De la encuesta al dato: un flujo reproducible con R para caracterizar estudiantes en educación superior

Francisco Javier León[1], Miguel Oswaldo Pérez Pulido[1, 2], Leonardo Andrés Pinto Guarguatí[1]

[1] Universidad de Santander. Grupo de Investigación Ciencias Básicas Aplicadas para la Sostenibilidad

2. Unidades Tecnológicas de Santander. Dirección de Educación Virtual.

Resumen

La caracterización de estudiantes de primer ingreso constituye una estrategia esencial para comprender la diversidad, las trayectorias educativas y las condiciones de inclusión en la educación superior. En América Latina, los procesos de recolección y análisis de información suelen ser fragmentarios y con escaso nivel de explotación analítica, lo que limita la toma de decisiones basadas en evidencia. Este trabajo presenta un flujo reproducible basado en R, tidyverse, ggplot2 y Quarto, diseñado para consolidar y analizar datos de caracterización estudiantil mediante un sitio web interactivo y de acceso abierto.

Palabras clave: Calidad de la educación, proceso de datos, tecnología de la información, mejora continua, visualización de datos

1. Introducción

La caracterización de los estudiantes de primer ingreso es una práctica esencial en la educación superior, pues permite diseñar estrategias de permanencia y acompañamiento acordes con las necesidades del entorno (Pérez, Mejía, & Serrano, 2021). Sin embargo, los métodos tradicionales de análisis suelen carecer de reproducibilidad, dificultando la comparación entre campus y limitando la toma de decisiones (Romero & Ventura, 2020). El uso de R y sus ecosistemas de paquetes (*tidyverse*, *ggplot2*, *patchwork*) facilita la trazabilidad y la transparencia de los procesos analíticos (Wickham et al., 2019). Además, herramientas de publicación como Quarto permiten la creación de narrativas dinámicas e interactivas (Xie, Dervieux, & Riederer, 2023), promoviendo una cultura del dato en línea con los avances institucionales en analítica educativa (Pérez, Mejía, Suescún, & León, 2023).

Objetivo. Construir un sistema reproducible de análisis y visualización que transforme los resultados de encuestas masivas en narrativas interactivas y accesibles, fortaleciendo la transparencia institucional y la gestión educativa.

2. Metodología

Se aplicó una encuesta institucional a 1.250 estudiantes de primer ingreso (periodo 2025-1) en tres campus universitarios, con más de 100 ítems sobre dimensiones sociodemográficas, académicas, económicas y de bienestar.

El flujo de trabajo se estructuró en tres fases:

1. Procesamiento y limpieza de datos: estandarización, manejo de ausentes y creación de variables derivadas mediante *tidyverse* (Wickham, 2019).

2. Análisis descriptivo y visualización: uso de *ggplot2* (Wickham, 2016) y *patchwork* (Pedersen, 2020) para producir gráficos comparativos entre sedes.

3. Publicación interactiva: generación de un sitio web con Quarto (Xie et al., 2023) que integra resultados, narrativas y tablas dinámicas, disponible en el repositorio *Alerta 1 – Caracterización 2025-1*.

3. Resultados

El sistema consolidó la información institucional evitando la generación de informes separados y reduciendo los tiempos de procesamiento. Se identificó una diversidad significativa de perfiles de ingreso en edad, género y procedencia, así como una alta proporción de estudiantes de contextos vulnerables. Las narrativas visuales interactivas mejoraron la comunicación de resultados entre las áreas académicas, administrativas y de bienestar, fortaleciendo la transparencia y la gestión basada en datos (Pérez et al., 2023). Este flujo permitió además documentar las variables de seguimiento y crear insumos para futuras estrategias de permanencia (Pérez, Mejía, Serrano, Suescún, & León, 2023). A continuación se muestran los enlaces https://udesanalitica.github.io/Alerta1_Cx_2025-1/ y https://udesanalitica.github.io/Alerta1_Cx_2025-2/

4. Discusión

El modelo evidencia el potencial de R y Quarto para la analítica educativa reproducible, garantizando trazabilidad, apertura y eficiencia (Wickham et al., 2019). Su arquitectura modular facilita la adaptación del flujo en otros contextos institucionales mediante el uso de código abierto y documentación en Quarto. El repositorio asociado (GitHub institucional) contiene scripts parametrizables, ejemplos de limpieza y plantillas de visualización que pueden ser reutilizadas por otras universidades sin requerir infraestructura especializada, promoviendo redes de aprendizaje y colaboración interinstitucional (Pérez et al., 2025). Como línea futura, se prevé incorporar analítica predictiva y personalización del sitio para distintos grupos de usuarios, ampliando su impacto institucional (Xie et al., 2023).

5. Conclusiones

El flujo reproducible propuesto transforma una encuesta tradicional en una plataforma abierta, escalable y replicable de analítica educativa, que puede adaptarse a otras instituciones latinoamericanas. Este trabajo contribuye a las buenas prácticas en la gestión y visualización de datos en la educación superior, fortaleciendo la cultura del dato, la transparencia institucional y la colaboración interinstitucional sustentada en principios de ciencia de datos abierta. Asimismo, promueve el desarrollo de competencias en alfabetización tecnológica y analítica, esenciales para los procesos de mejora continua y la toma de decisiones basadas en evidencia (Pérez et al., 2021, 2023, 2025).

Referencias

Pedersen, T. L. (2020). *Patchwork: The composer of plots* [R package version 1.1.1]. The R Foundation. https://CRAN.R-project.org/package=patchwork

Pérez, M., Mejía, O., & Serrano, C. (2021). Estrategias de seguimiento, monitoreo e intervención académica para la permanencia universitaria: Caso Universidad de Santander. En M. Meléndez-Domínguez (Ed.), *Avances en educación superior e investigación* (Vol. 2, p. 58). Dykinson S.L.

Pérez, M., Mejía, S., Suescún, S., & León, F. (2023). *Valor agregado intermedio: Estrategia de mejoramiento para el aprendizaje (Informe de analítica N.° 19)*. https://doi.org/10.13140/RG.2.2.16028.67204

Pérez, M., Mejía, O., Serrano, C., Suescún, S., & León, F. (2023). Estrategias de intervención preventiva para fomentar la permanencia y éxito estudiantil: Alertas tempranas. En F. A. de Almeida (Ed.), *Desafios de ensinar e educar na contemporaneidade: escola, família e professores em pesquisa* (pp. 100–124). Editorial Científica Digital. https://doi.org/10.37885/230914480

Pérez, M., Mejía, O., León, F., Bohórquez, L., Bolívar, H., & Rincón-Yánez, D. (2025). A knowledge database discovery approach for improving quality in higher education institutions. En *Data-driven insights and analytics for measurable sustainable development goals* (pp. 207–228). Elsevier. https://doi.org/10.1016/B978-0-443-33044-5.00017-6

Romero, C., & Ventura, S. (2010). *Educational data mining: A review of the state of the art. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 40* (6), 601–618. https://doi.org/10.1109/TSMCC.2010.2053532

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer. https://doi.org/10.1007/978-3-319-24277-4

Wickham, H., François, R., Henry, L., & Müller, K. (2019). *dplyr: A grammar of data manipulation* [R package version 1.0.0]. The R Foundation. https://CRAN.R-project.org/package=dplyr

Xie, Y., Dervieux, C., & Riederer, E. (2023). *R Markdown Cookbook*. Chapman and Hall/CRC. https://bookdown.org/yihui/rmarkdown-cookbook/