USO DE INTELIGENCIA ARTIFICIAL PARA LA CONFECCIÓN DE ÍTEMS PARA EVALUACIONES DIAGNÓSTICAS: PERCEPCIONES DOCENTES.

Camila Villanueva Morales, Pontificia Universidad Católica de Valparaíso, cami.villa@gmail.com
Pablo Cáceres, Pontificia Universidad Católica de Valparaíso, pablo.caceres@pucv.cl
Fernanda Alarcón, Pontificia Universidad Católica de Valparaíso, fernanda.alarcon@gmail.com
Martín Barra, Pontificia Universidad Católica de Valparaíso, martin.barra@sansano.usm.cl
Nicolás Irribarra, Universidad Técnica Federico Santa María, nicolas.irribarra.zu@gmail.com
Maximiliano Abarca, Pontificia Universidad Católica de Valparaíso, maximiliano.abarcam@sansano.usm.cl
Javier Aguayo, Pontificia Universidad Católica de Valparaíso, javier.aguayo@sansano.usm.cl

RESUMEN

Este estudio analiza las percepciones de docentes de matemática sobre el uso de inteligencia artificial (IA) en la confección de ítems para evaluaciones diagnósticas dirigidas a estudiantes de primer año de ingeniería. A través de la realización de pruebas de usuario de una plataforma se identificaron y analizaron conceptos clave, revelando tanto las ventajas como las limitaciones de esta tecnología. Los resultados muestran que, aunque la IA ofrece beneficios significativos en términos de eficiencia y automatización, persisten preocupaciones sobre la calidad de los ítems generados, especialmente en relación con errores matemáticos y la adecuación al nivel académico. Se sugiere la implementación de procesos de validación humana para garantizar la precisión y pertinencia de las evaluaciones.

PALABRAS CLAVES: inteligencia artificial, evaluación diagnóstica, calidad de ítems, educación superior.

INTRODUCCIÓN

Diversos factores han generado que el acceso a la Educación Superior se haya masificado y, con ello, aumentado la diversidad de realidades socioeducativas que coexisten en un mismo nivel académico. En este ámbito, son las carreras del área de Ciencias, Tecnología, Ingeniería y Matemáticas (STEM, por su definición en inglés), las que poseen aún más brechas debido a que, además del desempeño, se requiere poseer un amplio manejo de elementos disciplinares previos (Donoso et al., 2020). El fortalecimiento de estas habilidades es altamente complejo debido al elevado desarrollo del pensamiento abstracto que se requiere para obtener buenos resultados. Para mejorar las tasas de retención y aprobación en los primeros años las instituciones de Educación Superior han desarrollado estrategias para apoyar a los alumnos de primer año en su proceso de aprendizaje, siendo la implementación de cursos previos (generalmente de *nivelación*), la estrategia más utilizada (Miranda-Molina, 2022).

El término *nivelación* ha sido utilizado no solo para definir diversas políticas educativas o experiencias institucionales, sino que también acompaña la definición de propósitos que poseen programas y planes remediales en ciencias básicas siendo su intención lograr aprendizajes (conocimientos, habilidades y competencias) previos que son elementales como capital inicial antes del inicio del proceso de formación (Miranda-Molina, 2022).

Considerando lo anterior y en el marco del fortalecimiento de la eficiencia interna de las instituciones de educación superior, la Comisión Nacional de Acreditación (CNA, 2023), exige que *"La institución cuenta con procesos de enseñanza y aprendizaje que las condiciones necesarias para el logro efectivo de los aprendizajes y perfil de egreso por parte de los estudiantes"* , buscando como criterio de cumplimiento de primer nivel que las instituciones desarrollen *"procesos de enseñanza-aprendizaje que consideran el diagnóstico de brecha de sus estudiantes con el perfil de ingreso y genera mecanismos de apoyo a su progresión".* Formalizando con esto la exigencia de procesos diagnósticos en los primeros años.

En general, los modelos utilizados comúnmente para diagnosticar son unidimensionales (TIC, IRT) y apuntan a ordenar al grupo de estudiantes en torno a su desempeño, atribuyendo una nota que no indica, específicamente, qué conocimientos están descendidos o logrados.

En este marco, los Modelos de Diagnóstico Cognitivo ofrecen una alternativa interesante, ya que poseen multidimensionalidad en la información que arrojan, mostrando cuáles atributos cognitivos se deben trabajar, granulando los resultados que son presentados en perfiles cognitivos, información que es muy útil para los profesores en el diseño de sus cursos y para los estudiantes, quienes tienen la posibilidad de recibir retroalimentación. Las limitaciones de este modelo están, por un lado, en las capacidades técnicas que debe tener quien realiza el análisis, así como también en las grandes cantidades de muestras que se deben recopilar para hacer la validación. Todo esto implica que los tiempos y recursos requeridos no son de acceso universal.

Assessment System es una plataforma en desarrollo que propone automatizar estos procesos para que sean un real aporte en el proceso educativo de estudiantes de primer año de carreras STEM. La plataforma se sustenta en 2 grandes áreas para poder realizar esta automatización: cruce de información entre habilidades a medir y habilidades logradas, y por otra parte, el uso de inteligencia artificial para la creación de bancos de preguntas para generar evaluaciones diagnósticas.

La gran cantidad de estudiantes, la diversidad de apoyos que se necesitan, y la rapidéz con la que se implementan cambios, requiere de medidas tecnológicas que permitan avanzar a la misma velocidad, siendo la inteligencia artificial (IA) una herramienta que ha transformado la manera en que se desarrollan y administran las evaluaciones en el ámbito educativo.

El uso de IA para la creación de ítems en evaluaciones diagnósticas promete aumentar la eficiencia y reducir la carga de trabajo de los docentes, permitiéndoles centrarse en aspectos más estratégicos de la enseñanza (Zawacki-Richter et al., 2019). Sin embargo, la implementación de estas tecnologías plantea desafíos significativos, especialmente en lo que respecta a la calidad y precisión de los ítems generados (Vera, 2023).

Este estudio se origina en el marco de dicha plataforma, con el objetivo de identificar las percepciones docentes en torno a la creación automatizada de ítems y el levantamiento de áreas de mejora para potenciar el uso de esta tecnología en el diagnóstico para estudiantes que ingresan a la educación superior.

DESARROLLO

Descripción

La actividad principal consistió en una Prueba de Usuario (PU) del Mínimo Producto Viable de una plataforma diseñada para la confección de ítems de evaluación diagnóstica mediante inteligencia artificial. La prueba se realizó en dos jornadas presenciales,

organizadas en la Sala Creativa COIL de la Facultad de Filosofía de la Pontificia Universidad Católica de Valparaíso (PUCV). La primera jornada se llevó a cabo el 30 de julio de 2024, con la participación de profesores de educación media, y la segunda, el 31 de julio de 2024, con profesores de educación superior. Esta división se realizó para poder obtener la información lo más detalladamente posible de ambos perfiles.

El cronograma de cada jornada incluyó una introducción al proyecto y la plataforma, seguida de la confección de una prueba diagnóstica donde los docentes utilizaron la plataforma con el objetivo de crear evaluaciones. Posteriormente, se entregó una maqueta de resultados que mostraba la retroalimentación ideal que la plataforma podía ofrecer. Tras un breve receso, se realizó un grupo focal (focus group) para recoger las percepciones de los docentes sobre la plataforma. Fue en este punto donde los docentes pudieron entregar de forma abierta su opinión sobre el funcionamiento y sus aspectos principales, como el uso de IA para la generación de bancos de preguntas que da origen a las evaluaciones diagnósticas. Posteriormente se completó con una encuesta para capturar cualquier información adicional.

La actividad se registró mediante observación, audio, fotografía y video, lo que permitió un análisis posterior detallado de las interacciones y comentarios de los participantes mediante transcripciones de los diferentes formatos utilizados.

Figura Nº 1. Docentes interactuando con la plataforma.

En la Fig. 1 se puede apreciar cómo fue el proceso de la prueba de usuario. Se propició el trabajo colaborativo, así como la reflexión personal de cada docente para después discutir y elaborar en grupo un diagnóstico del estado actual de la plataforma.

Participantes

Los participantes de este estudio fueron docentes de matemática que imparten clases en diversas instituciones de educación media - en el último año - y en el primer año de educación superior en carreras de ingeniería. La muestra estuvo conformada por 24 docentes, entre los participantes, un 42% realiza clases en educación superior y un 58% en educación media. El grupo fue seleccionado por su experiencia en la enseñanza de matemáticas en contextos de ingreso a la universidad, lo que los convierte en usuarios clave de herramientas de evaluación diagnóstica.

Etapas
Las percepciones de los docentes fueron extraídas del focus group realizado posterior a la Prueba de Usuario de la plataforma Assessment System, en donde se realizó un testeo preliminar de su funcionamiento, permitiendo su revisión mediante la interacción controlada en tiempo real, de forma de identificar áreas de mejora, errores y potencialidades. Las etapas de este proceso se encuentran diagramadas en la Fig. 2.

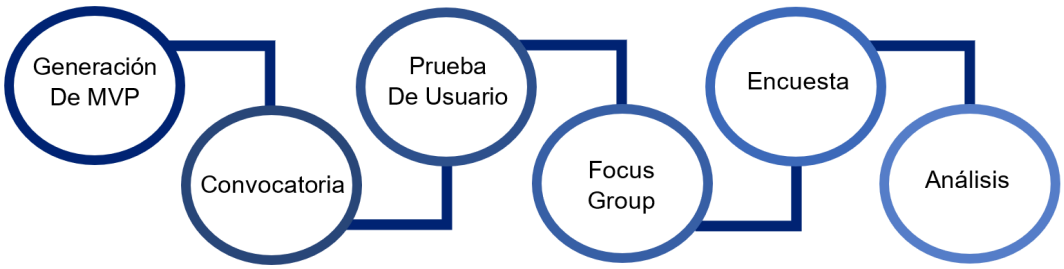

Figura Nº 2. Etapas de la actividad. Fuente: Elaboración propia

1. Generación de Producto Mínimo Viable (MVP):

Para la realización de este estudio, se confeccionó una versión preliminar de la plataforma con las características y requerimientos esenciales para la confección de evaluaciones diagnósticas. Para esto, se jerarquizaron las las funciones para lograr obtener información de calidad en cuanto a usabilidad y trasfondo teórico. El MVP consistió en:

- Visualización de Dashboard perfil docente.
- Función “Crear Evaluación”.
- Selección características de la evaluación: atributos/habilidades, cantidad de preguntas, tiempo, curso.
- Visualización de atributos/habilidades a evaluar.
- Generación de evaluación preliminar.
- Modificación de preguntas.
- Generación de evaluación final.

2. Convocatoria:

En este caso, la plataforma está diseñada para diagnosticar las habilidades de estudiantes de primer año de carreras STEM, por lo que se define que el grupo objetivo estará compuesto por docentes que imparten clases en dicho nivel en educación superior, así como también docentes que dicten clases en último año de educación media. De esta forma se buscó conformar un grupo con alta experiencia y preparación en el tema a tratar.

La selección de docentes comenzó un mes antes de la fecha estipulada para la prueba, identificando instituciones con trayectoria y experiencia en el área. El grupo fue convocado por medios digitales y presenciales, sosteniendo reuniones con las instituciones, así como también recibiendo recomendaciones de posibles participantes. Finalmente, el grupo fue convocado por distintos medios entre los cuales destacan reuniones y charlas, envío de correspondencia y contacto telefónico.

3. Prueba de Usuario (Testeo Plataforma):

En ambas jornadas, luego de la introducción al proyecto, los docentes confeccionaron una evaluación diagnóstica, interactuaron con la plataforma y conocieron su funcionamiento a nivel tecnológico. Es en este punto que se les muestra el resultado del uso de la IA en la

creación de los bancos de preguntas, su relación con los atributos a medir y su presentación en la evaluación final.

3. Grupo Focal (Focus Group):

Luego de interactuar con la plataforma, los docentes compartieron sus apreciaciones con el equipo. Esta actividad está dirigida a obtener la mayor cantidad de información posible en un tiempo acotado, en este caso correspondiente a media hora. Para eso, se diseñaron preguntas agrupadas para abarcar las temáticas que se consideraron relevantes para el desarrollo de la plataforma: Experiencia general del usuario, usabilidad, funcionalidad, expectativas y comparaciones, y finalmente, uso y comportamiento.

Durante la ejecución de la conversación guiada, los participantes comentaron sobre temas relacionados con educación, diseño de ítem, rigurosidad matemática, metodologías de evaluación actuales y uso de Inteligencia Artificial en la educación, dándose un espacio de conversación guiada pero en la que los y las asistentes podían elevar temáticas asociadas de forma libre.

4. Encuesta:

El Focus Group fue complementado con una encuesta abierta para maximizar la cantidad de información recolectada. La encuesta fue enviada a los docentes posterior a su participación en cada jornada. De esta forma, quedó registrada la opinión de la mayoría de los docentes individualmente.

5. Análisis:

A partir de las opiniones vertidas en la actividad y la información recopilada en las encuestas, se realiza el proceso de organizar, analizar e interpretar los testimonios. Se utilizó la estrategia de codificación abierta, axial y selectiva mediante el uso del software Atlas.ti para el análisis de ambos focus groups y un análisis descriptivo para la encuesta posterior a la actividad. Se generaron 3 grupos de códigos en base a las opiniones de los 24 docentes. Los grupos y códigos utilizados se diagraman la Fig. 3:

Figura Nº 3: Codificación Focus Group, elaboración propia.

Comentarios como *“Es que yo, con colegas, ya cuando tenemos nuestra pregunta le pedimos a la IA que genere alguna versión, que tiene mejores contextos de lo que uno podría crear.”*, *“Quedan hermosas las preguntas”,* se identifica que algunos de los docentes utilizan IA para generar su propia base de datos de preguntas como iniciativa propia paralela a la metodología que utilice la institución, destacando la capacidad de mejora constante de la herramienta y que, a pesar de no contar con la doble validación esperada, puede utilizarse exitosamente para mejorar los procesos.
En cuanto a la generación del banco de preguntas nos encontramos con observaciones como: *“Que alguien cree la pregunta y después le pidan a la IA generar otros contextos y otras cosas más ricas”;* centrados más en ser un complemento de preguntas ya generadas. Por otro lado, se identifica que la precisión en la formulación de las preguntas que se utilizan para la generación automatizada del banco de ítems, es decir, de entrenamiento, es un punto de mejora sustancial para los resultados a obtener.
Para todos los docentes es de gran importancia poder trabajar con gráficas para poder medir la capacidad de análisis alcanzada por los estudiantes independiente del nivel en que se encuentren. A su vez, la mayoría de los profesores cree que este sería el mayor desafío de la IA, pues tendría que alcanzar un avanzado nivel de entrenamiento para poder llegar a trabajar con el cambio entre registros y dar con gráficas correctas.

Otro punto a evaluar es el uso del lenguaje y nomenclatura matemática. Existen sutilezas en cómo se nombran algunos conceptos matemáticos en diferentes países. Esta información se puede inferir de comentarios como *“matemáticamente es muy curioso, pero el término debe ser el mismo pero no. A veces me ha costado entender a los estudiantes, porque es Uruguay y ellos le dicen a lo que nos decimos A, ellos le dicen B, entonces en esas cosas hay que tener cuidado”*; *“o quizás en el mismo sentido pueden tener errores [...] pero también errores de que la gente tiene otro formato más común*”; “*Hay error de edición del LATEX.*”, este tipo de opiniones nos invitan a reflexionar en torno al lenguaje matemático utilizado, que puede ser diverso incluso entre instituciones del mismo país. La IA al ser una herramienta que no conoce límites geográficos, puede tomar data de cualquier lugar, sin considerar estas diferencias de forma de expresar conceptos universales. Estas sutilezas deberán ser emparejadas en el proceso de entrenamiento de la IA para mantener la rigurosidad matemática en el resultado y que exista coherencia entre, al menos, las preguntas generadas de una misma plataforma.

En cuanto a la encuesta, un total de 14 participantes la completó, siendo un 57% de las respuestas de docentes de educación superior y un 43% de docentes de educación media. En esta área, se destaca la valorización de la herramienta en torno a la retroalimentación, se aprecia que la mayoría de los docentes (92,8%) consideran que esta información es valiosa tanto para ellos como para el estudiante que la recibe. Inferimos que la forma de mostrar la información es útil y fácil de entender para los implicados y realmente puede mejorar el proceso educativo. Por otro lado, se refuerza lo identificado en el Focus Group en torno a la calidad de los ítems, el uso de inteligencia artificial para procesos de evaluación y el funcionamiento de la plataforma.

RESULTADOS

1. Calidad de los Ítems Generados:

Una preocupación central que surgió de las entrevistas y grupos focales con los docentes fue la calidad de los ítems generados por la IA, especialmente en la precisión y formulación. Numerosos estudios han resaltado los desafíos asociados con la automatización de la creación de ítems, particularmente en disciplinas que requieren una alta precisión, como las matemáticas e ingeniería (Mitrović et al., 2017; Vera et al., 2023). En este caso, los docentes señalaron errores recurrentes en los ítems generados, incluyendo inexactitudes

matemáticas y ambigüedades en los enunciados, lo que compromete la validez de las evaluaciones. Esto podría llevar al error del estudiante por forma y no por fondo, o sea, comprender el concepto por el que se le pregunta, pero no comprender la forma en que se le pregunta.

Además, se observó que algunos ítems no estaban alineados con el nivel académico ni con los objetivos de aprendizaje esperado para estudiantes de primer año de ingeniería, lo que podría llevar a evaluaciones ineficaces.

2. Percepción de la IA en la Educación:

La percepción general de la IA en la educación es mixta entre los docentes. Por un lado, reconocen los beneficios de la automatización en términos de ahorro de tiempo y recursos. Por otro lado, persisten preocupaciones sobre la dependencia excesiva de la tecnología, especialmente sin una supervisión adecuada, lo que podría comprometer la calidad educativa (Luckin & Holmes, 2016; Tomalá et al., 2023).

Los docentes coinciden en que, aunque la IA tiene un gran potencial para mejorar la educación, su implementación exitosa depende en gran medida de cómo se combine con la experiencia y el juicio crítico de los educadores. La IA por sí sola no es capaz de reproducir los procesos complejos de toma de decisiones en el diseño de herramientas para evaluar si los estudiantes manejan y comprenden los conceptos que se están trabajando: la importancia de un correcto entrenamiento de la IA. Es fundamental que se manejen los parámetros claros desde un comienzo para que se genere información con validez técnica. Esta base sólida sólo puede ser desarrollada con la experiencia y criterio humano, formado a través la experiencia profesional a través de los años (Vera, 2023). Los y las participantes entregaron información valiosa respecto de los lineamientos a seguir en cuanto a este entrenamiento, mencionando referentes actuales y compartiendo su experiencia de metodologías que han tenido buenos resultados. Por ejemplo, es fundamental que la IA pueda discernir entre distractores de relevancia para la pregunta, pues de otra forma se transforman en opciones ilógicas y fácilmente descartables. Se identifica que este proceso de validación posee numerosos pasos y requiere de mucho tiempo, entrampando la eficiencia del mismo.

3. Funcionalidad y Uso de la Plataforma:

A pesar de los desafíos mencionados, los docentes valoraron positivamente la funcionalidad y la eficiencia de la plataforma. La literatura apoya la idea de que la automatización puede reducir significativamente la carga de trabajo docente, permitiendo una mayor focalización en el desarrollo pedagógico y en la personalización del aprendizaje (Tomalá et al., 2023). El grupo convocado estuvo de acuerdo con que la automatización de los procesos agilizaría la labor de construcción de herramientas evaluativas, mejorando los procesos de diagnóstico y, por consiguiente, los procesos de enseñanza.

Los participantes destacaron que la plataforma es intuitiva y fácil de usar, lo que facilita su integración en las actividades cotidianas de enseñanza. Sin embargo, subrayan la necesidad de una validación humana para corregir los errores introducidos por la IA y asegurar que los ítems sean adecuados y precisos.

CONCLUSIONES

El uso de inteligencia artificial en la confección de ítems de evaluaciones diagnósticas para estudiantes de primer año de ingeniería ofrece ventajas importantes, como la automatización y la eficiencia. No obstante, es esencial que los ítems generados sean

rigurosamente revisados y validados por expertos para garantizar su calidad y precisión para que puedan incidir positivamente en el proceso de inducción a la educación superior.

La IA ofrece un amplio campo de posibilidades para enriquecer el proceso de evaluación diagnóstica, dado que permitiría generación de evaluaciones personalizadas según grupo de estudiantes, temáticas y planes de estudio, así como también formatos de un mismo ítem gracias a la elevada capacidad de esta tecnología de compilar y procesar información en tiempo real. Este potencial de personalización y adaptación ofrece una ventaja significativa al proceso educativo. Sin embargo, para garantizar la efectividad de estas evaluaciones es necesario encontrar el punto de encuentro con la requerida rigurosidad matemática, aplicando técnicas y métodos precisos para asegurar que el nivel de dificultad del test se ajuste adecuadamente a los lenguajes locales, a los niveles de dificultad y a los planes y programas previos y venideros.

Por ahora y debido a que la selección de atributos/habilidades depende del usuario, establecer dichos contenidos depende de la pericia del profesor que esté utilizando la herramienta y su destreza al evaluar la pertinencia de los temas a diagnosticar. Estas habilidades son esenciales para generar evaluaciones y finalmente para todo el proceso educativo y definen, lo que la IA entregará como test final. En este sentido, hay que hacer la diferencia entre las habilidades a evaluar y su interacción con los planes de estudios de la institución que aplique la evaluación, existiendo diferencias entre el nivel de dificultad de las asignatura de ciencias básicas como matemática en los primeros años de educación superior, a pesar de tratarse de una misma carrera.

La percepción de los docentes sugiere que la IA debe ser vista como una herramienta complementaria, no como un sustituto completo de la intervención humana en la creación de evaluaciones. El rol docente es crucial para la comprensión del contexto, y difícilmente podrá ser superado por la IA, la cual viene a potenciar y agilizar proceso y no a reemplazarlos. Hay un enorme potencial en la combinación de tecnología avanzada con la experiencia humana en la creación y ajuste de evaluaciones diagnósticas, que propende a la implementación de procesos educativos más dinámicos y adaptativos, beneficiando así a todos los involucrados. A medida que la tecnología avanza, su éxito dependerá de cómo se integre con la experiencia y el juicio crítico de los educadores.

Es importante comunicar estas nuevas metodologías de uso de IA en los procesos tradicionales para que así los docentes las integren a su propia agenda y agilicen los procesos factibles de automatizar, para dar más tiempo al desarrollo de habilidades no transferibles a este tipo de tecnología, como las relaciones personales en el aprendizaje, identificación de capacidades en los estudiantes para la motivación del grupo, fomento del pensamiento crítico, adaptación de estrategias didácticas, entre otras.

AGRADECIMIENTOS

Agradecemos sinceramente a los docentes que participaron en el grupo de discusión y en las pruebas de usuario, quienes con su tiempo y experiencia han contribuido de manera invaluable a este estudio. Su disposición para compartir abiertamente sus percepciones ha sido fundamental para el desarrollo de este análisis.

También queremos agradecer al equipo de apoyo y gestión de la Dirección de Innovación de la PUCV por facilitar este proceso y poner todas las herramientas disponibles para la correcta ejecución de la prueba.

REFERENCIAS

- Comisión Nacional de Acreditación. (2023). Criterios y estándares de acreditación para universidades. Comisión Nacional de Acreditación Chile. https://www.cnachile.cl/SiteAssets/Paginas/consulta_criterios_y_estandares/universidades.pdf

- Donoso Osorio, E., Valdés Morales, R. A., Cisternas Núñez, P., & Cáceres Serrano, P. (2020). Enseñanza de la resolución de problemas matemáticos: un análisis de correspondencias múltiples. Diálogos Sobre Educación, 0(21). https://doi.org/10.32870/dse.v0i21.629

- Luckin, R., & Holmes, W. (2016, February). (PDF) Intelligence Unleashed: An argument for AI in Education. ResearchGate. https://www.researchgate.net/publication/299561597_Intelligence_Unleashed_An_argument_for_AI_in_Education

- Miranda-Molina, R. (2022). Brechas y desniveles: el problema representado en las iniciativas de “nivelación” en la Educación Superior Latinoamericana. Revista de Estudios Y Experiencias En Educación, 21(46), 292–311. https://doi.org/10.21703/0718-5162.v21.n46.2022.016

- Mitrovic, A., Martin, B., & Suraweera, P. (2007). Intelligent Tutors for All: The Constraint-Based Approach. IEEE Intelligent Systems, 22(4), 38–45. https://doi.org/10.1109/mis.2007.74

- Tomalá, M., Mascaró, E., Carrasco, C., & Aroni, E. (2023). Incidencias de la inteligencia artificial en la educación. RECIMUNDO, 7(2), 238–251. https://doi.org/10.26820/recimundo/7.(2).jun.2023.238-251

- Vera, F. (2023). Integración de la Inteligencia Artificial en la Educación superior: Desafíos y oportunidades. Transformar, 4(1), 17–34. https://www.revistatransformar.cl/index.php/transformar/article/view/84

- Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education – where are the educators? International Journal of Educational Technology in Higher Education, 16(1). Springeropen. https://doi.org/10.1186/s41239-019-0171-0