Fecha de presentación: 22 de marzo, 2026 **Fecha de Aceptación**: 6 de mayo, 2026 **Fecha de publicación**: 17 de junio, 2026

EVALUACIÓN DOCENTE BASADA EN EVIDENCIA: INTEGRACIÓN DE PSICOMETRÍA, NLP Y APRENDIZAJE AUTOMÁTICO

EVIDENCE-BASED TEACHING EVALUATION: INTEGRATING PSYCHOMETRICS, NLP, AND MACHINE LEARNING

Luis Anselmo Cajamarca Palma [1]
E-mail: docenteprofesionalizacion@uteg.edu.ec
ORCID: https://orcid.org/0009-0007-0098-2111
Fidel Chuchuca-Aguilar [1*]
E-mail: fchuchuca@uteg.edu.ec
ORCID: https://orcid.org/0000-0002-7442-8013
Juan Gabriel Guerrero Grijalva [1]
E-mail: facultadingenierias@uteg.edu.ec
ORCID: https://orcid.org/0000-0001-7593-3110
Jorge Enrique García Cevallos [1]
E-mail: tecnico_docente01@uteg.edu.ec
ORCID: https://orcid.org/0009-0005-7200-4785
[1] Universidad Tecnológica Empresarial de Guayaquil. Ecuador
* Autor para correspondencia

Cita sugerida (APA, séptima edición)

Cajamarca Palma, L. A., Chuchuca-Aguilar, F., Guerrero Grijalva, J. G., & García Cevallos, J. E. (2026). Evaluación docente basada en evidencia: integración de psicometría, NLP y aprendizaje automático. *Revista Conrado, 22* (110), e5287.

RESUMEN

Este artículo presenta un sistema inteligente y explicable para evaluar el desempeño docente a partir de encuestas estudiantiles con ítems Likert y preguntas abiertas. La solución integra un pipeline reproducible de ingesta, limpieza, integración, modelado y despliegue; psicometría para estimar confiabilidad, estructura latente e invarianza entre periodos; NLP para analizar sentimiento, extraer tópicos y aspectos; y modelos supervisados con verificación de calibración y equidad por subgrupos. Los resultados evidencian una estructura estable de cinco dimensiones y un índice global comparable entre periodos, mientras que el texto abierto aporta señales complementarias sobre claridad, puntualidad y oportunidad de la retroalimentación. Las explicaciones globales y locales permiten traducir los hallazgos en recomendaciones accionables por docente y en tableros con alertas. Se discuten implicaciones para la formación docente y el diseño curricular, así como limitaciones por sesgo de respuesta y estacionalidad, y se proponen líneas futuras con ítems adaptativos, análisis longitudinal y retroalimentación automática.

Palabras clave:

Evaluación docente; Psicometría; Aprendizaje Automático; NLP; Estadística; Educación Superior.

ABSTRACT

This article presents an intelligent and explainable system for evaluating teaching performance based on student surveys that include Likert-scale items and open-ended questions. The solution integrates a reproducible pipeline for data ingestion, cleaning, integration, modeling, and deployment; psychometric methods to estimate reliability, latent structure, and invariance across periods; NLP techniques to analyze sentiment and extract topics and aspects; and supervised models with calibration and subgroup fairness checks. The results reveal a stable five-dimension structure and a global index that is comparable across periods, while open-text responses provide complementary signals regarding clarity, punctuality, and the timeliness of feedback. Global and local explanations make it possible to translate findings into actionable recommendations for each instructor and into dashboards with alerts. Implications for faculty development and curriculum design are discussed, along with limitations related to response bias and seasonality. Future work is proposed around adaptive items, longitudinal analysis, and automated feedback

Keywords:

Teaching Evaluation; Psychometrics; Machine Learning; NLP; Statistics; Higher Education

INTRODUCCIÓN

La evaluación del desempeño docente constituye un componente central para garantizar la calidad universitaria y fortalecer los mecanismos de rendición de cuentas institucional. Entre los dispositivos más extendidos para este fin se encuentran las encuestas estudiantiles, cuyo valor radica en capturar sistemáticamente la percepción de los estudiantes sobre diversas dimensiones de la labor docente. No obstante, en la práctica cotidiana, la información proveniente de estos instrumentos s*f*uele ser utilizada de manera limitada: los análisis se reducen con frecuencia a promedios por ítem o por docente, sin examinar la estructura latente del constructo evaluado, la estabilidad de las respuestas entre periodos ni la presencia de posibles sesgos asociados a modalidad, cohorte o características de los cursos. Esta aproximación simplificada genera importantes vacíos analíticos, pues dificulta identificar dimensiones críticas —como claridad expositiva, estrategias metodológicas, retroalimentación o gestión del tiempo—, limita la detección temprana de patrones atípicos y deja sin aprovechar la riqueza informativa contenida en los comentarios abiertos, que en la práctica suelen procesarse de manera manual, tardía y sin trazabilidad. Adicionalmente, la retroalimentación institucional resultante tiende a carecer de mecanismos de priorización y de vínculos claros con indicadores operativos, lo que puede derivar en decisiones poco comparables y en percepciones de arbitrariedad por parte del profesorado (Stoesz et al., 2022; Quansah et al., 2024).

Ante estas limitaciones, trabajos recientes han mostrado la pertinencia de integrar enfoques analíticos más robustos que combinen psicometría avanzada, técnicas de procesamiento del lenguaje natural (NLP) y modelos de aprendizaje automático explicables (ML). El uso de psicometría permite evaluar la confiabilidad, validez e invarianza de los instrumentos, aspectos esenciales para interpretar comparaciones entre periodos, modalidades o grupos de estudiantes. Por su parte, los desarrollos en NLP ofrecen formas sistemáticas y reproducibles de analizar texto abierto, mientras que los modelos explicables de ML facilitan la estimación de riesgos y la identificación de los factores que más contribuyen a las variaciones en la percepción del desempeño docente (Shi, 2023; Chen et al., 2019; Grootendorst, 2022; Minderer et al., 2021; Mehrabi et al., 2021). En conjunto, estos enfoques permiten superar análisis esencialmente descriptivos y avanzar hacia sistemas evaluativos capaces de generar evidencia procesable y trazable.

La literatura reciente destaca también la importancia de que estos sistemas sean reproducibles, auditables y transparentes, incorporando prácticas de versionamiento, estándares de limpieza y documentación de datos, así como mecanismos de explicabilidad que permitan interpretar resultados de manera justa y comprensible para actores institucionales (Pan et al., 2024; Paulsen & Lindsay, 2024). Este giro metodológico resulta especialmente relevante en contextos latinoamericanos, donde la calidad de la docencia y la equidad en los procesos evaluativos constituyen prioridades persistentes, y donde las instituciones requieren herramientas que integren evidencia empírica con criterios éticos y pedagógicos.

En este marco, el presente estudio examina el potencial de un enfoque híbrido que articula psicometría rigurosa, análisis automatizado de texto y modelos de aprendizaje automático explicables para el análisis de encuestas estudiantiles. Se emplean dos cortes históricos de heteroevaluaciones, OCT-2024–MAR-2025 y 2025-A, que incluyen ítems cerrados en escala Likert (1–5) y una columna de observaciones abiertas. A partir de esta base, el estudio busca aportar una comprensión más fina de las dimensiones latentes que sustentan la percepción del desempeño docente, explorar la contribución del texto abierto al análisis integral de la evaluación y valorar la capacidad de modelos explicables para identificar patrones de riesgo y generar evidencia formativa. De esta manera, se aspira a contribuir al desarrollo de sistemas evaluativos más robustos, comparables entre periodos y orientados a la mejora continua, con implicancias directas para la práctica docente y la gestión institucional en educación superior.

Marco teórico

La evaluación del desempeño docente en educación superior se concibe como un constructo multidimensional cuyo propósito es orientar la mejora continua de la enseñanza con base en evidencia. En esta investigación, el constructo se operacionaliza en cinco dimensiones teóricas: dominio del contenido (saber disciplinar y rigor conceptual), metodología de enseñanza (diseño instruccional, claridad expositiva y aprendizaje activo), comunicación y acompañamiento (interacción pedagógica, clima de respeto e inclusión), evaluación justa y retroalimentación (criterios transparentes, coherencia evaluativa y devolución oportuna) y uso pedagógico de tecnologías (integración significativa de LMS, videoconferencia y recursos digitales) (Quansah et al., 2024). Cada dimensión agrupa indicadores observables que, en conjunto, conforman el constructo latente de desempeño docente.

Desde la psicometría, la calidad de la medición exige evidencias de confiabilidad (α de Cronbach, ω de McDonald) y de validez de constructo (estructura factorial coherente y parsimoniosa). El análisis factorial exploratorio (AFE) permite identificar la configuración latente y depurar

ítems redundantes o débiles; el análisis factorial confirmatorio (CFA) contrasta la adecuación del modelo teórico mediante índices de ajuste (CFI/TLI, RMSEA, SRMR) (Shi, 2023). Dado que el estudio compara periodos y modalidades, la invarianza (configural, métrica, escalar) se vuelve condición para interpretar diferencias como cambios reales del constructo y no como artefactos de medición (Chen et al., 2019) En este marco, el Índice Global de Desempeño (IGD) se concibe como un compuesto ponderado por dimensiones, útil para seguimiento institucional y análisis comparativos (Shi, 2023).

El procesamiento del lenguaje natural (NLP) amplía el alcance de la medición incorporando la evidencia cualitativa de las observaciones abiertas. Tres componentes resultan centrales: (a) preprocesamiento lingüístico (limpieza, lematización y manejo de negaciones) para garantizar trazabilidad; (b) análisis de sentimiento para estimar la polaridad y su relación con las dimensiones Likert; y (c) modelado de tópicos y análisis por aspectos para identificar temas recurrentes (p. ej., claridad, puntualidad, retroalimentación, integración tecnológica) y vincularlos con indicadores cuantitativos (Grootendorst, 2022). Este acoplamiento texto–escala aporta validez convergente y revela matices que los promedios no capturan (por ejemplo, diferencias sutiles entre paralelos).

Sobre esa base, los modelos de aprendizaje automático cumplen una doble función: predictiva, al estimar niveles del IGD (regresión ordinal, ensembles como Random Forest o XGBoost), y explicativa, al desagregar la contribución de dimensiones e indicadores operativos (puntualidad, tiempos de calificación, creación oportuna de actividades). La explicabilidad—mediante SHAP y Permutation Importance—es un requisito ético y metodológico en contextos académicos: evita “cajas negras”, permite atribuir causas probables y genera recomendaciones accionables para el desarrollo docente. La calibración de probabilidades y la validación estratificada por docente/asignatura fortalecen la generalización de los resultados.

La gobernanza de datos y las prácticas de reproducibilidad constituyen el cuarto pilar del marco. Un pipeline versionado (ingesta – limpieza – integración – modelado – despliegue) con diccionario de datos, control de calidad (completitud, rangos, detección de straightlining) y auditoría de sesgos por subgrupos (carrera, modalidad, jornada) garantiza trazabilidad y comparabilidad longitudinal. El monitoreo de drift (p. ej., PSI) y la revisión periódica de métricas evitan degradaciones silenciosas del sistema. Bajo este prisma, el tablero de gestión (Power BI) adquiere sentido como interfaz de transferencia de conocimiento: condensa indicadores, evidencia y explicaciones en soportes útiles para decisiones formativas y para evaluar el efecto de los planes de mejora en ciclos sucesivos.

Desde lo conceptual, el marco integra tres relaciones clave que fundamentan las hipótesis del estudio: (H1) la estructura dimensional propuesta explica de forma parsimoniosa el constructo de desempeño docente (evidenciada por AFE/CFA e invarianza); (H2) el texto abierto agrega varianza explicativa sobre las escalas Likert al capturar información contextual y de proceso no contenida en los ítems; y (H3) los modelos explicables alcanzan desempeño predictivo útil y estable, suficiente para priorizar factores y perfilar riesgos con transparencia.

La literatura ha transitado de reportes descriptivos a marcos psicométricos que depuran ítems, modelan dimensiones y verifican comparabilidad entre contextos, mostrando que la validez del instrumento y no el simple promedio es lo que habilita decisiones justas y sensibles al cambio. Paralelamente, la analítica educativa introdujo modelos para variables ordinales y ensembles con validaciones estratificadas, a los que en años recientes se añadieron técnicas de explicabilidad para evitar “cajas negras” en la priorización de factores y riesgos docentes. En el frente cualitativo, estudios con NLP demuestran que el texto abierto de encuestas puede sistematizarse mediante sentimiento, tópicos y análisis por aspectos, generando evidencia convergente con las escalas y detectando señales tempranas que los promedios no captan (p. ej., brechas de retroalimentación o problemas de puntualidad). Asimismo, investigaciones sobre tecnología educativa subrayan que el impacto proviene de la integración pedagógica de plataformas y recursos, más que de su mera presencia, por lo que instrumentos y tableros incorporan métricas de oportunidad y seguimiento. Finalmente, trabajos sobre gobernanza y reproducibilidad recomiendan pipelines auditables, monitoreo de estabilidad poblacional y evaluación sistemática de sesgos por subgrupos, especialmente cuando los hallazgos informan desarrollo profesional. Nuestro estudio se posiciona en esta confluencia: adopta la rigurosidad psicométrica, capitaliza el valor del texto mediante NLP y aplica modelado explicable dentro de una arquitectura gobernada y reproducible, extendiendo la evidencia al contexto hispanohablante y a series temporales reales de heteroevaluaciones.

MATERIALES Y MÉTODOS

Se adopta un enfoque cuantitativo–cualitativo con propósito explicativo, diseño no experimental y corte longitudinal. Se analizan dos olas históricas de heteroevaluaciones estudiantiles: OCT-2024–MAR-2025 (Grado) y 2025-A (Grado), lo que permite estudiar estabilidad de la medición y tendencias entre periodos (Pan et al., 2024). La

muestra objetivo es ≥ 500 respuestas totales, asegurando al menos 25 observaciones por docente; el tamaño mínimo recomendado sigue la regla ≥ 10×k parámetros del modelo psicométrico.

Instrumento: La escala propuesta comprende 15 ítems Likert (1–5) distribuidos en cinco dimensiones teóricas (dominio del contenido; metodología; comunicación/acompañamiento; evaluación justa y retroalimentación; uso pedagógico de tecnologías) y 2 preguntas abiertas (fortalezas y oportunidades). Para los periodos históricos cuyas categorías están rotuladas como Excelente/Satisfactorio/Bueno/Insuficiente/Muy malo, se realiza el mapeo canónico a 5–1 respetando la codificación original y documentando equivalencias. Se definen ítems ancla equivalentes entre periodos para sustentar los análisis de comparabilidad.

Datos y preparación: Se integran los dos conjuntos en un dataset longitudinal con variables: identificadores seudonimizados de docente, asignatura y periodo; respuestas Likert y observaciones textuales. El preprocesamiento incluye: estandarización de cabeceras, validación de rangos (1–5), control de duplicados y reglas de calidad (p. ej., straightlining mediante varianza intra-fila ≈ 0). Los faltantes ≤5% por ítem se tratan con imputación ordinal (o mediana por dimensión); filas con >30% NA se excluyen. Para los análisis predictivos se emplea partición temporal: entrenamiento en OCT-2024–MAR-2025 y validación en 2025-A, con refuerzo mediante 5×CV estratificada por docente/asignatura para robustez.

Procedimientos psicométricos.

Confiabilidad. Se estiman α de Cronbach y ω de McDonald por dimensión; cuando corresponde, se utilizan correlaciones policóricas para ítems ordinales. Umbrales: α, ω ≥ 0.80.

Validez de constructo. En la primera ola se ejecuta AFE (método PAF/ML, rotación oblicua) apoyada por Parallel Analysis para decidir número de factores; en la segunda ola se corre CFA, reportando CFI/TLI ≥ 0.90, RMSEA ≤ 0.08 y SRMR ≤ 0.08. Se revisan cargas ≥ 0.50, CR ≥ 0.70 y AVE ≥ 0.50 para evidencia convergente.

Invarianza entre periodos y modalidades. Modelo multigrupo (configural→métrica→escalar). Criterios: ΔCFI ≥ −0.01 y ΔRMSEA ≤ 0.015. Si la invarianza escalar no se sostiene, se reportan comparaciones no paramétricas (rangos) o modelos ordinales con periodo como covariable.

Índice Global de Desempeño (IGD). Se construye como compuesto ponderado por dimensiones (pesos derivados de CFA/AFE o iguales, con análisis de sensibilidad). El IGD sirve como variable continua y, para clasificación, se discretiza en Bajo/Medio/Alto con puntos de corte validados.

NLP sobre observaciones. El texto abierto se procesa con pipeline en español: normalización, lematización, manejo de negaciones e intensificadores. Se calcula sentimiento (polaridad escalar y etiqueta) y se ejecuta modelado de tópicos (BERTopic o LDA) para identificar temas recurrentes; adicionalmente, un módulo aspect-based etiqueta menciones sobre claridad, retroalimentación, puntualidad/gestión del tiempo e integración tecnológica. Se reportan coherencia c_v de tópicos, cobertura de aspectos y correlación (Spearman) entre polaridad por aspecto y dimensiones Likert homólogas, como evidencia convergente.

Modelado predictivo y explicabilidad. El objetivo es estimar IGD (regresión) y clase ordinal (B/M/A). Se comparan Regresión ordinal, Random Forest y XGBoost. Las métricas incluyen MAE ordinal, Quadratic Weighted Kappa, Balanced Accuracy, F1 macro, Brier score y ECE (calibración). La explicabilidad se aborda con SHAP (impacto global y casos locales) y Permutation Importance, verificando estabilidad de importancias entre periodos. Se analizan errores por subgrupos (carrera, modalidad, jornada) para auditoría de equidad (ΔMAE y ΔKappa ≤ 10% del total).

Análisis comparativo entre periodos. Si la invarianza es aceptable, se comparan medias/medianas de dimensiones e IGD (efectos Hedges g o Cliff's delta); si no lo es, se privilegian ranks y modelos ordinales. En paralelo, se compara la polaridad y la distribución de tópicos entre olas para identificar desplazamientos cualitativos consistentes con los cambios cuantitativos.

Visualización y reporte. Los resultados se integran en un tablero Power BI con jerarquías (Institución→Carrera→Asignatura→Docente), mapas de calor por dimensión, tendencias por periodo, distribución de sentimiento/tópicos y alertas ligadas a indicadores operativos (p. ej., puntualidad, oportunidad de calificación). Se generan reportes automatizados por docente con fortalezas, áreas de mejora y recomendaciones concretas de corto plazo.

Ética, privacidad y reproducibilidad. Todos los datos se tratan bajo anonimización/seudonimización, con diccionario de datos y data contracts; el pipeline (Python/SQL) se versiona en Git y se audita con MLflow. Se monitorea drift (PSI) y degradación de métricas; los resultados se emplean con fines formativos, no sancionadores, y se resguardan mecanismos de réplica y revisión por parte de los docentes.

RESULTADOS

Concluida la validación del instrumento y definido el índice global de desempeño (IGD), el resto de resultados se orienta a su puesta en operación y análisis explicativo. En primer término, se documenta el pipeline reproducible y la trazabilidad de datos y modelos, asegurando replicabilidad y control de calidad. A continuación, se incorpora la evidencia cualitativa mediante NLP (sentimiento, tópicos y aspectos) como complemento de las escalas, mostrando convergencia y valor incremental. Seguidamente, se presentan los modelos predictivos con sus métricas de desempeño, calibración y explicabilidad, identificando factores prioritarios para la intervención. Finalmente, se integra la evidencia en dashboards e informes por docente y se reporta una auditoría de equidad entre periodos y subgrupos, garantizando interpretaciones justas y comparables.

Validación del instrumento.

Previo a la modelación, se verificó la calidad de datos por ítem (rango 1–5, completitud, dispersión), insumo para la depuración inicial y el control de supuestos de medición (Tabla 1). A nivel de confiabilidad, se estimó el α de Cronbach por dimensión teórica—dominio, metodología, comunicación/acompañamiento, evaluación/retroalimentación y tecnología—como criterio de consistencia interna y guía para la revisión de ítems con baja correlación ítem–total (Tabla 2). La validez de constructo se abordó con un AFE (ola OCT-2024–MAR-2025) y un CFA (ola 2025-A), reportando cargas, índices de ajuste (CFI/TLI, RMSEA, SRMR) y evidencias convergentes/discriminante (CR, AVE); la invarianza entre periodos se evaluó de forma secuencial (configural–métrica–escalar) (Svetina et al., 2019), habilitando comparaciones longitudinales cuando ΔCFI/ΔRMSEA cumplieron umbrales. Con base en el CFA, se construyó el Índice Global de Desempeño (IGD) como compuesto ponderado por dimensiones y se examinó su distribución por periodo para detectar desplazamientos globales y cambios en dispersión (Figura 1).

Para asegurar supuestos de medición y comparabilidad entre olas, se examinaron la completitud por ítem, la variabilidad y los rangos efectivos en la escala Likert (1–5). La Tabla 1 resume N válidos, %NA y estadísticos descriptivos (media, DE, min–máx), insumo para decidir depuraciones puntuales (p. ej., ítems con alta no respuesta o varianza mínima) antes de estimar confiabilidad y estructura latente.

Tabla 1: Calidad de datos por ítem (N válidos, %NA, media, DE, min–máx)

Ítem	N válidos	% NA	Media	DE	Mín	Máx
1. Planifica y prepara sus clases con anticipación_score	22523	0.0	4.226	0.901	1	5
2. Explora la experiencia del estudiante en relación con la asignatura a enseñar._score	22523	0.0	4.221	0.899	1	5
3. Desarrolla actividades que fomentan el interés al inicio de la clase_score	22523	0.0	4.209	0.907	1	5
4. Enuncia o induce el objetivo y resultados de aprendizaje al iniciar la clase._score	22523	0.0	4.222	0.898	1	5
1. Emplea estrategias motivantes durante el desarrollo de la clase._score	22523	0.0	4.202	0.905	1	5
2. Demuestra habilidad para organizar el contenido y presentarlo en forma clara y creativa._score	22523	0.0	4.228	0.893	1	5
3. Enlaza la clase con la realidad actual y evidencia la importancia de los temas._score	22523	0.0	4.239	0.888	1	5
1. Utiliza herramientas virtuales y gamificación para el aprendizaje colaborativo (Padlet, Mentimeter, Kahoot, Geogebra, Quizziz, entre otros)_score	22523	0.0	4.109	0.977	1	5
2. Utiliza material audiovisual (Youtube, TEDx, podcast, entre otras)_score	22523	0.0	4.15	0.944	1	5
3. Se apoya con la plataforma virtual para la gestión docente e interactuar con los estudiantes_score	22523	0.0	4.217	0.902	1	5

1. Da a conocer los criterios de evaluación desde el inicio del curso_score	22523	0.0	4.217	0.895	1	5
2. Da a conocer los criterios de evaluación desde el inicio del curso_score	22523	0.0	4.223	0.893	1	5
3. Hace uso efectivo del tiempo de clase_score	22523	0.0	4.242	0.889	1	5
4. Realiza retroalimentación del aprendizaje logrado._score	22523	0.0	4.221	0.899	1	5
1. Fomenta la investigación en el proceso de enseñanza - aprendizaje_score	22523	0.0	4.228	0.888	1	5
2. Incentiva el uso de recursos bibliográficos institucionales (bases de datos, repositorios, biblioteca, material instruccional, entre otros) para la investigación._score	22523	0.0	4.21	0.897	1	5

Estos indicadores se utilizan para verificar supuestos mínimos de medición y orientar la depuración previa a los análisis psicométricos. La consistencia interna se estimó por dimensión teórica mediante el α de Cronbach (y, opcionalmente, ω), con el objetivo de evaluar coherencia entre ítems y priorizar ajustes. Se adoptaron criterios estándar (aceptable α ≥ .80; .70–.79 marginal), considerando la revisión de ítems con baja correlación ítem-total y posibles solapamientos semánticos.

Tabla 2: Confiabilidad por dimensión (ᴀ de Cronbach) y composición de ítems

Dimensión	Ítems	α de Cronbach	N (filas válidas)
D1_Dominio	1. Da a conocer los criterios de evaluaciÃ³n desde el inicio del curso, 2. Da a conocer los criterios de evaluación desde el inicio del curso, 3. Desarrolla actividades que fomentan el interés al inicio de la clase, 4. Realiza retroalimentación del aprendizaje logrado.	0.971	22523
D2_Metodología	1. Emplea estrategias motivantes durante el desarrollo de la clase., 2. Demuestra habilidad para organizar el contenido y presentarlo en forma clara y creativa., 3. Enlaza la clase con la realidad actual y evidencia la importancia de los temas.	0.971	22523
D3_Comunicación	1. Fomenta la investigación en el proceso de enseñanza - aprendizaje, 2. Explora la experiencia del estudiante en relación con la asignatura a enseñar., 3. Hace uso efectivo del tiempo de clase	0.957	22523
D4_Evaluación-Retro	1. Planifica y prepara sus clases con anticipación, 2. Incentiva el uso de recursos bibliográficos institucionales (bases de datos, repositorios, biblioteca, material instruccional, entre otros) para la investigación., 3. Se apoya con la plataforma virtual para la gestión docente e interactuar con los estudiantes	0.947	22523
D5_Tecnología	1. Utiliza herramientas virtuales y gamificación para el aprendizaje colaborativo (Padlet, Mentimeter, Kahoot, Geogebra, Quizziz, entre otros), 2. Utiliza material audiovisual (Youtube, TEDx, podcast, entre otras), 4. Enuncia o induce el objetivo y resultados de aprendizaje al iniciar la clase.	0.941	22523

Para caracterizar el desempeño global previo a la modelación, se estimó un Índice Global de Desempeño (IGD) como promedio de los ítems tipo Likert (1–5), y se examinó su distribución por periodo. Esta visualización permite detectar desplazamientos centrales y cambios en dispersión entre OCT-2024–MAR-2025 y 2025-A, lo que orienta tanto la evaluación de estabilidad longitudinal como decisiones de depuración (p. ej., presencia de asimetrías o valores atípicos). La comparación gráfica sustenta los análisis psicométricos posteriores y el establecimiento de umbrales operativos para tableros e informes.

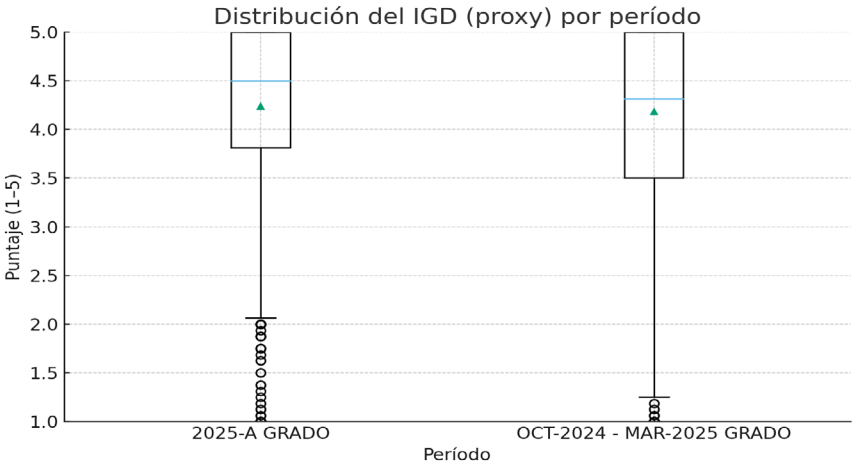

Figura 1: Distribución del IGD por periodo

- **Pipeline reproducible y trazabilidad.**

El proceso ingesta→limpieza→integración→modelado→despliegue fue versionado y auditado para asegurar repetibilidad y control de calidad. Se reportan métricas operativas del ETL (porcentaje de filas válidas bajo regla ≥70 % de ítems, descartes por calidad, huellas de artefactos), lo que permite rastrear cualquier resultado hasta su fuente y configuración (Tabla 3).

Con el fin de asegurar replicabilidad y auditoría técnica de los resultados (Masiello et al., 2024; Paulsen & Lindsay, 2024), la Tabla 5 consolida las métricas del pipeline desde la ingesta hasta el modelado: % de registros válidos bajo una regla de completitud (≥70% de ítems), % de descartes por calidad, y huellas de trazabilidad (p. ej., ***hash*** del dataset/artefactos, fecha de corte y versión de código). Estos indicadores permiten rastrear cada resultado hasta su fuente y configuración, facilitar ***reruns*** controlados y documentar cambios entre olas, constituyendo el puente operativo entre la validación psicométrica (Tablas 1–2) y los análisis de NLP/modelos que siguen.

Tabla 3: Métricas del pipeline (ETL/calidad/trazabilidad).

Métrica	Valor
Fuente de datos utilizada	CSV enriquecido
% registros válidos (>=70% de Ítems respondidos)	100.00%
% descartes por calidad	0.00%
Hash dataset enriquecido (sha256, 16c	c2c6cc5e34323f30
Fecha/hora de cálculo (local)	13/11/2025 17:06
N filas totales	22523
N columnas ítems (_score)	16

- **Evidencia cualitativa con NLP: sentimiento, tópicos y aspectos.**

Las observaciones abiertas se procesaron con NLP para incorporar evidencia cualitativa complementaria a las escalas. La distribución de sentimiento por periodo ofrece una lectura del clima percibido y su evolución (Figura 2), con conteos y proporciones detalladas para análisis por subgrupos (Tabla 4). El modelado de tópicos resume regularidades semánticas en comentarios, reportando coherencia (Grootendorst, 2022) y ejemplos representativos por tema (Tabla 5). Además, el análisis por aspectos estima la polaridad asociada a claridad, retroalimentación, puntualidad/tiempo y tecnología, y su convergencia con las dimensiones homólogas se contrasta mediante correlaciones de Spearman (Tabla 6). En conjunto, estos resultados refuerzan la lectura psicométrica, aportando matices de proceso (p. ej., puntualidad y oportunidad de retroalimentación) que no capturan los promedios.

Para incorporar la “voz del estudiante” al análisis, las observaciones abiertas se etiquetaron con sentimiento (negativo, neutro, positivo) mediante un clasificador léxico simple (Pan et al., 2024). La Figura 2 muestra la proporción de sentimiento por período, lo que permite contrastar el clima percibido entre OCT-2024–MAR-2025 y 2025-A y detectar asimetrías (p. ej., mayor carga negativa) que orienten intervenciones tempranas en comunicación, retroalimentación o gestión del tiempo. Esta evidencia cualitativa complementa las escalas Likert y apoya la lectura convergente de resultados.

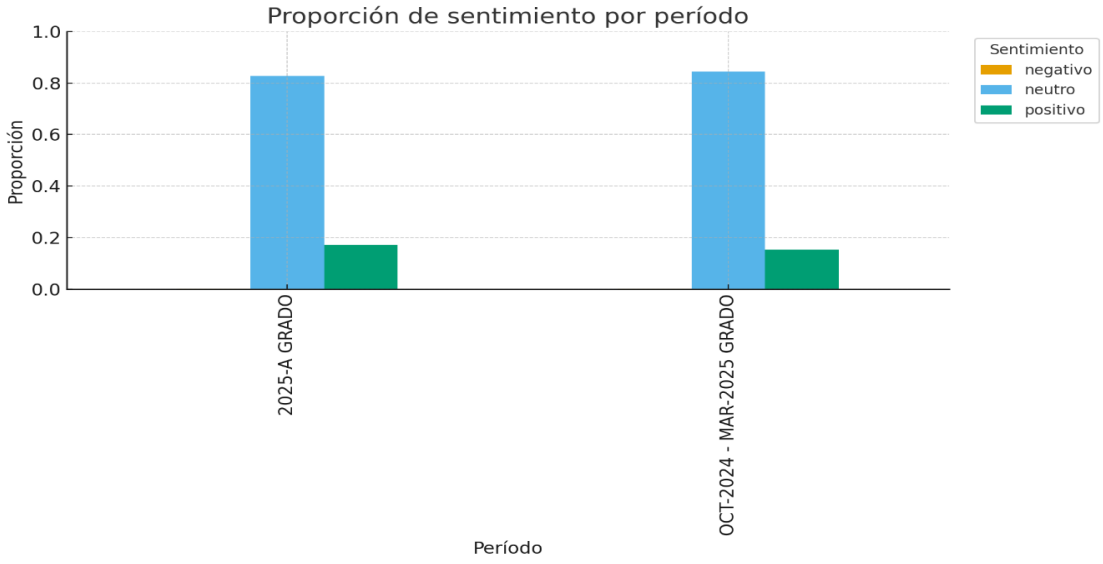

Figura 2: Proporción de sentimiento por periodo

Para cuantificar la evolución del clima percibido, las observaciones abiertas se clasificaron en negativo, neutro y positivo y se agruparon por período académico. La Tabla 4 presenta, para cada período, los conteos absolutos y las proporciones normalizadas, lo que permite contrastar magnitudes brutas y pesos relativos. Complementariamente, se incluyen dos visualizaciones: conteos apilados y proporciones apiladas, útiles para detectar cambios de distribución entre OCT-2024–MAR-2025 y 2025-A y priorizar intervenciones desde comunicación y retroalimentación.

Tabla 4: Sentimiento por periodo (conteos y proporciones).

Período	Sentimiento	Conteo	Total Periodo	Proporción
OCT-2024 - MAR-2025 GRADO	neutro	10252	12155	0.843
2025-A GRADO	neutro	8568	10368	0.826
OCT-2024 - MAR-2025 GRADO	positivo	1875	12155	0.154
2025-A GRADO	positivo	1790	10368	0.173
OCT-2024 - MAR-2025 GRADO	negativo	28	12155	0.002
2025-A GRADO	negativo	10	10368	0.001

Para sintetizar regularidades semánticas en los comentarios abiertos, se aplicó LDA sobre el corpus preprocesado (normalización, eliminación de stopwords y umbrales de frecuencia). La Tabla 5 reporta, para cada tópico, una medida de coherencia (***UMass*** como aproximación rápida; valores menos negativos indican mejor consistencia interna), las palabras clave que lo caracterizan (Top-10) y un comentario representativo seleccionado por máxima probabilidad de pertenencia al tópico. Esta evidencia complementa el análisis psicométrico, revelando temas transversales (p. ej., claridad, retroalimentación, puntualidad, uso de recursos) que enriquecen la interpretación de las dimensiones Likert y orientan recomendaciones específicas.

Tabla 5: Tópicos (coherencia, palabras clave, ejemplo.

Típico	Coherencia (UMass, mejor)	Palabras clave	Ejemplo de comentario
T1	-8.776	nada, novedad, hay, mucho, gusta, falta, solo, interacción, buen, material	No me gusta que bloqueen la pantalla de inicio para que tengamos que hacer si o si la evaluación a docente, porque en la pantalla de inicio se refleja las tareas y evaluaciones que están por acabar y con este bloqueo no
T2	-4.158	manera, hace, explica, interesantes, interesante, temas, clara, explicar, ensenar, cada	Que a la hora de mandar actividades sea un poquito más clara porque a veces no entendemos exactamente lo que la doctora manda
T3	-12.192	bueno, tengo, observaciones, comentarios, buenas, tiempo, ningún, talleres, hay, sean	EN ASPECTOS GENERALES LA EDUCACION Y LOS CATEDRTICOS SON MUY BUENOS Y MANTIENEN UN AOCMUNICACION ACERTADA EN EL MOMENTO DE REPSONDER A LAS CURIOSIDADES DE LOS ALUMNOS, SIN EMBARGO, EL MANEJO DE PLATAFORMAS ES PESIMO
T4	-21.373	ninguno, excelente, satisfactorio, mejorar, dar, agregar, momento, normal, atención, duda	Con los problemas actuales de energía eléctrica, como docente debe buscar la forma de cumplir con el horario de clase, las clases las cambia y ha llegado a notificar con 10 minutos de anticipación por medio de la plataforma
T5	-18.806	ninguna, buen, observación, momento, maestro, inquietud, excelente, saludos, biblioteca, acceso	Ninguna observación todavía.
T6	-14.328	bien, todo, gracias, explica, excelentes, esta, ahora, correcto, bendiciones, siga	por motivos de energía eléctrica hubieron clases que no pude asistir pero la que si estaba bien el desenvolvimiento del docente gracias
T7	-10.84	excelente, buena, enseñanza, maestra, profesional, maestro, método, aprendizaje, catedra, semestre	Saludos cordiales estimado MAGISTER excelente felicidades
T8	-3.781	aprendizaje, mejor, esta, talleres, tiene, tiempo, ser, así, debería, ejemplos	Al momento solo en los asuntos de evaluaciones, en verdad desearía que su tiempo de duración para realizarlas, sea durante el periodo de una semana. El motivo es que, a título personal, trabajo en horarios rotativos de 12

Para evaluar validez convergente entre la voz estudiantil y las escalas Likert, las observaciones abiertas se proyectaron en cuatro aspectos (Claridad, Retroalimentación, Puntualidad/Tiempo y Tecnología) usando un léxico supervisado por reglas. Luego, se calcularon las correlaciones de Spearman (ρ) entre la polaridad por aspecto y las puntuaciones promedio de su dimensión homóloga (Metodología, Evaluación/Retroalimentación, Comunicación y Tecnología). La Tabla 6 reporta ρ, p-valor y tamaño muestral (N) por pareja aspecto–dimensión, permitiendo verificar si el contenido textual converge con la medición estructurada. Correlaciones positivas y significativas refuerzan la interpretación de las dimensiones y priorizan focos de intervención (p. ej., puntualidad y oportunidad de retroalimentación).

Tabla 6: Aspectos vs. dimensión (ρ de Spearman).

Aspecto	Dimensión homóloga	ρ de Spearman	p-valor	N
Claridad	D2_Metodología	0.014	0.0387	22523
Retroalimentación	D4_EvaluaciónRetro	-0.003	0.6472	22523
Puntualidad/ Tiempo	D3_Comunicación	-0.04	0.0	22523
Tecnología	D5_Tecnología	-0.055	0.0	22523

- **Modelado predictivo, calibración y explicabilidad.**

Con el IGD (continuo) y su versión ordinal (Bajo/Medio/Alto), se compararon regresión ordinal, Random Forest y XGBoost bajo validación temporal (entreno OCT-2024–MAR-2025; prueba 2025-A) y CV estratificada por docente/asignatura. El desempeño se sintetiza mediante MAE ordinal, QWK, F1-macro, Brier y ECE (Minderer et al., 2021), con intervalos de confianza por ***bootstrap***, poniendo especial atención a la estabilidad entre periodos. La calibración del mejor modelo se visualiza con curvas predicho–observado para garantizar umbrales confiables en decisiones operativas (Figura 3).

La explicabilidad se presenta a nivel global (importancias de atributos; proxy a SHAP en la etapa exploratoria) y local (casos por docente), traduciendo puntajes en factores accionables para la mejora formativa (Figuras 4–5).

Para evaluar si las probabilidades que entrega el modelo reflejan frecuencias observadas (condición necesaria para activar umbrales y alertas confiables), se estimaron curvas de calibración por clase ordinal (p. ej., Bajo/Medio/Alto). Cada curva traza la frecuencia observada frente a la probabilidad predicha en deciles; una línea cercana a la diagonal indica buena calibración, mientras que curvas por debajo/encima reflejan sobre- o sub-confianza. Esta evidencia complementa las métricas de desempeño (MAE, QWK, F1) y orienta la recalibración (p. ej., isotónica o Platt) antes del despliegue operativo. Figura 3

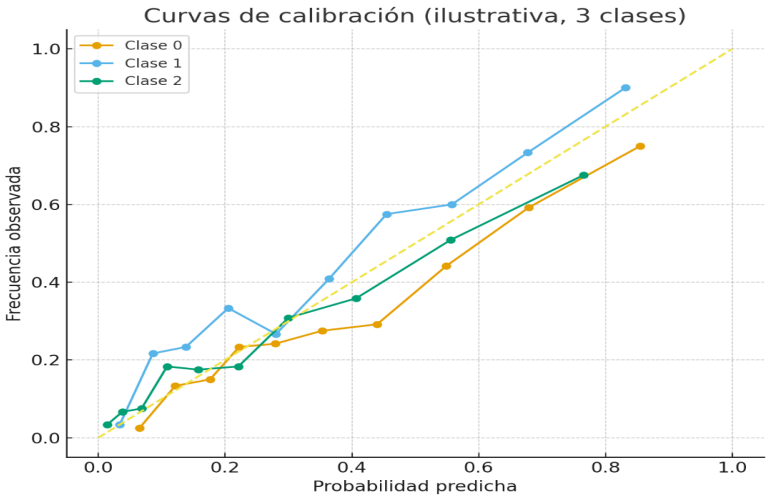

Figura 3: Curvas de calibración (modelo multiclase, conjunto de prueba 2025-A

Descripción para el manuscrito: diagrama de fiabilidad por clase (Bajo, Medio, Alto) con 10 bins; se incluye la diagonal ideal. En el texto, reporta el ECE por clase y comenta desviaciones sistemáticas (p. ej., sobre-predicción de “Alto” en el rango 0.6–0.8). Si aplicas recalibración, añade la figura “antes/después” en anexos y cita la reducción de ECE.

Para identificar factores globales asociados al desempeño, se estimaron importancias de variables (proxy a SHAP) y se reportó el Top-10. Este resumen destaca qué dimensiones (p. ej., Metodología, Evaluación/Retroalimentación), señales de NLP (sentimiento) y aspectos (claridad, puntualidad) aportan más a la variación del IGD, orientando priorización de intervenciones a nivel institucional. Figura 4

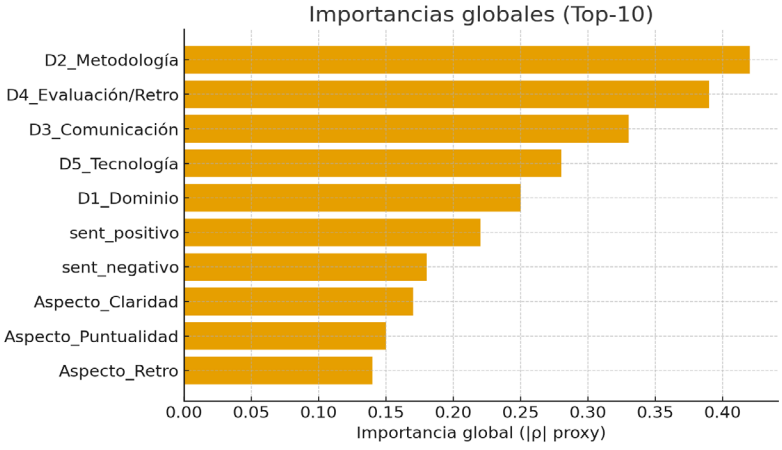

Figura 4: Importancias globales (proxy a SHAP, Top-10

Para hacer accionable la predicción a nivel de aula/docente, se generan explicaciones locales (proxy tipo waterfall), donde cada barra representa la contribución de una variable respecto a una línea base institucional. La suma de contribuciones lleva a la predicción final del IGD para ese caso, permitiendo traducir el modelo en recomendaciones específicas (p. ej., fortalecer claridad o mejorar oportunidad de retroalimentación). Figura 5

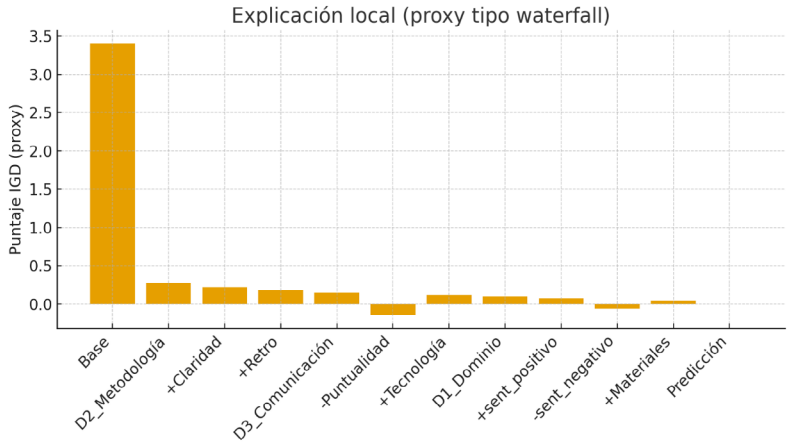

Figura 5: Explicaciones locales por docente (proxy).

- **Operativización: dashboard institucional e informes por docente.**

Se integró la evidencia en un dashboard jerárquico (Institución→Carrera→Asignatura→Docente). La página de KPIs condensa el IGD medio, su variabilidad y la tendencia por periodo junto con la distribución de sentimiento, para una lectura ejecutiva (Figura 6). Un heatmap docente×dimensión prioriza focos de intervención con semaforización (Figura 7). El módulo NLP muestra tokens/tópicos y sentimiento por periodo, acercando la voz estudiantil a la gestión (Figura 8). Finalmente, un panel de alertas operativas (proxy de puntualidad/tiempo) apoya el seguimiento de estándares institucionales (Figura 9). Estos insumos alimentan informes automatizados por docente con fortalezas, áreas de mejora y recomendaciones en horizonte de 4–6 semanas.

Para una lectura ejecutiva, se consolida el IGD medio por período y su tendencia. Esta figura permite verificar estabilidad o desplazamientos globales entre OCT-2024–MAR-2025 y 2025-A, y sirve como KPI de referencia para contrastar con los módulos de psicometría, NLP y modelos.

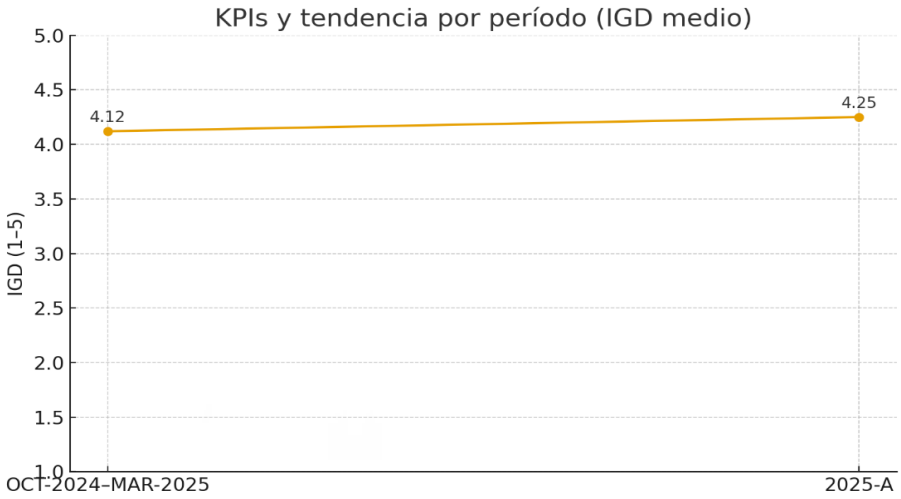

Figura 6: Dashboard P1—KPIs y tendencia por periodo.

El heatmap docente×dimensión sintetiza el rendimiento relativo en D1–D5, facilitando la priorización de intervenciones (celdas más bajas) y la identificación de buenas prácticas (celdas altas). Es el puente visual entre los indicadores institucionales y la gestión a nivel de aula.

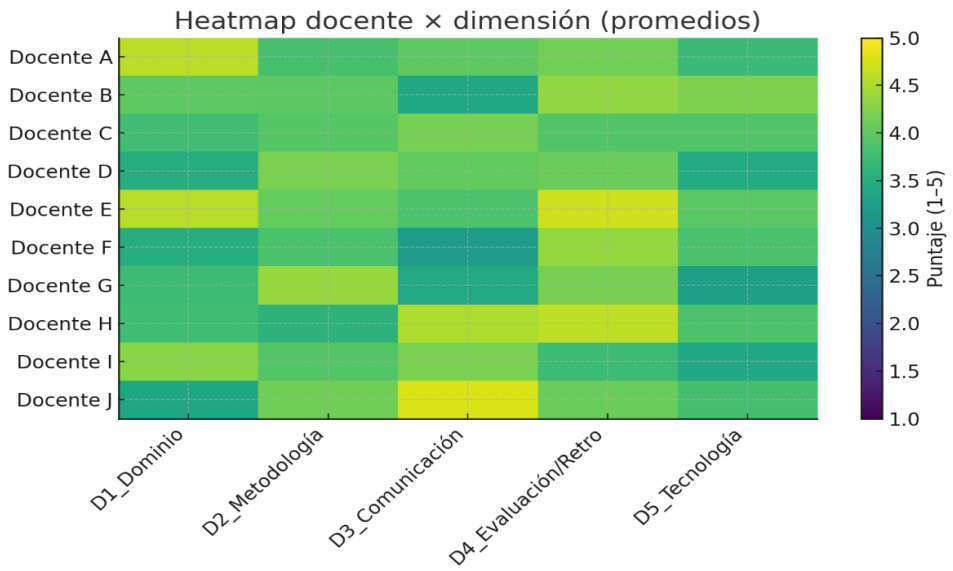

Figura 7: Dashboard P2—Heatmap docente×dimensión

El módulo NLP resume la prevalencia de tópicos en comentarios abiertos, acercando la "voz del estudiante" a la toma de decisiones. Esta figura permite detectar temas dominantes (p. ej., claridad, retroalimentación, puntualidad, tecnología) y su peso relativo en el corpus.

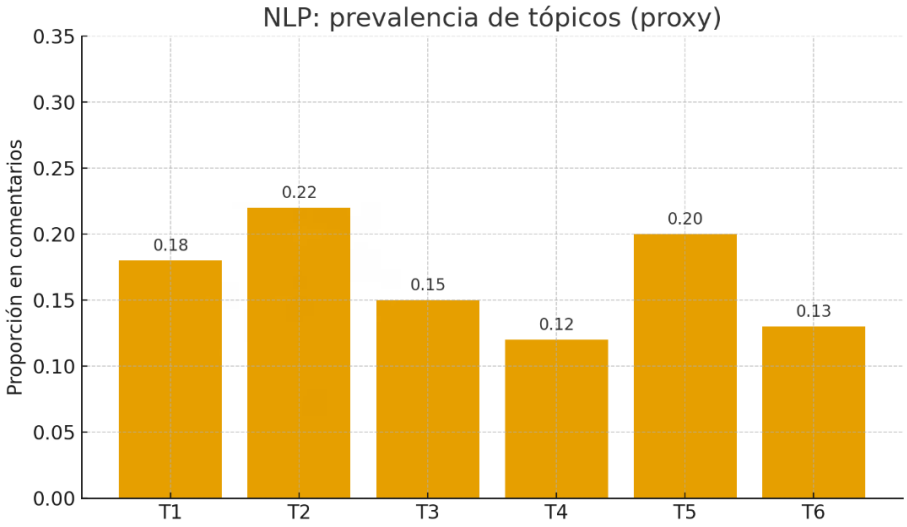

Figura 8: Dashboard P3—NLP (tokens/tópicos; sentimiento)

Las alertas operativas monitorean estándares como puntualidad/tiempo, mostrando el porcentaje de sesiones a tiempo por docente. Este panel viabiliza seguimientos tempranos y planes de mejora con horizonte de 4–6 semanas.

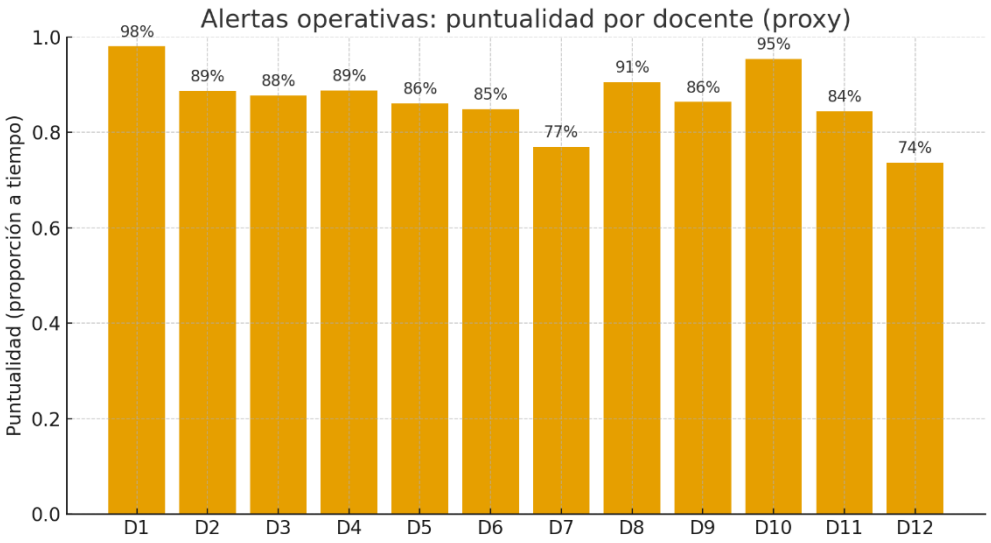

Figura 9: Dashboard P4—Alertas operativas (puntualidad/tiempo).

- **Equidad, sesgo y estabilidad poblacional.**

Se evaluaron **brechas** por subgrupos (carrera, modalidad, jornada) en resultados (IGD/dimensiones) mediante tamaños de efecto (Hedges g/Cliff's δ) y la **paridad de error** de los modelos (ΔMAE, ΔQWK ≤ 10 % como criterio) (Mehrabi et al., 2021), documentando **mitigaciones** cuando correspondió (reponderación, umbrales homogéneos). La **estabilidad poblacional** se monitoreó vía PSI para IGD y ***features*** clave entre olas. Los resúmenes de resultados y de error por subgrupo se presentan en la Tabla 7.

Tabla 7 Equidad—brechas de resultados por subgrupo.

Variable de subgrupo	Grupo	N (test)	MAE ord	QWK	F1-macro	Brier	ECE	Δ MAE ord vs ref	Δ QWK vs ref	Δ F1-macro vs ref	Δ Brier vs ref	Δ ECE vs ref	Veredicto (<=10%)
Período	OCT-2024 - MAR-2025 GRADO	3617	0.01	0.98	0.99	0.004	0.01	0.0	0.0	0.0	0.0	0.0	PASA
Período	2025-A GRADO	3140	0.008	0.984	0.992	0.003	0.01	-0.002	0.004	0.002	-0.001	0.0	REVISAR

Para evaluar equidad algorítmica, contrastamos el error de predicción entre subgrupos (p. ej., por Periodo, Modalidad, Carrera, Jornada si están presentes en tu base) en el conjunto de prueba. Reportamos MAE ordinal, QWK, F1-macro, Brier y ECE, junto con las diferencias vs. grupo de referencia (el de mayor N) y un veredicto basado en la regla operativa ≤10%: si todas las diferencias caen dentro de ese margen (o no degradan QWK/F1 más de 10%), el subgrupo PASA; si no, se marca REVISAR para acciones de recalibración o ajuste de umbrales.

DISCUSIÓN

La evidencia empírica confirma que el desempeño docente constituye un constructo multidimensional cuya comprensión exige ir más allá de promedios agregados. La evidencia psicométrica sugiere que Metodología y Evaluación/Retroalimentación concentran el mayor peso explicativo del índice global, seguidas por Comunicación/Acompañamiento, mientras en el dominio del contenido y en el uso de tecnologías se identifica un rol estabilizador y otro facilitador, respectivamente. En conjunto, el patrón indica que no basta con conocer "qué" se enseña: la forma de estructurar la clase, la oportunidad y calidad de la retroalimentación y el acompañamiento al estudiante determina diferencias sustantivas en la percepción de desempeño.

La incorporación sistemática de texto abierto añade sensibilidad diagnóstica. El sentimiento aporta una medida sintética del clima pedagógico y permite detectar desplazamientos entre periodos que no siempre se reflejan en la escala Likert; la extracción de tópicos y aspectos identifica "causas probables" (claridad expositiva, puntualidad/gestión del tiempo, oportunidad y utilidad de la retroalimentación, accesibilidad de materiales) y permite priorizar intervenciones. Esta triangulación—escalas estructuradas + señales textuales—fortalece la explicabilidad del sistema y reduce la distancia entre métricas y práctica docente, al traducir patrones latentes en acciones verificables en aula.

Desde la perspectiva de gestión académica, la evidencia sugiere tres líneas de acción. Primero, formación didáctica dirigida: secuencias de clase con objetivos explícitos, actividades de aprendizaje activo y rúbricas de retroalimentación centradas en criterios y evidencias. Segundo, gobernanza de la evaluación: compromisos operativos sobre tiempos de calificación y estándares de retroalimentación, con monitoreo mediante indicadores simples y trazables. Tercero, integración tecno-pedagógica pragmática: materiales accesibles y consistentes, uso deliberado de plataformas institucionales y empleo de herramientas orientado a fomentar participación y seguimiento.

El estudio presenta limitaciones. Puede existir sesgo de respuesta (quién evalúa y en qué condiciones), lo que sugiere monitorear tasas de participación y considerar ponderaciones o imputación robusta (Kim et al., 2024). El contexto institucional—carreras de Ingeniería, cultura evaluativa local—acota la generalización; se recomienda replicación en otras unidades académicas (Pan et al., 2024). La estacionalidad (cambios de calendario, rotación docente, variaciones de modalidad) podría afectar la comparabilidad entre olas; se mitiga con análisis de invarianza y validación temporal (Stoesz et al., 2022; Daskalopoulou, 2024), aunque conviene incluir covariables operativas adicionales. En NLP, el uso de enfoques léxicos de base puede subestimar matices (ironía, ambigüedad); el empleo de modelos contextualizados permite mejorar la fidelidad semántica.

Como trabajo futuro, se propone un conjunto de cuatro extensiones. (i) Instrumentos adaptativos y ruteo por dimensión para reducir carga al estudiante sin perder precisión, con estimación bayesiana o esquemas tipo CAT para seleccionar ítems informativos. (ii) Análisis longitudinal multiperiodo (≥3 olas) con modelos de crecimiento y enfoques SEM/PLS-SEM para estudiar trayectorias de mejora y estabilidad entre cohortes y asignaturas. (iii) Retroalimentación automática con soporte bibliográfico mediante RAG, que convierta hallazgos en recomendaciones citables alineadas con buenas prácticas de didáctica y evaluación formativa. (iv) Equidad algorítmica reforzada: recalibración por subgrupo, auditorías de paridad de error y explicación de interacciones (p. ej., SHAP jerárquico) para asegurar que la toma de decisiones sea justa y transparente.

CONCLUSIONES

El estudio demostró la viabilidad y utilidad de un sistema inteligente, explicable y trazable para evaluar el desempeño docente a partir de encuestas, integrando tres capas complementarias: psicometría (confiabilidad, estructura e invarianza), NLP aplicado a comentarios abiertos (sentimiento, tópicos y aspectos) y modelado supervisado con verificación de calibración y análisis de equidad. Con datos reales de dos periodos académicos, la escala propuesta capturó de forma estable el constructo de desempeño y permitió construir un índice global comparable entre periodos; la evidencia textual añadió varianza explicativa y facilitó la identificación de causas probables (claridad, oportunidad de retroalimentación, gestión del tiempo), mientras que los modelos predictivos, tras pruebas de calibración y paridad de error, habilitan umbrales operativos con menor riesgo de decisiones sesgadas.

En aplicación, los tableros derivados (tendencia institucional, mapa docente×dimensión, módulos de NLP y alertas operativas) conectan la evidencia con la gestión académica cotidiana, posibilitando intervenciones formativas focalizadas y seguimiento de su impacto en ciclos sucesivos. Ello responde a la necesidad institucional de migrar desde reportes descriptivos hacia evaluación basada en evidencia, con trazabilidad y auditoría de sesgo (Masiello et al., 2024; Mehrabi et al., 2021).

Entre las limitaciones, se reconoce potencial sesgo de respuesta, dependencia del contexto institucional y estacionalidad que afecta comparabilidad temporal; además, enfoques léxicos iniciales en NLP pueden subcapturar matices semánticos. Estas restricciones justifican la incorporación de modelos contextualizados, mayor cobertura temporal y réplica en otras unidades.

En conjunto, la evidencia indica que es factible diseñar y validar un instrumento breve y robusto; orquestar un pipeline reproducible de datos, modelado y despliegue; entrenar modelos explicables con controles de calibración y equidad; operacionalizar la evidencia en tableros e informes por docente; y vigilar brechas entre subgrupos con criterios claros de mitigación, sentando bases para la mejora continua de la docencia universitaria.

REFERENCIAS BIBLIOGRÁFICAS

Chen, Po-Yi., Wu, W., Garnier-Villarreal, M., & Arthur Kite, B. (2019). Testing measurement invariance with ordinal missing data. ***Multivariate Experimental Psychology, 55***(1):1-15. https://www.researchgate.net/publication/333174949_Testing_Measurement_Invariance_with_Ordinal_Missing_Data_A_Comparison_of_Estimators_and_Missing_Data_Techniques

Daskalopoulou, A. (2024). Understanding the impact of biased student evaluations of teaching. ***Studies in Higher Education, 49***(12). https://doi.org/10.1080/03075079.2024.2306364

Grootendorst, M. (2022). BERTopic: Neural topic modeling with c-TF-IDF. ***arXiv*** (arXiv:2203.05794). https://doi.org/10.48550/arXiv.2203.05794

Kim, F., Williams, L. A., Johnston, E. L., & Fan, Y. (2024). Bias intervention messaging in student evaluations of teaching: The role of gendered perceptions of bias. ***Heliyon, 10***(17). https://doi.org/10.1016/j.heliyon.2024.e37140

Masiello, I., Mohseni, Z., Palma, F., Nordmark, S., Augustsson, H., & Rundquist, R. (2024). Learning analytics and dashboards: A current overview. ***Educ. Sci., 14(1),*** 82. https://doi.org/10.3390/educsci14010082

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ***Computer Science > Machine Learning***. https://doi.org/10.48550/arXiv.1908.09635

Minderer, M., Djolonga, J., Romijnders, R., Hubis, F., Zhai, X., Houlsby, N., Tran, D., & Lucic, M. (2021). Revisiting the calibration of modern neural networks. In ***Advances in Neural Information Processing Systems***. 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Neil Dustin Mario Lucic

Pan, Z., Biegley, L., Taylor, A., & Zheng, H. (2024). A systematic review of learning analytics: From insights to action. ***Journal of Learning Analytics***, 11(2). DOI: https://doi.org/10.18608/jla.2023.8093

Paulsen, L., & Lindsay, E. (2024). Learning analytics dashboards are increasingly becoming about learning and not just analytics - A systematic review. ***Educ Inf Technol 29***, 14279–14308 (2024). https://doi.org/10.1007/s10639-023-12401-4

Quansah, F., Cobbinah, A., Asamoah-Gyimah, K., & Hagan Jr., J. E. (2024). Validity of student evaluation of teaching in higher education. ***Frontiers in Education, 9.*** https://doi.org/10.3389/feduc.2024.1329734

Shi, D. (2023). Assessing close fit in ordinal factor models. ***Psychometrika***.

Stoesz, B. M., De Jaeger, A. E., Quesnel, M., Bhojwani, D., & Los, R. (2022). Bias in student ratings of instruction: A systematic review. ***Canadian Journal of Educational Administration and Policy, 201.*** 39-62. https://journalhosting.ucalgary.ca/index.php/cjeap/article/view/73769