CAPÍTULO 1

INVESTIGACIÓN EN EDUCACIÓN A LA LUZ DE LA CIENCIA DE DATOS

Data de aceite: 01/11/2024

José González Campos

Felipe Lillo Viedma

Catherine Araya Pérez

CONSIDERACIONES INICIALES

Este capítulo ha sido organizado en tres secciones, la primera se identifica como un espacio de provocación al lector e investigador en educación, así como precisar en algunos conceptos metodológicos, y discusiones filosóficas relativa a posibles áreas de investigación. La segunda sección nos presenta una serie de herramientas de análisis desde la ciencia de datos pertinentes a investigaciones en educación, que cierra con el uso de IA en revisiones sistemáticas. Finalmente, en la tercera sección, se presenta una breve discusión relativa a la ética en investigación en educación a la luz de la ciencia del dato.

SECCIÓN I

Para introducir esta sección, reflexionaremos sobre la internacionalización y la relevancia de este libro como un mecanismo para su operacionalización. La internacionalización implica que las universidades se comprometan de manera conjunta, colaborativa y abierta a abordar el desafío del aprendizaje de nuestros estudiantes. Entre las estrategias para poner en práctica este enfoque se encuentran los espacios de intercambio de experiencias, tanto exitosas como no tanto, en el proceso de enseñanza y aprendizaje.

Eso nos permitirá unir fuerzas en la búsqueda de un objetivo común. La internacionalización de la educación es un proceso que amplía los horizontes de los estudiantes a través de experiencias globales, fomentando un entorno educativo solidario y colaborativo. Este enfoque no debe ser un privilegio exclusivo, sino una oportunidad accesible para todos, enmarcando la inclusión como un principio

fundamental. La colaboración internacional debe centrarse en eliminar las barreras que dificultan el acceso de ciertos grupos, adaptando prácticas y recursos para ser culturalmente inclusivos. Además, es esencial crear espacios que aborden temáticas emergentes, facilitando el intercambio de ideas y la construcción de estrategias que enriquezcan el aprendizaje. Integrar estos temas en los diálogos internacionales asegura que nuestras pedagogías se mantengan relevantes y efectivas ante los nuevos desafíos. En última instancia, la promoción de estos espacios contribuye a una educación dinámica y proactiva, donde la innovación y la colaboración se unen para enfrentar los retos educativos del presente y del futuro.

En función de este contexto, traemos a discusión una primera temática que dice relación con la necesidad de la ciencia de datos en la investigación educativa. En la actualidad, los datos se han convertido en un recurso esencial en la educación, transformando la forma en que abordamos y entendemos los problemas educativos. La inmensa cantidad de datos generados por sistemas y dispositivos educativos ofrece oportunidades valiosas para mejorar el aprendizaje. En un entorno académico competitivo, las instituciones que no adopten la ciencia de datos corren el riesgo de quedar rezagadas, ya que la capacidad de analizar información puede ser crucial para obtener fondos y publicar investigaciones. La investigación educativa debe integrar herramientas de ciencia de datos, como análisis estadístico avanzado y aprendizaje automático, para abordar problemas con mayor precisión.

Esas técnicas permiten identificar patrones y predecir tendencias, mejorando la calidad de la investigación y optimizando la toma de decisiones. Al basarse en evidencia empírica, los investigadores pueden respaldar sus recomendaciones, fortaleciendo su impacto en las políticas educativas. Sin embargo, la implementación de la ciencia de datos enfrenta desafíos, como la necesidad de habilidades técnicas, la gestión de grandes volúmenes de datos y la protección de la privacidad estudiantil. Superar estos obstáculos requiere inversión en capacitación y tecnología. A pesar de estos retos, la ciencia de datos abre nuevas oportunidades para la innovación, permitiendo el desarrollo de soluciones personalizadas y la colaboración con expertos en tecnología. La ciencia de datos es una necesidad estratégica para avanzar en un entorno educativo cada vez más complejo y competitivo.

Para seguir avanzando en este capítulo, haremos algunos alcances metodológicos o precisiones. La investigación requiere una clara distinción entre las estrategias de revisión y los enfoques metodológicos. Estrategias como el método PRISMA, que significa “Preferred Reporting Items for Systematic Reviews and Meta-Analyses”, son herramientas específicas que organizan y justifican revisiones sistemáticas de la literatura. PRISMA asegura que estas revisiones sean exhaustivas y replicables, pero no define el enfoque metodológico, que es un marco teórico más amplio que guía el proceso investigativo. Este enfoque puede ser cualitativo, cuantitativo o mixto, orientando cómo se

plantean las preguntas de investigación, se recopilan y analizan los datos, y se interpretan los resultados.

El enfoque cuantitativo, en particular, va más allá de la simple aplicación de métodos de revisión. Se basa en una perspectiva teórica que guía la recopilación y análisis de datos, centrándose en la cuantificación de variables y en el uso de técnicas estadísticas para identificar patrones y relaciones. Este enfoque también implica la formulación de hipótesis y la aplicación de modelos teóricos, proporcionando un contexto empírico que enriquece la interpretación de los hallazgos. Es fundamental definir cómo se operacionalizan las variables en la investigación cuantitativa, especificando indicadores, métodos de recolección y técnicas de análisis estadístico. Además, el enfoque metodológico no se limita a la naturaleza cualitativa o cuantitativa de las variables. Una investigación puede emplear un enfoque cuantitativo para analizar patrones en datos cualitativos o viceversa, dependiendo de la pregunta de investigación y los objetivos del estudio. Por ejemplo, al investigar el impacto de un programa educativo en el rendimiento académico, un enfoque cuantitativo implicaría definir variables como calificaciones y asistencia, recolectar datos mediante encuestas, y aplicar análisis estadísticos como ANOVA. Este enfoque proporciona evidencia empírica sólida y permite una evaluación objetiva. El enfoque metodológico define la estructura y ejecución de los estudios, siendo esencial para la investigación. Mientras que estrategias como PRISMA ayudan en la presentación de estudios, el enfoque metodológico proporciona el marco teórico que guía todo el proceso, sin estar determinado por la naturaleza de las variables, sino por la pregunta de investigación y los objetivos del estudio.

Basado en este contexto y la precisión metodológica referida al enfoque, damos paso a la primera provocación que dice relación con la integración de constructos teóricos y la invitación a reflexionar sobre realidades complejas en investigación educativa. La identidad personal es un collage de experiencias, interacciones y contextos que nos conforman. Analizarla solo a través de un aspecto aislado, como la historia familiar o los logros profesionales, limita nuestra comprensión. Reconocer la complejidad de la identidad implica considerar la multiplicidad de influencias, lo que refleja un enfoque más holístico en la investigación.

La evolución de la investigación contemporánea ha cambiado su perspectiva teórica, reconociendo que fenómenos complejos, como el aprendizaje, no pueden entenderse mediante un único constructo. Tradicionalmente, los estudios se basaban en un solo constructo, asumiendo que estos eran independientes. Sin embargo, ahora se busca integrar múltiples constructos teóricos para ofrecer una visión más completa. Por ejemplo, al estudiar el rendimiento académico, resulta útil combinar teorías sobre motivación, estrategias de aprendizaje y contexto socioeconómico, permitiendo así un análisis más profundo y matizado. Un desafío importante es cómo operacionalizar el diálogo entre diferentes constructos teóricos.

La integración requiere desarrollar modelos que muestren las interacciones entre estos constructos. En estudios sobre intervenciones educativas, por ejemplo, es necesario conectar teorías de capital cultural y motivación académica, definiendo indicadores claros para medir sus efectos. La representación gráfica de estas interacciones mediante diagramas de ruta facilita la comprensión de la complejidad de las teorías involucradas. Además, la integración de constructos complica el análisis tradicional. Por ello, es esencial aplicar estrategias analíticas avanzadas que permitan explorar las interacciones de manera efectiva. Métodos como modelos de ecuaciones estructurales o análisis de redes son útiles para examinar cómo los constructos se relacionan, extrayendo información significativa de estas interacciones. Superar la idea de que los constructos deben ser independientes permitirá un enfoque más integrado y comprensivo. Dado que los fenómenos educativos son sistemas complejos, se requiere el uso de herramientas que manejen esta complejidad. La teoría de grafos es una opción poderosa para modelar relaciones entre constructos, ayudando a visualizar y analizar patrones de interacción que no son evidentes con métodos tradicionales. Esa aproximación permite una comprensión más matizada de la realidad investigada. Asimismo, la realidad educativa es un sistema interconectado donde el aprendizaje depende de múltiples factores, como el entorno familiar, social y cultural.

Un enfoque integral es esencial para captar las conexiones que afectan a los estudiantes, evitando análisis simplistas que ignoren la complejidad de sus experiencias. Las trayectorias educativas están determinadas por interacciones entre diversos factores. Predecir estos trayectos requiere un entendimiento detallado de cómo se relacionan. Aunque la predicción exacta puede ser difícil debido a la naturaleza compleja de los sistemas, los modelos bien fundamentados pueden ofrecer estimaciones más cercanas a la realidad. Es fundamental utilizar métodos avanzados como sistemas dinámicos y teoría de grafos para representar y analizar las interacciones en la realidad educativa. Los enfoques analíticos deben ser adaptables, integrando nuevos datos a medida que surgen para mantener su precisión y relevancia. Así, la investigación educativa puede obtener visiones más precisas y efectivas, mejorando los procesos formativos y los resultados de aprendizaje.

Una perspectiva que toma fuerza en esta mirada compleja, es la denominada estadística a posteriori, que levanta constructos latentes y no imposiciones previas a los datos. En la investigación, el enfoque tradicional a priori define variables y factores antes de la recolección de datos, basándose en categorías y constructos teóricos preexistentes. Este método proporciona estructura y la posibilidad de aplicar técnicas estadísticas consolidadas, pero presenta limitaciones significativas, especialmente en contextos complejos como el educativo. Las caracterizaciones a priori pueden restringir la perspectiva, ya que imponen un marco rígido que a menudo no captura la diversidad y complejidad de los fenómenos educativos. Además, la falta de flexibilidad en estas categorías puede llevar a la pérdida de patrones emergentes y relaciones inesperadas, limitando así la interpretación de datos relevantes. La estadística a posteriori emerge como una herramienta poderosa para superar

estas limitaciones. Este enfoque permite que los datos revelen patrones y estructuras latentes no anticipadas, utilizando técnicas como el análisis de conglomerados y el análisis factorial. Esto facilita la identificación de perfiles emergentes y relaciones subyacentes que enriquecen la comprensión del fenómeno estudiado. Al ofrecer flexibilidad en el análisis, la estadística a posteriori permite explorar nuevas relaciones sin las restricciones de las categorías predefinidas, lo que es especialmente valioso en campos heterogéneos como la educación.

Ese enfoque también contribuye al refinamiento de los constructos teóricos existentes, proporcionando evidencia empírica sobre cómo se organizan y relacionan los datos. Así, se posibilita la revisión o creación de nuevos constructos que se alineen mejor con la realidad observada. Al romper con los esquemas teóricos rígidos, la estadística a posteriori promueve la actualización continua de las teorías, adaptándolas a la realidad emergente identificada en los datos. En el ámbito educativo, este enfoque no solo enriquece el conocimiento sobre los factores que influyen en el aprendizaje y el rendimiento, sino que también facilita una comprensión más dinámica y matizada de las interacciones entre variables.

La capacidad de la estadística *a posteriori* para revelar nuevas estructuras y relaciones convierte este enfoque en un componente esencial para abordar la complejidad de los fenómenos investigados. La estadística a posteriori se presenta como una alternativa fundamental a las caracterizaciones a priori, permitiendo una visión más rica y precisa de la realidad estudiada. Este enfoque no solo ayuda a desvelar patrones ocultos y a redefinir constructos teóricos, sino que también proporciona una base para una comprensión más profunda de las interacciones en contextos educativos. Al adaptar las teorías a los datos emergentes, se abre la posibilidad de expandir nuestras concepciones teóricas, reflejando mejor la complejidad del mundo real.

Otro aspecto que busca provocar, dice relación con una emergente discusión en torno a determinismo y libre albedrío en el proceso de enseñanza y aprendizaje. En el ámbito educativo, el determinismo plantea que los factores externos, como el contexto socioeconómico y las políticas educativas, son determinantes clave en el éxito y las trayectorias de los estudiantes. Por ejemplo, aquellos que provienen de comunidades desfavorecidas enfrentan barreras como la falta de recursos y apoyo académico, lo que limita sus oportunidades educativas y, por ende, su futuro. Estas limitaciones son más que simples circunstancias; son el resultado de un entramado de influencias históricas que modelan su realidad. La persona que eres es el resultado de todas las interacciones entre biología y entorno que vinieron antes, y cada influencia anterior proviene sin interrupción de los efectos de sus propias influencias. Como tal, no hay ningún punto en la secuencia en la que puedas insertar una libertad de voluntad.

Por otro lado, el concepto de libre albedrío sostiene que los individuos pueden tomar decisiones que moldean su propio destino, a pesar de las condiciones externas (Sapolsky,

2024). Esto se traduce en la autonomía del estudiante, quien puede influir activamente en su aprendizaje mediante elecciones sobre sus estrategias de estudio y metas académicas. La motivación y la autodisciplina son esenciales en este contexto, al igual que el acceso a recursos y apoyo que facilitan la toma de decisiones informadas. Es crucial reconocer que determinismo y libre albedrío no son excluyentes, sino que interactúan de maneras complejas en el proceso educativo. Aunque el entorno puede influir en las oportunidades, los estudiantes aún pueden ejercer su libre albedrío dentro de esos límites. Las estrategias educativas deben, por lo tanto, abordar ambas dimensiones: trabajar para eliminar barreras estructurales y al mismo tiempo fomentar la autonomía de los estudiantes. La flexibilidad en el diseño de programas educativos permite que los estudiantes ejerciten su libre albedrío en contextos diversos. En este sentido, se pueden desarrollar estrategias adaptativas que reconozcan las circunstancias individuales de los estudiantes, facilitando un equilibrio entre las limitaciones externas y las decisiones personales. Tanto el determinismo como el libre albedrío juegan roles interrelacionados en el proceso educativo. Mientras que el determinismo subraya las restricciones impuestas por el contexto, el libre albedrío enfatiza la capacidad de los estudiantes para decidir y moldear su camino. Un enfoque integral en la educación debe considerar estas dos perspectivas, proporcionando el apoyo necesario para superar obstáculos y, al mismo tiempo, fomentar el empoderamiento y la autonomía, maximizando así las oportunidades de aprendizaje y éxito para todos los estudiantes.

Para ilustrar, presentamos una fábula. En una selva vibrante y llena de vida, un brillante autobús verde esperaba a los animales para llevarlos a un gran evento. Este autobús podía transportar a treinta pasajeros, y todos estaban emocionados por la aventura que les esperaba. El primero en llegar fue un astuto puma llamado Pedro. Al subirse, vio que el autobús estaba vacío y eligió el asiento más cercano a la ventana. Su libertad era absoluta; podía sentarse donde quisiera. Desde allí, observaba el paisaje con satisfacción, disfrutando de la sensación de tener todas las opciones a su alcance. Poco después, una elegante gacela se acercó al autobús. Al subir, notó que el asiento junto a la ventana ya estaba ocupado por Pedro. Aunque aún había otros asientos libres, su libertad se vio limitada. Sabía que no podía sentarse al lado del puma, famoso por ser un rival en la selva. Así, su elección se complicó desde el principio. A medida que la fila avanzaba, una majestuosa elefanta llegó al autobús. Con dos asientos ya ocupados, su libertad se redujo aún más. Pedro y la gacela no eran amigos, y la elefanta debía evitar acercarse a ellos. Con cada nuevo pasajero, las opciones disponibles se esfumaban, y la presión por encontrar un lugar adecuado aumentaba.

Con el paso del tiempo, más animales subieron al autobús. Cada nuevo pasajero encontraba menos asientos libres, y algunos incluso se dieron cuenta de que ya no podían subir, pues los espacios disponibles estaban situados en un entorno irrealizable para ellos. La libertad de elección, que al principio había sido amplia, se convirtió en una serie de restricciones dictadas por las decisiones de los que llegaron primero. Esta historia

demuestra que las decisiones de quienes llegan primero pueden restringir las opciones de los que siguen, reflejando cómo las circunstancias pueden influir en las oportunidades futuras. La fábula ilustra cómo, a medida que avanzamos, nuestras elecciones se ven condicionadas por factores externos. La libertad no es absoluta; está sujeta a restricciones impuestas por el entorno.

Las decisiones iniciales de los primeros pasajeros afectan significativamente a los que vienen después, simbolizando cómo las circunstancias pueden modelar las oportunidades futuras, especialmente en la educación. Ante un escenario donde las opciones son limitadas, los animales deben ser flexibles y adaptarse para encontrar una solución viable, subrayando la importancia de ser resilientes ante cambios. La fábula del autobús en la selva nos recuerda que determinismo y libre albedrío están entrelazados. Pedro, al elegir su asiento, parecía tener total libertad, pero su decisión estuvo influenciada por su carácter y experiencias previas. Este escenario nos enseña que a menudo lo que percibimos como libertad es, en realidad, una ilusión condicionada por múltiples factores. Así, la historia invita a una reflexión profunda sobre cómo nuestras elecciones se ven limitadas por el contexto y las decisiones de los demás. Reconocer estas influencias no implica renunciar a nuestra capacidad de decisión, sino entender la complejidad de nuestras vidas en un mundo donde ambos conceptos coexisten. La verdadera libertad puede ser más una cuestión de percepción que de pura elección.

SECCIÓN II - CIENCIA DE DATOS E INVESTIGACIÓN EDUCATIVA

La Asociación de Ciencia de Datos[1] define la ciencia de datos como un campo interdisciplinario que combina métodos científicos, estadísticos y computacionales para extraer conocimiento y generar predicciones a partir de grandes volúmenes de datos. En esta sección, se abordará cómo dicha definición se manifiesta en el estudio científico de la educación, considerando aspectos metodológicos y áreas de aplicación en el ámbito educativo. En primer lugar, se comentará sobre el advenimiento de grandes volúmenes de datos en educación y los desafíos metodológicos asociados. Posteriormente, se presentarán problemáticas educativas en las que la ciencia de datos puede constituir un apoyo relevante.

ASPECTOS METODOLÓGICOS EN LA CIENCIA DE DATOS

Si en Física la unidad de análisis o sujeto de estudio son las partículas que conforman el átomo, en la ciencia de datos este rol lo asume el dato. La Data Science Association presenta al dato como un registro tangible o electrónico de información cruda (fáctica o no fáctica) utilizada como base para el razonamiento, la discusión o el cálculo, y que debe ser procesado o analizado para que adquiera significado. En esta era de grandes

1 https://www.datascienceassn.org/

volúmenes de datos, el procesamiento y análisis pueden ser desafiantes. Por ello, la ciencia de datos se apoya en tecnologías informáticas que intervienen en puntos clave del proceso metodológico, facilitando la extracción, estructuración y limpieza de datos en bruto (raw data). La Figura 1 muestra la secuencia metodológica asociada a la ciencia de datos aspectos presentados por Hayashi (1998).

En una primera etapa, el foco está en transformar datos no estructurados (raw data) en datos estructurados (structured data) o "rectangulares", que inician simultáneamente los procesos de preprocesamiento (data preprocessing) y exploración (exploratory data analysis). Estos dos procesos se caracterizan por generar posibles hipótesis que pueden ser abordadas de manera confirmatoria mediante el desarrollo de modelos (modelling) y su correspondiente validación (validation).

Cabe destacar que los primeros dos pasos están relacionados con los conceptos de levantamiento primario y secundario de datos, respectivamente. Los aspectos tecnológicos y de gestión de datos son fundamentales para apoyar estos pasos. Además, las tendencias organizacionales hacia la gobernanza de datos están impulsando la creación de unidades internas destinadas a la recolección y estructuración de datos con el fin de mejorar la toma de decisiones. En el contexto educativo, plataformas como Moodle[2] y Canvas[3] han contribuido significativamente a sistematizar y estructurar datos en las etapas iniciales, facilitando los procesos de implementación en gobernanza.

En la etapa de modelamiento es donde el aprendizaje automático (*machine learning*) ha tomado un rol destacado. El aprendizaje automático es una disciplina que pertenece principalmente al campo de la estadística y se enfoca en aplicar métodos estadísticos para extraer conocimiento de los datos, desarrollando nuevos modelos y algoritmos que permiten comprender, interpretar y analizar la información de manera detallada, sin necesidad de programación explícita (Mitchell, 1997). Dependiendo de si se buscan patrones en los datos (aprendizaje no supervisado) o predecir un valor o etiqueta (aprendizaje supervisado), es posible seleccionar un modelo entre una amplia lista de opciones. La Figura 1 proporciona una lista de modelos de *machine learning*, según el tipo de aprendizaje.

2 https://moodle.org/
3 https://www.instructure.com/

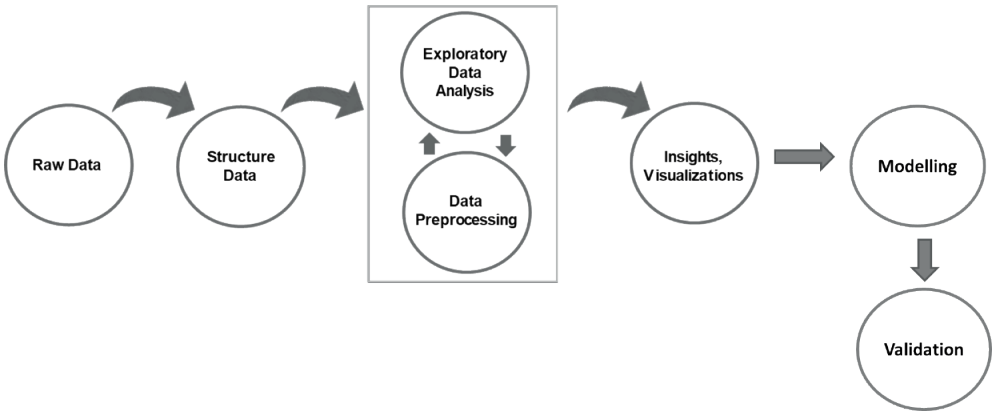

Figura 1: Proceso metodológico en la Ciencia de Datos

Fuente: elaborada por los autores

La validación de modelos es un paso crucial en el desarrollo de cualquier sistema de aprendizaje automático. Esta etapa garantiza que el modelo funcione como se espera y pueda manejar datos no vistos previamente. Sin una validación adecuada, no se puede tener plena confianza en su capacidad de generalización sobre datos nuevos. Además, la validación permite identificar el mejor modelo, los parámetros óptimos y las métricas de precisión más adecuadas para la tarea en cuestión (Mitchell, 1997). En el ámbito del aprendizaje no supervisado, las estrategias de validación se basan principalmente en evaluar la cohesión y separación de los grupos formados a partir de los datos. Sin embargo, también deben considerarse aspectos de interpretabilidad, ya que los grupos (clusters) identificados deben tener sentido en el contexto de aplicación en el que se encuentren (Shutaywi; Kachouie, 2021). En el caso de modelos supervisados, la validación se lleva a cabo mediante enfoques que buscan establecer la diferencia entre los valores predichos y los valores reales. Es aquí donde los términos “conjunto de entrenamiento” y “conjunto de testeo” adquieren relevancia. Las métricas utilizadas cuantifican esta diferencia, ya sea para la predicción de valores (regresión) o de etiquetas (clasificación). Ejemplos bien conocidos en estos casos son el Error Cuadrático Medio para regresión y la Entropía Cruzada Binaria para clasificación (Chicco, 2023).

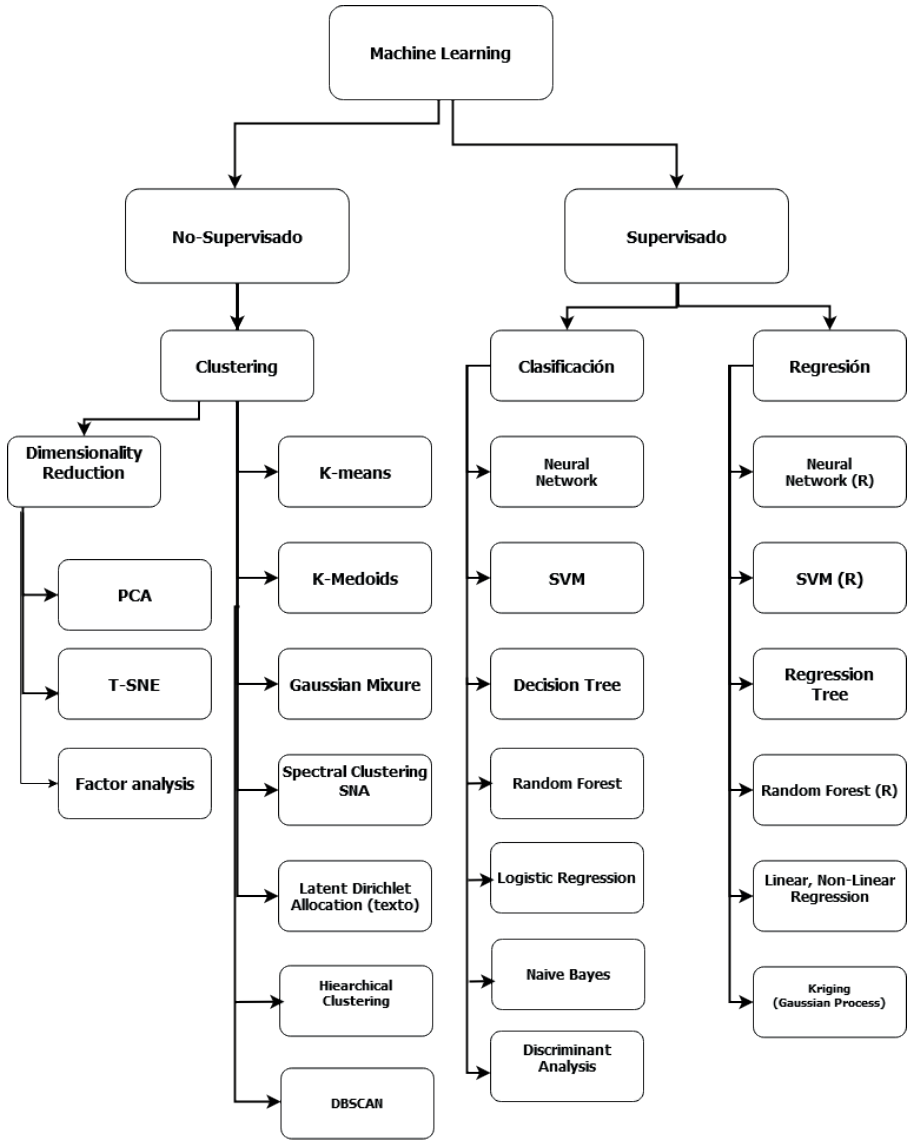

Figura 2: Algunas alternativas de modelos en machine learning

Fuente: los autores (2024).

MODELAMIENTO DE DATOS EN EDUCACIÓN

Según Ersozlu, Taheri y Koch (2024), la integración de modelos de aprendizaje automático en la investigación educativa puede transformar significativamente la enseñanza, el aprendizaje y la evaluación. Esto se logra mediante el uso de dichos modelos en el aprendizaje personalizado, evaluaciones adaptativas y la generación de información valiosa sobre el rendimiento estudiantil, su progreso y los patrones de aprendizaje.

La literatura en investigación educativa evidencia el uso de modelos supervisados y no supervisados para abordar diversas problemáticas. En el caso de los modelos no supervisados, se identifican comúnmente cuatro líneas de análisis:

- Segmentación de estudiantes (Ahmad; Arshad; Sarlan, 2021; Mohd Talib; Abd Majid; Sahran, 2023): se busca agrupar a los estudiantes con características similares (como perfiles de comportamiento o rendimiento), lo que ayuda a identificar grupos con necesidades específicas de intervención educativa.
- Detección de patrones de aprendizaje (Maier; Czibula; Oneț-Marian, 2021; Vermunt; Vermetten, 2004): las técnicas de reducción de dimensionalidad, como PCA o t-SNE, pueden identificar patrones ocultos en los datos de actividades en línea o interacciones en plataformas de aprendizaje. Esos patrones contribuyen a mejorar los diseños curriculares o la interfaz de las plataformas educativas.
- Análisis de redes sociales (Saqr *et al.*, 2018; Han; Ellis, 2021): En un contexto de aprendizaje colaborativo, algoritmos como el Análisis de Grafos permiten entender cómo los estudiantes interactúan entre sí, identificando líderes de opinión, estudiantes aislados o la estructura de colaboración entre grupos.
- Descubrimiento de conceptos en texto (Chatterjee *et al.*, 2021; Chang *et al.*, 2021): Algoritmos como Latent Dirichlet Allocation (LDA) facilitan la identificación de temas subyacentes en grandes volúmenes de texto, como en respuestas a encuestas abiertas, detectando preocupaciones o intereses recurrentes entre estudiantes o profesores.

En cuanto a los modelos supervisados, la literatura destaca su aplicación en tres áreas educativas principales:

- Predicción del rendimiento estudiantil (Guanin-Fajardo *et al.*, 2024; Mohd Talib *et al.*, 2023): usando datos como calificaciones anteriores, hábitos de estudio y demografía, los modelos de regresión o clasificación (como la Regresión Lineal, Árboles de Decisión o Redes Neuronales) pueden predecir el rendimiento de los estudiantes y ayudar a identificar aquellos en riesgo de fracaso escolar.
- Detección temprana de deserción escolar (Albreiki *et al.*, 2021; Hassan *et al.*, 2024): Algoritmos como Support Vector Machines (SVM) pueden predecir la probabilidad de deserción escolar utilizando variables como asistencia, participación en actividades y desempeño en evaluaciones.
- Personalización del aprendizaje (Salinas-Chipana *et al.*, 2023; Shawky y Badawi, 2019; Ross *et al.*, 2013; Yekollu *et al.*, 2024): los modelos de clasificación supervisada permiten identificar estilos de aprendizaje de los estudiantes y recomendar recursos o intervenciones personalizadas, mejorando la eficiencia del aprendizaje.

IMPLICACIONES DE LA CIENCIA DE DATOS EN EDUCACIÓN

Actualmente, se vive una revolución de los datos, y la investigación en educación no debe quedar al margen. En este contexto, los métodos de aprendizaje automático se utilizan principalmente para predecir el rendimiento estudiantil, analizar las preferencias de aprendizaje y evaluar la efectividad de la enseñanza. Esto no solo facilita a los educadores diseñar experiencias de aprendizaje más personalizadas y efectivas, sino que también permite a estadísticos e investigadores obtener resultados precisos de los datos educativos, maximizando el potencial de los modelos de datos para mejorar la toma de decisiones basada en evidencia (Hilbert et al., 2021). Ese enfoque proporciona un soporte fundamental a los stakeholders del ámbito educativo, especialmente a quienes participan en el diseño de políticas y estrategias focalizadas en educación. Sin duda, la aplicación de la ciencia de datos en la educación está en pleno crecimiento y tiene el potencial de transformar tanto la enseñanza como el aprendizaje.

Finalmente, queremos indicar que ya han aparecido innovadoras aplicaciones que integran inteligencia artificial y Python, capaz de leer artículos en formato PDF desde una carpeta en Google Drive para generar revisiones sistemáticas. Este proceso, que ha pasado por numerosas iteraciones de mejora, enfatiza la importancia de definir buenos prompts para maximizar la efectividad del modelo. Utilizando versiones gratuitas de herramientas de IA, se ha desarrollado una solución que cumple con todos los requisitos y características necesarias para una revisión sistemática robusta. Este avance aborda un desafío significativo en la investigación en educación, donde en 2010 se realizaban apenas 60 revisiones anuales, frente a las 2000 actuales. Con esta aplicación, la inteligencia artificial nos permite redirigir nuestras fuerzas investigativas hacia la generación de nuevo conocimiento, evitando el desvío de atención que ha supuesto centrarse excesivamente en las revisiones sistemáticas.

SECCIÓN III - IA EN LA ACTIVIDAD ACADÉMICA CIENTÍFICA E INVESTIGATIVA, RIESGOS ÉTICOS DE SU USO

La frase ***la ética no pasa de moda*** puede incomodar a más de un investigador o profesional en la actualidad, ya sea por las complejas implicancias que esta tiene o bien por lo difícil de mantenerse ecuánime y leal a sí mismos y a los principios que subyacen en ella. La amalgama “ética, investigación, datos y escritura” se encuentran en la bisectriz en donde los profesionales conjugan con algunas coordenadas que agregan tensión al ejercicio, como el tiempo de dedicación a la escritura, el tiempo que toma la búsqueda de datos, la búsqueda de datos en sí, el tiempo de dedicación al procesamiento de la información, entre otros temas contextuales que se encuentran en el marco de este proceso. La gran mayoría de los procesos que requieren los profesionales e investigadores de hoy se encuentran asistidos por la Inteligencia Artificial (IA), agregando valor a los factores

temporales, de diseño, de síntesis o de presentación, entre otros, todos ellos tensionando el proceso investigativo.

Entre las interrogantes que surgen asociadas a los límites visibles del uso de la IA en el campo de la investigación académica, el atractivo que esta representa en términos de optimización del tiempo, la diversificación de herramientas, amplitud de recopilación y procesamiento de ideas, con énfasis en la simplificación de múltiples tareas, a simple vista parece más atractivo que riesgoso. Este escenario ha impregnado de entusiasmo a quienes la utilizan, alcanzando altas y amplias expectativas de desarrollo en esferas organizacionales, aunque a la misma velocidad que se desarrolla, una parte importante de los seres humanos no imagina ni dimensiona los alcances ni la velocidad en que se están produciendo estos cambios. En otras palabras, la selección de personas que hace uso de la IA también está inmersa en el cúmulo de factores contextuales que evidencia desigualdad, cabe preguntarse si este hecho no es más que una muestra clara del tipo consecuencia del vertiginoso avance tecnológico (MINISTERIO DE CIENCIA Y TECNOLOGÍA, 2024) o es otro desafío ético frente a un acelerado analfabetismo digital de los grupos más desfavorecidos.

A este paso, la diversificación del uso de la IA en la academia ocupa espacios importantes e interesantes de observar y acompañar. La formación profesional exige considerar capacidades reflexivas de alto nivel, tanto crítica como propositiva, donde incorporan este tipo de herramientas en sus procesos de formación, investigación junto a sus procesos académicos. Actualmente, estudiantes universitarios y sus académicos se han desafiado en esta área, sin embargo, los límites difusos de su uso ponen en riesgo el prestigio de ambos actores, incluyendo el de las instituciones formadoras (Antaki, 2000) El atractivo de las herramientas IA atrapan ante sus cualidades, cuyo campo de funcionamiento queda al arbitrio de códigos de ética formulados por cada institucionalidad, donde se instalan comités de ética que sesionan para dar garantías de humanidad en todo su sentido, reglamentos o protocolos de integridad que las Instituciones de Educación Superior están requiriendo, ante un espacio que ha ocupado la IA con todas sus herramientas. Estos protocolos no son más que la expresión necesaria de prevención de múltiples riesgos que permanecen aún latentes como el plagio, la fidelización de datos y sus fuentes, la anonimización en pruebas, la adulteración de autorías o autorías fantasmas (Salazar Raymond; Icaza Guevara; Machado, 2018) u otras fórmulas cuestionadas cuyos hallazgos pueden comprometer la verdad científica.

Es preciso tener en cuenta que en la actualidad existe una gran crisis de creencias mientras que se entrega un gran respaldo a la razón, con resabios del siglo anterior, entendiendo por esta última los datos que respaldan las creencias, es decir, creeremos en una premisa siempre y cuando esta tenga sus datos de respaldo. Para confirmar la veracidad y el buen uso o uso responsable de la información circulante se amplía el alcance de la ética, proponiéndose una ética de la responsabilidad social y ética aplicada a distintas disciplinas del conocimiento que se conectan a través de un diálogo permanente (García;

González, 2014) En este sentido, se puede pensar que el conocimiento y la creatividad son ámbitos propiamente humanos, nada hace pensar que esto no será alcanzado por la IA, en esos momentos, la comprensión del conocimiento adquirido y la expresión de emociones impredecibles provocadas por situaciones complejas, son acciones humanas irremplazables aún en este momento, aunque todo parece indicar que estas mismas características limitantes de un robot o herramienta IA también serán prontamente superadas (Roubini, 2023).

Será que la tensión se encuentra entre la amenaza que plantea Roubini (2023) de reemplazo avanzado de una masiva mano de obra local por una reducida presencia artificial que automatiza labores y el desarrollo de capacidades que manifiesta Sen (2009). De cualquiera de las dos formas, todo parece indicar que en el campo de la formación profesional y de la investigación en particular seguirán manteniéndose esquemas de riesgo ético que la institucionalidad universitaria en particular y científica, en general deben resguardar. A nivel internacional se han plasmado esfuerzos notables por regular cada vez más el uso de estas herramientas de IA tanto en la academia como en la investigación científica con fines preventivos y de erradicación de malas prácticas que aún permanecen.

En este sentido, la Unión Europea y su consejo, entendiendo la importancia de la investigación y de sus investigadores, propone en su Carta al Investigador (2005) integrar una serie de recomendaciones y exigencias que van de la mano con el sentido ético de esta labor, sugerencias del tipo preventivas en relación a la IA que se han replicado en algunos países de avanzada en la temática como Reino Unido, donde se ha desarrollado un organismo que monitorea la diversidad y el comportamiento temático de algoritmos; en tanto algunos países de América Latina como México, con su Comité Ético de Inteligencia Artificial va a la vanguardia del análisis de la ciencia de datos. Cualquiera sea la forma que estén adoptando los países y universidades, este es un tema que aún presenta grandes desafíos, los sistemas de protección de datos desde los fireware o cortafuegos, encriptaciones, almacenamiento u otras formas, no han sido suficientes a la hora atender los riesgos que tiene la IA en la investigación científica, es necesario se avance en términos de prevención de conflictos éticos (Reyes; Audave, 2022).

Por otro lado, en los últimos cinco años, el aumento de la actividad científica del tipo revisión sistemática en educación ha llamado fuertemente la atención para los riesgos de plagio, utilización perversa de herramientas IA para procesos de investigación científica y escritura; y de autorías fantasmas u honorarias, entre otros problemas éticos que toman un lugar relevante. Junto con ello, la supervigilancia desde los comités de ética principalmente, han tomado la agenda de cambio, reconociendo la resistencia actual debido a los riesgos y discrepancias éticas que presentan las revisiones sistemáticas en estos ámbitos de investigación científica. Esta agenda de cambio no las exime del uso de las herramientas de la IA, sino más bien les insta a los investigadores a explorar y trabajar apoyándose en

las técnicas herramientas IA tomando en cuenta las alertas y riesgos de uso con el fin de eliminar las malas prácticas (Calvo, 2022).

REFERENCIAS

AHMAD, M.; ARSHAD, N. I. B.; SARLAN, A. B. An analysis of students' academic performance using K-means clustering algorithm. In: **International Conference of Reliable Information and Communication Technology**, 2021, Cham. Proceedings... Cham: Springer International Publishing, 2021. p. 309-318.

ALBREIKI, B.; ZAKI, N.; ALASHWAL, H. A systematic literature review of student performance prediction using machine learning techniques. **Education Sciences**, v. 11, n. 9, p. 552, 2021.

ANTACKI, A. **El Manual del Ciudadano Contemporáneo**. Editorial Planeta Mexicana. S. A. C. V., 2000.

CALVO, P. Una ética de la investigación en el marco de las éticas aplicadas. **VERITAS**, n. 52, p. 29-51, ago. 2022.

CHANG, I. C.; YU, T. K.; CHANG, Y. J.; YU, T. Y. Applying text mining, clustering analysis, and latent Dirichlet allocation techniques for topic classification of environmental education journals. **Sustainability**, v. 13, n. 19, p. 10856, 2021.

CHATTERJEE, R.; MUKHERJEE, C.; CHATTERJEE, S.; NATH, B. Latent Dirichlet allocation for topic modeling and intelligent document classification. **Innovations in Data Analytics**, p. 71, 2021.

CHICCO, D. The ABC recommendations for validation of supervised machine learning results in biomedical sciences. **Frontiers in Artificial Intelligence**, v. 6, art. 703737, 2023. DOI: https://doi.org/10.3389/frai.2023.703737.

COMUNIDAD EUROPEA. **Carta Europea del Investigador:** Código de conducta para la contratación de investigadores. Printed in Belgium, 2005.

ERSOZLU, Z.; TAHERI, S.; KOCH, I. A review of machine learning methods used for educational data. **Education and Information Technologies**, p. 1-21, 2024.

GARCÍA, D.; GONZÁLEZ, E. **Ética**. Publicacions de la Universitat Jaume I. Servei de Comunicació i Publicacions, 1. ed., 2014.

GUANIN-FAJARDO, J. H.; GUAÑA-MOYA, J.; CASILLAS, J. Predicting academic success of college students using machine learning techniques. **Data**, v. 9, n. 4, p. 60, 2024.

HAN, F.; ELLIS, R. A. Patterns of student collaborative learning in blended course designs based on their learning orientations: a student approaches to learning perspective. **International Journal of Educational Technology in Higher Education,** v. 18, n. 1, p. 66, 2021.

HASSAN, M. A.; MUSE, A. H.; NADARAJAH, S. Predicting student dropout rates using supervised machine learning: insights from the 2022 National Education Accessibility Survey in Somaliland. **Applied Sciences**, v. 14, n. 17, p. 7593, 2024.

HAYASHI, C. What is data science? Fundamental concepts and a heuristic example. In: International Federation of Classification Societies. Data Science, Classification, and Related Methods: **Proceedings of the Fifth Conference,** Kobe, Japan, 1996. Tokyo: Springer Japan, 1998. p. 40-51.

HILBERT, S. *et al*. **Machine learning for the educational sciences**. 2021. DOI: https://doi.org/10.31234/osf.io/3hnr6.

MAIER, M.-I.; CZIBULA, G.; ONEŢ-MARIAN, Z.-E. Towards using unsupervised learning for comparing traditional and synchronous online learning in assessing students' academic performance. **Mathematics**, v. 9, p. 2870, 2021. DOI: https://doi.org/10.3390/math9222870.

MITCHELL, T. M. **Machine learning**. New York: McGraw-Hill, 1997.

MINISTERIO DE CIENCIAS, TECNOLOGÍA, CONOCIMIENTO E INNOVACIÓN. **Política Nacional de Inteligencia Artificial:** Actualización 2024. Gobierno de Chile.

MOHD TALIB, N. I.; ABD MAJID, N. A.; SAHRAN, S. Identification of student behavioral patterns in higher education using k-means clustering and support vector machine. **Applied Sciences**, v. 13, n. 5, p. 3267, 2023.

REYES, S.; AUDAVE, D. Conductas no éticas en la investigación científica: prevalencia, causas asociadas y estrategias de prevención. Una revisión sistemática. **Revista Innovaciones Educativas**, v. 24, n. Especial, p. 1-18, oct. 2022. DOI: https://doi.org/10.22458/ie.v24iespecial.4312.

ROBINI, N. Megathreats. **Ten Dangerous trends That Imperil our Future and how to survive them.** Little, Brown and Company Hachette Book Group, 1st ebook ed., 2022.

ROSS, M.; GRAVES, C. A.; CAMPBELL, J. W.; KIM, J. H. Using support vector machines to classify student attentiveness for the development of personalized learning systems. In: International Conference on Machine Learning and Applications, 2013, Washington, DC. **Proceedings...** Washington, DC: IEEE, 2013. v. 1, p. 325-328.

SALAZAR RAYMOND, M. B.; ICAZA GUEVARA, M. F.; MACHADO, O. A. La importancia de la ética en la investigación. **Universidad y Sociedad**, v. 10, n. 1, p. 305-311, 2018. Recuperado de: http://rus.ucf.edu.cu/index.php/rus.

SALINAS-CHIPANA, J.; OBREGON-PALOMINO, L.; IPARRAGUIRRE-VILLANUEVA, O.; CABANILLAS-CARBONELL, M. Machine learning models for predicting student dropout—a review. In: International Congress on Information and Communication Technology, 2023, Singapore. **Proceedings...** Singapore: Springer Nature Singapore, 2023. p. 1003-1014.

SAPOLSKY, R. **Decidido**: una Ciencia de la Vida Sin libre Albedrío. 1st ed. Capitán Swing, 2024.

SAQR, M.; FORS, U.; TEDRE, M.; NOURI, J. How social network analysis can be used to monitor online collaborative learning and guide an informed intervention. **PloS One**, v. 13, n. 3, e0194777, 2018.

SHAWKY, D.; BADAWI, A. Towards a personalized learning experience using reinforcement learning. In: Machine Learning Paradigms: **Theory and Application**, p. 169-187, 2019.

SHUTAYWI, M.; KACHOUIE, N. N. Silhouette analysis for performance evaluation in machine learning with applications to clustering. **Entropy**, v. 23, n. 6, p. 759, 2021. DOI: https://doi.org/10.3390/e23060759.

VERMUNT, J. D.; VERMETTEN, Y. J. Patterns in student learning: relationships between learning strategies, conceptions of learning, and learning orientations. **Educational Psychology Review**, v. 16, p. 359–384, 2004. DOI: https://doi.org/10.1007/s10648-004-0005-y.

YEKOLLU, R. K.; BHIMRAJ GHUUGE, T.; SUNIL BIRADAR, S.; HALDIKAR, S. V.; FAROOK MOHIDEEN ABDUL KADER, O. AI-driven personalized learning paths: enhancing education through adaptive systems. In: International Conference on Smart Data Intelligence, 2024, Singapore. **Proceedings...** Singapore: Springer Nature Singapore, 2024. p. 507-517.