International Journal of Business Analytics
Volume 13 • Issue 1 • January-December 2026

REPROPREP:

Reproducible Preprocessing Validation Framework - A Systematic Framework for Preprocessing Validation in Business Analytics

Miguel Angel Jimenez Garcia
http://orcid.org/0009-0007-1495-1746
Universidad Americana de Europa, Mexico

Richard De Jesus Gil Herrera
http://orcid.org/0000-0003-4481-7808
Universidad Internacional de la Rioja, Spain
Received: August 4th, 2025 | **Accepted:** March 25th, 2026

ABSTRACT

Preprocessing strategy selection in business analytics typically relies on convention rather than systematic evidence, despite consuming 60–80% of project effort. This study introduces REPROPREP (v1.0), a methodological framework for validating preprocessing effectiveness assumptions through statistical analysis and cost-benefit assessment. The framework applies Benjamini-Hochberg false discovery rate correction, quality degradation protocols, and cost-effectiveness evaluation. A demonstration across 10 UCI datasets, three preprocessing strategies, and gradient boosting classifiers with 5-fold stratified cross-validation yielded no statistically significant performance differences after multiple comparisons correction (mean effect size: 0.001 AUC), with implementation cost differences ranging from $150–$800. Focused on numeric preprocessing, REPROPREP provides organizations with a rigorous, context-specific methodology for evaluating preprocessing assumptions. Generalizability requires validation beyond tested conditions. Reproducible code is publicly available.

KEYWORDS

Customer Segmentation, Data Management, Business Analytics, Marketing ROI, Open Science, Data Preprocessing, Machine Learning

1. INTRODUCTION

Methodological rigor in data science research faces challenges that have implications for academic inquiry and business practice. Recent systematic reviews have identified concerns about reproducibility in computational research, including questions about validating analytical assumptions that underpin published findings (Kapoor and Narayanan, 2023; Gundersen and Kjensmo, 2018). These methodological considerations extend to business analytics, where preprocessing decisions represent significant resource investments and potential sources of analytical variability.

Data preprocessing typically consumes 60–80% of analytical effort in business analytics projects (Dasu and Johnson, 2003; Kandel et al., 2011), yet strategy selection often relies on conventional practices rather than systematic validation within specific organizational contexts.

DOI: 10.4018/IJBAN.406288

Organizations frequently implement preprocessing approaches without comprehensively evaluating their effectiveness relative to alternatives, representing a gap between resource allocation and evidence-based decision-making. Understanding when simpler approaches may achieve comparable results could inform more efficient resource utilization (Domingos, 2012).

The limited availability of systematic frameworks for preprocessing validation creates challenges for researchers and practitioners. Current preprocessing research exhibits several characteristics that may limit practical application. First, most studies focus on technique development rather than comparative evaluation across diverse contexts (García et al., 2016). Second, existing comparisons typically examine methods within specific settings, which may limit generalizability across different data characteristics and business domains. Third, statistical validation practices in preprocessing research vary considerably, with inconsistent application of appropriate controls for multiple comparisons (Ioannidis, 2005).

The machine learning community has recognized data leakage as a significant concern for research validity (Kaufman et al., 2012), yet preprocessing-specific validation methodologies require further development. Traditional approaches to preprocessing evaluation may not consistently incorporate rigorous statistical validation practices, including proper train-test separation during preprocessing decisions and appropriate statistical controls for multiple testing scenarios.

Business analytics practitioners face additional constraints that influence preprocessing decisions, including computational resources, implementation timelines, regulatory requirements, and the need for transparent cost-benefit analysis. However, existing research provides limited guidance for translating preprocessing research findings into actionable business decisions with explicit consideration of resource allocation trade-offs.

This study addresses these methodological gaps by introducing REPROPREP (Reproducible Preprocessing Validation Framework), a systematic approach designed to enable rigorous validation of preprocessing effectiveness assumptions.

1.1 Framework Scope and Research Contributions

REPROPREP follows agile development principles (Schröer et al., 2021), positioning this work as the initial release (v1.0) of an adaptive framework designed for progressive enhancement as new evidence emerges. This approach enables systematic framework development where each version incorporates organizational feedback, emerging preprocessing techniques, and validation evidence from diverse analytical contexts. The adaptive development philosophy treats findings as context-specific evidence requiring organizational validation rather than universally prescriptive recommendations, and organizations implementing REPROPREP should conduct pilot studies with their own data to assess framework applicability before operational deployment.

This release deliberately delimits scope to establish rigorous methodological foundations while enabling future extension. The current version encompasses three preprocessing strategies representing different complexity levels (minimal, standard, and advanced), with a focus on numeric preprocessing including imputation and scaling techniques. Validation employs a single algorithm family (gradient boosting classifiers) across public benchmark datasets from the UCI repository, addressing binary classification tasks with controlled quality degradation. This focused scope supports thorough validation within defined boundaries and provides extensible architecture for systematic enhancement based on accumulated evidence and organizational experience.

The framework is designed for progressive enhancement across subsequent releases. Future versions will systematically expand coverage to include categorical encoding strategies, followed by feature engineering validation, deep learning and neural network algorithms, and time-series specific preprocessing approaches. More comprehensive extensions addressing text preprocessing, NLP pipelines, multi-class classification, and regression task validation are planned for later releases. This iterative enhancement approach balances immediate practical utility with long-term framework development, enabling REPROPREP to evolve systematically as the business analytics community