 y 2 → Correlación muy baja**
  Existe una relación mínima, pero no afecta el modelo.
- **VIF entre 2 y 5 → Correlación moderada**
  La variable empieza a compartir información con otras; conviene revisarla.
- **VIF entre 5 y 10 → Correlación alta**
  La multicolinealidad podría alterar los coeficientes y su estabilidad.
- **VIF > 10 → Multicolinealidad severa** En este caso, la variable debe eliminarse o transformarse, pues compromete la validez del análisis.

En el análisis realizado, los valores de VIF fueron los siguientes:

Variable	VIF
fami_tieneautomovil	1.065
fami_tienecomputador	1.065
punt_lectura_critica	1.009
punt_sociales_ciudadanas	1.009

Estos resultados permiten concluir que:

- Todos los VIF están muy cerca de 1, lo cual representa el escenario óptimo.
- Las variables presentan una correlación extremadamente baja entre sí.
- Se confirma que los predictores evaluados son prácticamente independientes.
- No existe ningún indicio de multicolinealidad que pueda afectar al modelo.

Tarea 11	1. Genere un gráfico de correlaciones tipo *heatmap*, el código es el siguiente: library(reshape2) cor_melt <- melt(cor_matrix) ggplot(cor_melt, aes(Var1, Var2, fill = value)) + geom_tile() + geom_text(aes(label = round(value, 2))) + scale_fill_gradient2(limits=c(-1,1)) + labs(title = "Mapa de calor de correlaciones entre predictores") + theme_minimal() 2. Responde las siguientes preguntas: - ¿Qué par de variables muestra mayor asociación? - ¿Qué implicaciones tendría una correlación alta en la regresión logística? - ¿Coinciden los resultados con los valores del VIF? ¿Por qué?

Paso 10: prueba chi-cuadrado de independencia para variables categóricas

Este bloque se emplea para cumplir tres propósitos fundamentales en el análisis:

- Confirmar la existencia de relación entre las variables categóricas y la variable objetivo *alto_rendimiento*. Esto permite identificar si las características familiares evaluadas muestran asociación estadística con el desempeño académico.
- Determinar si las variables deben mantenerse en el modelo logístico. Solo es pertinente incluir en la regresión aquellas variables que presentan alguna relación con la respuesta; de lo contrario, no aportarían información útil al modelo.
- Justificar estadísticamente la relevancia de estas variables en el contexto del ICFES.
  El análisis proporciona evidencia de que ciertos factores familiares están efectivamente vinculados con el nivel de rendimiento en la prueba, lo cual respalda su inclusión en el estudio.

```
variables_prueba <- c( "fami_tieneautomovil", "fami_tienecomputador")

for (var in variables_prueba)

{print(chisq.test(table(datosicfes$alto_rendimiento, datosicfes[[var]])))}

#Como todos los valores p son menores a 0.05, rechazamos la hipótesis nula de independencia.

#Por tanto, sí existe una asociación estadísticamente significativa entre alto_rendimiento y cada una de las variables categóricas analizadas.
```

```
        Pearson's Chi-squared test with Yates' continuity correction

data:  table(datosicfes$alto_rendimiento, datosicfes[[var]])
X-squared = 200.7, df = 1, p-value < 2.2e-16


        Pearson's Chi-squared test with Yates' continuity correction

data:  table(datosicfes$alto_rendimiento, datosicfes[[var]])
X-squared = 490.75, df = 1, p-value < 2.2e-16
```

La regla usual para la interpretación de los resultados de la prueba chi cuadrado es mediante el p-valor obtenido:

- Si p-valor < α, se rechaza H_0 → existe evidencia de asociación entre las variables.
- Si p-valor ≥ α , no se rechaza H_0 → no hay evidencia suficiente de asociación.

Tarea 12	Elabore una breve reflexión escrita en la que explique: • ¿Qué tan coherentes son los resultados de la correlación con tus expectativas iniciales sobre la relación entre el logit y la variable evaluada? • ¿La asociación encontrada (el coeficiente de correlación obtenido) es fuerte, moderada o débil en términos prácticos? Explica. • ¿Qué aportan al análisis del rendimiento académico las variables 'fami_tieneautomovil' y 'fami_tienecomputador'?"

Paso 11: promedios y desviaciones estándar por grupo

Este procedimiento permite verificar si los estudiantes con alto y bajo rendimiento presentan diferencias claras en los puntajes de Lectura Crítica y Sociales y Ciudadanas antes de interpretar por completo la regresión logística. Al comparar los promedios y la dispersión de cada grupo, se evalúa si las variables continuas realmente distinguen entre niveles de desempeño.

Además, este análisis sirve para:

- Identificar diferencias entre grupos antes del modelado.
- Validar que estas variables tienen sentido como predictores.
- Explorar la consistencia y variabilidad de los puntajes dentro de cada grupo.
- Comprobar si la diferencia entre medias es lo suficientemente grande como para ser estadísticamente relevante.

A continuación, el código para el cálculo de los promedios y las desviaciones por grupos.

```
datosicfes %>%
  group_by(alto_rendimiento) %>%
  summarise(across(c(punt_lectura_critica),
                   list(media = mean, sd = sd),
                   na.rm = TRUE)) %>%
  print()
```

```
datosicfes %>%
  group_by(alto_rendimiento) %>%
  summarise(across(punt_sociales_ciudadanas, list(media = mean, sd = sd), na.rm = TRUE))
```

Interpretación variable "punt_lectura_critica "

```
# A tibble: 2 × 3
  alto_rendimiento punt_lectura_critica_media punt_lectura_critica_sd
             <dbl>                      <dbl>                    <dbl>
1                0                       45.3                     7.21
2                1                       59.9                     6.67
```

Figura 30. Promedio y desviación estándar Puntaje Lectura Critica

Como se observa en la Figura 30, los estudiantes de alto rendimiento (codificados con 1) presentan un promedio de 59.9, mientras que el grupo de bajo rendimiento (codificados con 0) alcanza 45.3, mostrando una diferencia cercana a 15 puntos. Esta brecha amplia indica una relación fuerte entre el desempeño en Lectura Crítica y la probabilidad de obtener un rendimiento global alto. La dispersión es similar entre grupos ($SD \approx 6.7 - 7.2$), lo que refleja patrones consistentes y estables en esta competencia.

Interpretación variable "punt_sociales_ciudadanas "

```
# A tibble: 2 × 3
  alto_rendimiento punt_sociales_ciudadanas_media punt_sociales_ciudadanas_sd
             <dbl>                          <dbl>                        <dbl>
1                0                           39.7                         7.11
2                1                           58.0                         8.35
```

Figura 31. Promedio y desviación estándar Puntaje sociales

En esta área también se observa una diferencia marcada entre los grupos: como se ilustra con la Figura 31 el puntaje promedio del grupo de alto rendimiento supera al del grupo de bajo rendimiento en aproximadamente 20 puntos. Aunque la desviación estándar es ligeramente mayor en el grupo de mejor desempeño, la diferencia sustancial entre los promedios evidencia una tendencia clara: puntajes más altos en Sociales y Ciudadanas se asocian fuertemente con un mayor rendimiento global.

Tarea 13	Consulta y escribe un resumen acerca de: * ¿Qué estudios o experiencias en otros países respaldan la relación entre tener recursos como computador o auto y el rendimiento académico? ¿Qué implicaciones educativas y de política pública se pueden derivar de las diferencias en rendimiento asociadas a la disponibilidad de recursos familiares?

Paso 12: ajuste de modelo logístico

Después de ajustar el modelo logístico final "modelo_resultante", es necesario evaluar qué tan bien explica la variabilidad en la probabilidad de alto rendimiento. Para ello, se

utilizan medidas conocidas como pseudo-R^2, que permiten evaluar el desempeño del modelo. En R se emplea la librería pscl y la función pR2 sobre el modelo ajustado, como se observa enseguida:

```
library(pscl)

pR2(modelo_resultante)
```

Rangos, interpretación de valores y criterios:

- **Mcfadden**

Valor de McFadden (x)	Interpretación
$\mathbf{0.02 - 0.09}$	Malo
$\mathbf{0.10 - 0.30}$	Aceptable
$\mathbf{0.30 - 0.40}$	Bueno
$\boldsymbol{x > 0.40}$	Excelente

Tabla 1. Mcfadden

- **r2ML**

Valores por encima de 0.50 ya apuntan a modelos con gran poder explicativo.

- **r2CU (Nagelkerke)**

Valor (x)	Calidad del modelo

$\mathbf{0.20 - 0.40}$	Moderado
$\mathbf{0.40 - 0.60}$	Fuerte
$\mathbf{0.60 - 0.80}$	Muy fuerte
$\boldsymbol{x > 0.80}$	Excelente

Tabla 2. R2CU (Nahelkerke)

Valores resultantes de la investigación:

Indicador	Valor	Interpretación
McFadden	0.7255	Indica que el modelo explica cerca del 72.6% de la variabilidad en la log-verosimilitud, lo cual es un rendimiento excepcional.
r2ML	0.6338	Indica excelente capacidad explicativa del modelo.
r2CU (Nagelkerke)	0.8455	Indica que la combinación de estas variables, junto con computador y automóvil, explica casi el 85% de la variación en la probabilidad de ser alto rendimiento.

Tabla 3. Resultados de la investigación

Paso 13: evaluación del modelo con el 30% (test)

Hasta este punto ya:

- Se entrenó un modelo de regresión logística con el 70% de los datos (training).
- Se seleccionó un modelo resultante depurado, sin multicolinealidad y con variables significativas.
- El objetivo ahora es evaluar si el modelo generaliza bien a datos NO vistos usando el 30% restante (test), que fue creado en el Paso 3 del procedimiento.

```
# Crear variable binaria en el conjunto de prueba

test$alto_rendimiento <- ifelse(test$punt_global >= 250, 1, 0)

# Probabilidades predichas con el modelo ya entrenado

prob_test <- predict(modelo_resultante, newdata = test, type = "response")

#  Clasificación según punto de corte 0.5

pred_test <- ifelse(prob_test > 0.5, 1, 0)

#  Matriz de confusión

tabla_test <- table(Predicho = pred_test, Real = test$alto_rendimiento)

print(tabla_test)
```

```
> #  Matriz de confusión
> tabla_test <- table(Predicho = pred_test, Real = test$alto_rendimiento)
> print(tabla_test)
        Real
Predicho    0    1
       0 1475  126
       1  147 1252
```

```
# Exactitud del modelo en el conjunto de prueba (30%)
```

```
exactitud_test <- mean(pred_test == test$alto_rendimiento)

cat("Exactitud (30%):", round(exactitud_test * 100, 2), "%\n")

# Curva ROC con el 30%

library(pROC)

roc_test <- roc(test$alto_rendimiento, prob_test)

plot(roc_test, col = "darkgreen", main = "Curva ROC (30% de datos)")

auc_valor <- auc(roc_test)

cat("AUC (30%):", round(auc_valor, 3), "\n")
```

Para evaluar el desempeño del modelo de predicción de alto rendimiento académico, se utilizó un conjunto de prueba correspondiente al 30% de los datos. Se generó una matriz de confusión que permite comparar las predicciones del modelo con los resultados reales, y se calcularon métricas como exactitud, sensibilidad y especificidad, se obtiene las siguientes interpretaciones de los resultados:

- TN (True Negative) = 1475: estudiantes correctamente predichos como no alto rendimiento.
- FN (False Negative) = 126: estudiantes que realmente tuvieron alto rendimiento, pero el modelo los clasificó como no alto rendimiento.
- FP (False Positive) = 147: estudiantes clasificados por el modelo como alto rendimiento, pero que no lo fueron.

- TP (True Positive) = 1252: estudiantes correctamente clasificados como alto rendimiento.

En el proceso de evaluación del modelo de clasificación para predecir el alto rendimiento académico, los errores de tipo falso positivo adquieren especial relevancia. Estos casos corresponden a estudiantes que fueron clasificados por el modelo como de alto rendimiento, cuando en realidad no lo fueron.

Aunque el modelo muestra una alta capacidad predictiva general, los 147 falsos positivos identificados representan situaciones en las que se sobreestimó el desempeño estudiantil, lo que podría tener implicaciones en decisiones pedagógicas, asignación de recursos o seguimiento académico. Este tipo de error, junto con los 126 falsos negativos estudiantes que sí tuvieron alto rendimiento, pero no fueron reconocidos por el modelo, evidencia la necesidad de revisar los criterios de clasificación y considerar ajustes que mejoren la sensibilidad sin comprometer la especificidad.

Aun así, las cifras son bajas en comparación con los aciertos, lo que refuerza la consistencia del modelo

```
> cat("Exactitud (30%):", round(exactitud_test * 100, 2), "%\n")
Exactitud (30%): 90.9 %
```

Figura 32. Exactitud (30%)

Como se observa en la Figura 32, el modelo clasifica correctamente cerca del 91% de los casos, lo que indica un buen poder predictivo sobre datos no vistos (conjunto de prueba).

Síntesis interpretativa de los factores que explican el alto rendimiento en el ICFES

La evidencia obtenida permite responder directamente a la pregunta de investigación: ¿Qué factores del contexto socioeconómico y escolar permiten predecir la probabilidad de obtener un alto rendimiento en las pruebas ICFES? Los resultados muestran que esta probabilidad está explicada por una combinación de factores escolares y factores socioeconómicos, los cuales actúan de manera conjunta para influir en el desempeño académico.

En primer lugar, los factores escolares, especialmente los puntajes en Lectura Crítica y Sociales y Ciudadanas se consolidan como los predictores más influyentes. Estas áreas representan habilidades transversales de comprensión, argumentación y análisis que impactan directamente el rendimiento global en el examen.

En segundo lugar, ciertos factores socioeconómicos, como disponer de computador en el hogar y contar con automóvil, también aportan a la predicción del rendimiento. Estos elementos reflejan mejores condiciones de acceso a recursos tecnológicos, culturales y de apoyo académico.

El modelo final que integra simultáneamente los factores escolares y socioeconómicos alcanzó un 91% de precisión al predecir el rendimiento en datos nuevos, lo que confirma la solidez y relevancia de los predictores identificados.

Tarea 14	1. Repite el proceso de partición de datos, pero esta vez divide en 80% entrenamiento y 20% prueba. 2. Ajusta nuevamente el modelo con las mismas variables seleccionadas previamente. 3. Evalúa en el 20%: - Matriz de confusión - Exactitud 4. Compara todos estos resultados con los obtenidos en la validación 70/30 (91% de exactitud) 5. Contesta a las preguntas: ● ¿El modelo es estable o sus métricas cambian demasiado cuando se altera la partición? ● ¿Es más conveniente entrenar con 70% o con 80% en este caso? Justifica con evidencia numérica y conceptual: sesgo–varianza, tamaño del test, etc.

Cierre del ciclo de datos en la regresión lineal: interpretación final

Con este análisis se completa la última fase del ciclo de datos, correspondiente a la interpretación, luego de haber desarrollado

rigurosamente las etapas previas del proceso. En primer lugar, se formuló la pregunta de investigación que *guio* todo el trabajo. A partir de ella, se avanzó a la segunda etapa, centrada en la manipulación, preparación y transformación de los datos. En esta fase se limpiaron variables, se crearon nuevas categorías como *a*lto_rendimiento, se seleccionaron predictores mediante métodos como LASSO y se aplicaron procedimientos para garantizar la calidad, coherencia y utilidad analítica del conjunto de datos.

Posteriormente, el análisis multivariado y las comparaciones entre grupos revelaron patrones consistentes entre el desempeño académico y las variables del contexto escolar y socioeconómico.

Finalmente, en la etapa de análisis estadístico e interpretación, se integraron modelos, métricas, gráficos y contrastes de hipótesis para extraer conclusiones sólidas y fundamentadas.

Entre los hallazgos más relevantes, variables como tener computador o tener automóvil, indicadores de capital económico y acceso a recursos, mostraron una asociación significativa con el alto rendimiento, evidenciando brechas educativas persistentes.

Por otro lado, el fortalecimiento de competencias académicas, como lectura crítica y pensamiento social, junto con políticas que aumenten el acceso a recursos tecnológicos, se identifican como estrategias clave para mejorar la equidad y el rendimiento académico.

5.2. Análisis en Componentes Principales

Antes de avanzar con el análisis en componentes principales (ACP), es importante señalar que se repetirán los primeros cuatro

pasos de la regresión logística: preparación de los datos, selección de variables, codificación de las categorías y estimación inicial del modelo.

Aunque tanto el ACP como la regresión logística pueden involucrar múltiples variables independientes, sus propósitos y fundamentos estadísticos son distintos. El ACP es una técnica de aprendizaje no supervisado, lo que significa que no utiliza una variable objetivo para guiar el análisis; su objetivo es descubrir patrones internos en los datos, reducir la dimensionalidad y sintetizar la información en componentes que explican la mayor parte de la variabilidad.

En contraste, la regresión logística es una técnica de aprendizaje supervisado, diseñada para modelar la probabilidad de ocurrencia de una variable dependiente dicotómica (por ejemplo, alto rendimiento académico: sí/no) en función de un conjunto de predictores. Mientras el ACP busca estructura latente sin necesidad de etiquetas, la regresión logística busca relaciones explícitas entre variables para hacer predicciones. Por tanto, aunque ambas técnicas pueden usar las mismas variables como insumo, su lógica, aplicación y tipo de inferencia son fundamentalmente diferentes.

La pregunta guía que orienta este análisis es: *"¿Qué relaciones existen entre los puntajes por área y cómo se agrupan las competencias evaluadas en función de su aporte al rendimiento global?"*

En el marco del ciclo de datos, actualmente nos encontramos en la fase de formulación de la pregunta y manipulación de los datos. Una vez se ejecuten los códigos correspondientes y se realice el análisis del ACP, procederemos a la etapa de análisis e interpretación de los datos, para finalmente completar el ciclo con conclusiones basadas en evidencia y respaldadas por los resultados obtenidos.

Paso 5: seleccionar variables continuas

Se trabaja con los puntajes ICFES de cinco áreas: inglés, lectura crítica, matemáticas, sociales/ciudadanas y puntaje global, seleccionando únicamente aquellas variables continuas que serían utilizadas en el ACP.

Se eligieron punt_ingles, punt_lectura_critica, punt_matematicas, punt_sociales_ciudadanas y punt_global debido a su naturaleza de rendimiento académico continuo y a que presentan correlaciones moderadas a altas entre sí, lo que justifica su inclusión y permite obtener componentes significativos en el análisis.

```
datos_cont <- datosicfes[, c("punt_ingles", "punt_lectura_critica",
                             "punt_matematicas",      "punt_sociales_ciudadanas",
"punt_global")]
```

Paso 6: matriz de correlaciones

El cálculo de la matriz de correlaciones entre los puntajes de inglés, lectura crítica, matemáticas, sociales/ciudadanas y el puntaje

global constituye un paso esencial para la aplicación del ACP. Esto se debe a que el ACP se fundamenta en identificar patrones de relación entre variables y sintetizar la variabilidad compartida en componentes principales.

> Si las variables no estuvieran correlacionadas, cada una aportaría información independiente y el ACP perdería sentido, pues no habría estructura común que reducir. En cambio, cuando existen correlaciones significativas, el ACP puede captar esa variabilidad conjunta y transformarla en factores subyacentes[8] que explican mejor el rendimiento académico. Por ello, la matriz de correlaciones no solo permite evaluar la fuerza y dirección de las relaciones lineales, sino que también justifica la pertinencia de aplicar el ACP como técnica de reducción de dimensionalidad

```
cor_matrix <- cor(datos_cont)

print(round(cor_matrix, 2))
```

[8] Los factores subyacentes son dimensiones ocultas o comunes que explican las relaciones entre un conjunto de variables observadas.

```
> cor_matrix <- cor(datos_cont)
> print(round(cor_matrix, 2))
                          punt_ingles punt_lectura_critica
punt_ingles                      1.00                 0.58
punt_lectura_critica             0.58                 1.00
punt_matematicas                 0.58                 0.71
punt_sociales_ciudadanas         0.61                 0.78
punt_global                      0.72                 0.88
                          punt_matematicas punt_sociales_ciudadanas
punt_ingles                           0.58                     0.61
punt_lectura_critica                  0.71                     0.78
punt_matematicas                      1.00                     0.71
punt_sociales_ciudadanas              0.71                     1.00
punt_global                           0.88                     0.91
                          punt_global
punt_ingles                      0.72
punt_lectura_critica             0.88
punt_matematicas                 0.88
punt_sociales_ciudadanas         0.91
punt_global                      1.00
```

Figura 33. Matriz de correlación

Interpretación de la matriz de correlaciones de pearson

Los resultados muestran que:

- La correlación de inglés con las demás áreas varía entre 0.58 y 0.72, indicando relaciones moderadas a fuertes.
- Entre Lectura Crítica, Matemáticas y Sociales, las correlaciones oscilan entre 0.71 y 0.78, lo que representa relaciones fuertes.
- El puntaje global se correlaciona con todas las demás variables entre 0.88 y 0.99, mostrando asociaciones muy fuertes.

Existe una estructura común clara entre las variables, lo que justifica la aplicación del ACP y sugiere que el primer componente principal probablemente explicará la mayor parte de la varianza del conjunto de datos.

Tarea 15	Consulta qué significan las siguientes condiciones previas del ACP y explica si se cumplen en este caso: - Correlaciones moderadas–fuertes entre variables - Posible colinealidad - Variabilidad compartida - Asociación lineal **Nota**: no necesitas calcular nada adicional: solo interpreta con base en lo que ya se observó.

Paso 7: pruebas de adecuación del ACP

Antes de continuar con el estudio, se aplica el estadístico **Kaiser-Meyer-Olkin (KMO)** como medida de adecuación muestral, con el fin de evaluar si los datos son apropiados para realizar un ACP o un análisis factorial. El KMO compara la magnitud de las correlaciones observadas entre las variables con la de las correlaciones parciales, es decir, aquellas que permanecen una vez controlada la influencia de las demás.

Un valor global cercano a 1 indica que las correlaciones son sustanciales y que existe una estructura común que puede ser sintetizada mediante componentes principales; en cambio, valores bajos (por debajo de 0.5) sugieren que las variables no comparten

suficiente variabilidad y que el ACP no sería recomendable. Por ello, el cálculo del KMO constituye un paso previo indispensable, pues permite verificar la pertinencia de reducir la dimensionalidad de los datos y garantiza que el análisis se apoye en relaciones lineales significativas.

```
# KMO (Kaiser-Meyer-Olkin)

kmo_result <- KMO(cor_matrix)

print(kmo_result)
```

```
> # KMO (Kaiser-Meyer-Olkin)
> kmo_result <- KMO(cor_matrix)
> print(kmo_result)
Kaiser-Meyer-Olkin factor adequacy
Call: KMO(r = cor_matrix)
Overall MSA =  0.62
MSA for each item =
             punt_ingles     punt_lectura_critica
                    0.70                     0.67
        punt_matematicas punt_sociales_ciudadanas
                    0.59                     0.63
            punt_global
                    0.56
```

Figura 34. KMO resultado

```
bartlett_result <- cortest.bartlett(cor_matrix, n = nrow(datos_cont))

print(bartlett_result)
```

Rango MSA Global	Interpretación
0.90 – 1.00	Excelente

0.80 – 0.89	Muy bueno
0.70 – 0.79	Bueno
0.60 – 0.69	Aceptable
0.50 – 0.59	Malo
< 0.50	Inadecuado

Tabla 4. Interpretación resultados

MSA[9] global = 0.62 → Aceptable:

La medida global sugiere que, en conjunto, la matriz de correlaciones tiene suficiente estructura común para aplicar ACP, aunque no es excelente.

MSA por variable:

- punt_ingles = 0.70 → Bueno: La variable aporta información consistente y comparte correlación suficiente con otras variables.
- punt_lectura_critica = 0.67 → Aceptable: La variable es adecuada para ACP, aunque no destaca tanto como inglés.

[9] El MSA (*Measure of Sampling Adequacy*), o medida de adecuación muestral, es un indicador que se deriva del estadístico Kaiser-Meyer-Olkin (KMO) y que evalúa hasta qué punto las variables de un conjunto de datos comparten una estructura común de correlaciones

- punt_matematicas = 0.59 → Mediocre: La correlación parcial con otras variables es baja, por lo que su contribución al componente común es limitada.
- punt_sociales_ciudadanas = 0.63 → Aceptable: Adecuada para ACP, con correlaciones moderadas con otras variables.
- punt_global = 0.56 → Mediocre: Esta variable aporta menos estructura común relativa, indicando que parte de su varianza es única o específica.

```
> bartlett_result <- cortest.bartlett(cor_matrix, n = nrow(datos_cont))
> print(bartlett_result)
$chisq
[1] 43439.9

$p.value
[1] 0

$df
[1] 10
```

Figura 34. pchisq resultado

La Figura 34 presenta el resultado de la prueba de esfericidad de Bartlett aplicada a la matriz de correlaciones entre los puntajes por área. Esta prueba permite evaluar si las correlaciones observadas son suficientemente significativas como para justificar la aplicación del ACP. En este caso, el valor del estadístico chi-cuadrado fue de 43,439.9 con un valor p igual a 0, lo que indica que la matriz de correlaciones es significativamente diferente de una matriz identidad. Esto confirma que existe una estructura correlacional fuerte entre las variables, lo cual valida la pertinencia de aplicar el ACP para sintetizar la información.

p-valor	Interpretación
$p < \alpha = 0.05$	Rechazar $H_0 \rightarrow$ hay correlaciones significativas
$p \geq \alpha = 0.05$	No rechazar $H_0 \rightarrow$ no hay estructura correlacional

Tabla 5. Interpretación de hipótesis

La Tabla 6 complementa la figura al ofrecer una guía interpretativa clara sobre los posibles resultados de la prueba de Bartlett. En ella se establece que si el valor p es menor al nivel de significancia ($\alpha = 0.05$), se debe rechazar la hipótesis nula y concluir que hay correlaciones significativas entre las variables. Por el contrario, si el valor p es igual o superior a 0.05, no se rechaza la hipótesis nula y se considera que no existe una estructura correlacional adecuada. Esta tabla facilita la comprensión del criterio de decisión y refuerza la interpretación estadística del resultado obtenido en la figura.

El resultado de la prueba de esfericidad de Bartlett arrojó un estadístico chi-cuadrado de 43,439.9 con 10 grados de libertad y un valor p que se aproxima a cero. Este valor tan pequeño indica que la matriz de correlaciones es significativamente diferente de una matriz identidad, lo que significa que las relaciones entre las variables no son aleatorias. En consecuencia, se confirma la existencia de correlaciones sustanciales entre los puntajes evaluados, lo cual justifica y respalda la

aplicación ACP como técnica adecuada para sintetizar la información y explorar la estructura subyacente de los datos.

En consecuencia, los resultados evidencian una estructura subyacente sólida en la matriz de correlaciones, lo que respalda la pertinencia de aplicar el ACP como técnica adecuada para identificar los factores subyacentes y sintetizar los puntajes académicos en componentes que reflejen de manera más clara las competencias evaluadas.

Tarea 16	A partir de los valores observados, responde: - ¿Qué implicaciones tiene que Matemáticas y Puntaje Global tengan un MSA menor que 0.60? - ¿Este resultado justificaría eliminar alguna variable, o simplemente interpretar con cautela?

Paso 8: estandarizar las variables

Ahora, basándonos en la matriz de correlaciones entre variables, si estas tienen escalas diferentes, las variables con mayor varianza pueden dominar el análisis, distorsionando los resultados. Para evitarlo, se realiza una estandarización de las variables:

- Centrar: se resta la media de cada variable, de modo que todas queden con media cero.
- Escalar: se divide por la desviación estándar, logrando que todas tengan desviación estándar igual a uno.

Este procedimiento garantiza que cada variable contribuya de manera equitativa al ACP, evitando que algunas dominen por su escala y permitiendo identificar correctamente la estructura latente común. En R se la estandarización se realiza con la función scale, así:

```
# (ACP se realiza sobre variables centradas y escaladas)
datos_scaled <- scale(datos_cont)
```

Tarea 17	Explica qué efectos tendría NO estandarizar en un conjunto como el del ICFES, donde las áreas tienen distribuciones de puntajes diferentes: - ¿Qué variable dominaría el primer componente? - ¿Qué relación tendría con la varianza? - ¿Cómo se distorsionarían las cargas o pesos?

Paso 9: realizar el ACP

En este paso se procede a la ejecución formal del ACP sobre los datos previamente estandarizados. El objetivo es identificar la estructura que subyace en las variables evaluadas y sintetizar la información en un conjunto reducido de componentes que concentren la mayor parte de la variabilidad.

```
acp_result <- principal(datos_scaled, nfactors = ncol(datos_scaled), rotate = "none")

# Ver resultados principales

print(acp_result)
```

```
> # Ver resultados principales
> print(acp_result)
Principal Components Analysis
Call: principal(r = datos_scaled, nfactors = ncol(datos_scaled), rotate = "none")
Standardized loadings (pattern matrix) based upon correlation matrix
                          PC1   PC2   PC3   PC4   PC5 h2       u2 com
punt_ingles              0.77  0.63 -0.01  0.03  0.01  1 0.0e+00 1.9
punt_lectura_critica     0.89 -0.19 -0.25  0.33  0.03  1 3.4e-15 1.5
punt_matematicas         0.88 -0.15  0.46  0.02  0.04  1 2.0e-15 1.6
punt_sociales_ciudadanas 0.91 -0.13 -0.21 -0.33  0.04  1 2.3e-15 1.4
punt_global              0.99 -0.07  0.02 -0.03 -0.11  1 2.0e-15 1.0

                       PC1  PC2  PC3  PC4  PC5
SS loadings           3.97 0.48 0.31 0.22 0.02
Proportion Var        0.79 0.10 0.06 0.04 0.00
Cumulative Var        0.79 0.89 0.95 1.00 1.00
Proportion Explained  0.79 0.10 0.06 0.04 0.00
Cumulative Proportion 0.79 0.89 0.95 1.00 1.00

Mean item complexity =  1.5
Test of the hypothesis that 5 components are sufficient.

The root mean square of the residuals (RMSR) is  0
 with the empirical chi square  0  with prob <  NA

Fit based upon off diagonal values = 1
```

Figura 35. Resultados principales

El ACP revela que un único componente, el PC1, concentra la mayor parte de la varianza explicada (79%). Este componente actúa como un factor general de rendimiento académico, ya que presenta cargas muy elevadas en todas las áreas evaluadas, lo que indica que resume de manera eficaz el desempeño global de los estudiantes. Los demás componentes aportan una proporción mínima de variabilidad y carecen de relevancia práctica, lo que simplifica la interpretación: el

rendimiento académico puede representarse adecuadamente en una sola dimensión subyacente que integra las competencias analizadas.

Tarea 18	Explica qué problemas podrían surgir si no se escalan las variables antes de realizar el ACP.

Paso 10: analizar autovalores

Tras confirmar la adecuación de los datos y estandarizar las variables, se procede al análisis de los autovalores obtenidos del ACP (eigenvalues <- acp_result$values) con el fin de determinar la proporción de varianza explicada por cada componente principal (var_explicada <- acp_result$Vaccounted). Cada autovalor indica la cantidad de información o variabilidad que captura su componente correspondiente; así, los autovalores de mayor magnitud señalan los componentes que concentran más información del conjunto de datos. Por ejemplo, si el primer componente tiene un autovalor de 3.5, esto significa que explica una gran proporción de la variabilidad total. Este examen resulta fundamental para seleccionar los componentes más relevantes, garantizando que el ACP resuma de forma eficiente la variabilidad y permita identificar las dimensiones que mejor representan el rendimiento académico. De esta manera, se priorizan los componentes que aportan más información y se descartan aquellos que aportan poco, optimizando el análisis y facilitando la interpretación de los resultados.

```
eigenvalues <- acp_result$values

eigenvalues

# Varianza explicada (%)

var_explicada <- acp_result$Vaccounted

var_explicada
```

```
> #========================================
> # 10. Analizar autovalores (Eigenvalues)
> #========================================
> eigenvalues <- acp_result$values
> eigenvalues
[1] 3.97004420 0.48163356 0.31295008 0.22009839 0.01527377
> # Varianza explicada (%)
> var_explicada <- acp_result$Vaccounted
> var_explicada
                             PC1         PC2         PC3         PC4          PC5
SS loadings            3.9700442 0.48163356 0.31295008 0.22009839 0.015273768
Proportion Var         0.7940088 0.09632671 0.06259002 0.04401968 0.003054754
Cumulative Var         0.7940088 0.89033555 0.95292557 0.99694525 1.000000000
Proportion Explained   0.7940088 0.09632671 0.06259002 0.04401968 0.003054754
Cumulative Proportion  0.7940088 0.89033555 0.95292557 0.99694525 1.000000000
```

Figura 36. Varianza Explicada

Tras confirmar la adecuación de los datos y estandarizar las variables, se procede al análisis de los autovalores con el fin de determinar la proporción de varianza explicada por cada componente principal. Los resultados muestran que el primer componente (PC1) posee un autovalor de 3.97 y explica el 79.4 % de la varianza total, lo que indica que concentra la mayor parte de la información contenida en las variables originales. Los componentes restantes presentan autovalores significativamente menores y explican proporciones reducidas de varianza (PC2: 9.6 %, PC3: 6.3 %, PC4: 4.4 %, PC5: 0.3 %), lo que sugiere que su aporte es marginal. Este patrón justifica

la selección de PC1 como componente relevante, ya que permite sintetizar el rendimiento académico en una sola dimensión representativa.

Además, este análisis prepara la base para el siguiente paso con el criterio de Kaiser, que ayuda a decidir la cantidad componentes a retener y que se utilizarán para interpretar la estructura latente y construir las representaciones gráficas del ACP.

Tarea 19	Responde las siguientes preguntas - ¿Cómo se relaciona la tabla de varianza explicada con la elección del número de componentes? - ¿Qué pasaría si interpretaras todos los componentes sin considerar los autovalores? - ¿En qué afectaría esto la calidad y claridad del análisis?

Paso 11: determinar el número de componentes significativos

```
# 1. Criterio de Kaiser (autovalores > 1)
which(eigenvalues > 1)
eigenvalues <- acp_result$values
plot(eigenvalues, type = "b",
     pch = 19, frame = FALSE,
```

```
    xlab = "Componente Principal",
    ylab = "Autovalor",
    main = "Gráfico de Sedimentación (Scree Plot)")
abline(h = 1, col = "red", lty = 2) # línea de Kaiser
```

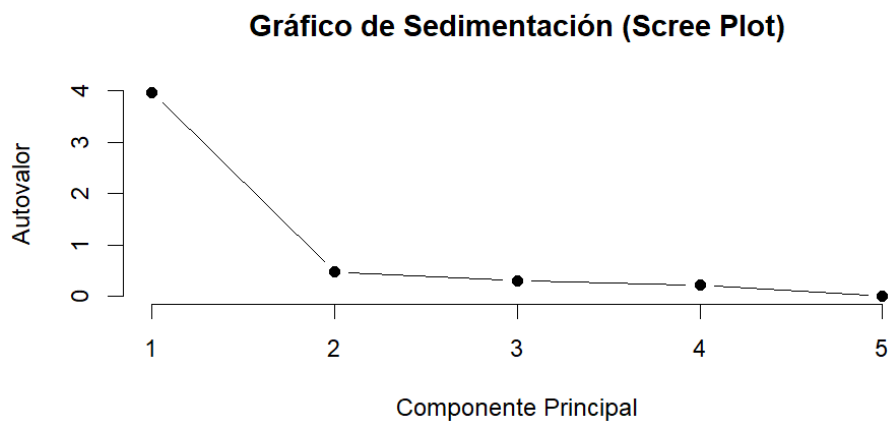


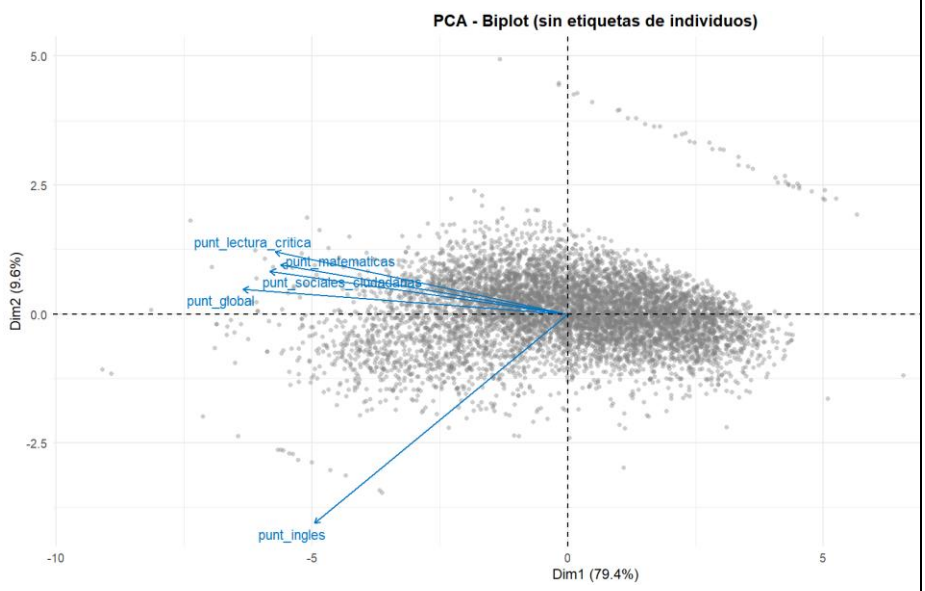
Figura 37. Gráfico de Sedimentación

El criterio de Kaiser y el scree plot confirman que solo el primer componente (PC1) es significativo, con autovalor = 3.97 (> 1). Los demás componentes tienen autovalores < 1 y no aportan información relevante. El ACP se puede interpretar como un factor único de rendimiento académico.

```
#Representación gráfica.
library(factoextra)
```

```
# Convierte a un objeto tipo PCA para usar fviz
acp_for_plot <- prcomp(scale(datosicfes[, c("punt_ingles",
                                            "punt_lectura_critica",
                                            "punt_matematicas",
                                            "punt_sociales_ciudadanas",
                                            "punt_global")]))
# Biplot sin los valores individuales, con mejor estética
fviz_pca_biplot(acp_for_plot,
                axes = c(1, 2),
                geom.ind = "point",       # solo puntos para individuos
                col.ind = "gray50",       # color tenue para los puntos
                alpha.ind = 0.4,          # transparencia
                pointshape = 16,          # forma de los puntos
                pointsize = 1.5,          # tamaño de los puntos
                label = "var",            # solo etiquetas de variables
                col.var = "#0073C2FF",    # color de las variables
                repel = TRUE,             # evita superposición de textos
                title = "PCA - Biplot (sin etiquetas de individuos)") +
```

```
theme_minimal() +
theme(
  plot.title = element_text(hjust = 0.5, face = "bold", size = 14),
  axis.title = element_text(size = 12),
  axis.text = element_text(size = 10)  )
```



Paso 12: análisis del Biplot PCA – Componentes y estructura

El biplot del ACP permite visualizar cómo las variables y los estudiantes se distribuyen en un plano reducido a dos dimensiones, capturando la mayor parte de la varianza del conjunto de datos.

1. Varianza explicada

- Dim1 (PC1): 79.4%
- Dim2 (PC2): 9.6%
- Acumulado (Dim1 + Dim2): 89%

Esto indica que el plano formado por Dim1 y Dim2 resume casi toda la variabilidad, logrando una reducción dimensional eficaz sin pérdida relevante de información.

2. Dimensión 1 (79.4%)

- Representa un factor general de rendimiento académico.
- Las variables punt_lectura_critica, punt_matematicas, punt_sociales_ciudadanas y punt_global tienen flechas largas y alineadas, mostrando alta asociación con Dim1.
- Estudiantes con valores altos en Dim1 presentan buen desempeño global y equilibrado en estas áreas.

3. Dimensión 2 (9.6%)

- Aporta información secundaria, especialmente relacionada con punt_ingles.
- El vector de inglés apunta en una dirección distinta, reflejando habilidades lingüísticas que no siguen el patrón general del rendimiento académico.

4. Relación entre variables

- Flechas cercanas y alineadas → alta correlación positiva (ej. Lectura critica, matemáticas, sociales, punt_global).

- Flechas divergentes → correlación débil o negativa (ej. inglés).
- Esto confirma que las áreas centrales están altamente correlacionadas, mientras que inglés tiene un comportamiento más independiente, útil para análisis específicos por área.

5. Dispersión de los estudiantes

- La nube central indica rendimiento promedio.
- Valores positivos en Dim1 → estudiantes sobresalientes.
- Valores negativos en Dim1 → estudiantes con bajo desempeño.

- El biplot permite identificar perfiles y segmentar estudiantes según su rendimiento, proporcionando información visual complementaria al análisis numérico.

Relaciones que se pueden deducir entre áreas evaluadas a través de componentes principales

Tarea 20	1. A partir de la matriz de correlaciones, redacta tres conclusiones adicionales sobre las relaciones entre las variables: dirección, fuerza de asociación, agrupamientos o posibles estructuras subyacentes no mencionadas en el texto. 2. Con la base de datos de Pokémon del apartado "Atrápalos con R", realiza un ACP (estandarización, autovalores, criterio de Kaiser y

	scree plot). Luego, escribe conclusiones sobre: - Los componentes significativos. - La estructura subyacente que revelan las variables. - Una breve comparación con el ACP del ICFES, señalando similitudes o diferencias en la distribución de la varianza y el comportamiento de los componentes.

- Los dos primeros componentes explican el 89% de la variabilidad total, lo que permite una representación bidimensional altamente fiel del rendimiento académico.
- Las variables punt_global, punt_lectura_critica, punt_matematicas y punt_ciencias están fuertemente correlacionadas entre sí y alineadas con la dimensión principal (Dim1), lo que sugiere un patrón común de desempeño.
- La variable punt_ingles se orienta en una dirección distinta, indicando que su comportamiento no se ajusta completamente al patrón general de rendimiento académico.
- Los individuos ubicados en el extremo positivo de Dim1 representan estudiantes con alto rendimiento global, mientras que los del extremo negativo reflejan desempeños más bajos.

- La dispersión moderada en Dim2 revela que el componente relacionado con inglés introduce una variabilidad secundaria que no está presente en las demás áreas.

Conclusión de análisis del ACP

El ACP resume eficazmente el rendimiento académico en un componente dominante (Dim1) y una dimensión secundaria (Dim2) que aporta información específica sobre inglés. La estructura subyacente se refleja claramente en las correlaciones y la dispersión de los estudiantes.

Con respecto al ciclo de datos, el análisis del ACP se sitúa dentro de la fase de interpretación de resultados, que constituye la etapa final del ciclo tras la recolección, limpieza, transformación y preparación de los datos. En esta fase, se integran modelos estadísticos, métricas numéricas y visualizaciones, como el biplot, para comprender de manera profunda cómo las variables académicas se relacionan entre sí y cómo influyen en el desempeño global de los estudiantes.

El ACP permitió identificar un componente dominante (Dim1) que resume la mayor parte de la variabilidad del rendimiento académico, mostrando que variables como lectura crítica, matemáticas, sociales y puntaje global están fuertemente correlacionadas. Esta dimensión refleja un factor general de rendimiento académico, útil para comparar perfiles de estudiantes y

detectar patrones de desempeño equilibrado. La Dimensión 2 (Dim2) aporta información secundaria, especialmente relacionada con el desempeño en inglés, evidenciando que ciertas habilidades específicas no siguen el patrón general y pueden requerir análisis o intervenciones particulares.

El biplot también permitió visualizar la dispersión de los estudiantes, identificando grupos de alto rendimiento y bajo rendimiento, y evaluando la consistencia de los patrones observados. Esta combinación de análisis cuantitativo y visual proporciona evidencia sólida sobre la estructura latente del conjunto de datos, confirmando que la reducción dimensional realizada no pierde información relevante.

En conclusión, la interpretación del ACP cierra el ciclo de datos, ya que traduce los resultados estadísticos en información útil y accionable, permite comprender la relación entre variables y desempeño, y brinda insumos para la toma de decisiones educativas, como la identificación de áreas que requieren refuerzo o el diseño de estrategias pedagógicas más efectivas. Este análisis demuestra cómo un enfoque sistemático del Ciclo de Datos puede generar conocimiento profundo y fundamentado, integrando preparación de datos, modelado y visualización para obtener conclusiones significativas y aplicables.

Bibliografía

Agresti, A. (2018). *Statistical methods for the social sciences* (5th ed.). Pearson.

Anderson, D. R., Sweeney, D. J., Williams, T. A., Camm, J. D., & Cochran, J. J. (2019). *Estadística para administración y economía* (13.ª ed.). Cengage Learning.

Allison, P. D. (2001). *Missing Data*. Sage Publications.

A. Mood, F. A. Graybill y D. Boes, Introduction to the Theory of Statistics. Third Edi tion, McGraw Hill. 1974. [En línea]. Disponible en: https://amzn.to/32v5MPK Freund, J. E., & Perles, B. M. (2016). *Estadística matemática con aplicaciones* (8.ª ed.). Pearson Educación.

Davenport, T. H., & Harris, J. G. (2007). *Competing on Analytics: The New Science of Winning*. Harvard Business School Press.

Freund, J. E. (2014). *Estadística matemática con aplicaciones* (7.ª ed.). Pearson Educación.

Hernández, R., Fernández, C., & Baptista, P. (2014). *Metodología de la investigación* (6.a ed.). McGraw-Hill.

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied Logistic Regression* (3rd ed.). Wiley.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning: With applications in R* (2nd ed.). Springer. https://doi.org/10.1007/978-1-0716-1418-1

Johnson, R. A., & Wichern, D. W. (2018). *Applied multivariate statistical analysis* (7th ed.). Pearson.

Montgomery, D. C., & Runger, G. C. (2018). *Applied statistics and probability for engineers* (7th ed.). Wiley.

Montgomery, D. C. (2019). *Design and analysis of experiments* (10th ed.). Wiley.

Little, R. J. A., & Rubin, D. B. (2019). *Statistical Analysis with Missing Data* (3rd ed.). Wiley.

Pokémon Database. (s. f.). *Pokémon Database*. https://pokemondb.net/

Peña, D. (2013). *Fundamentos de estadística* (2.ª ed.). Alianza Editorial.

P. Dalgaard, Introductory Statistics with R. Second Edition, Springer, 2008. [En línea]. Disponible en: http://bit.ly/30yYw3

Ross, S. M. (2010). *Introducción a la probabilidad y estadística para ingeniería y ciencias* (4.ª ed.). Academic Press.

Stowell, Using R for Statistics. Apress, 2014. [En línea]. Disponible en: https:// amzn.to/2XGIoQx

Triola, M. F. (2020). *Estadística* (13.ª ed.). Pearson Educación.

Wooldridge, J. M. (2020). *Introducción a la econometría: Un enfoque moderno* (7.ª ed.). Cengage Learning.