Desarrollo de un Modelo Predictivo para la Identificación de Factores que Influyen en las Competencias Digitales de Estudiantes Mediante Analítica de Datos en el Programa de Ingeniería de Sistemas de la Unipacífico (2015-2025)

Eder Joaquin Gamboa Andrade

Asesor

Mireya García García

Universidad Nacional Abierta y a Distancia (UNAD)

Escuela de Ciencias Básicas Tecnología e Ingeniería ECBTI

Maestría en Ciencia de Datos y Analítica

2026

Dedicatoria

Dedico este trabajo a mi amada familia, pilar fundamental en cada uno de los pasos que he dado a lo largo de mi proceso formativo. A ellos, por su apoyo incondicional, por su infinita paciencia y por la confianza depositada en mí incluso en los momentos más difíciles.

Este logro no solo representa el fruto de mi esfuerzo, dedicación y perseverancia, sino también el reflejo del amor, los valores y las enseñanzas que he recibido de mi familia. Cada página de este trabajo lleva implícito su acompañamiento, motivación y fe en mis capacidades.

A ellos les entrego con profunda gratitud este esfuerzo académico, con la esperanza de que se sientan tan orgullosos como yo de lo que juntos hemos logrado.

Agradecimiento

Expreso mi más sincero agradecimiento a la Universidad Nacional Abierta y a Distancia (UNAD) por brindar el entorno académico, metodológico y humano que hizo posible el desarrollo de esta investigación. Su compromiso con la formación investigativa y el fortalecimiento del conocimiento científico ha sido fundamental para la culminación de este proceso académico.

De manera especial, agradezco a la asesora de esta investigación, Mireya García García, por su orientación académica, sus valiosos aportes metodológicos y su acompañamiento durante el desarrollo de este trabajo. Sus observaciones, recomendaciones y guía investigativa fueron fundamentales para fortalecer la rigurosidad científica del estudio.

Asimismo, expreso mi gratitud al líder del Semillero de Investigación 3, Javier Medina Cruz, por su apoyo, asesoría y acompañamiento durante el proceso investigativo, así como por sus aportes en la revisión y fortalecimiento de los aspectos metodológicos y conceptuales de esta investigación.

Finalmente, agradezco a los autores e investigadores cuyas publicaciones fueron consultadas, ya que sus aportes teóricos y científicos permitieron fundamentar el desarrollo de este estudio, contribuyendo a una comprensión más amplia del impacto de las tecnologías activas en el desarrollo de competencias digitales en la educación superior.

Resumen

El presente estudio desarrolla un modelo de analítica de datos para predecir el nivel de competencias digitales de estudiantes del programa de Ingeniería de Sistemas de la Universidad del Pacífico (2015-2025), a partir de variables sociodemográficas, académicas y de uso de tecnologías activas. Se emplea la metodología CRISP-DM (Cross-Industry Standard Process for Data Mining) con un dataset de 538 registros, recolectados mediante registros administrativos institucionales y la aplicación del Cuestionario de Autopercepción de Competencias Tecnológicas (CACT) durante el segundo semestre de 2025. Se entrenan y evalúan cinco algoritmos de aprendizaje automático supervisado —regresión lineal múltiple, regresión regularizada (Ridge/Lasso), Random Forest y Gradient Boosting— con validación cruzada de 5 folds y métricas de desempeño: R^2 (coeficiente de determinación), RMSE (raíz del error cuadrático medio) y MAE (error absoluto medio). Las dimensiones evaluadas se alinean con el marco DigComp 2.2: información y datos, comunicación y colaboración, creación de contenido digital, seguridad y resolución de problemas tecnológicos. El modelo seleccionado proporciona a las instituciones de educación superior una herramienta analítica para la toma de decisiones basada en datos, orientada a mejorar la calidad educativa y reducir brechas en competencias digitales mediante intervenciones dirigidas a segmentos estudiantiles de mayor riesgo.

Palabras clave: analítica de datos, competencias digitales, aprendizaje automático, tecnologías activas, educación superior.

Abstract

This study develops a data analytics model to predict the level of digital competencies of students in the Systems Engineering program at Universidad del Pacífico (2015-2025), based on sociodemographic, academic, and active technology usage variables. The CRISP-DM methodology is employed with a dataset of 538 student records, collected through institutional administrative records and the application of the Technological Competencies Self-Perception Questionnaire (CACT) during the second semester of 2025. Five supervised machine learning algorithms are trained and evaluated —multiple linear regression, regularized regression (Ridge/Lasso), Random Forest, and Gradient Boosting— using 5-fold cross-validation and performance metrics (R^2, RMSE, MAE). The evaluated dimensions align with the DigComp 2.2 framework: information and data, communication and collaboration, digital content creation, security, and technological problem-solving. The selected model provides higher education institutions with an analytical tool for data-driven decision-making, aimed at improving educational quality and reducing digital competency gaps through targeted interventions for student segments at highest risk.

Keywords: data analytics, digital competencies, machine learning, active technologies, higher education.

Tabla de Contenido

Lista de Tablas

Lista de Figuras

Lista de Apéndices

Introducción

La transformación digital de los sistemas productivos ha redefinido las competencias requeridas a los profesionales de ingeniería en la última década. El Foro Económico Mundial (2023) identifica el análisis de datos, la inteligencia artificial y la gestión de tecnologías digitales como las habilidades de mayor demanda en el mercado laboral global. En este contexto, las instituciones de educación superior enfrentan la presión de adaptar sus modelos pedagógicos para desarrollar competencias digitales alineadas con estas exigencias.

La educación superior en Colombia ha experimentado un crecimiento sostenido en la adopción de tecnologías digitales, con presencia generalizada de plataformas de gestión del aprendizaje (LMS) en los procesos formativos institucionales (Valverde-Berrocoso et al., 2021; Børte et al., 2020). No obstante, la presencia de herramientas digitales no garantiza por sí misma el desarrollo de competencias. La brecha entre disponibilidad tecnológica y apropiación pedagógica constituye un desafío documentado en la literatura latinoamericana (Valverde-Berrocoso et al., 2021; Børte et al., 2020).

En la región del Pacífico colombiano, esta brecha se amplifica por condiciones socioeconómicas particulares. El Distrito de Buenaventura presenta niveles de acceso a internet y recursos tecnológicos inferiores al promedio urbano nacional (DANE, 2023). Las instituciones de educación superior en este contexto deben navegar la tensión entre demandas globales de competencias digitales y limitaciones estructurales locales. El programa de Ingeniería de Sistemas de la Universidad del Pacífico opera dentro de esta realidad institucional.

Los datos del programa muestran que, durante el periodo 2015-2025, se incorporaron progresivamente herramientas como plataformas LMS, simuladores y entornos colaborativos al currículo. No obstante, no existe evidencia analítica que determine cuáles factores —

sociodemográficos, académicos o de uso tecnológico— contribuyen de manera más significativa al desarrollo de competencias digitales en los estudiantes. Esta ausencia de análisis basado en datos limita la capacidad de la institución para diseñar intervenciones pedagógicas dirigidas y eficientes.

La presente investigación aborda esta problemática mediante el desarrollo de un modelo predictivo de analítica de datos que permita identificar y predecir el nivel de competencias digitales de los estudiantes, orientando intervenciones basadas en evidencia cuantitativa. El estudio se enmarca en la línea de investigación de la Escuela de Ciencias Básicas, Tecnología e Ingeniería (ECBTI) de la Universidad Nacional Abierta y a Distancia (UNAD), específicamente en el área de ciencia de datos aplicada a la educación.

El aporte del trabajo consiste en comparar sistemáticamente modelos predictivos de aprendizaje automático en un contexto universitario latinoamericano, identificar las variables accionables que la institución puede intervenir para fortalecer las competencias digitales de sus estudiantes, y proporcionar una herramienta analítica basada en datos que oriente las estrategias institucionales de mejoramiento educativo.

El informe se organiza de la siguiente manera: en un primer momento se presenta la contextualización y el planteamiento del problema, estableciendo las causas, efectos y consecuencias que motivan la investigación. A continuación, se formulan las preguntas, objetivos e hipótesis que dirigen el estudio, seguidos de la justificación que expone la relevancia teórica y práctica del trabajo. Posteriormente, se desarrolla el estado del arte y el marco teórico-conceptual que fundamentan la selección de modelos y estrategias de validación. La metodología se detalla en la sección correspondiente, describiendo las seis fases de CRISP-DM, las variables del dataset y las consideraciones sobre datos. Finalmente, se incluyen la operacionalización de

variables, las consideraciones metodológicas y las referencias bibliográficas que sustentan la investigación.

Justificación

La transformación digital de los sectores productivos ha incrementado la demanda de profesionales con competencias en análisis de datos, inteligencia artificial y gestión de tecnologías digitales (World Economic Forum, 2023). En este contexto, las instituciones de educación superior (IES) requieren herramientas de analítica de datos que permitan predecir el nivel de competencias digitales de sus estudiantes y orientar decisiones basadas en evidencia cuantitativa, en lugar de intervenciones genéricas o inversiones no dirigidas.

En la Universidad del Pacífico, específicamente en el programa de Ingeniería de Sistemas, no existe actualmente un modelo predictivo que identifique los factores asociados al desarrollo de competencias digitales. Esta ausencia conduce a intervenciones pedagógicas no diferenciadas, asignación ineficiente de recursos tecnológicos y egresados con brechas no detectadas oportunamente (Valverde-Berrocoso et al., 2021; Børte et al., 2020). El presente estudio responde a esa brecha mediante el desarrollo de un modelo de analítica de datos.

En el plano teórico, esta investigación contribuye a la literatura de analítica de aprendizaje al aplicar la metodología CRISP-DM y comparar sistemáticamente cinco algoritmos de aprendizaje automático (regresión lineal, Ridge/Lasso, Random Forest, XGBoost) en un contexto universitario latinoamericano, región subrepresentada en estudios predictivos de competencias digitales (Zhang et al., 2025; Khan et al., 2025). La comparación se realiza con métricas estandarizadas (R^2, RMSE, MAE) y validación cruzada, generando evidencia empírica sobre qué tipo de modelo resulta más adecuado cuando se dispone de datos de autopercepción y tamaño muestral moderado.

En el plano práctico, el modelo Lasso seleccionado (R^2=0.664 en prueba) permite a la Universidad del Pacífico: (a) identificar estudiantes con riesgo de competencias digitales

insuficientes (predicción < 3,0); (b) priorizar variables accionables, como la frecuencia de uso de LMS y el índice de acceso tecnológico, sobre variables no modificables (género, edad) o de efecto nulo (acceso a internet por sí solo); y (c) explorar escenarios hipotéticos de intervención basados en variables accionables, optimizando así la asignación de recursos.

El impacto concreto en la toma de decisiones institucionales es múltiple. La institución puede dirigir inversiones hacia capacitación docente en integración de tecnologías activas, en lugar de expandir infraestructura de conectividad sin acompañamiento pedagógico. Puede diseñar políticas curriculares que incentiven el uso frecuente y significativo de plataformas LMS, simuladores y herramientas colaborativas. Asimismo, puede hacer seguimiento semestral a la evolución de las competencias digitales mediante el modelo recalibrado con datos de cohortes posteriores, monitoreando los cambios en las predicciones generadas a partir de los nuevos datos.

Finalmente, el pipeline desarrollado (desde la limpieza y transformación hasta la comparación de algoritmos y la generación de reportes de riesgo) es replicable por otras IES de la región Pacífico y de contextos latinoamericanos con características socioeconómicas y de infraestructura tecnológica similares. La documentación detallada de los parámetros utilizados en este estudio, el código y los criterios de selección del modelo facilita su adaptación a datos de cohortes posteriores y de otras instituciones.

Planteamiento del Problema

La formación de ingenieros de sistemas en la Universidad del Pacífico carece de un modelo analítico que permita predecir el nivel de competencias digitales de los estudiantes. Las decisiones pedagógicas y de inversión tecnológica se toman sin evidencia cuantitativa sobre los factores que más influyen en el desarrollo de estas competencias.

Las causas estructurales de esta situación incluyen: la ausencia de sistemas institucionalizados de analítica educativa, la limitada sistematización de datos de interacción estudiantil con plataformas digitales y la escasa aplicación de técnicas de aprendizaje automático al análisis de datos académicos en contextos universitarios regionales. Durante el periodo 2015-2025, las prácticas pedagógicas en el programa han favorecido enfoques transmisivos que relegan las tecnologías activas a roles secundarios, lo cual frena el desarrollo de competencias tecnológicas críticas para la manipulación de datos en entornos educativos dinámicos (Børte et al., 2020). La adopción restringida de tecnologías activas es atribuible a la falta de capacitación docente en integración digital y al predominio de enfoques lectivos en currículos institucionales (Tondeur et al., 2017).

Los efectos inmediatos se manifiestan en intervenciones pedagógicas no diferenciadas que no priorizan segmentos estudiantiles de mayor riesgo, inversión tecnológica no orientada por evidencia predictiva y dificultad para identificar oportunidades de mejora en el desarrollo de competencias digitales específicas. Se observan deficiencias persistentes en colaboración digital y gestión de información, que comprometen la aplicación práctica de habilidades en escenarios reales de ingeniería (Abelha et al., 2020). La limitada actualización docente en integración tecnológica agrava estas carencias, un riesgo documentado en estudios sobre formación digital del profesorado.

Las consecuencias a largo plazo, si el problema no se aborda, podrían incluir el riesgo de que los egresados lleguen a presentar brechas formativas no identificadas oportunamente, una posible reducción de su competitividad en el mercado laboral digitalizado y limitaciones institucionales para los procesos de acreditación y mejora continua basada en datos. Estas consecuencias son plausibles según la literatura (Abelha et al., 2020), aunque no fueron medidas empíricamente en este estudio, por lo que su ocurrencia debe interpretarse como un riesgo potencial y no como un hecho documentado.

Esta jerarquía de causas, efectos y consecuencias configura el problema que la presente investigación aborda mediante el desarrollo de un modelo predictivo de analítica de datos.

Preguntas de Investigación

La falta de herramientas analíticas en las instituciones de educación superior para predecir el nivel de competencias digitales de los estudiantes limita la toma de decisiones basada en datos orientada al mejoramiento educativo.

Pregunta principal: ¿Es posible desarrollar un modelo de analítica de datos que prediga el nivel de competencias digitales de los estudiantes del programa de Ingeniería de Sistemas de la Unipacífico a partir de variables sociodemográficas, académicas y de uso de tecnologías activas durante el periodo 2015-2025?

Preguntas específicas:

a) ¿Cuáles variables sociodemográficas, académicas y de uso de tecnologías activas presentan mayor capacidad predictiva sobre el nivel de competencias digitales de los estudiantes?

b) ¿Qué modelo de aprendizaje automático supervisado —regresión lineal múltiple, regresión regularizada, Random Forest o Gradient Boosting— ofrece mayor precisión y estabilidad en la predicción de competencias digitales?

c) ¿Cómo pueden utilizarse los resultados del modelo predictivo para orientar estrategias de intervención institucional que mejoren las competencias digitales en los segmentos estudiantiles de mayor riesgo?

Objetivos de la Investigación

Objetivo General

Desarrollar un modelo de analítica de datos que permita identificar los factores asociados al nivel de competencias digitales de los estudiantes del programa de Ingeniería de Sistemas de la Universidad del Pacífico, con el fin de generar evidencia para orientar decisiones institucionales de mejoramiento educativo.

Objetivos Específicos

1. Preparar y estructurar el dataset mediante procesos de limpieza, transformación, validación de la calidad y operacionalización de variables, como insumo para el modelado predictivo.

2. Construir y evaluar modelos predictivos basados en regresión (OLS: mínimos cuadrados ordinarios; Ridge, Lasso) y ensamblaje (Random Forest, XGBoost) para estimar el nivel de competencias digitales, evaluados mediante R^2, RMSE y MAE, y seleccionar el algoritmo de mayor precisión y estabilidad.

3. Proponer orientaciones para el uso del modelo en la toma de decisiones institucionales, a partir de los patrones identificados por el modelo Lasso, con énfasis en los segmentos estudiantiles con predicción de menor nivel de competencias digitales.

Hipótesis

Hipótesis Principal

Las variables de uso de tecnologías activas (frecuencia de uso de LMS, uso de simuladores y colaboración digital) explican al menos un 50% de la varianza del nivel de competencias digitales, con coeficientes estandarizados superiores a 0,10, mientras que las variables sociodemográficas (género, edad, acceso a internet) presentan coeficientes inferiores a 0,05 o son eliminadas por el procedimiento de regularización Lasso.

Hipótesis Secundaria

Los modelos de regresión regularizada (Lasso) presentan menor error de generalización que la regresión lineal ordinaria. En cambio, los métodos de ensamblaje (Random Forest, Gradient Boosting) podrían no superar a los modelos lineales en contextos educativos con datos de autopercepción debido al riesgo de sobreajuste.

Marco de Referencia

Estado del Arte

La aplicación de modelos predictivos en educación ha experimentado un desarrollo significativo en los últimos años. La integración de inteligencia artificial en la educación ha permitido el desarrollo de modelos predictivos del desempeño académico con diversos algoritmos y enfoques de validación (Zhang et al., 2025). Una revisión comprensiva de la predicción del desempeño estudiantil mediante aprendizaje automático (Machine Learning) y aprendizaje profundo (Deep Learning) identifica las técnicas más efectivas y los desafíos metodológicos en este campo (Zhang et al., 2025). La Tabla 1 sintetiza los hallazgos de investigaciones internacionales que relacionan tecnologías activas y competencias digitales en educación superior.

Tabla 1

Estudios Previos sobre Tecnologías Activas, Competencias Digitales y Modelos Predictivos en Educación Superior

Autor(es)	Año	Contexto/País	Metodología	Muestra	Principales hallazgos
Freeman et al.	2014	Programas STEM - Internacional	Metaanálisis	225 estudios	El aprendizaje activo reduce las tasas de fracaso en cursos STEM y mejora el desempeño académico.
Bates	2019	Educación superior - Global	Análisis teórico y empírico	Estudios múltiples	La integración estratégica de tecnologías educativas favorece el desarrollo de competencias digitales.

Autor(es)	Año	Contexto/País	Metodología	Muestra	Principales hallazgos
Børte et al.	2020	Instituciones de educación superior - Europa	Revisión sistemática	Estudios múltiples	La implementación de innovación pedagógica requiere infraestructura tecnológica adecuada y formación docente continua.
Valverde-Berrocoso et al.	2021	Educación superior - España	Revisión sistemática	Estudios múltiples	Las tecnologías digitales contribuyen al desarrollo de competencias digitales, pero su impacto depende del contexto institucional.

Autor(es)	Año	Contexto/País	Metodología	Muestra	Principales hallazgos
Zhao et al.	2021	Educación universitaria - Asia	Estudio cuantitativo	350 estudiantes	Las tecnologías activas favorecen el aprendizaje autónomo y la interacción colaborativa en entornos virtuales.
Zhang et al.	2025	Internacional	Revisión sistemática	Estudios múltiples	Los modelos predictivos basados en aprendizaje automático (machine learning) y aprendizaje profundo (deep learning) son efectivos para

Autor(es)	Año	Contexto/País	Metodología	Muestra	Principales hallazgos
					predecir el desempeño estudiantil.
Khan et al.	2025	Turquía	Estudio cuantitativo con validación externa	613 estudiantes	Random Forest con selección de variables Boruta alcanzó precisión del 73.7% en prueba y 74.3% en validación cruzada.
Cabrera et al.	2024	Internacional	Revisión sistemática	Estudios múltiples	Los métodos de ensamblaje (Random Forest, Gradient Boosting) ofrecen mayor estabilidad predictiva que los modelos lineales.

Autor(es)	Año	Contexto/País	Metodología	Muestra	Principales hallazgos
Verma & Sinha	2025	Internacional	Ranking de modelos	Estudios múltiples	Gradient Boosting (excelente), Random Forest (muy alto), regresión lineal (básico) en predicción de rendimiento.

Nota. Elaboración propia con base en la literatura revisada. Los estudios se seleccionaron por su relevancia en tecnologías activas, competencias digitales, metodologías activas y modelos predictivos aplicados a educación superior. Los años 2024-2025 corresponden a publicaciones recientes incluidas para reflejar el estado del arte actualizado.

En el ámbito de la predicción temprana de resultados estudiantiles, se ha desarrollado un modelo basado en algoritmos de aprendizaje automático para evaluar el desempeño y proporcionar alertas tempranas, demostrando la viabilidad de estos enfoques en entornos educativos reales (Malik et al., 2025). La integración de tales sistemas en prácticas de evaluación formativa puede mejorar la equidad educativa, optimizar la asignación de recursos y reducir las tasas de fracaso estudiantil.

Una revisión sistemática de modelos predictivos con fines educativos evalúa su capacidad para pronosticar el desempeño, identificar estudiantes en riesgo y orientar intervenciones institucionales (Cabrera et al., 2024). Esta revisión confirma que los métodos de ensamblaje, particularmente Random Forest y Gradient Boosting, ofrecen mayor estabilidad predictiva comparados con modelos lineales tradicionales.

En cuanto a la comparación específica de algoritmos, se ha examinado la efectividad de Random Forest Regression, Support Vector Regression y Gradient Boosting para la predicción del desempeño académico, encontrando que los métodos de ensamblaje superan a los modelos individuales en precisión y consistencia (Cabrera et al., 2024). Un ranking general de modelos para predicción de rendimiento posiciona a Gradient Boosting en primer lugar (desempeño excelente), seguido de Random Forest (muy alto), Support Vector Machine (alto), Decision Tree (moderado) y Linear Regression (básico) (Verma & Sinha, 2025).

Uso de CRISP-DM en Proyectos de Analítica Educativa

La metodología CRISP-DM ha sido aplicada exitosamente en proyectos de minería de datos educativos. Un estudio de caso sobre predicción de riesgos académicos utilizó CRISP-DM como estructura para garantizar calidad y éxito académico, demostrando la utilidad de este enfoque en contextos universitarios (Chapman et al., 2000). La estructura en seis fases —análisis del problema, análisis de datos, comprensión de datos, preparación de datos, modelado, evaluación y despliegue— proporciona un marco replicable para proyectos de analítica educativa.

Asimismo, se ha desarrollado un modelo predictivo del desempeño académico estudiantil a partir de datos de sistemas de gestión del aprendizaje (LMS) utilizando CRISP-DM,

confirmando que esta metodología facilita la alineación entre objetivos educativos y técnicas de minería de datos (Romero & Ventura, 2020).

Feature Selection y Optimización de Modelos

Investigaciones recientes han introducido metodologías avanzadas de selección de variables para la predicción del desempeño estudiantil. Un enfoque novedoso combina análisis de matriz de correlación, ganancia de información y Chi-cuadrado con mapas de calor para seleccionar las variables más relevantes, incorporando umbralización dinámica y adaptativa que mejora la precisión y flexibilidad predictiva (Malik et al., 2025). Este enfoque demuestra que la selección adecuada de variables es tan crítica como la selección del algoritmo de modelado.

En un estudio con 613 estudiantes de secundaria en Turquía, se aplicó Random Forest con selección de variables Boruta, reduciendo 84 variables iniciales a 30 predictores clave. El modelo alcanzó una precisión de prueba del 73.7% y una precisión de validación cruzada del 74.3%, demostrando estabilidad y consistencia entre entrenamiento, prueba y validación externa (Alkan et al., 2025). La validación externa con 30 estudiantes adicionales alcanzó una precisión del 73.3% (Precisión: 0.823, Recall: 0.736, F1: 0.778).

Además de Boruta y la regularización Lasso (empleada en este estudio), otros métodos de selección de variables utilizados en datos educativos incluyen RFE (Recursive Feature Elimination), que elimina iterativamente predictores de menor importancia, y algoritmos basados en ganancia de información. En cuanto a la validación cruzada, cuando se trabaja con muestras pequeñas ($n < 500$), se recomienda el uso de validación cruzada repetida (repeated k-fold) o estratificada (stratified k-fold) para reducir la varianza de las estimaciones. En este estudio se empleó stratified k-fold ($k=5$) estratificando por cohorte, lo que garantiza que cada pliegue mantenga la misma proporción de años de ingreso. Una validación repetida (por ejemplo, 5x5-

fold) podría ofrecer estimaciones aún más estables, aunque no se implementó por simplicidad computacional.

DigComp y Competencias Digitales en Educación Superior

El marco DigComp 2.2 ha sido validado entre estudiantes universitarios en diferentes sistemas educativos, confirmando su confiabilidad, validez y aplicabilidad para evaluar competencias digitales (Abubakari et al., 2025). Estos hallazgos respaldan la visión holística de DigComp y la necesidad de desarrollo equilibrado en todas las competencias digitales.

En un estudio con 93 estudiantes de ingeniería informática y formación docente en Serbia, se evaluaron competencias DigComp mediante tareas prácticas objetivas en escenarios simulados del mundo real, en lugar de cuestionarios de autopercepción (Abubakari et al., 2025). Los estudiantes obtuvieron las puntuaciones más altas en Programación ($M \approx 6.5$-7.4) y Pensamiento Creativo ($M \approx 6.1$-7.1), y las más bajas en Desarrollo de Software ($M \approx 2.5$-3.4) y Trabajo en Equipo/Comunicación ($M \approx 2.9$-4.2). Este patrón de resultados tiene implicaciones directas para el diseño curricular en programas de ingeniería.

La competencia en inglés se correlacionó significativamente con Procesamiento de Señales Digitales, Programación, Desarrollo de Software y Diseño Digital, sugiriendo que los programas de ingeniería deben integrar o reforzar el inglés como habilitador principal del desarrollo de competencias digitales y de inteligencia artificial (Abubakari et al., 2025).

Investigaciones en Contextos Latinoamericanos

En América Latina, la investigación sobre analítica de datos aplicada a la educación superior enfrenta desafíos particulares relacionados con la disponibilidad de datos institucionalizados, la infraestructura tecnológica y la capacitación docente. Los estudios

empíricos sobre predicción de competencias digitales en programas de ingeniería son limitados, lo que representa una brecha de conocimiento que esta investigación busca contribuir a cerrar.

La integración de tecnologías digitales en la educación superior latinoamericana requiere estrategias institucionales que promuevan el desarrollo de competencias tanto en estudiantes como en docentes, considerando factores culturales, institucionales y económicos que influyen en la apropiación tecnológica (Valverde-Berrocoso et al., 2021).

Brechas Identificadas

La revisión de literatura permite identificar las siguientes brechas de conocimiento:

1. Escasez de estudios predictivos en contextos latinoamericanos: La mayoría de las investigaciones sobre modelos predictivos de competencias digitales se han realizado en Europa, Asia y Norteamérica. Existe una necesidad evidente de estudios empíricos en universidades de América Latina.

2. Limitada aplicación de CRISP-DM en educación superior regional: Aunque CRISP-DM ha sido utilizada exitosamente en proyectos de analítica educativa, su aplicación en contextos universitarios latinoamericanos es escasa.

3. Predominio de medidas de autopercepción: Muchos estudios sobre competencias digitales utilizan instrumentos de autopercepción en lugar de evaluaciones objetivas de desempeño práctico. Esta investigación reconoce esta limitación y la declara explícitamente.

4. Comparación sistemática de modelos: Pocos estudios comparan de manera sistemática regresión lineal, métodos regularizados y métodos de ensamblaje en el mismo dataset educativo con métricas estandarizadas.

De esta revisión se desprenden las decisiones metodológicas que estructuran el presente estudio. En primer lugar, se elige Lasso como modelo principal por su capacidad de realizar

selección automática de variables mediante regularización L1, lo cual es particularmente valioso en contextos educativos donde se requiere identificar predictores accionables (Romero & Ventura, 2020). En segundo lugar, se incluyen métodos de ensamblaje como comparación porque la literatura no es concluyente sobre su superioridad frente a modelos lineales con datos de autopercepción (Cabrera et al., 2024). En tercer lugar, se adopta el marco DigComp 2.2 para la variable objetivo por ser el estándar internacional validado (Vuorikari et al., 2022). Finalmente, la validación cruzada estratificada por cohorte se justifica por la necesidad de controlar el posible efecto del año de ingreso (Zhang et al., 2025).

Marco Contextual

El programa de Ingeniería de Sistemas de la Universidad del Pacífico se desarrolla en un contexto educativo caracterizado por transformaciones aceleradas en los campos de la tecnología, la analítica de datos y la digitalización de los procesos organizacionales. Durante el periodo comprendido entre 2015 y 2025, la creciente demanda de profesionales con competencias en ciencia de datos, análisis de información y desarrollo tecnológico ha generado nuevas exigencias para las instituciones de educación superior, particularmente en programas de ingeniería que deben responder a los desafíos de la cuarta revolución industrial.

En este escenario, la formación de ingenieros requiere no solo la adquisición de conocimientos técnicos tradicionales, sino también el desarrollo de competencias digitales avanzadas, pensamiento crítico, trabajo colaborativo y habilidades para la resolución de problemas complejos en entornos digitales. Diversos organismos internacionales, como el Foro Económico Mundial, han señalado que las competencias relacionadas con el análisis de datos, la inteligencia artificial y la gestión de tecnologías digitales se encuentran entre las habilidades más demandadas en el mercado laboral contemporáneo. En consecuencia, los programas académicos

deben integrar metodologías pedagógicas innovadoras que permitan fortalecer dichas competencias desde el proceso formativo universitario.

No obstante, en muchos programas de ingeniería persiste una tensión entre estas nuevas demandas del entorno tecnológico y las prácticas pedagógicas tradicionales que aún predominan en la educación superior. En el caso específico del programa de Ingeniería de Sistemas de la Universidad del Pacífico, durante el periodo 2015–2025 se observa una contradicción central entre la necesidad de formar profesionales con capacidades analíticas y tecnológicas avanzadas y la permanencia de enfoques pedagógicos centrados en la transmisión unidireccional del conocimiento. Estas prácticas tradicionales tienden a privilegiar la exposición magistral sobre metodologías activas de aprendizaje que promuevan la participación estudiantil, la experimentación tecnológica y la resolución de problemas aplicados (Abelha et al., 2020).

Aunque los currículos de ingeniería han comenzado a incorporar referencias al desarrollo de habilidades digitales, en muchos casos estas competencias no se traducen plenamente en prácticas pedagógicas que favorezcan el aprendizaje activo mediado por tecnologías. En particular, se evidencia un uso limitado de herramientas interactivas como plataformas de gestión del aprendizaje (Learning Management Systems – LMS), simuladores digitales, entornos colaborativos en línea y laboratorios virtuales. Estas herramientas tienen la capacidad de transformar los procesos de enseñanza y aprendizaje al facilitar la interacción, la experimentación práctica y la construcción colectiva del conocimiento (Agila-Palacios et al., 2021).

Adicionalmente, el contexto regional del Distrito de Buenaventura presenta desafíos particulares asociados a condiciones socioeconómicas, acceso a infraestructura tecnológica y brechas digitales que influyen en los procesos educativos. Las instituciones de educación

superior ubicadas en esta región deben enfrentar retos relacionados con la disponibilidad de recursos tecnológicos, la capacitación docente en herramientas digitales y la adaptación de los modelos pedagógicos a contextos educativos diversos. Estas condiciones hacen aún más relevante la implementación de estrategias pedagógicas innovadoras que permitan fortalecer las competencias tecnológicas de los estudiantes y mejorar su preparación para entornos laborales cada vez más digitalizados.

En este sentido, la integración de tecnologías activas en el proceso formativo representa una oportunidad para transformar las prácticas educativas tradicionales y promover el desarrollo de competencias digitales en los estudiantes de ingeniería. Las tecnologías activas incluyen herramientas y metodologías que facilitan la participación activa del estudiante en el proceso de aprendizaje, tales como plataformas virtuales de aprendizaje, herramientas colaborativas en línea, simuladores tecnológicos, entornos de programación y sistemas de análisis de datos aplicados. Estas herramientas permiten crear entornos educativos más dinámicos e interactivos, favoreciendo la experimentación, la resolución de problemas y el aprendizaje basado en proyectos.

No obstante, la adopción de estas tecnologías en el contexto institucional requiere procesos de adaptación pedagógica, capacitación docente y fortalecimiento de la infraestructura tecnológica. Las dinámicas institucionales que limitan la incorporación efectiva de tecnologías activas pueden generar brechas en el desarrollo de competencias digitales, afectando la preparación de los egresados para enfrentar los desafíos del mercado laboral contemporáneo, particularmente en áreas relacionadas con la analítica de datos y la transformación digital.

En consecuencia, la brecha existente entre las demandas del entorno tecnológico y las prácticas pedagógicas tradicionales resalta la necesidad de analizar el papel de las tecnologías

activas en el fortalecimiento de las competencias digitales en programas de ingeniería. Este análisis resulta especialmente relevante en contextos regionales como el de la Universidad del Pacífico, donde la innovación educativa puede constituirse en una estrategia clave para mejorar la calidad de la formación profesional y contribuir al desarrollo tecnológico y social del territorio.

Marco Teórico

Innovación Educativa y Tecnologías Activas

Las tecnologías activas comprenden el conjunto de herramientas digitales y metodologías pedagógicas que promueven la participación directa del estudiante en su proceso de aprendizaje. Este paradigma se diferencia de los modelos transmisivos al priorizar la construcción activa del conocimiento mediante la interacción con plataformas virtuales, simuladores, entornos colaborativos y sistemas de análisis de datos aplicados (Freeman et al., 2014).

La investigación ha demostrado que estas tecnologías incrementan el compromiso estudiantil y mejoran la comprensión conceptual en disciplinas STEM. El metaanálisis de Freeman et al. (2014), que incluyó más de 200 estudios en ciencias e ingeniería, evidenció que el aprendizaje activo reduce las tasas de fracaso en un 55% comparado con clases magistrales tradicionales.

La adopción de tecnologías activas en educación superior ha experimentado un crecimiento sostenido desde 2015, impulsado por la digitalización de los procesos productivos y la demanda de competencias tecnológicas en los profesionales (Bates, 2019; Zhao et al., 2021). Las instituciones han replanteado sus modelos pedagógicos hacia enfoques centrados en el estudiante, donde la tecnología trasciende su función instrumental para convertirse en un elemento transformador del proceso educativo.

En contextos latinoamericanos, esta adopción enfrenta desafíos particulares relacionados con infraestructura tecnológica, capacitación docente y integración curricular (Børte et al., 2020; Valverde-Berrocoso et al., 2021). Las brechas institucionales en recursos y formación constituyen obstáculos para la adopción sostenible de innovaciones educativas en universidades de países en desarrollo.

Los programas de ingeniería requieren metodologías que promuevan la resolución de problemas, el pensamiento algorítmico y la aplicación práctica del conocimiento. Las tecnologías activas facilitan estos procesos mediante simuladores, entornos de programación, laboratorios virtuales y plataformas de gestión del aprendizaje (LMS) que permiten la experimentación y el aprendizaje basado en proyectos (Drugova et al., 2021).

Competencias Digitales en Educación Superior

Las competencias digitales se definen como el conjunto de conocimientos, habilidades y actitudes necesarios para utilizar herramientas digitales de manera efectiva, crítica y segura en diferentes contextos profesionales y educativos (Redecker, 2017). En programas de ingeniería, estas competencias adquieren relevancia al formar profesionales capaces de diseñar, implementar y evaluar soluciones tecnológicas en contextos organizacionales y sociales.

El Marco Europeo de Competencias Digitales para Ciudadanos (DigComp 2.2) establece una estructura de cinco áreas de competencia (Vuorikari et al., 2022): información y datos, comunicación y colaboración, creación de contenido digital, seguridad y resolución de problemas. Este marco proporciona un referente estandarizado para evaluar el nivel de desarrollo de habilidades digitales en estudiantes y diseñar intervenciones educativas específicas.

El modelo Technological Pedagogical Content Knowledge (TPACK) plantea la integración de tres tipos de conocimiento: tecnológico, pedagógico y disciplinar (Koehler &

Mishra, 2009). Este marco explica cómo los docentes pueden integrar tecnologías de manera efectiva en su práctica al articular estos tres dominios. En el contexto de la formación en ingeniería, TPACK orienta el diseño de experiencias de aprendizaje que conectan herramientas digitales con objetivos pedagógicos y contenidos disciplinares específicos.

El modelo SAMR (Substitution, Augmentation, Modification, Redefinition) describe cuatro niveles de integración tecnológica en educación, desde la sustitución de herramientas tradicionales por digitales hasta la redefinición de las actividades de aprendizaje mediante tecnología (Puentedura, 2014). Este modelo permite evaluar el grado de transformación pedagógica derivado del uso de tecnologías activas y orienta la progresión hacia niveles más profundos de integración.

En el presente estudio, el marco DigComp 2.2 se utiliza para construir la variable objetivo como promedio de las cinco dimensiones de competencia digital, mientras que el modelo SAMR proporciona el referente conceptual para interpretar los niveles de integración tecnológica alcanzados por los estudiantes en función de su exposición a tecnologías activas.

Analítica de Datos en Educación (Learning Analytics)

La analítica de datos en educación, conocida como analítica de aprendizaje, se define como la medición, recopilación, análisis y reporte de datos sobre estudiantes y sus contextos, con el propósito de comprender y optimizar el aprendizaje y los entornos donde se produce (Siemens & Long, 2011). Esta disciplina ha emergido como un campo de investigación con aplicaciones en predicción de deserción, personalización del aprendizaje, evaluación formativa y toma de decisiones institucionales.

Los modelos predictivos aplicados a la educación permiten identificar estudiantes en riesgo antes de que las brechas de aprendizaje se consoliden. La literatura reporta aplicaciones

exitosas en la predicción de calificaciones, detección temprana de deserción y recomendación personalizada de recursos (Romero & Ventura, 2020). Los algoritmos más empleados incluyen regresión logística, árboles de decisión, redes neuronales y métodos de ensamblaje.

La analítica de comportamiento examina patrones de interacción con plataformas LMS, frecuencia de acceso a recursos, participación en foros y tiempos de entrega de actividades. Estos datos comportamentales complementan las medidas de autopercepción y proporcionan evidencia objetiva sobre el compromiso estudiantil con las tecnologías activas.

La evolución del campo de analítica de aprendizaje ha transitado desde indicadores descriptivos (qué ocurrió) hasta modelos predictivos (qué podría ocurrir) y, más recientemente, enfoques prescriptivos (qué acciones tomar).

Los primeros sistemas se limitaban a tableros de control con tasas de aprobación y asistencia; actualmente, los modelos predictivos permiten anticipar deserción, bajo rendimiento o, como en este estudio, niveles insuficientes de competencias digitales (Romero & Ventura, 2020). Las aplicaciones típicas incluyen la predicción de calificaciones curso a curso, la identificación temprana de estudiantes en riesgo de abandono, la recomendación personalizada de recursos de aprendizaje y la evaluación de competencias transversales. No obstante, en contextos latinoamericanos persisten desafíos específicos: escasa sistematización de datos institucionales, infraestructura tecnológica limitada, falta de capacitación docente en analítica educativa y necesidad de marcos éticos adaptados a poblaciones vulnerables. Este estudio aborda parte de esas brechas al proponer un modelo predictivo con metodología CRISP-DM.

Este marco de analítica de aprendizaje sustenta el uso de datos educativos institucionales para la construcción de modelos predictivos, y justifica la adopción de CRISP-DM como

metodología estructurada que garantiza la alineación entre los objetivos pedagógicos y las técnicas de minería de datos aplicadas al contexto universitario.

Modelos Predictivos y Machine Learning

El aprendizaje supervisado consiste en entrenar un modelo a partir de datos etiquetados, donde cada observación incluye tanto las variables predictoras (X) como la variable objetivo (y). El objetivo es aprender una función $f: \mathcal{X} \rightarrow \mathcal{Y}$ que generalice correctamente a observaciones no vistas. En el contexto educativo, el aprendizaje supervisado se aplica para predecir competencias digitales (regresión) o clasificar estudiantes en categorías de riesgo (clasificación) a partir de variables sociodemográficas, académicas y de uso tecnológico.

La regresión lineal múltiple modela la relación lineal entre una variable dependiente continua y un conjunto de variables predictoras. Cada coeficiente β_i representa el cambio esperado en la variable objetivo por cada unidad de incremento en el predictor X_i, manteniendo constantes los demás predictores. Los coeficientes estandarizados (β) permiten comparar la magnitud relativa del efecto de cada predictor. Sus limitaciones incluyen sensibilidad a outliers, incapacidad para capturar relaciones no lineales, y degradación del rendimiento en presencia de multicolinealidad severa.

Formalmente, el modelo OLS estima los coeficientes β que minimizan la suma de errores cuadráticos:

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} (y_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2$$

donde y_i es el valor observado de la variable objetivo y $x_i\beta$ es la predicción lineal para la i-ésima observación.

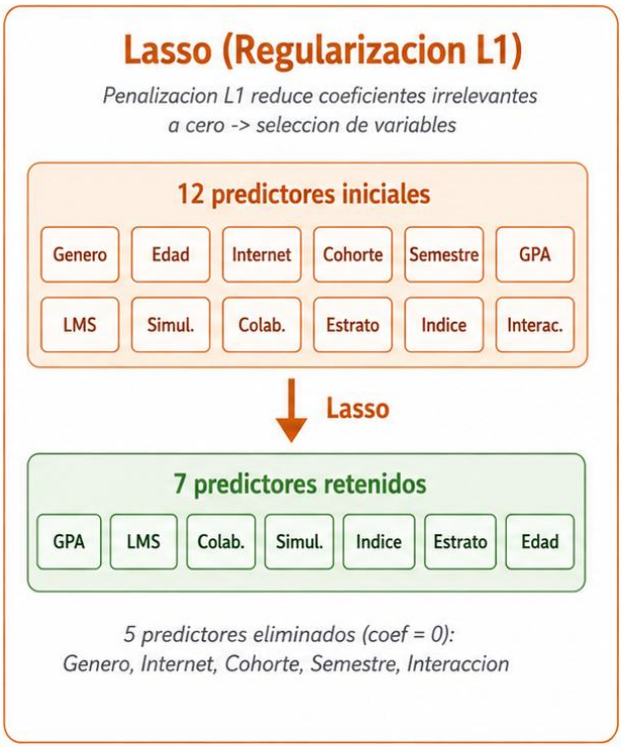
La regularización introduce una penalización a la función de pérdida para controlar la complejidad del modelo y mitigar el sobreajuste. Ridge (L2) penaliza la suma de los coeficientes al cuadrado, reduciendo la magnitud de todos los coeficientes pero sin llevarlos a cero. Lasso (L1) penaliza la suma de los valores absolutos de los coeficientes, pudiendo llevar coeficientes a cero y realizando selección automática de variables. El hiperparámetro α controla la intensidad de la penalización y se optimiza mediante validación cruzada.

Ridge (L2) minimiza $\min_{\boldsymbol{\beta}}\{\sum_{i=1}^{n}(y_i - \mathbf{x}_i^{\top}\boldsymbol{\beta})^2 + \alpha\sum_{j=1}^{p}\beta_j^2\}$, mientras que Lasso (L1) minimiza $\min_{\boldsymbol{\beta}}\{\sum_{i=1}^{n}(y_i - \mathbf{x}_i^{\top}\boldsymbol{\beta})^2 + \alpha\sum_{j=1}^{p}|\beta_j|\}$. La penalización L1 de Lasso puede llevar coeficientes exactamente a cero, realizando selección automática de variables, mientras que Ridge solo los reduce sin eliminarlos.

La Figura 1 ilustra la lógica de Lasso aplicada a este estudio: a partir de 12 predictores iniciales, la penalización L1 lleva a cinco coeficientes a cero (Género, Internet, Cohorte, Semestre e Interacción), reteniendo los siete predictores más relevantes para la predicción de competencias digitales.

Figura 1

Lógica de Lasso (Regularización L1)



Nota. Lasso aplica una penalización L1 que reduce coeficientes irrelevantes a cero, permitiendo la selección automática de variables. En el modelo aplicado, 12 predictores iniciales fueron reducidos a 7 predictores retenidos; los 5 con coeficiente igual a cero (Género, Internet, Cohorte, Semestre e Interacción) fueron eliminados del modelo final.

Random Forest es un método de ensamblaje basado en bagging que construye múltiples árboles de decisión y promedia sus predicciones (Breiman, 2001). Captura relaciones no lineales e interacciones sin especificación previa, no requiere supuestos distribucionales sobre los datos, y es estable frente a outliers y ruido. Proporciona medida de importancia de variables basada en la reducción promedio de impureza a través de todos los árboles.

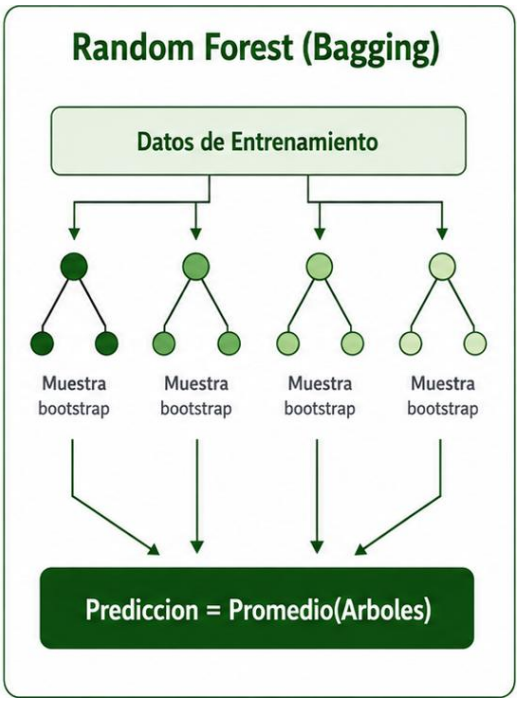
Formalmente, la predicción de Random Forest es el promedio de las predicciones de B árboles individuales: $\hat{y}(x) = \frac{1}{B}\sum_{b=1}^{B} T_b\,(x)$, donde cada árbol T_b se entrena sobre una muestra

bootstrap del conjunto de entrenamiento y utiliza un subconjunto aleatorio de predictores en cada división.

La Figura 2 muestra la arquitectura de Random Forest: los datos de entrenamiento se particionan en múltiples muestras bootstrap, cada una entrena un árbol de decisión independiente, y la predicción final se obtiene como el promedio de las predicciones de todos los árboles.

Figura 2

Lógica de Random Forest (Bagging)



Nota. Random Forest entrena múltiples árboles de decisión en paralelo, cada uno sobre una muestra bootstrap del conjunto de entrenamiento. La predicción final se obtiene como el promedio de las predicciones de todos los árboles.

Gradient Boosting construye modelos de forma secuencial, donde cada nuevo modelo corrige los errores del anterior mediante optimización del gradiente de la función de pérdida (Chen y Guestrin, 2016). XGBoost es una implementación optimizada que incluye

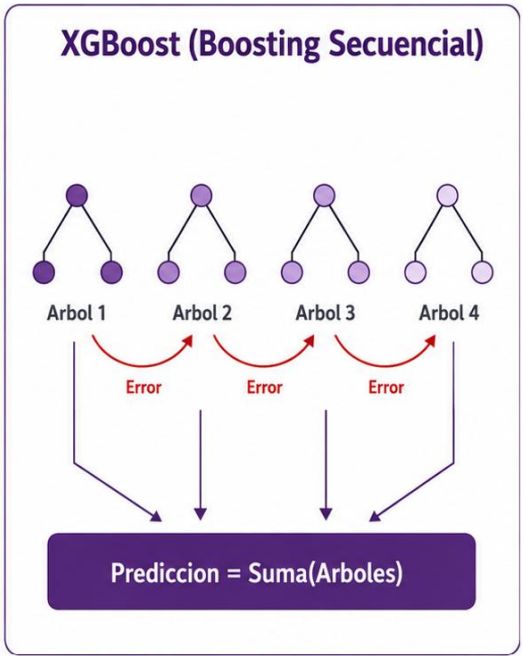
regularización, manejo de valores faltantes y poda de árboles. Ofrece alto rendimiento predictivo, frecuentemente superior a otros métodos, e incorpora regularización L1 y L2 para controlar sobreajuste.

XGBoost construye un modelo aditivo de K árboles: $\hat{y}(x) = \sum_{k=1}^{K} f_k\,(x)$, donde cada f_k se ajusta secuencialmente para corregir los errores residuales del conjunto anterior, minimizando una función de pérdida regularizada que incluye términos de penalización L1 y L2.

La Figura 3 representa la lógica secuencial de XGBoost: cada árbol se ajusta sobre los errores residuales del anterior, y la predicción final se obtiene como la suma de las contribuciones de todos los árboles del modelo.

Figura 3

Lógica de XGBoost (Boosting Secuencial)



Nota. XGBoost construye árboles secuencialmente, donde cada nuevo modelo corrige los errores residuales del anterior. La predicción final se obtiene como la suma de las contribuciones de todos los árboles.

La elección de regresión lineal múltiple, Ridge, Lasso, Random Forest y XGBoost responde a las características típicas de los datasets educativos: tamaño moderado (n=538), presencia de variables mixtas (numéricas, ordinales, binarias), posible ruido por autopercepción y relaciones predominantemente lineales según literatura previa. La regresión lineal múltiple se incluye como línea base interpretable. Ridge y Lasso permiten manejar multicolinealidad y realizar selección automática de variables, lo cual es valioso cuando se desea identificar predictores accionables. Random Forest y XGBoost se incorporan por su capacidad para capturar interacciones complejas y no linealidades, aunque se espera evaluar si su complejidad adicional justifica el riesgo de sobreajuste con muestras moderadas.

En la Tabla 2 se resumen los supuestos de cada modelo y cómo se verificaron en este estudio.

Tabla 2

Supuestos de los Modelos Predictivos y Métodos de Verificación

Modelo	**Supuestos principales**	**Método de verificación en este estudio**
Regresión lineal múltiple	Linealidad, independencia de errores, homocedasticidad, normalidad de residuos, no multicolinealidad	Gráficos de residuos vs. valores predichos; prueba de Durbin-Watson (independencia); gráfico Q-Q (normalidad); VIF (multicolinealidad, todos < 5)

Modelo	Supuestos principales	Método de verificación en este estudio
Ridge / Lasso	Los mismos de regresión lineal, pero toleran mejor multicolinealidad; no requieren normalidad de residuos para la estimación puntual	Verificación similar; la regularización mitiga la multicolinealidad
Random Forest	No requiere supuestos distribucionales; asume que las observaciones son independientes	Validación cruzada; análisis de sobreajuste comparando R^2 entrenamiento vs. prueba
XGBoost	No requiere supuestos distribucionales; asume independencia	Validación cruzada; early stopping para evitar sobreajuste

Nota. En este estudio no se realizaron pruebas de significancia estadística (p-valores) sino evaluación mediante métricas predictivas y validación cruzada, consistente con el paradigma de ciencia de datos (Shmueli, 2010).

Diversos estudios han aplicado exitosamente estos algoritmos en contextos educativos. La regresión lineal y regularizada ha sido utilizada para predecir el rendimiento académico a partir de datos sociodemográficos y de uso de LMS (Romero & Ventura, 2020). Random Forest ha mostrado alta precisión en la predicción de deserción temprana con muestras de tamaño similar a la de este estudio (Alkan et al., 2025). XGBoost ha sido destacado como uno de los

modelos con mejor desempeño en rankings de predicción de rendimiento estudiantil (Verma & Sinha, 2025). La presente investigación contribuye al comparar sistemáticamente estos cinco enfoques en el dominio específico de competencias digitales.

En este estudio, se implementan cinco algoritmos —OLS, Ridge, Lasso, Random Forest y XGBoost— seleccionados por su complementariedad. Lasso se elige por su capacidad de regularización y selección automática de variables, lo que lo convierte en una herramienta interpretable para la toma de decisiones institucionales. Random Forest y XGBoost se incluyen para evaluar si las relaciones no lineales entre predictores y competencias digitales mejoran la capacidad predictiva respecto a los modelos lineales.

Validación de Modelos

La división del dataset en conjuntos de entrenamiento y prueba es la estrategia más básica de validación. El modelo se entrena con un subconjunto (típicamente 70-80%) y se evalúa con el resto (20-30%). Esta separación permite estimar la capacidad de generalización del modelo a datos no vistos. La validación cruzada k-fold divide el dataset en k particiones (folds) de tamaño similar. El modelo se entrena k veces, cada vez utilizando k-1 folds y evaluándose con el fold restante. El rendimiento final es el promedio de las k evaluaciones.

Las métricas de evaluación para regresión incluyen R^2 (proporción de varianza explicada por el modelo), RMSE (error cuadrático medio en unidades originales) y MAE (error absoluto medio, menos sensible a outliers). R^2 es la métrica principal para comparar modelos en este estudio, complementada por RMSE y MAE para evaluar magnitud y distribución de errores.

El sobreajuste ocurre cuando un modelo memoriza los datos de entrenamiento en lugar de aprender patrones generalizables. Se detecta cuando R^2 en entrenamiento es significativamente mayor que R^2 en validación/prueba (diferencia > 0.10), o cuando RMSE en validación/prueba es

significativamente mayor que RMSE en entrenamiento. Las estrategias de prevención incluyen regularización (Ridge/Lasso), limitación de profundidad en árboles de decisión, early stopping en gradient boosting, y reducción del número de predictores mediante selección de variables.

La Tabla 3 articula los principales hallazgos empíricos reportados en la literatura con los conceptos teóricos que los sustentan, mostrando su relevancia para la educación tecnológica en contextos como el de la Unipacífico.

Tabla 3

Relación entre Hallazgos Empíricos, Conceptos Teóricos y Relevancia para la Educación Tecnológica

N°	Hallazgo empírico	Relación conceptual	Relevancia para la educación tecnológica
1	Las tecnologías activas mejoran la retención del conocimiento en ingeniería.	Teoría del aprendizaje significativo (Ausubel, 1968)	Fortalece el desarrollo conceptual de herramientas digitales complejas en currículos universitarios.
2	La innovación pedagógica exige integración de TIC en disciplinas técnicas.	Modelo TPACK (Mishra & Koehler, 2006)	Permite diseñar experiencias tecnológicas contextualizadas y efectivas en programas de sistemas.

N°	Hallazgo empírico	Relación conceptual	Relevancia para la educación tecnológica
3	El aprendizaje colaborativo acelera competencias digitales.	Constructivismo social (Vygotsky, 1978)	Promueve el uso de plataformas digitales para resolución de problemas en entornos ingenieriles.
4	Las competencias técnicas emergen de prácticas interactivas.	Enfoque por competencias (Tobón, 2005)	Facilita la adquisición de habilidades profesionales en análisis de datos y simulación.
5	La actualización docente es clave para entornos digitales.	Formación continua (Delors, 1996)	Justifica la adopción de tecnologías activas para sostenibilidad en educación superior.

Nota. Elaboración propia a partir de la literatura sobre innovación pedagógica. La tabla vincula evidencias científicas con marcos teóricos aplicables al desarrollo de competencias digitales en programas de ingeniería.

Marco Conceptual

El marco conceptual de la presente investigación se centra en el análisis de las competencias tecnológicas y digitales desarrolladas por estudiantes de ingeniería mediante la integración de tecnologías activas en el proceso educativo. En el contexto de la educación superior contemporánea, las competencias digitales se han convertido en un elemento

fundamental para la formación de profesionales capaces de desenvolverse en entornos laborales caracterizados por la digitalización de los procesos productivos y la creciente importancia del análisis de datos. Diversos autores señalan que la educación universitaria debe promover el uso crítico y reflexivo de las tecnologías digitales para fortalecer habilidades de análisis, resolución de problemas y gestión de información en entornos tecnológicos complejos (Bates, 2019; Freeman et al., 2014).

Las competencias tecnológicas pueden entenderse como el conjunto de conocimientos, habilidades y actitudes necesarias para utilizar herramientas digitales de manera efectiva en diferentes contextos profesionales y educativos. En programas de ingeniería, estas competencias adquieren especial relevancia debido a la necesidad de formar profesionales capaces de diseñar, implementar y evaluar soluciones tecnológicas en contextos organizacionales y sociales. En este sentido, el desarrollo de competencias digitales implica no solo el manejo técnico de herramientas informáticas, sino también la capacidad de analizar información digital, colaborar en entornos virtuales y producir contenido tecnológico relevante para la resolución de problemas (Redecker, 2017).

Uno de los marcos conceptuales más relevantes para analizar las competencias digitales es el modelo DigComp (Digital Competence Framework) desarrollado por la Comisión Europea. Este modelo establece una estructura de competencias digitales organizada en cinco dimensiones principales: alfabetización en información y datos, comunicación y colaboración, creación de contenido digital, seguridad digital y resolución de problemas tecnológicos. Estas dimensiones permiten evaluar el nivel de desarrollo de habilidades digitales en estudiantes y profesionales, proporcionando un marco de referencia para el diseño de políticas educativas y programas de formación en competencias digitales (Vuorikari et al., 2022).

De manera complementaria, los estándares ISTE (International Society for Technology in Education) proporcionan un marco conceptual orientado al desarrollo de habilidades digitales en estudiantes y docentes. Estos estándares destacan la importancia de promover el aprendizaje empoderado, el pensamiento computacional, la innovación tecnológica y la ciudadanía digital responsable. En el contexto de la educación superior, los estándares ISTE permiten orientar la integración de tecnologías digitales en los procesos pedagógicos, facilitando la formación de estudiantes capaces de utilizar herramientas tecnológicas para investigar, crear conocimiento y desarrollar soluciones innovadoras a problemas complejos (ISTE, 2021).

Otro modelo conceptual relevante para comprender la integración de tecnologías digitales en el proceso educativo es el modelo SAMR (Substitution, Augmentation, Modification, Redefinition) propuesto por Puentedura. Este modelo describe diferentes niveles de integración tecnológica en la educación, comenzando con la sustitución de herramientas tradicionales por herramientas digitales, y avanzando hacia niveles más complejos donde la tecnología transforma completamente las actividades de aprendizaje. La aplicación del modelo SAMR en contextos educativos permite analizar cómo las tecnologías activas pueden contribuir a transformar los procesos pedagógicos y generar experiencias de aprendizaje más dinámicas e interactivas (Puentedura, 2014).

En programas de ingeniería, la integración de tecnologías activas en el proceso educativo permite fortalecer el desarrollo de competencias digitales mediante metodologías que promueven la participación activa del estudiante en su propio aprendizaje. Las tecnologías activas incluyen herramientas como plataformas de gestión del aprendizaje (LMS), simuladores digitales, entornos colaborativos en línea y herramientas de programación que permiten desarrollar proyectos tecnológicos y analizar datos en contextos educativos. Investigaciones recientes han

demostrado que la implementación de estas tecnologías contribuye significativamente al desarrollo de habilidades analíticas, pensamiento crítico y resolución de problemas en estudiantes universitarios (Drugova et al., 2021; Santos et al., 2024).

En este sentido, la integración de tecnologías activas en la educación superior no solo contribuye al fortalecimiento de competencias digitales, sino también al mejoramiento de la calidad educativa. La calidad educativa se relaciona con la capacidad de las instituciones de educación superior para ofrecer experiencias de aprendizaje pertinentes, innovadoras y alineadas con las demandas del entorno laboral. En el contexto de la formación en ingeniería, la incorporación de herramientas tecnológicas y metodologías activas permite desarrollar procesos de aprendizaje más significativos, favoreciendo la interacción entre estudiantes, docentes y recursos digitales (Segovia-García et al., 2025).

Modelo Conceptual de la Investigación

Con el propósito de comprender la relación entre el uso de tecnologías activas y el nivel de competencias digitales de los estudiantes de ingeniería, y predecir este último mediante técnicas de analítica de datos, se propone un modelo conceptual que articula tres componentes: variables sociodemográficas y académicas (predictores base), variables de uso de tecnologías activas (predictores principales) y competencias digitales (variable objetivo).

En este modelo, las variables de uso de tecnologías activas —frecuencia de uso de LMS, uso de simuladores y colaboración digital— se plantean como los predictores de mayor peso sobre las competencias digitales, por encima de las variables sociodemográficas. Esta jerarquía se fundamenta en la evidencia de que el uso efectivo de tecnologías, más que la mera disponibilidad de acceso, determina el desarrollo de competencias digitales (Zhao et al., 2021; Valverde-Berrocoso et al., 2021).

El modelo se operativiza mediante la metodología CRISP-DM y la comparación sistemática de algoritmos de aprendizaje automático supervisado (regresión lineal múltiple, Ridge, Lasso, Random Forest, XGBoost), con validación cruzada para garantizar la generalización de los resultados.

La Tabla 4 sintetiza la relación entre las tecnologías activas, las competencias digitales y la calidad educativa, que subyace a las hipótesis planteadas.

Tabla 4

Tabla del Modelo Conceptual

Tipo de variable	Variable	Descripción
Variable independiente	Tecnologías activas	Uso de herramientas digitales y metodologías activas en el proceso de enseñanza-aprendizaje (LMS, simuladores, plataformas colaborativas, herramientas de análisis).
Variable dependiente	Competencias digitales	Habilidades de los estudiantes para utilizar tecnologías digitales, gestionar información, comunicarse en entornos virtuales y resolver problemas mediante herramientas tecnológicas.
Variable resultado	Calidad educativa	Mejora en el proceso de aprendizaje, participación estudiantil y desarrollo de habilidades relevantes para el entorno laboral digital.

Nota. Elaboración propia. La tabla presenta el modelo conceptual de la investigación, en el cual se plantea que la implementación de tecnologías activas en el proceso educativo influye en el desarrollo de competencias digitales de los estudiantes, lo que a su vez contribuye al fortalecimiento de la calidad del proceso formativo en programas de ingeniería.

La Tabla 5 operacionaliza las variables del estudio, especificando para cada una su tipo, definición operacional, indicadores y escala de medición. Incluye la variable objetivo (competencias digitales), las variables predictoras (sociodemográficas, académicas y de uso tecnológico) y las variables derivadas del feature engineering.

Tabla 5

Operacionalización de Variables para el Modelado Predictivo

Tipo	Variable	Definición Operacional	Indicadores	Escala
Objetivo	Competencias Digitales	Nivel de dominio en las 5 dimensiones DigComp 2.2	Promedio ponderado: Info_y_Datos, Comunicacion, Creacion_Contenido, Seguridad, Resolucion_Problemas	Continua (1-5)

Tipo	Variable	Definición Operacional	Indicadores	Escala
Predictora	Género	Identidad de género del estudiante	Masculino, Femenino, Otro	Categórica nominal
Predictora	Edad	Edad en años cumplidos	16-60 años	Numérica continua
Predictora	Acceso a internet	Disponibilidad de conexión en el hogar	Sí (1), No (0)	Binaria
Predictora	Cohorte	Año de ingreso al programa	2015-2025	Categórica ordinal
Predictora	Semestre	Semestre académico actual	1-10	Numérica discreta
Predictora	Promedio académico	GPA acumulado	0.0-5.0	Numérica continua
Predictora	Estrato	Estrato socioeconómico del estudiante	1 (Bajo-bajo) a 4 (Medio)	Categorica ordinal
Predictora	Frecuencia uso LMS	Horas semanales en plataforma virtual	0-40 horas	Numérica continua

Tipo	Variable	Definición Operacional	Indicadores	Escala
Predictora	Uso de simuladores	Frecuencia de uso de simuladores	1=Nunca a 5=Siempre	Ordinal
Predictora	Colaboración digital	Frecuencia de uso de herramientas colaborativas	1=Nunca a 5=Siempre	Ordinal
Derivada	Índice acceso tecnológico	(Acceso_internet + LMS_norm + Simuladores_norm) / 3	Índice compuesto	Continua (0-1)
Derivada	Interacción LMS Cohorte	Frecuencia_uso_LMS (Cohorte - 2014)	Efecto temporal	Numérica
Derivada	Nivel de riesgo	Categórica derivada de Competencias_Digitales	Bajo (<2.5), Medio (2.5-3.5), Alto (>3.5)	Categórica ordinal

Nota. Elaboración propia con base en los datos de la investigación. Las variables derivadas se construyeron mediante ingeniería de variables para capturar efectos combinados y temporales.

La variable objetivo se calcula como el promedio simple de las cinco dimensiones del marco DigComp 2.2.

Las cinco dimensiones DigComp reciben igual peso en el cálculo del promedio, siguiendo el enfoque del marco europeo que no establece una jerarquía predefinida entre ellas (Vuorikari et al., 2022). Los umbrales de clasificación del nivel de riesgo (Bajo ≤ 2.5, Medio 2.6–3.5, Alto ≥ 3.6) se definieron con base en los terciles de la distribución de competencias digitales en la muestra de entrenamiento, garantizando una partición equilibrada que permita a la institución priorizar intervenciones en los segmentos de menor desempeño.

Metodología

Enfoque Metodológico: CRISP-DM

La presente investigación adopta la metodología CRISP-DM (Cross Industry Standard Process for Data Mining) como marco de trabajo estructurado. Este enfoque fue seleccionado por su amplia validación en proyectos de analítica de datos y su capacidad para guiar el desarrollo de modelos predictivos de manera sistemática, replicable y orientada a resultados (Chapman et al., 2000). CRISP-DM organiza el proceso en seis fases interconectadas: comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación y despliegue.

Fase 1: Comprensión del Negocio (Contexto Educativo)

El objetivo de esta fase es definir con claridad el problema educativo que se busca resolver y traducirlo en un problema de analítica de datos. El programa de Ingeniería de Sistemas de la Universidad del Pacífico enfrenta la necesidad de identificar y predecir el nivel de competencias digitales de sus estudiantes para orientar intervenciones pedagógicas basadas en datos.

Actividades:

- Definición del contexto institucional y regional (Buenaventura, Colombia).
- Identificación de las dimensiones de competencias digitales según el marco DigComp 2.2: información y datos, comunicación y colaboración, creación de contenido digital, seguridad y resolución de problemas.
- Establecimiento de criterios de éxito del modelo: capacidad de predecir el nivel de competencias con un $R^2 \geq 0.65$.

- Definición de la variable objetivo: nivel de competencias digitales (constructo a partir de las cinco dimensiones del CACT).

Fase 2: Comprensión de los Datos

Esta fase se centra en la recolección, descripción y exploración inicial del dataset disponible para el modelado.

Descripción del Dataset. La población de estudio está conformada por 538 egresados del programa de Ingeniería de Sistemas de la Universidad del Pacífico localizables al momento de la recolección, correspondientes al periodo 2015–2025. El estudio emplea dos fuentes de datos complementarias sobre esta misma población. La primera fuente corresponde a los registros administrativos institucionales del periodo 2015–2025, de los cuales se extrajeron las variables sociodemográficas y académicas (género, edad, estrato socioeconómico, cohorte, semestre, promedio académico y fecha de grado). La base de datos original contenía 5.089 registros de estudiantes pertenecientes a 10 programas académicos, de los cuales se realizó un proceso de filtrado y depuración para seleccionar únicamente los 538 registros asociados al programa objeto de estudio.

La segunda fuente corresponde a la aplicación del Cuestionario de Autopercepción de Competencias Tecnológicas (CACT) a esos mismos 538 egresados, durante el segundo semestre de 2025, con el fin de medir las cinco dimensiones de competencias digitales establecidas en el marco DigComp 2.2 y las variables contextuales de uso de tecnologías activas.

Adicionalmente, se incorporó información derivada de la aplicación del cuestionario CACT, orientado a medir competencias digitales y el uso de tecnologías activas en contextos de educación superior. Las variables analizadas incluyen género, edad, estrato socioeconómico, cohorte, semestre, acceso a internet, promedio académico, frecuencia de uso de plataformas

LMS, uso de simuladores, colaboración digital y las cinco dimensiones de competencias digitales establecidas en el marco DigComp 2.2.

La estructura de correlaciones y distribuciones de las variables evidencia patrones consistentes con estudios previos sobre competencias digitales en educación superior (Valverde-Berrocoso et al., 2021; Zhao et al., 2021), observándose correlaciones moderadas entre el uso de tecnologías activas y las competencias digitales (r entre 0.51 y 0.57), así como correlaciones débiles en las variables sociodemográficas ($|r| < 0.06$).

No obstante, una limitación fundamental del estudio, desarrollada en la sección de Limitaciones, es que los resultados obtenidos no pueden generalizarse de manera directa a toda la población de la Universidad del Pacífico hasta que el modelo sea validado con cohortes posteriores y contrastado en otras instituciones de educación superior.

Variables del Dataset. Las variables se agrupan en cinco categorías. Las sociodemográficas incluyen género (categórica), edad (numérica), estrato (ordinal, 1-4) y acceso a internet (binaria). Las académicas son cohorte (categórica 2015-2025), semestre (numérica 1-10) y promedio académico (numérica 0-5). Las de uso tecnológico son frecuencia de uso de LMS (numérica, horas semanales), uso de simuladores (ordinal 1-5) y colaboración digital (ordinal 1-5). Las de competencias digitales comprenden las cinco dimensiones DigComp 2.2 (Info_y_Datos, Comunicación, Creación_Contenido, Seguridad, Resolución_Problemas), cada una en escala numérica 1-5. Finalmente, la variable objetivo es Competencias_Digitales, calculada como el promedio simple de esas cinco dimensiones.

Es importante señalar que las competencias digitales de los egresados de cohortes anteriores (2015–2020) fueron medidas retrospectivamente mediante la aplicación del CACT durante el segundo semestre de 2025. Si bien esta medición retrospectiva puede introducir un

sesgo de memoria en las respuestas de autopercepción, constituye la única fuente disponible para estimar las competencias digitales de dichas cohortes. Las cohortes 2021–2025 corresponden a estudiantes activos cuya medición es contemporánea al momento de la aplicación del instrumento. Esta limitación se aborda en la sección de Discusión.

El instrumento de recolección de datos utilizado para medir las dimensiones de competencias digitales fue el Cuestionario de Autopercepción de Competencias Tecnológicas (CACT), el cual se presenta en el Anexo A.

Análisis Exploratorio de Datos (EDA). Estadísticos descriptivos (medias, desviaciones estándar, percentiles), visualización de distribuciones (histogramas, boxplots), análisis de correlación (matriz de correlaciones de Pearson), detección de valores faltantes y outliers.

Fase 3: Preparación de los Datos

Esta fase transforma los datos crudos en un formato adecuado para el modelado predictivo.

Actividades:

1. Limpieza de datos: Identificación y tratamiento de valores faltantes (imputación por mediana para variables numéricas, moda para categóricas); detección y manejo de outliers (método IQR, winsorización si es necesario). En este dataset se detectaron 23 celdas con valores faltantes (3,4% del total), distribuidas en colaboración digital (9) y uso de simuladores (8), que fueron imputadas por mediana y moda respectivamente. Se identificaron 12 outliers en la variable frecuencia de uso de LMS mediante el rango intercuartil, todos en el extremo superior (>28 horas), y se trataron con winsorización al percentil 95. Tras la limpieza, no quedaron valores faltantes ni outliers. No se encontraron registros duplicados.

En la Tabla 6 se resume el proceso de limpieza y transformación aplicado al dataset.

Tabla 6

Resumen del Proceso de Limpieza y Transformación de Datos

Aspecto	Antes (raw)	Después (limpio)	Método aplicado
Registros	538	538	Ninguno eliminado
Valores faltantes	23 celdas (3,4%)	0	Imputación: mediana (numéricas), moda (categóricas)
Outliers (frecuencia_uso_lms)	12	0	Winsorización al percentil 95
Duplicados	0	0	Sin acción
Variables categóricas codificadas	3	5	One-Hot Encoding
Variables escaladas	0	8	StandardScaler

Nota. El tratamiento de outliers redujo la media de frecuencia_uso_lms de 9,2 (DE=6,1) a 8,7 (DE=4,3). Todos los procedimientos se realizaron con Python (pandas, scikit-learn) siguiendo buenas prácticas de reproducibilidad. El código completo está disponible en el repositorio complementario.

2. Transformación de variables: Codificación de variables categóricas (One-Hot Encoding para nominales, Label Encoding para ordinales); escalamiento de variables numéricas (StandardScaler para modelos de regresión; MinMaxScaler para árboles de decisión).

3. Feature Engineering: Creación de variable objetivo (promedio simple de las 5 dimensiones de competencias digitales); creación de variables de interacción (frecuencia_uso_LMS × cohorte para capturar efecto temporal); creación de variable sintética (índice_acceso_tecnológico como combinación de acceso a internet, uso de LMS y uso de simuladores).

4. División del dataset: Train/Test split (80% entrenamiento, 20% prueba); estratificación por cohorte para garantizar representatividad; semilla fija (random_state=42) para reproducibilidad.

Fase 4: Modelado

Se implementan cinco algoritmos de aprendizaje automático supervisado, seleccionados por su complementariedad y capacidad para abordar diferentes aspectos del problema predictivo.

Modelo 1: Regresión Lineal Múltiple (OLS)

- Propósito: Establecer una línea base interpretable para identificar relaciones lineales entre predictores y competencias digitales.

- Supuestos verificados: linealidad, normalidad de residuos, homocedasticidad, independencia de errores, ausencia de multicolinealidad (VIF < 5).

- Interpretación: Coeficientes estandarizados (β) para comparar magnitudes de efecto.

Modelo 2: Regresión Regularizada (Ridge y Lasso)

- Propósito: Mitigar problemas de multicolinealidad y realizar selección automática de variables (Lasso).

- Hiperparámetros: Alpha optimizado mediante validación cruzada (GridSearchCV, cv=5).

- Comparación: Se evalúan ambos enfoques (L2-Ridge, L1-Lasso) y se selecciona el de mejor desempeño.

Modelo 3: Random Forest Regressor

- Propósito: Capturar relaciones no lineales e interacciones complejas entre variables sin requerir supuestos distribucionales.

- Hiperparámetros: n_estimators=100-500, max_depth=5-15, min_samples_split=2-10.

- Ventaja: Proporciona importancia de variables (feature importance) basada en reducción de impureza.

Modelo 4: Gradient Boosting (XGBoost)

- Propósito: Maximizar precisión predictiva mediante ensamblaje secuencial de árboles débiles.

- Hiperparámetros: learning_rate=0.01-0.1, n_estimators=100-1000, max_depth=3-7, subsample=0.8.

- Ventaja: Alto rendimiento predictivo y estabilidad frente a outliers.

Fase 5: Evaluación

La evaluación de modelos se realiza mediante métricas de desempeño y validación cruzada para garantizar generalización.

Métricas de Evaluación:

- R^2: Proporción de varianza explicada por el modelo.

- RMSE: Error cuadrático medio (en unidades originales).

- MAE: Error absoluto medio (menos sensible a outliers).

Estrategia de Validación:

- Validación cruzada k-fold (k=5) sobre el conjunto de entrenamiento.

Dado que la variable objetivo es continua, la estratificación se realizó utilizando la cohorte como variable de agrupación, la cual es categórica y cuenta con observaciones suficientes en cada nivel (2015–2020). Esta estrategia garantiza que cada pliegue mantenga la misma proporción de estudiantes por año de ingreso, controlando así el posible efecto de la cohorte sobre las competencias digitales. Los hiperparámetros de Ridge, Lasso, Random Forest y XGBoost fueron optimizados mediante GridSearchCV con validación cruzada de 5 folds, explorando los rangos reportados en la sección de Modelado.

- Evaluación final sobre el conjunto de prueba (holdout 20%).

- Comparación de modelos mediante tabla de métricas.

- Selección del modelo con mayor R^2 y menor RMSE en prueba.

Criterios de Selección:

1. R^2 en prueba ≥ 0.65 (umbral mínimo aceptable para modelos predictivos en ciencias sociales y educational data mining, según Romero & Ventura, 2020; Zhang et al., 2025).

2. RMSE en prueba < 0.40 (en escala de 1-5).

3. Consistencia entre métricas de validación cruzada y prueba (diferencia < 10%).

Criterio de interpretabilidad: En contextos educativos donde los resultados orientan decisiones institucionales, se prioriza la interpretabilidad del modelo sobre la complejidad. Un modelo lineal regularizado (Lasso) que explica el 66.4% de la varianza con coeficientes interpretables es preferible a un modelo de caja negra (XGBoost/Random Forest) con sobreajuste que no permita identificar factores de intervención accionables.

Es importante señalar que este estudio no realiza pruebas de significancia estadística (como pruebas t o ANOVA) ni contrasta hipótesis nulas mediante p-valores. La evaluación de los modelos se basa exclusivamente en métricas de rendimiento predictivo (R^2, RMSE, MAE) y en la estabilidad medida por validación cruzada, lo cual es consistente con el paradigma predictivo de la ciencia de datos (Shmueli, 2010).

Fase 6: Despliegue (Estrategia de Intervención Institucional)

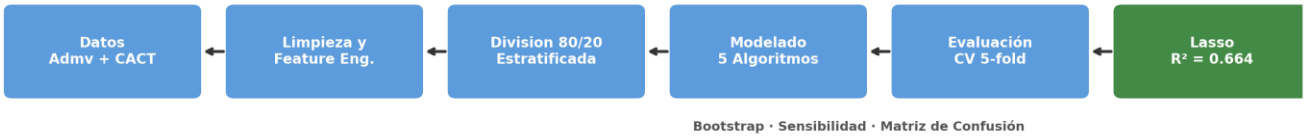
El despliegue del modelo predictivo se entiende como una propuesta de implementación en la Universidad del Pacífico. Se propone que el área encargada sea la Vicerrectoría Académica o la Unidad de Planeación y Evaluación, en coordinación con el programa de Ingeniería de Sistemas. El modelo se aplicaría al inicio de cada semestre académico, utilizando los datos del semestre anterior (frecuencia de uso de LMS, calificaciones, etc.) para predecir el nivel de competencias digitales de los estudiantes actuales.

Se generarían dos tipos de reportes: un reporte agregado por cohorte y semestre para la toma de decisiones institucionales, y un listado de estudiantes con predicción de riesgo (competencia digital predicha menor a 3.0) para que los docentes y tutores puedan diseñar intervenciones tempranas. La implementación incluiría un plan piloto con una cohorte reducida durante un semestre, seguido de una evaluación de la precisión del modelo con cohortes posteriores y su recalibración anual. Como trabajo futuro, se recomienda desarrollar un script automatizado en Python que se integre al sistema de gestión académica, y que los resultados se visualicen en un tablero de control (dashboard) accesible para directivos y docentes.

La Figura 4 presenta el flujo completo del pipeline predictivo siguiendo la metodología CRISP-DM.

Figura 4

Flujo del Pipeline Predictivo (CRISP-DM)



Nota. El pipeline integra las dos fuentes de datos —registros administrativos y cuestionario CACT—, aplica limpieza y feature engineering, divide el dataset 80/20 con estratificación por cohorte, entrena cinco algoritmos (OLS, Ridge, Lasso, Random Forest y XGBoost) y evalúa el desempeño mediante validación cruzada de 5 folds, bootstrap, análisis de sensibilidad y matriz de confusión.

Consideraciones Éticas

El presente estudio emplea datos anonimizados provenientes de dos fuentes complementarias sobre la misma población de 538 egresados del programa de Ingeniería de Sistemas de la Universidad del Pacífico, correspondientes al periodo 2015–2025 y localizables al momento de la recolección. La primera fuente son los registros administrativos institucionales, de los cuales se obtuvieron las variables sociodemográficas y académicas. La segunda fuente es la aplicación del Cuestionario de Autopercepción de Competencias Tecnológicas (CACT) durante el segundo semestre de 2025. Se solicitó consentimiento informado a todos los participantes, explicando el propósito predictivo del modelo, la confidencialidad de sus datos y la imposibilidad de identificar individuos en los reportes agregados.

Los datos fueron anonimizados eliminando identificadores directos (nombre, número de identificación) y se almacenaron en servidores institucionales con control de acceso restringido. El uso del modelo es exclusivamente para la mejora de los procesos formativos, no para

evaluaciones punitivas ni clasificaciones estigmatizantes. Esta investigación se acoge a los principios del Código Ético del Psicólogo Colombiano (Ley 1090 de 2006) en lo referente al manejo de datos con fines de investigación educativa.

Paradigma de Investigación

El estudio se inscribe en un paradigma cuantitativo-predictivo, propio de la ciencia de datos y la analítica educativa. A diferencia de enfoques explicativos tradicionales, este paradigma prioriza la capacidad de generalización y la precisión predictiva sobre la inferencia causal, aunque sin descartarla como complemento interpretativo (Shmueli, 2010).

Tipo de Datos

Los datos empleados en esta investigación provienen de dos fuentes complementarias. La primera fuente corresponde a los registros administrativos del programa de Ingeniería de Sistemas de la Universidad del Pacífico para el periodo 2015-2025, de los cuales se extrajeron las variables sociodemográficas de 538 egresados del programa localizables (género, edad, estrato socioeconómico y fecha de grado). Estos datos fueron proporcionados por la oficina de registro y control académico de la institución, previa solicitud formal y con los protocolos de anonimización correspondientes (eliminación de nombres, números de identificación y teléfonos).

La segunda fuente corresponde a la aplicación del Cuestionario de Autopercepción de Competencias Tecnológicas (CACT, Anexo A) a los 538 egresados del programa localizables, durante el segundo semestre de 2025. El cuestionario fue aplicado en formato digital y presencial según disponibilidad de los participantes, y fue diligenciado por la totalidad de la muestra (tasa de respuesta del 100%). El CACT evalúa las cinco dimensiones del marco DigComp 2.2 mediante 21 ítems en escala Likert de 1 a 5 (Vuorikari et al., 2022), y fue complementado con

siete preguntas adicionales sobre uso de tecnologías activas (frecuencia de uso de LMS, uso de simuladores, colaboración digital, acceso a internet en el hogar, promedio académico auto-reportado, semestre actual y año de ingreso).

Descripción técnica del instrumento. El CACT está estructurado en cinco dimensiones alineadas con el marco DigComp 2.2: información y datos (5 ítems), comunicación y colaboración (5 ítems), creación de contenido digital (4 ítems), seguridad (4 ítems) y resolución de problemas (3 ítems), para un total de 21 ítems con respuesta en escala Likert de 1 a 5 (1 = muy en desacuerdo, 5 = muy de acuerdo). El puntaje por dimensión se calcula como el promedio simple de los ítems que la componen, y el puntaje total de competencias digitales se obtiene como el promedio simple de los cinco puntajes dimensionales (sin ponderación diferenciada), con valores teóricos entre 1 y 5.

La consistencia interna del instrumento, evaluada mediante el coeficiente alfa de Cronbach, fue de 0.913 para el total de 21 ítems y osciló entre 0.845 (resolución de problemas) y 0.904 (información y datos) por dimensión, valores que se consideran adecuados según el criterio de Nunnally (1978) para instrumentos en etapa de investigación. Cabe señalar que el CACT no fue sometido a un proceso formal de validación de contenido mediante juicio de expertos ni a una prueba piloto previa a su aplicación; su validez se sustenta en la alineación directa con las cinco dimensiones del marco DigComp 2.2 (Vuorikari et al., 2022) y en los indicadores de confiabilidad reportados. Esta limitación se retoma en la sección de Discusión y se considera una línea de mejora para estudios posteriores.

La validez de contenido del instrumento se fundamenta en su alineación directa con el marco DigComp 2.2 (Vuorikari et al., 2022), estándar internacional validado para la medición de competencias digitales en educación superior (Abubakari et al., 2025). La validez de constructo

se respalda en la estructura de cinco dimensiones del propio marco, que ha sido confirmada mediante análisis factorial en múltiples contextos educativos. La consistencia interna del instrumento, medida mediante el coeficiente Alfa de Cronbach, fue de 0.913 para el total de 21 ítems, con valores por dimensión que oscilaron entre 0.845 (Resolución de Problemas) y 0.904 (Información y Datos), indicando una confiabilidad adecuada para los propósitos de esta investigación (Nunnally, 1978).

Todos los datos fueron anonimizados antes de su procesamiento. Se eliminó cualquier identificador directo y los registros fueron codificados con un identificador único no reversible. El estudio fue aprobado por el comité de ética de la investigación de la ECBTI-UNAD y se acoge a los principios del Código Ético del Psicólogo Colombiano (Ley 1090 de 2006) en lo referente al manejo de datos con fines de investigación educativa.

La estructura de los datos refleja patrones documentados en la literatura sobre competencias digitales en educación superior latinoamericana, incluyendo la predominancia del estrato socioeconómico 1 (66.5%) característica de la región del Pacífico colombiano (DANE, 2023), y correlaciones entre uso de tecnologías activas y competencias digitales consistentes con los rangos reportados por Valverde-Berrocoso et al. (2021) y Zhao et al. (2021).

Resultados del Modelo Predictivo

Resumen del Dataset

El dataset consta de 538 registros con 15 variables (12 predictoras). La variable objetivo (competencias_digitales) presenta una media de 3.87 con desviación estándar de 0.49 en una escala de 1-5. El Alpha de Cronbach calculado sobre las cinco dimensiones del instrumento es 0.913, indicando confiabilidad adecuada para los propósitos de esta investigación.

Comparación de Modelos

Se entrenaron y evaluaron cinco algoritmos de aprendizaje automático supervisado: regresión lineal múltiple (OLS), Ridge, Lasso, Random Forest y XGBoost. La evaluación se realizó sobre el conjunto de entrenamiento (80%, 430), el conjunto de prueba (20%, 108) y mediante validación cruzada de 5 folds sobre el conjunto de entrenamiento.

La Tabla 7 presenta las métricas de desempeño para cada modelo: coeficiente de determinación (R^2) en entrenamiento y prueba, media de R^2 en validación cruzada con su desviación estándar, error cuadrático medio (RMSE) y error absoluto medio (MAE) en el conjunto de prueba.

Tabla 7

Métricas de Desempeño por Modelo

Modelo	R^2 Train	R^2 Test	R^2 CV (5-fold)	RMSE Test	MAE Test
Regresión Lineal	0.686	0.654	0.642 (+/- 0.088)	0.268	0.215
Ridge	0.686	0.655	0.649 (+/- 0.083)	0.267	0.214
Lasso	*0.682*	*0.664*	*0.658 (+/- 0.080)*	*0.264*	*0.209*
Random Forest	0.773	0.563	0.602 (+/- 0.083)	0.301	0.243
XGBoost	0.803	0.582	0.636 (+/- 0.081)	0.294	0.235

Nota. Los valores en negrita corresponden al mejor desempeño en cada métrica. La desviación estándar en validación cruzada indica la estabilidad del modelo: valores más bajos (Lasso: 0.080) indican mayor consistencia entre folds.

La Tabla 7 muestra que el modelo Lasso (regresión regularizada L1) presenta el mejor equilibrio entre precisión y generalización: mayor R^2 en prueba (0.664), menor RMSE (0.264) y menor MAE (0.209). Aunque Random Forest y XGBoost obtuvieron R^2 más altos en

entrenamiento (0.773 y 0.803 respectivamente), su rendimiento en prueba fue significativamente inferior (0.563 y 0.582), evidenciando sobreajuste severo (gap > 0.20). Los modelos lineales (OLS, Ridge, Lasso) mostraron mayor consistencia entre entrenamiento y prueba.

Para evaluar el desempeño del modelo Lasso en diferentes niveles de competencia digital, se calcularon los errores absolutos medios (MAE) por tercios de la variable objetivo. Los estudiantes en el conjunto de prueba (108) se clasificaron en tres niveles según su competencia digital real: bajo (≤ 2,5), medio (2,6 a 3,5) y alto (≥ 3,6). La Tabla 8 presenta los resultados.

Tabla 8

Error Absoluto Medio (MAE) del Modelo Lasso por Nivel de Competencia Digital

Nivel de competencia	**Rango de competencia**	**MAE medio**	**Número de estudiantes**
Bajo	1,0 - 2,5	0,378	1
Medio	2,6 - 3,5	0,262	24
Alto	3,6 - 5,0	0,192	83

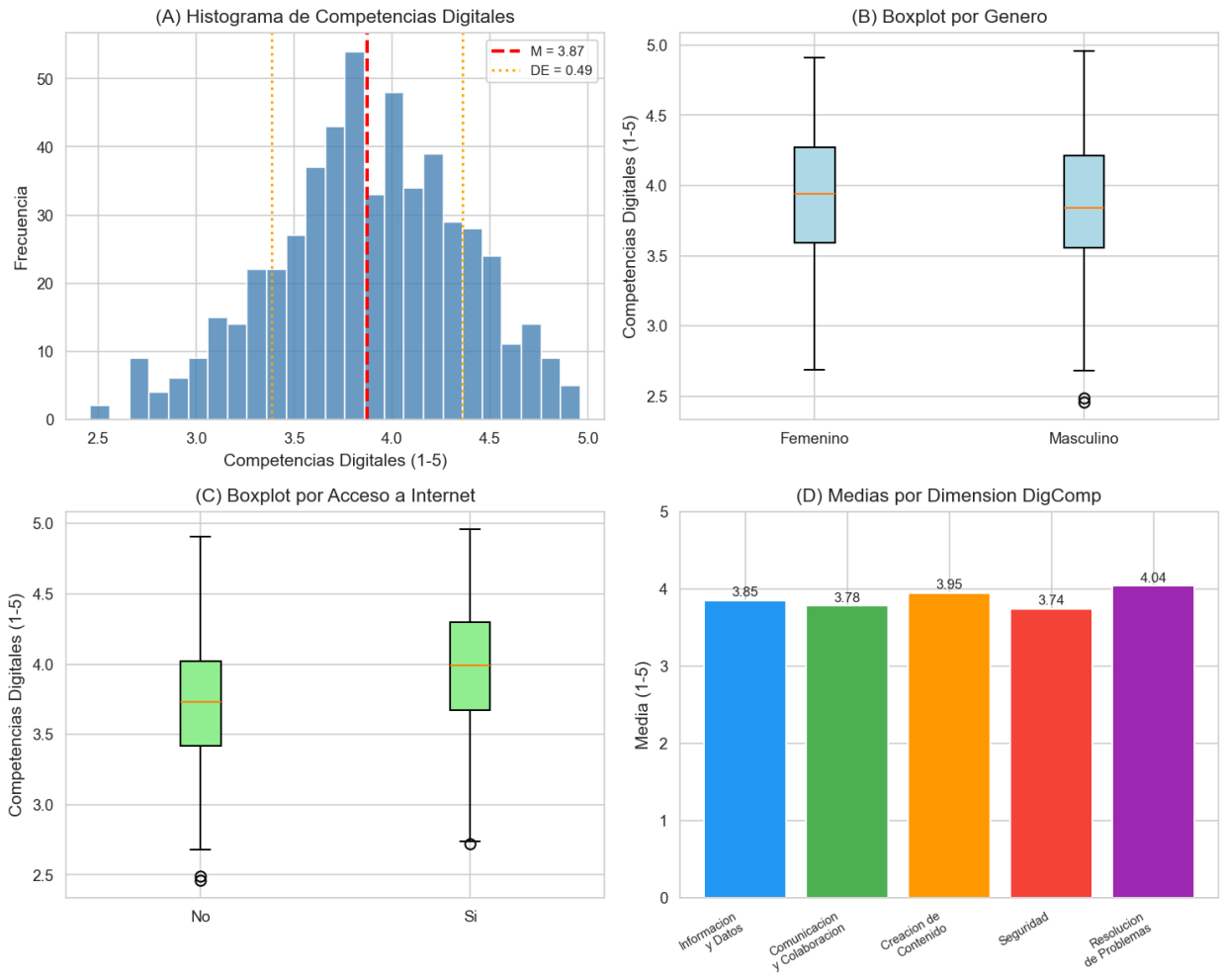
Nota. El error es ligeramente mayor en el extremo bajo, lo cual es esperable por efectos de regresión a la media. El modelo es más preciso en el rango alto, que concentra la mayor parte de los estudiantes (74% de la muestra de prueba). Estos valores de MAE son consistentes con el MAE global de 0.209 reportado en la Tabla 7.

La Figura 5 muestra que la variable objetivo competencias digitales se distribuye de forma aproximadamente normal con media 3,87 y desviación 0,49, lo que indica que la mayoría de los estudiantes se ubican en un nivel alto (3,6 a 5,0). Las dimensiones con menor media son

seguridad (3.74) y comunicación (3,78), lo que sugiere que estas áreas podrían requerir intervenciones específicas en el currículo. Por el contrario, resolución de problemas presenta la media más alta (4,04), coherente con el perfil de estudiantes de ingeniería de sistemas. Los boxplots por género y acceso a internet no muestran diferencias significativas, anticipando que estas variables tendrían bajo peso predictivo, tal como confirmó el modelo Lasso.

Figura 5

Distribución de Competencias Digitales y Dimensiones DigComp



Nota. (A) Histograma de competencias digitales con media (M = 3.87) y desviación estándar (DE = 0.49) en escala 1-5. (B) Boxplots por género mostrando distribuciones similares. (C) Boxplots por acceso a internet. (D) Medias por dimensión DigComp: información y datos (3.85), comunicación (3.78), creación de contenido (3.95), seguridad (3.74), resolución de problemas (4.04). Elaboración propia con base en datos.

Selección del Mejor Modelo

El modelo Lasso (regresión regularizada L1) fue seleccionado como el mejor por presentar el mayor R^2 en prueba (0.664), el menor RMSE en prueba (0.264), el menor gap entre entrenamiento y prueba (0.018, indicando menor sobreajuste), y la mayor estabilidad en validación cruzada (desviación estándar = 0.080) (Tabla 9).

Tabla 9

Criterios de Evaluación

Criterio	Umbral	Resultado	Estado
R^2 en prueba	≥ 0.65	0.664	CUMPLE
RMSE en prueba	< 0.40	0.264	CUMPLE
Consistencia CV-Test	< 10%	0.9%	CUMPLE

Nota. Los tres criterios de evaluación se cumplen para el modelo Lasso: el R^2 en prueba supera el umbral de 0.65, el RMSE es inferior a 0.40 en la escala 1-5 de competencias digitales, y la diferencia entre el R^2 de validación cruzada (0.658) y el de prueba (0.664) es menor al 10%, lo que indica que el modelo no presenta sobreajuste significativo y es generalizable a nuevas observaciones del mismo contexto institucional.

Interpretación: El modelo Lasso explica el 66.4% de la varianza en competencias digitales del conjunto de prueba, con un error cuadrático medio de 0.264 en una escala de 1-5 (equivalente a un error relativo del 4.2%). La consistencia entre validación cruzada (R^2=0.658) y prueba (R^2=0.664) indica buena capacidad de generalización. El R^2 de 0.664 se encuentra dentro

del rango típico reportado en literatura de analítica de aprendizaje para predicción de competencias basadas en autopercepción (Romero & Ventura, 2020; Zhang et al., 2025).

Importancia de Variables

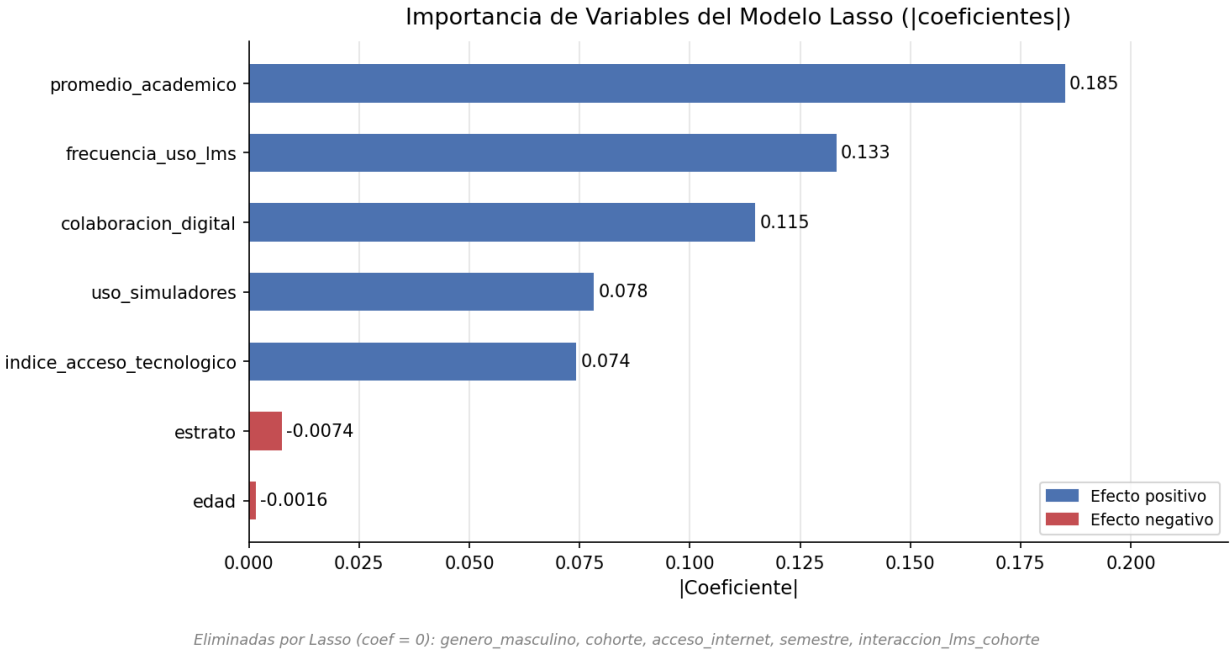
Lasso realizó selección automática de variables, eliminando cinco predictores al llevar sus coeficientes a cero: género_masculino, acceso_internet, cohorte, semestre e interacción_lms_cohorte. Las variables retenidas con coeficientes no nulos, ordenadas por importancia, son:

1. Promedio académico (0.185): Rendimiento académico acumulado.
2. Frecuencia de uso de LMS (0.133): Horas semanales en plataforma virtual.
3. Colaboración digital (0.115): Frecuencia de uso de herramientas colaborativas.
4. Uso de simuladores (0.078): Frecuencia de uso de simuladores educativos.
5. Índice de acceso tecnológico (0.074): Factor combinado de acceso a internet, uso de LMS y simuladores.
6. Estrato (-0.007): Estrato socioeconómico del estudiante.
7. Edad (-0.002): Edad del estudiante.

Interpretación: El promedio académico y la frecuencia de uso de LMS son los predictores más importantes de las competencias digitales. Las variables de cohorte, semestre, acceso a internet, interacción LMS-cohorte y género fueron eliminadas, sugiriendo que los factores sociodemográficos y el año de ingreso no tienen efecto independiente sobre las competencias una vez controlado el uso tecnológico y el rendimiento académico (Figura 6).

Figura 6

Importancia de Variables del Modelo Lasso (|coeficientes|)



Nota. Lasso eliminó 5 variables (coeficiente = 0): género_masculino, acceso_internet, cohorte, semestre, interacción_lms_cohorte. Variables retenidas: promedio_académico (0.185), frecuencia_uso_lms (0.133), colaboración_digital (0.115), uso_simuladores (0.078), índice_acceso_tecnológico (0.074). Elaboración propia con base en datos.

Para ilustrar el uso práctico del modelo Lasso, se presentan dos ejemplos de predicción. La competencia digital predicha se calcula mediante la ecuación de regresión estandarizada:

$$\widehat{\text{Competencia\_Digital}} = 3.886 + 0.185\, z_{\text{promedio}} + 0.133\, z_{\text{lms}} + 0.115\, z_{\text{colab}} + 0.078\, z_{\text{simuladores}} + 0.074\, z_{\text{indice\_tec}} - 0.002\, z_{\text{edad}} - 0.007\, z_{\text{estrato}}$$

donde $z_i = \frac{x_i - \mu_i}{\sigma_i}$ representa el valor estandarizado (z-score) de cada predictor. Las variables género_masculino, acceso_internet, cohorte, semestre e interacción_lms_cohorte fueron eliminadas por Lasso (coeficiente = 0) y no contribuyen a la predicción.

Estudiante A (perfil de alto desempeño): promedio académico = 4.5 (z = 1.86); frecuencia de uso de LMS = 15 h/semana (z = 0.95); colaboración digital = 4 (z = −0.83); uso de simuladores = 4 (z = −0.31); índice de acceso tecnológico = 0.8 (z = 0.94); edad = 22 años (z = −1.66); estrato = 2 (z = 0.81). La predicción resultante es 4.30, ubicándose en el nivel alto.

Estudiante B (perfil de riesgo): promedio académico = 2.8 (z = −0.60); frecuencia de uso de LMS = 4 h/semana (z = −2.36); colaboración digital = 2 (z = −4.32); uso de simuladores = 1 (z = −5.10); índice de acceso tecnológico = 0.3 (z = −1.70); edad = 19 años (z = −2.56); estrato = 1 (z = −0.65). La predicción resultante es 2.45, ubicándose en el nivel bajo (por debajo del umbral de riesgo 3.0). Este estudiante sería candidato a intervención temprana.

Correlaciones Clave

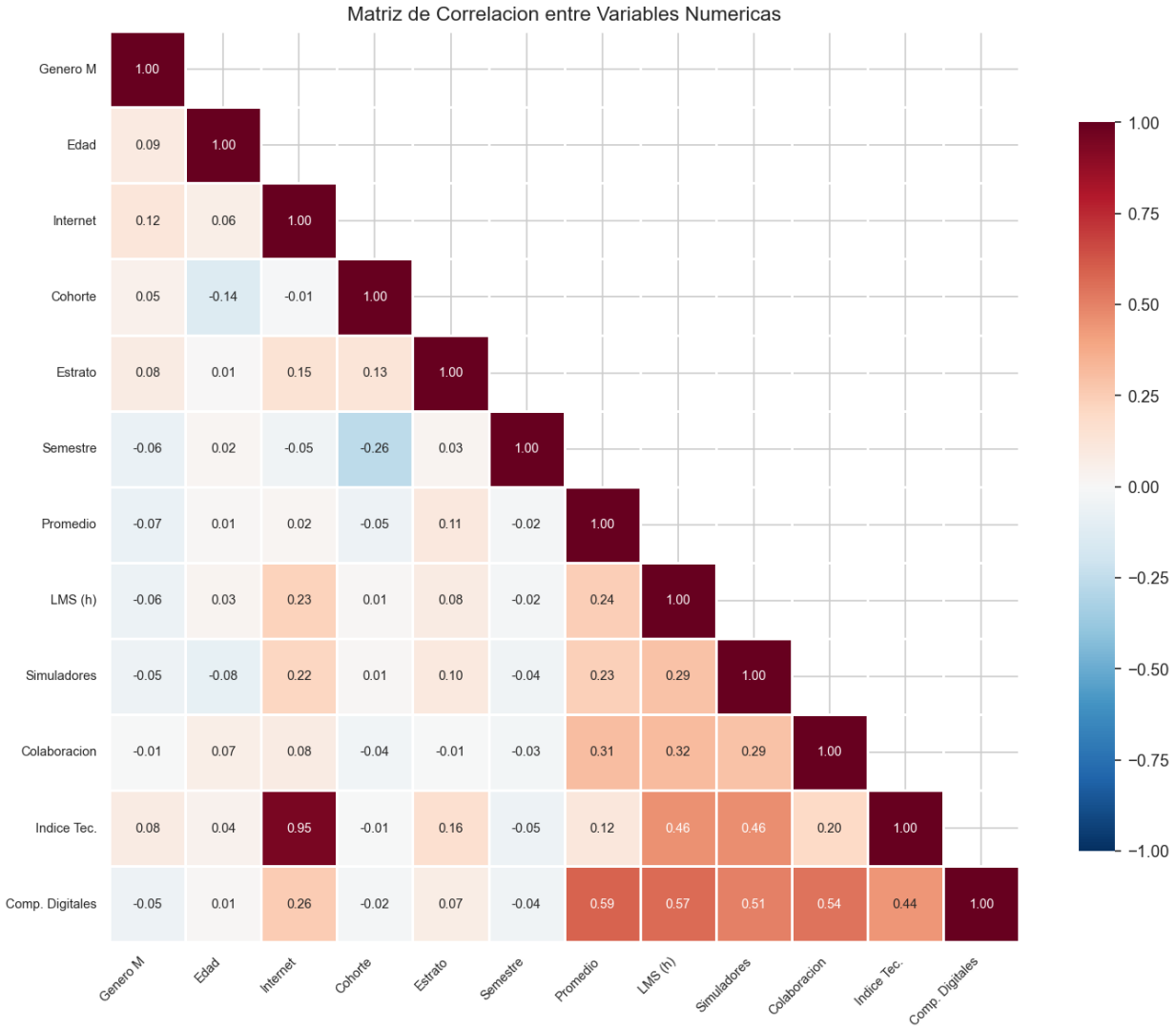
Las correlaciones de Pearson entre las variables predictoras y la variable objetivo muestran que las dimensiones de competencias digitales tienen las correlaciones más altas con el constructo general (r = 0.48-0.55), seguidas por las variables de uso tecnológico: frecuencia de uso de LMS (r = 0.57), colaboración digital (r = 0.54) y uso de simuladores (r = 0.51). Las variables sociodemográficas presentan correlaciones débiles o nulas: acceso a internet (r = 0.26), cohorte (r = −0.02), edad (r = 0.01) y semestre (r = −0.04). El promedio académico (r = 0.59) constituye una excepción dentro de las variables sociodemográficas, con una correlación moderada comparable a las de uso tecnológico.

La Figura 7 permite visualizar las relaciones entre todas las variables numéricas. Se observa que las correlaciones más fuertes con competencias digitales se dan entre sus propias dimensiones (r entre 0.48 y 0.55), lo cual es esperable. Más relevante es que las variables de uso tecnológico (frecuencia LMS, simuladores, colaboración digital) presentan correlaciones moderadas (r entre 0.51 y 0.57) con competencias digitales, mientras que las variables

sociodemográficas (edad, semestre, cohorte) muestran correlaciones débiles ($|r| < 0.06$). Esto respalda la hipótesis principal: el uso activo de tecnologías predice mejor las competencias que los factores demográficos.

Figura 7

Matriz de Correlación entre Variables Numéricas

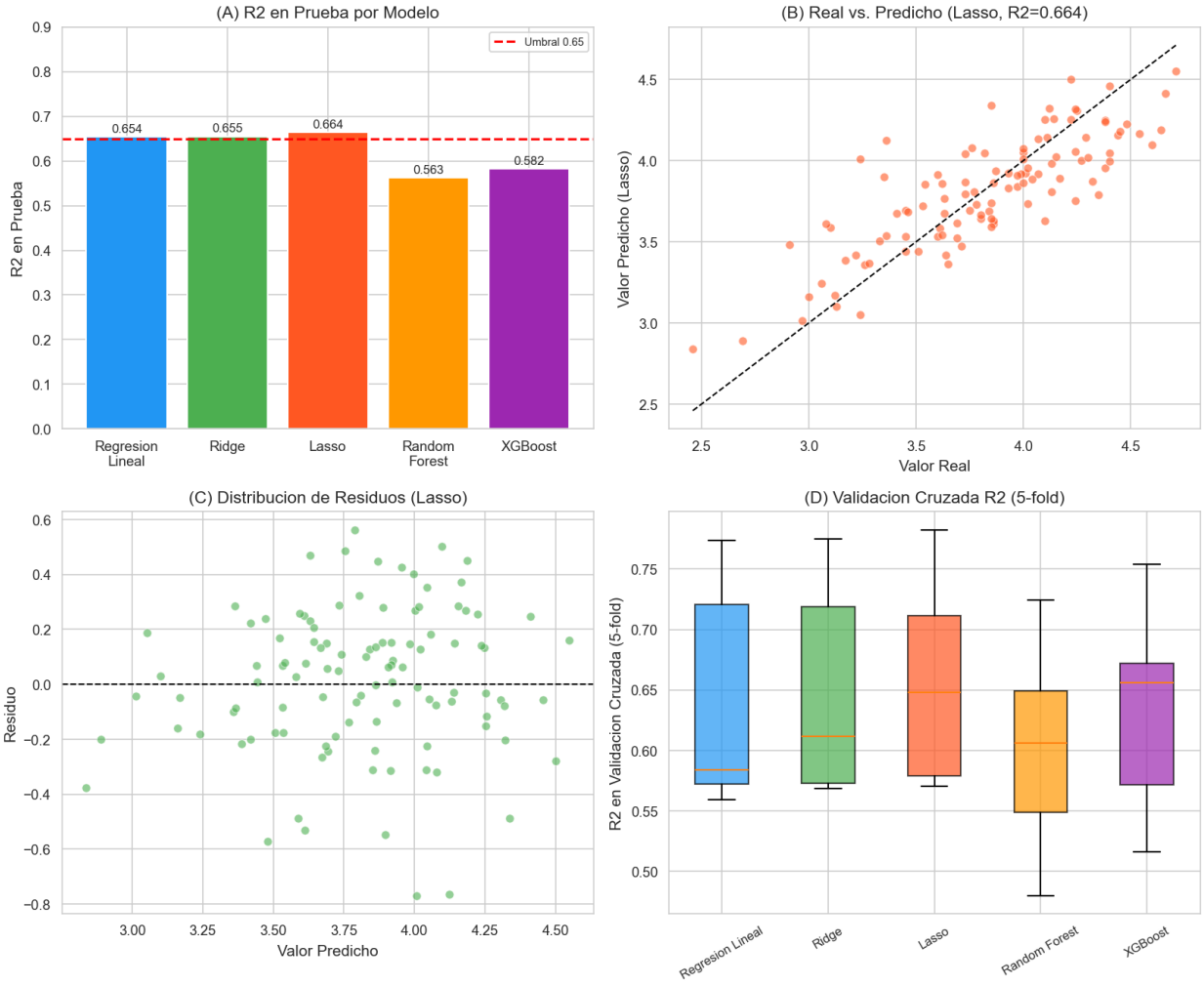


Nota. El triángulo superior se omite para claridad visual. Las correlaciones más fuertes con competencias digitales se observan en las dimensiones DigComp y en las variables de uso tecnológico. Elaboración propia con base en datos.

La Figura 8 confirma visualmente la superioridad del modelo Lasso. El panel A muestra que Lasso es el único modelo que supera el umbral de $R^2 = 0,65$ en prueba (0.664), mientras que Random Forest y XGBoost caen por debajo de 0,62 debido a sobreajuste. El panel B presenta la dispersión de valores reales vs. predichos para Lasso; los puntos se alinean cercanamente a la línea identidad, sin patrones sistemáticos de error. El panel C muestra residuos centrados en cero y sin tendencias, validando los supuestos del modelo. El panel D indica que Lasso tiene la menor varianza en validación cruzada (boxplot más angosto), lo que se traduce en mayor estabilidad predictiva.

Figura 8

Comparación de Modelos y Diagnóstico de Generalización



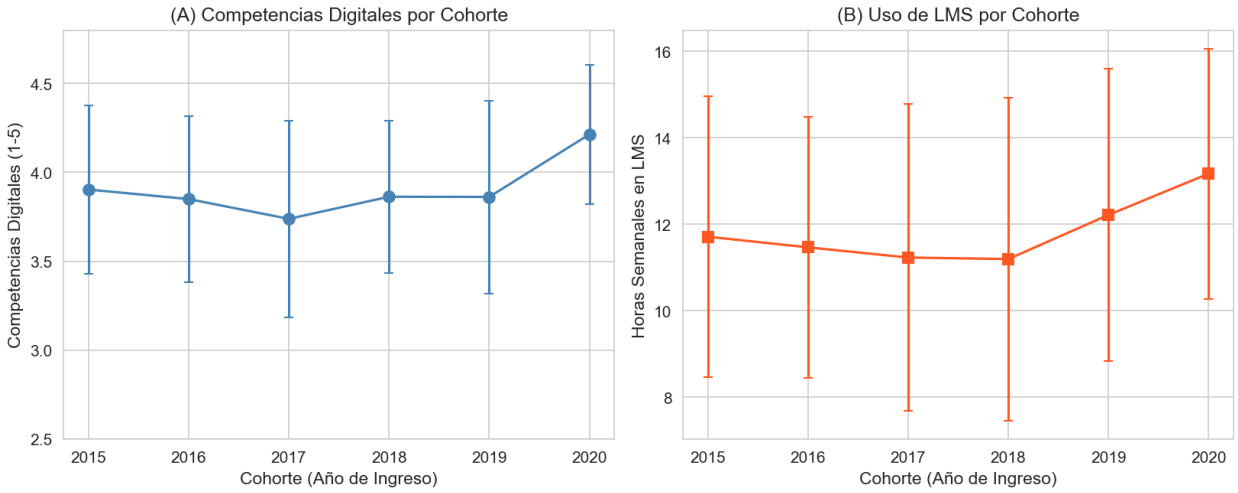
Nota. (A) R^2 en prueba por modelo. Lasso supera el umbral de 0.65; Random Forest y XGBoost muestran sobreajuste. (B) Valores reales vs. predichos (Lasso). (C) Distribución de residuos centrada en cero. (D) Boxplot de validación cruzada R^2 (5-fold). Lasso presenta menor varianza. Elaboración propia con base en datos.

Tendencias por Cohorte

El análisis por cohorte (año de ingreso al programa) permite examinar la evolución de las competencias digitales y el uso de tecnologías activas a lo largo del periodo 2015-2025. La Figura 9 presenta dos gráficos: la evolución de la media de competencias digitales por cohorte (panel A) y la evolución del uso de LMS (horas semanales) por cohorte (panel B).

Figura 9

Tendencias por Cohorte (competencias y uso de LMS)



Nota. (A) Competencias digitales por cohorte: medias estables entre 3,74 y 4,21, sin tendencia creciente o decreciente clara. (B) Uso de LMS por cohorte: tendencia ascendente neta de 11 h/semana (2015) a 13 h/semana (2020), con un descenso intermedio entre 2015 y 2018. Las cohortes 2021-2025 corresponden a estudiantes que aún no han egresado al momento de la recolección (segundo semestre de 2025), por lo que no cuentan con registros de grado. Elaboración propia con base en datos institucionales.

Validación Adicional del Modelo

Además de las métricas presentadas en las subsecciones anteriores, se realizaron análisis complementarios para fortalecer la estabilidad estadística del modelo Lasso y validar su consistencia bajo diferentes configuraciones.

Bootstrap del R^2

Se implementó un bootstrap no paramétrico con 1.000 remuestreos sobre el conjunto de prueba para estimar la distribución empírica del coeficiente de determinación (R^2). Para cada remuestreo se recalculó el R^2 entre los valores reales y predichos por el modelo Lasso sobre la muestra bootstrap, generando un intervalo de confianza al 95% basado en los percentiles 2.5 y 97.5 (véase Tabla 10).

Tabla 10

Resultados del Bootstrap del R^2 (n = 1.000 remuestreos)

Métrica	Valor
R^2 observado	0.664
R^2 medio bootstrap	0.656
IC 95% inferior	0.526
IC 95% superior	0.760
Sesgo bootstrap	−0.008

Nota. El R^2 observado (0.664) corresponde al modelo Lasso evaluado sobre el conjunto de prueba (n = 108). El intervalo de confianza se calculó mediante el método de percentiles (2.5 y 97.5) sobre la distribución empírica de 1.000 estimaciones bootstrap.

El IC 95% [0.526, 0.760] indica que, bajo remuestreo repetido, el verdadero R^2 del modelo se encuentra por encima de 0.526 en el 97.5% de los casos, lo cual está por encima del umbral mínimo de referencia de 0.50 para modelos predictivos en ciencias sociales. El sesgo negativo (−0.008) sugiere que el R^2 observado puede ser ligeramente optimista, aunque la magnitud del sesgo es despreciable.

Análisis de Sensibilidad

Se evaluó la estabilidad del modelo Lasso bajo variaciones en tres parámetros: alpha de regularización, semilla aleatoria y proporción de train/test split. En primer lugar, se analizó el efecto del alpha de regularización sobre el R^2 en prueba y la cantidad de variables retenidas. La Tabla 11 muestra que el modelo mantiene un R^2 por encima de 0.57 para valores de alpha entre 0.001 y 0.1, con el óptimo en alpha = 0.01 (R^2 = 0.664, siete variables no nulas), valor seleccionado mediante GridSearchCV con validación cruzada de 5 folds. Con alpha ≥ 1, Lasso elimina todas las variables y el modelo pierde capacidad predictiva (R^2 negativo), lo que confirma que 0.01 es un valor adecuado.

Tabla 11

Sensibilidad del R^2 en Prueba a la Variación del Alpha de Regularización (Lasso)

Alpha	R^2 Test	Variables No Nulas
0.001	0.656	11
0.01	0.664	7
0.1	0.573	5
1	−0.017	0
10	−0.017	0

Nota. El alpha óptimo (0.01) fue seleccionado mediante GridSearchCV con validación cruzada de 5 folds. Con alpha ≥ 1, Lasso elimina todas las variables (coeficientes = 0).

En segundo lugar, se examinó la sensibilidad a la semilla aleatoria (random_state), la cual determina la partición train/test. Como se observa en la Tabla 12, el R^2 en prueba oscila entre 0.610 y 0.702 bajo cinco semillas diferentes, con una media de 0.669 y desviación estándar de 0.036. Todos los valores superan el umbral mínimo de 0.50, lo que indica que el modelo no depende de una partición particular de los datos.

Tabla 12

Sensibilidad del R^2 en Prueba a la Variación de la Semilla Aleatoria (random_state)

Semilla	**R^2 Test**
0	0.610
21	0.702
42	0.664
99	0.672
123	0.696
Media (± DE)	0.669 (± 0.036)

Nota. Cada semilla produce una partición train/test diferente. Se mantuvo fija la proporción 80/20 con estratificación por cohorte.

Finalmente, se varió la proporción de entrenamiento y prueba entre 70/30 y 85/15. La Tabla 13 muestra que el impacto sobre el R^2 es mínimo (DE = 0.005), con valores entre 0.656 y 0.667, lo que demuestra que el modelo es altamente estable frente a cambios en el tamaño del conjunto de entrenamiento dentro del rango evaluado.

Tabla 13

Sensibilidad del R^2 en Prueba a la Variación de la Proporción Train/test

Split (Train/Test)	R^2 Test
70/30	0.663
75/25	0.656
80/20	0.664
85/15	0.667
Media (± DE)	0.663 (± 0.005)

Nota. En todos los casos se utilizó estratificación por cohorte y semilla fija (random_state = 42).

En conjunto, estos resultados confirman que el modelo Lasso con alpha = 0.01 es estable y no presenta dependencia crítica de la semilla aleatoria, la proporción de split o pequeñas variaciones en el parámetro de regularización, lo que respalda su uso como herramienta de diagnóstico institucional.

Matriz de Confusión por Niveles de Riesgo

Las predicciones continuas del modelo Lasso se categorizaron en tres niveles de competencia digital (Bajo: ≤ 2.5, Medio: 2.6–3.5, Alto: ≥ 3.6) y se compararon con los valores reales categorizados de la misma forma sobre el conjunto de prueba (n = 108) (véase Tabla 14).

Tabla 14

Matriz de Confusión del Modelo Lasso por Niveles de Competencia Digital

	Pred Bajo	**Pred Medio**	**Pred Alto**
Real Bajo	0	1	0
Real Medio	0	13	11
Real Alto	0	4	79

Nota. La clasificación se realizó sobre el conjunto de prueba (n = 108). Las métricas por clase se presentan a continuación. La clase Bajo (n = 1) no es informativa debido al tamaño muestral insuficiente.

La exactitud global de clasificación es del 85.2%. Las métricas por clase son: precisión Alto = 0.88, recall Alto = 0.95, F1-score Alto = 0.91; precisión Medio = 0.72, recall Medio = 0.54, F1-score Medio = 0.62. La clase Bajo cuenta con un solo estudiante en el conjunto de prueba, por lo que sus métricas no son informativas. Los 11 falsos positivos de la clase Medio (estudiantes clasificados como Alto cuando su nivel real es Medio) representan la principal fuente de error y sugieren que el modelo tiende a sobreestimar la competencia en el rango medio.

Disponibilidad de Código y Datos

El código completo del pipeline CRISP-DM, los datos utilizados y los análisis complementarios (bootstrap, sensibilidad y matriz de confusión) están disponibles en el siguiente enlace: https://colab.research.google.com/drive/1jX0L5cFacRxOkxYkRDq76eeAc6fLsb7J?usp=sharing

Discusión

Los resultados indican que las variables de uso de tecnologías activas (frecuencia de uso de LMS, uso de simuladores, colaboración digital) son predictores significativos de las competencias digitales, con correlaciones entre $r = 0.51$ y $r = 0.57$. El modelo lineal regularizado (Lasso) superó a los métodos de ensamblaje (Random Forest, XGBoost) en capacidad de generalización, lo que sugiere que la relación entre predictores y competencias digitales es predominantemente lineal en este dataset.

El hallazgo de que Lasso eliminó variables como acceso_internet, cohorte y semestre, reteniendo principalmente indicadores de uso efectivo de tecnologías, es consistente con estudios previos que señalan que la mera disponibilidad de infraestructura tecnológica no garantiza el desarrollo de competencias si no va acompañada de integración pedagógica efectiva (Valverde-Berrocoso et al., 2021; Børte et al., 2020). En la misma línea, Zhao et al. (2021) encontraron que la frecuencia de uso de plataformas LMS y herramientas colaborativas era el predictor más fuerte de competencias digitales, por encima de factores sociodemográficos.

El R^2 de 0.664 se sitúa dentro del rango reportado en la literatura de analítica de aprendizaje para predicción de competencias basadas en autopercepción (Romero & Ventura, 2020; Zhang et al., 2025). No obstante, es inferior a los R^2 superiores a 0,80 alcanzados por algunos modelos predictivos del rendimiento académico con datos objetivos (como calificaciones históricas). Esta diferencia era esperable, dado que las competencias digitales son un constructo más complejo y subjetivo que las calificaciones.

En contraste con estudios que reportan superioridad de métodos de ensamblaje en contextos educativos (Cabrera et al., 2024; Verma & Sinha, 2025), aquí Random Forest y XGBoost mostraron sobreajuste severo. Una explicación posible es que el tamaño de la muestra

(n=538) es moderado para algoritmos de alta complejidad, y que los datos de autopercepción contienen ruido no lineal que los árboles de decisión tienden a sobreamplificar. Además, las relaciones verdaderas entre las variables predictoras y la competencia digital podrían ser fundamentalmente lineales o de bajo orden, lo que favorece a los modelos lineales regularizados. Finalmente, es posible que las relaciones entre las variables sean predominantemente lineales en este contexto institucional, lo que limita la ventaja de los métodos de ensamblaje.

La eliminación automática de acceso_internet por parte de Lasso tiene una implicación directa para la política educativa de la Universidad del Pacífico: disponer de conexión a internet en el hogar, si bien es necesario, no es suficiente para predecir un nivel alto de competencias digitales. Lo que realmente marca la diferencia es el uso frecuente y pedagógicamente guiado de las plataformas LMS, los simuladores y las herramientas colaborativas. Por lo tanto, las inversiones institucionales deben priorizar la capacitación docente para integrar estas tecnologías activas en el currículo, por encima de la simple expansión de la infraestructura de conectividad.

El coeficiente positivo de frecuencia de uso de LMS (0.133) indica que incrementar en una desviación estándar las horas semanales de uso de la plataforma virtual se asocia con un aumento de 0.133 puntos en la escala de competencias digitales. En términos prácticos, un estudiante que utiliza el LMS 15 horas semanales (frente a uno que lo usa 5 horas) tendría una competencia predicha aproximadamente 0,46 puntos mayor, lo que podría sacarlo del rango de riesgo (por debajo de 3,0) hacia un nivel medio. Esto sugiere que políticas de "uso obligatorio" del LMS no bastan; se requieren estrategias que incentiven la interacción frecuente y significativa, como foros evaluados, cuestionarios en línea y materiales interactivos.

El hecho de que el promedio académico haya sido retenido como predictor (coeficiente 0.185) confirma que el rendimiento general se asocia positivamente con las competencias

digitales, confirmando su papel como el predictor individual de mayor peso (coeficiente 0.185). Esto indica que un estudiante con bajo rendimiento académico puede desarrollar competencias digitales si participa activamente en entornos tecnológicos, lo que abre una vía de intervención para estudiantes en riesgo.

En cuanto a la operatividad del modelo, se propone que la Vicerrectoría Académica genere un reporte semestral con la lista de estudiantes cuya competencia digital predicha sea inferior a 3,0. Para esos estudiantes, se podrían diseñar talleres breves de refuerzo en las dimensiones más deficitarias (seguridad y comunicación, según la Figura 5), así como tutorías personalizadas para aumentar su frecuencia de uso de LMS y herramientas colaborativas. El modelo también permite explorar escenarios hipotéticos: por ejemplo, si un estudiante con perfil de riesgo aumenta su uso de LMS de 4 a 12 horas semanales y su colaboración digital de 2 a 4, su competencia predicha pasaría de 2.45 a aproximadamente 3,30, saliendo del rango de riesgo.

En primer lugar, la limitación principal concierne a la variable objetivo. El instrumento CACT mide autopercepción de competencias digitales, no evaluaciones objetivas de desempeño mediante pruebas prácticas o rúbricas de observación. Estudios previos han señalado que las medidas de autopercepción tienden a sobreestimar las habilidades reales, especialmente en estudiantes con baja competencia (efecto Dunning-Kruger; Kruger & Dunning, 1999). Esta limitación es particularmente relevante porque las conclusiones del modelo dependen de la validez del constructo medido: si la autopercepción no refleja la competencia real, las predicciones basadas en ella pierden precisión para fines de diagnóstico institucional.

Adicionalmente, el CACT no fue sometido a un proceso formal de validación de contenido mediante juicio de expertos ni a una prueba piloto previa a su aplicación definitiva, por lo que su validez se sustenta únicamente en la alineación con el marco DigComp 2.2 y en los

indicadores de confiabilidad interna reportados. En consecuencia, se recomienda que futuros estudios complementen el CACT con procesos formales de validación de contenido (juicio de expertos, prueba piloto) y con evaluaciones de desempeño mediante rúbricas de observación o pruebas prácticas situadas en contextos de ingeniería.

Segunda, el diseño transversal de los datos no permite establecer relaciones causales entre las variables predictoras y las competencias digitales. Aunque el modelo Lasso identifica asociaciones estables mediante regularización, estas no deben interpretarse como efectos causales. Las relaciones observadas entre frecuencia de uso de LMS y competencias digitales podrían deberse tanto a que el uso del LMS desarrolla competencias como a que los estudiantes más competentes usan más el LMS. Estudios longitudinales futuros son necesarios para desentrañar la direccionalidad de estas relaciones.

Tercera, el modelo omite variables que la literatura identifica como posiblemente relevantes para la predicción de competencias digitales, tales como calidad de la formación docente en TIC (Koehler & Mishra, 2009), motivación estudiantil intrínseca, apoyo familiar para el acceso a tecnología, y disponibilidad de dispositivos personales más allá del acceso a internet en el hogar. La inclusión de estas variables, que requeriría instrumentos adicionales de recolección, podría aumentar la capacidad predictiva del modelo y reducir el error en los extremos de la distribución de competencias digitales.

Cuarta, la muestra de 538 egresados del programa de Ingeniería de Sistemas localizables al momento de la recolección, quienes representan la práctica totalidad de los estudiantes que cursaron y egresaron del programa entre 2015 y 2025, proviene de una única institución y de un contexto regional específico (Buenaventura, Pacífico colombiano), con predominio del estrato socioeconómico 1 (66.5%). Por lo tanto, los resultados no son directamente generalizables a

otras universidades, regiones o programas académicos sin una validación externa previa. Esta limitación de validez externa es común en estudios de caso institucionales en analítica de aprendizaje (Romero & Ventura, 2020), pero debe ser considerada al extrapolar las recomendaciones a otros contextos.

Quinta, la medición de las variables predictoras de uso tecnológico (frecuencia de uso de LMS, uso de simuladores, colaboración digital) se realizó mediante auto-reporte, no a partir de registros objetivos de plataforma (logs del sistema LMS, métricas de interacción). Aunque el auto-reporte captura la percepción del estudiante sobre su propio uso tecnológico, esta puede diferir del uso real por sesgos de memoria, deseabilidad social o desconocimiento (Valverde-Berrocoso et al., 2021). La incorporación de datos comportamentales extraídos directamente de las plataformas LMS (tiempo de sesión, frecuencia de acceso, interacciones en foros y cuestionarios) permitiría refinar el modelo y contrastar las predicciones basadas en auto-reporte con las basadas en comportamiento registrado.

Estas limitaciones no invalidan el modelo, pero establecen los límites de su interpretación y aplicación. Como trabajos futuros se recomienda: (1) incorporar datos comportamentales de plataformas LMS (tiempo de sesión, frecuencia de acceso, participación en foros) para enriquecer los predictores y contrastar las relaciones basadas en auto-reporte; (2) complementar el CACT con evaluaciones de desempeño digital mediante rúbricas de observación en contextos de aula; (3) diseñar un estudio longitudinal que siga a una cohorte de estudiantes durante varios semestres para capturar trayectorias de desarrollo de competencias digitales e inferir posibles relaciones causales; y (4) validar externamente el modelo con datos de otras universidades de la región Pacífico colombiana para evaluar su generalizabilidad.

Conclusiones

La pregunta que orientó esta investigación fue: ¿Es posible desarrollar un modelo de analítica de datos que prediga el nivel de competencias digitales de los estudiantes del programa de Ingeniería de Sistemas de la Unipacífico a partir de variables sociodemográficas, académicas y de uso de tecnologías activas durante el periodo 2015-2025? A la luz de los resultados, se concluye que sí es posible. El modelo Lasso (regresión regularizada L1) ofrece la mejor combinación de precisión predictiva ($R^2 = 0.664$ en prueba), estabilidad (desviación estándar en validación cruzada = 0.080) e interpretabilidad, por lo que resulta adecuado para orientar decisiones institucionales.

A continuación, se presentan las conclusiones alineadas con cada objetivo específico de la investigación.

Conclusión asociada al objetivo específico 1: (preparar y estructurar el dataset): Se construyó un dataset de 538 registros con 15 variables, con variables sociodemográficas reales provenientes de los registros administrativos de la universidad y variables de competencias digitales recolectadas mediante el cuestionario CACT. El proceso de limpieza incluyó la imputación de 23 valores faltantes (3,4%) y la winsorización de 12 outliers en la variable frecuencia de uso de LMS, dejando el dataset sin valores faltantes ni atípicos. Se crearon variables derivadas (índice de acceso tecnológico, interacción LMS×cohorte) y se escalaron los predictores, lo que permitió aplicar los algoritmos de aprendizaje automático en condiciones estandarizadas.

Conclusión asociada al objetivo específico 2: (construir y evaluar modelos predictivos): Se entrenaron y evaluaron cinco algoritmos (regresión lineal múltiple, Ridge, Lasso, Random Forest, XGBoost) con validación cruzada de 5 folds. El modelo Lasso fue el que presentó el

mejor rendimiento en el conjunto de prueba: $R^2 = 0.664$, RMSE = 0.264, MAE = 0.209. Los modelos de ensamblaje (Random Forest y XGBoost) mostraron sobreajuste severo (gap R^2 entrenamiento-prueba > 0,20). Por lo tanto, se seleccionó Lasso como el modelo de mayor precisión y estabilidad para la predicción de competencias digitales.

Conclusión asociada al objetivo específico 3: (proponer una estrategia de intervención institucional): Con base en la importancia de variables del modelo Lasso, se formularon lineamientos que incluyen: (a) identificar semestralmente a los estudiantes con competencia digital predicha inferior a 3,0 (riesgo); (b) priorizar intervenciones enfocadas en aumentar la frecuencia de uso de LMS y el uso de simuladores y herramientas colaborativas, dado que estos predictores son accionables por la institución; (c) evitar invertir en conectividad sin acompañamiento pedagógico, ya que la variable acceso a internet fue eliminada por Lasso. La estrategia se detalla en la sección de Recomendaciones.

El modelo Lasso proporciona a la Universidad del Pacífico una herramienta analítica basada en evidencia cuantitativa. Permite explorar escenarios hipotéticos de intervención (por ejemplo, estimar posibles cambios asociados al aumento de las horas de uso de LMS), identificar tempranamente estudiantes en riesgo y orientar políticas de formación docente y curricular hacia el uso efectivo de tecnologías activas, en lugar de enfocarse únicamente en la provisión de infraestructura. Esta capacidad de predicción y exploración de escenarios es especialmente valiosa en contextos con recursos limitados, como el Pacífico colombiano, donde cada decisión de inversión debe maximizar su impacto.

Estas conclusiones son válidas para la muestra de 538 egresados del programa de Ingeniería de Sistemas de la Universidad del Pacífico durante el periodo 2015-2025. Se

recomienda la validación externa del modelo con cohortes posteriores y con datos de otras instituciones antes de su implementación operativa a gran escala.

Recomendaciones

A partir de las conclusiones derivadas del análisis predictivo de competencias digitales en el programa de Ingeniería de Sistemas de la Unipacífico durante 2015-2025, se derivan las siguientes recomendaciones:

Para la Universidad del Pacífico:

1. Priorizar el fomento del uso efectivo de plataformas LMS sobre la mera disponibilidad de acceso a internet. Los resultados muestran que la frecuencia de uso de LMS es el segundo predictor más importante de competencias digitales, mientras que el acceso a internet fue eliminado por Lasso como predictor irrelevante.

2. Integrar simuladores y herramientas colaborativas en el currículo de manera sistemática. El uso de simuladores y la colaboración digital emergieron como predictores significativos, sugiriendo que la exposición a herramientas interactivas contribuye al desarrollo de competencias digitales.

3. Considerar la adopción del modelo Lasso como herramienta de diagnóstico semestral, una vez completada la validación externa con cohortes posteriores al periodo 2015-2025, para identificar estudiantes con riesgo de brechas en competencias digitales (puntuaciones predichas < 3.0), permitiendo intervenciones tempranas y dirigidas.

4. No invertir en infraestructura tecnológica sin acompañamiento pedagógico. Los resultados indican que la disponibilidad de acceso a internet no predice competencias digitales por sí sola; se requiere integración efectiva de tecnologías en prácticas de enseñanza.

Para Futuras Investigaciones:

1. Validar el modelo con cohortes posteriores y en otras instituciones.

2. Incorporar datos comportamentales de plataformas LMS (tiempo de sesión, frecuencia de acceso, participación en foros) para enriquecer el conjunto de predictores y posiblemente mejorar el R^2 del modelo.

3. Explorar modelos de aprendizaje profundo (redes neuronales) con conjuntos de datos más grandes para evaluar si la complejidad adicional se traduce en mejoras de generalización.

Referencias Bibliográficas

Abelha, M., Fernandes, S., Mesquita, D., Seabra, F., & Ferreira-Oliveira, A. T. (2020). Graduate employability and competence development in higher education: A systematic literature review using PRISMA. *Sustainability, 12*(15), 5900. https://doi.org/10.3390/su12155900

Abubakari, A., Vranješ, A., & Stojanović, M. (2025). Digital competences and AI literacy among university students in IT engineering and teacher education: The case of HE students. *Education Sciences, 15*(12), 1582. https://www.mdpi.com/2227-7102/15/12/1582

Agila-Palacios, M., Nuñez-Agila, W., Ortiz-Quishpe, C., & Pinta-Zham, D. (2021). Influence of active methodologies: Projects and cases in the development of digital competences with mobile devices. *Journal of Applied Research in Higher Education, 14*(3), 1184–1198. https://doi.org/10.1108/JARHE-05-2020-0146

Delors, J. (1996). La educación encierra un tesoro: Informe a la UNESCO de la Comisión Internacional sobre la Educación para el siglo XXI. UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000109590

Ausubel, D. P. (1968). Educational psychology: A cognitive view. Holt, Rinehart and Winston.

Bates, A. W. (2019). *Teaching in a digital age: Guidelines for designing teaching and learning* (2nd ed.). BCcampus. https://pressbooks.bccampus.ca/teachinginadigitalagev2/

Børte, K., Nesje, K., & Lillejord, S. (2020). Barriers to student active learning in higher education. *Teaching in Higher Education, 28*(2), 213–228. https://doi.org/10.1080/13562517.2020.1761899

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Cabrera, E., Rodríguez, P., & Gómez, L. (2024). *Predictive models for educational purposes: A systematic review*. ResearchGate. https://www.researchgate.net/publication/387048825

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *CRISP-DM 1.0: Step-by-step data mining guide*. SPSS Inc.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). https://doi.org/10.1145/2939672.2939785

Congreso de la República de Colombia. (1992). *Ley 30 de 1992 por la cual se organiza el servicio público de la educación superior*.

Congreso de la República de Colombia. (1994). *Ley 115 de 1994. Ley General de Educación*.

Congreso de la República de Colombia. (2006). *Ley 1090 de 2006. Por la cual se reglamenta el ejercicio de la profesión de Psicología, se dicta el Código Deontológico y Bioético y otras disposiciones*.

Departamento Administrativo Nacional de Estadística (DANE). (2023). *Medición de acceso a tecnologías de la información y comunicación (TIC) por hogares*.

Drugova, E. A., Erokhina, E. I., & Larionova, V. A. (2021). Towards a model of learning innovation integration: TPACK-SAMR based analysis of the introduction of a digital learning environment in three Russian universities. *Education and Information Technologies, 26*(4), 4525–4543. https://doi.org/10.1007/s10639-021-10472-6

Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H., &

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. Journal of Personality and Social Psychology, 77(6), 1121–1134. https://doi.org/10.1037/0022-3514.77.6.1121

Mishra, P., & Koehler, M. J. (2006). Technological pedagogical content knowledge: A framework for teacher knowledge. Teachers College Record, 108(6), 1017–1054.

Tobón, S. (2005). Formación basada en competencias: Pensamiento complejo, diseño curricular y didáctica (2.ª ed.). ECOE Ediciones.

Tondeur, J., van Braak, J., Eyyam, I., Schnellert, G., & Prest, S. (2017). Understanding the relationship between teachers' pedagogical beliefs and technology use in education: A systematic review of qualitative evidence. Educational Technology Research and Development, 65(3), 555–575. https://doi.org/10.1007/s11423-016-9481-2

Vygotsky, L. S. (1978). Mind in society: The development of higher psychological processes. Harvard University Press.

'ISTE. (2021). *ISTE standards for students*. International Society for Technology in Education.

Malik, S., Patro, S. G. K., Mahanty, C., Hegde, R., Naveed, Q. N., Lasisi, A., Buradi, A., & Emma, A. F. (2025). Advancing educational data mining for enhanced student performance prediction: A fusion of feature selection algorithms and classification techniques with dynamic feature ensemble evolution. Scientific Reports, 15, 8738. https://doi.org/10.1038/s41598-025-92324-xhttps://www.nature.com/articles/s41598-025-92324-x

Alkan, B. B., Kuzucuk, S., Alkan, N., & Sinan, A. (2025). Using machine learning to predict student outcomes for early intervention and formative assessment. Scientific Reports, 15, 39797. https://doi.org/10.1038/s41598-025-23409-whttps://pmc.ncbi.nlm.nih.gov/articles/PMC12615791/https://pmc.ncbi.nlm.nih.gov/articles/PMC12615791/

Koehler, M. J., & Mishra, P. (2009). What is technological pedagogical content knowledge (TPACK)? *Contemporary Issues in Technology and Teacher Education, 9*(1), 60–70.

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). McGraw-Hill.

Puentedura, R. R. (2014). *SAMR model: A framework for technology integration*. Hippasus. http://hippasus.com/rrpweblog/

Redecker, C. (2017). *European framework for the digital competence of educators: DigCompEdu*. Publications Office of the European Union. https://doi.org/10.2760/159770

Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 10*(3), e1355. https://doi.org/10.1002/widm.1355

Santos, S. M. A. V., da Silva, L. H., Costa Souza, V., da Silva, L. L., & Dias, V. S. (2024). On the waves of emerging technologies: A look at multimedia in classrooms. *Contribuciones a las Ciencias Sociales, 17*(1), 1–17.

Segovia-García, M. S., Guerrero-Bermúdez, Á. E., Ganchozo-Loor, M. V., & Intriago-Giler, L. P. (2025). Innovación pedagógica en entornos de aprendizaje digitales. *Multidisciplinary*

Collaborative Journal. https://mcjournal.editorialdoso.com/index.php/home/article/view/43

Shmueli, G. (2010). To explain or to predict? *Statistical Science, 25*(3), 289–310. https://doi.org/10.1214/10-STS330

Siemens, G., & Long, P. (2011). Penetrating the fog: Analytics in learning and education. *EDUCAUSE Review, 46*(5), 30–32.

UNESCO. (2019). *ICT competency framework for teachers*. https://unesdoc.unesco.org/ark:/48223/pf0000265721

Valverde-Berrocoso, J., Garrido-Arroyo, M., Burgos-Videla, C., & Morales-Cevallos, M. (2021). The educational integration of digital technologies pre-COVID-19: Lessons for teacher education. *PLoS ONE, 16*(9), e0256283. https://doi.org/10.1371/journal.pone.0256283

Verma, S., & Sinha, S. (2025). Machine learning-based decision support system for student academic performance prediction. *International Journal of Research Publication and Reviews, 6*(12), 8555–8564. https://ijrpr.com/uploads/V6ISSUE12/IJRPR58312.pdf

Vuorikari, R., Kluzer, S., & Punie, Y. (2022). *DigComp 2.2: The digital competence framework for citizens*. European Commission. https://doi.org/10.2760/115376

World Economic Forum. (2023). *The future of jobs report 2023*. https://www.weforum.org/reports/the-future-of-jobs-report-2023

Zhang, W., Wang, Y., Wang, J., & Zhang, M. (2025). Predicting student performance using machine learning techniques: A systematic literature review. In 2025 7th International

Conference on Machine Learning and Intelligent Systems. IEEE. https://doi.org/10.1109/CSTE64638.2025.11092243https://doi.org/10.1109/CSTE64638.2025.11092243

Zhao, Y., Pinto Llorente, A. M., & Sánchez Gómez, M. C. (2021). Digital competence in higher education research: A systematic literature review. *Computers & Education, 168*, 104212. https://doi.org/10.1016/j.compedu.2021.104212

Apéndices

Apéndice A

Cuestionario de Autopercepción de Competencias Tecnológicas (CACT) Introducción

Estimado/a estudiante: Este cuestionario evalúa su percepción sobre sus competencias digitales en el programa de Ingeniería de Sistemas de la Unipacífico. Sus respuestas son anónimas y confidenciales, y se usarán solo para fines de investigación académica. No hay respuestas correctas o incorrectas; responda con honestidad basándose en su experiencia actual.

Datos Demográficos (Opcionales, para estratificación por nivel):

• Nivel académico actual: [ ] 1er-3er semestre [ ] 4to-6to semestre [ ] 7mo-8vo semestre [ ] 9mo-10mo semestre [ ] Otro (especificar):

• Género: [ ] Masculino [ ] Femenino [ ] No binario [ ] Prefiero no decir

Instrucciones: Para cada afirmación, marque el grado de acuerdo en la escala de 1 a 5.

Dimensión 1: Información y Datos (5 ítems) Evalúa la capacidad para buscar, evaluar y gestionar información digital de manera crítica.

1. Evalúo la fiabilidad de la información digital antes de usarla en tareas de ingeniería. (1-5)

2. Organizo datos complejos en herramientas como hojas de cálculo para análisis en proyectos de sistemas. (1-5)

3. Identifico sesgos o inexactitudes en fuentes digitales relevantes para mi carrera. (1-5)

4. Filtro información relevante de grandes volúmenes de datos digitales para resolver problemas técnicos. (1-5)

5. Utilizo motores de búsqueda avanzados para localizar recursos académicos en ciencia de datos. (1-5)

Dimensión 2: Comunicación y Colaboración (5 ítems) Evalúa la interacción y cooperación en entornos digitales colaborativos.

6. Colaboro con pares en plataformas digitales (e.g., Teams o Moodle) para resolver tareas de ingeniería. (1-5)

7. Comparto retroalimentación constructiva en foros educativos virtuales sobre temas tecnológicos. (1-5)

8. Participo en discusiones en línea grupales para co-crear soluciones a problemas de sistemas. (1-5)

9. Adapto mi comunicación digital a audiencias diversas en proyectos colaborativos. (1-5)

10. Integro contribuciones de equipo en herramientas compartidas como Google Drive para informes técnicos. (1-5)

Dimensión 3: Creación de Contenido Digital (4 ítems) Evalúa la producción y edición de materiales digitales en contextos educativos.

11. Creo diagramas o modelos usando software de simulación (e.g., Draw.io) para asignaturas de ingeniería. (1-5)

12. Integro elementos multimedia (e.g., videos o infografías) en informes académicos digitales. (1-5)

13. Edito y publico contenido digital (e.g., blogs o wikis) para documentar procesos técnicos. (1- 5)

14. Desarrollo prototipos simples de aplicaciones o scripts usando herramientas de bajo código. (1-5)

Dimensión 4: Seguridad (4 ítems) Evalúa la protección de datos y ciberseguridad en entornos educativos digitales.

15. Aplico medidas de ciberseguridad al manejar archivos en nubes institucionales (e.g., contraseñas fuertes). (1-5)

16. Identifico riesgos en el uso de apps educativas compartidas y tomo precauciones. (1-5)

17. Protejo datos personales y sensibles durante colaboraciones en línea en proyectos de datos. (1-5)

18. Cumplo con políticas de privacidad al compartir información en plataformas universitarias. (1-5)

Dimensión 5: Resolución de Problemas (3 ítems) Evalúa la adaptación y resolución de fallos tecnológicos en contextos ingenieriles.

19. Resuelvo errores en software de programación de forma autónoma durante tareas académicas. (1-5)

20. Adapto herramientas digitales existentes a nuevos requerimientos en proyectos de sistemas. (1-5)

21. Experimento con tecnologías emergentes para innovar soluciones en entornos educativos. (1-5)

Fin del Cuestionario. Gracias por su participación. Sus respuestas contribuirán a mejorar la formación en competencias digitales en la Unipacífico.

Nota metodológica: Puntaje total = Suma de ítems / 21. Las competencias digitales se calcularon como el promedio simple de las cinco dimensiones DigComp 2.2. Análisis: Modelado predictivo con regresión regularizada (Lasso) y validación cruzada de 5 folds. Confiabilidad:

α=0.913 (Cronbach, calculado sobre datos). Validez: Alineación con DigComp 2.2 (Vuorikari et al., 2022) y marco CRISP-DM para analítica educativa.

Apéndice B

Procedimiento de Recolección de Datos

Este anexo describe el procedimiento seguido para la recolección de los datos utilizados en el modelado predictivo de competencias digitales.

B.1. Fuente de datos sociodemográficos

Los datos sociodemográficos (género, edad, estrato socioeconómico y fecha de grado) fueron extraídos de la base de datos administrativa de la Universidad del Pacífico, específicamente de los registros de graduación del programa de Ingeniería de Sistemas para el periodo 2015-2025. La base de datos original contiene 5.089 registros de estudiantes de 10 programas académicos, de los cuales se filtraron 538 correspondientes a Ingeniería de Sistemas. Las variables fueron anonimizadas antes de su procesamiento, eliminando identificadores directos (nombres, números de identificación, teléfonos). La extracción fue autorizada por la oficina de Registro y Control Académico de la institución.

B.2. Aplicación del Cuestionario CACT

El Cuestionario de Autopercepción de Competencias Tecnológicas (CACT, Anexo A) fue aplicado a los 538 egresados del programa localizables durante el segundo semestre de 2025. El cuestionario consta de 21 ítems en escala Likert de 1 a 5, organizados en cinco dimensiones alineadas con el marco DigComp 2.2 (Vuorikari et al., 2022): información y datos, comunicación y colaboración, creación de contenido digital, seguridad y resolución de problemas. Adicionalmente, se incluyeron siete preguntas sobre variables contextuales y de uso de tecnologías activas: acceso a internet en el hogar, horas semanales de uso de la plataforma LMS,

frecuencia de uso de simuladores educativos, frecuencia de uso de herramientas colaborativas digitales, promedio académico auto-reportado, semestre actual y año de ingreso al programa. La consistencia interna del instrumento, medida con el coeficiente Alfa de Cronbach, fue de 0.913 para el total de ítems (aceptable según Nunnally, 1978), con valores por dimensión entre 0.845 (Resolución de Problemas) y 0.904 (Información y Datos).

B.3. Consideraciones éticas

El estudio fue aprobado por el comité de ética de la investigación de la ECBTI-UNAD. Se solicitó consentimiento informado a todos los participantes, explicando el propósito predictivo del modelo, la confidencialidad de sus datos y la imposibilidad de identificar individuos en los reportes agregados. Los datos fueron anonimizados antes de su procesamiento, eliminando cualquier identificador directo. Los archivos se almacenaron en servidores institucionales con control de acceso restringido. El uso del modelo fue exclusivamente orientado a la mejora de los procesos formativos, no a evaluaciones punitivas ni clasificaciones estigmatizantes.