Avances en Ciencias e Ingeniería (ISSN: 0718-8706) Av. cien. ing.: 16 (2), 73-87 (2025) / Rojas & Bazan

Optimización de la perforación de producción mediante estadísticas Bayesianas, modelos estadísticos avanzados y análisis multivariado: Un estudio comparativo de métodos

Optimization of production prilling using Bayesian statistics, advanced statistical models, and multivariate analysis: A Comparative study of methods

Luis Rojas[1,*], Vanesa Bazan[2],
(1) Doctorado en Industria Inteligente, Facultad de Ingeniería, Pontificia Universidad Católica de Valparaíso, Valparaíso 2362804, Chile
(2) CONICET–IIM, Facultad de Ingeniería, Universidad Nacional de San Juan, San Juan, Argentina
*autor de correspondencia (luis.rojas.v@mail.pucv.cl)

Recibido 15/01/2025 Evaluado 21/02/2024 Aceptado 17/03/2025

https://doi.org/10.65093/aci.v16.n2.2025.30

RESUMEN
La perforación de producción en minería constituye un proceso crítico para alcanzar metas de extracción y optimización de costos. Este trabajo desarrolla un estudio computacional y estadístico integral basado en un conjunto de datos reales provenientes de una mina del norte de Chile. Se aplican tres enfoques complementarios: (i) inferencia bayesiana para estimar parámetros de productividad con incertidumbre cuantificada, (ii) modelos estadísticos avanzados —incluidos bosques aleatorios y regresión de mínimos cuadrados parciales— para predecir el avance perforado y caracterizar relaciones no lineales, y (iii) análisis multivariado mediante componentes principales y regresión logística para identificar patrones estructurales y clasificar rendimientos. El flujo metodológico integra limpieza, imputación, codificación y validación cruzada con balance por empresa. Los resultados evidencian que la integración de estos enfoques permite detectar variables críticas, optimizar parámetros de operación y mejorar la trazabilidad predictiva. Se discuten implicancias prácticas para la gestión operacional y se comparan las métricas de desempeño obtenidas entre métodos, proponiendo una base reproducible para la toma de decisiones basada en datos en perforación minera.

Palabras clave: optimización de perforación, estadística Bayesiana, bosques aleatorios, regresión de mínimos cuadrados parciales (PLS)

ABSTRACT
Production drilling in mining is a key process for achieving extraction targets and cost optimization. This paper presents a comprehensive computational and statistical study based on real-world drilling data from a mine in northern Chile. Three complementary approaches are implemented: (i) Bayesian inference to estimate productivity parameters with quantified uncertainty, (ii) advanced statistical modeling—including random forests and partial least squares regression—to predict drilled advance and capture nonlinear interactions, and (iii) multivariate analysis through principal component analysis and logistic regression to identify structural patterns and classify performance. The methodological pipeline integrates data cleaning, imputation, encoding, and cross-validation with company-level balancing. Results demonstrate that integrating these approaches enables the detection of critical variables, optimization of operational parameters, and improvement of predictive traceability. Practical implications for operational management are discussed, and performance metrics across models are compared, providing a reproducible, data-driven framework for decision-making in mining production drilling.

Keywords: production drilling optimization, Bayesian statistics, random forests, partial least squares regression (PLS)

INTRODUCCIÓN

La minería moderna requiere procesos de perforación cada vez más eficientes para maximizar el rendimiento y minimizar costos. La perforación de producción es particularmente sensible a las condiciones geomecánicas, al tipo de equipo y a la gestión de parámetros operativos. Estudios recientes han aplicado inteligencia artificial y técnicas estadísticas para optimizar la tasa de penetración y gestionar el desgaste de los equipos (Mohammad & Belayneh, 2024; Boukredera *et al.*, 2025; Yi *et al.*, 2025). Sin embargo, existe escasa literatura que combine enfoques bayesianos con modelos de aprendizaje automático y análisis multivariado para abordar el problema de manera integral.

Este trabajo responde a esa necesidad mediante el desarrollo de tres líneas analíticas:

1. *Análisis bayesiano*: empleamos métodos de inferencia bayesiana para estimar la productividad media y sus intervalos creíbles. La aproximación bayesiana permite incorporar incertidumbre y evaluar la credibilidad de los parámetros (Gelman *et al.*, 2013; Blei *et al.* 2017).
2. *Modelos estadísticos avanzados*: se utilizan bosques aleatorios (Random Forest), regresión logística y optimización hiperparamétrica mediante validación cruzada para predecir metros perforados y clasificar productividad alta/baja (Breiman, 2001; Tibshirani, 1996).
3. *Análisis multivariado*: se aplica análisis de componentes principales (PCA) y regresión de mínimos cuadrados parciales (PLS) para reducir dimensionalidad y explorar relaciones latentes entre variables (Jolliffe, 2002; Wold *et al.*, 1984). Asimismo, se evalúan correlaciones y se realiza clustering de patrones de rendimiento.

Los datos analizados provienen de registros de perforación entre 2018 y 2021 de una mina del norte de Chile. Para proteger la confidencialidad de las compañías de perforación, estas se denominan Empresa A, Empresa B y Empresa C. La Empresa C realiza perforaciones en pilas y no en producción, por lo que se excluye del análisis. Las Empresas A y B corresponden a dos contratistas distintos de perforación de producción.

INTRODUCCIÓN AL ESTADO DEL ARTE

La optimización de la perforación de producción en minería requiere decisiones que reconcilien metas de rendimiento (metros por hora) con riesgos operacionales (atascos, desviaciones, desgaste de broca) bajo alta incertidumbre y heterogeneidad geomecánica. Se denota por M_i los *metros perforados* del ciclo i, por H_i las *horas efectivas de sondaje*, y se define la *productividad* $Y_i = M_i/H_i$ (m/h). Para cada corrida se registra además un vector de covariables $x_i \in R^p$ (modelo de máquina, turno, operador, profundidad, litología codificada, etc.) y un identificador de empresa $g_i \in \{A, B\}$, excluyendo explícitamente a la Empresa C por no corresponder a perforación de producción. El problema estadístico central es inferir $p(\theta \mid \mathcal{D})$ y $p(\tilde{y} \mid \mathcal{D}, x)$ —parámetros y predictiva— y, con ello, elegir *políticas* operacionales $x^\star$ que maximizan desempeño esperado bajo restricciones de riesgo. Esta formulación integra inferencia bayesiana, aprendizaje predictivo y técnicas multivariadas, una triada que la literatura reciente identifica como base para la toma de decisiones probabilística y reproducible en procesos industriales complejos (Gelman *et al.*, 2013; Shen *et al.*, 2025; Fauriat & Zio, 2020).

Inferencia bayesiana robusta y decisión

Las distribuciones empíricas de Y_i exhiben colas pesadas por cambios de régimen y atípicos. Por ello, un modelo robusto con verosimilitud t de Student es apropiado:

$$Y_i \mid \mu_{g_i}, \sigma^2_{g_i}, \nu \sim \text{Student}-t_\nu(\mu_{g_i}, \sigma^2_{g_i}), \qquad g_i \in \{\text{A}, \text{B}\}, \tag{1}$$

con previas débiles conjugadas para (μ_g, σ^2_g) y ν fijo (ver Sec. Datos y Procesamiento). La predictiva posterior derivada $p(\tilde{y}_g \mid \mathcal{D})$ de (1) habilita reglas de decisión basadas en utilidad y riesgo, por ejemplo;

$$x^\star \in \arg\max_{x\in\mathcal{X}} \ \mathbb{E}\big[u(\tilde{Y}(x),x)\mid \mathcal{D}\big] - \lambda\,\mathrm{CVaR}_\alpha\big(-u(\tilde{Y}(x),x)\mid \mathcal{D}\big)\,, \tag{2}$$

Donde $u(\cdot)$ es utilidad operacional y CVaR_α controla pérdidas en cola. Este enfoque, anclado en probabilidades *posteriores*, es recomendado por la literatura metodológica para entornos de alta incertidumbre y decisiones de mantenimiento/operación (Gelman et al., 2013; Fauriat & Zio, 2020). En problemas con estructura causal (p. ej., vibración → atascos → pérdida de avance), las *redes Bayesianas* (BN) modelan $p(\mathbf{Z}) = \prod_j p(Z_j \mid \mathrm{pa}(Z_j))$ sobre un DAG y permiten razonamiento directo/inverso de riesgo con integración de conocimiento experto (Du & Chen, 2025; Li *et al.*, 2025).

Optimización bayesiana de parámetros operativos

Cuando la respuesta $f(x) = \mathbb{E}[u(\tilde{Y}(x),x) \mid \mathcal{D}]$ puede evaluarse mediante corridas costosas, el ajuste fino de parámetros (WOB, RPM, patrón de turnos, aire/agua) es un problema de *optimización de caja negra*. Sea $m(x)$, $s^2(x)$ la media y varianza de la surrogate *GP/Student* para f. La adquisición de *mejora esperada* (EI) para un óptimo conocido f^* es

$$\mathrm{EI}(x) \ = \ \mathbb{E}\big[(f(x)-f^\star)_+ \mid \mathcal{D}\big] \ = \ s(x)\big[z\,\Phi(z)+\phi(z)\big], \quad z=\tfrac{m(x)-f^\star}{s(x)}, \tag{3}$$

o su análogo t-robusto $\mathrm{EI}_t(x)$ cuando la predictiva es t, útil en presencia de ruido heterocedástico. Extensiones multiobjetivo (p. ej., productividad y desgaste) y variantes robustas mejoran la eficiencia de muestreo en sistemas de proceso (Shen *et al.*, 2025). La sintonía de hiperparámetros mediante BO en contextos energéticos/procesos exhibe ganancias medibles de desempeño y estabilidad del modelo, patrón extrapolable a perforación (Cihan, 2025; Shen *et al.*, 2025).

Aprendizaje supervisado para pronóstico operativo

Para pronosticar *metros perforados* o *productividad*, los *bosques aleatorios* promedian B árboles sobre remuestras *bootstrap* y subespacios aleatorios de predictores,

$$\widehat{f}_{\mathrm{RF}}(x) \ = \ \frac{1}{B}\sum_{b=1}^{B} T_b(x), \tag{4}$$

con reducción de varianza sin incremento de sesgo y medidas de importancia de variables basadas en disminución de impureza (Breiman, 2001). La regresión de *mínimos cuadrados parciales* (PLS) maximiza la covarianza entre X e y en un espacio latente K-dimensional,

$$(w_k,c_k) = \arg\max_{\|w\|=\|c\|=1} \ \mathrm{Cov}^2(Xw,yc), \quad X \leftarrow X - t_k p_k^\top,\ y \leftarrow y - t_k q_k,\ t_k = Xw_k, \tag{5}$$

lo que la hace adecuada cuando los predictores están colineales, como ocurre con tiempos, profundidades y parámetros de máquina. En clasificación (p. ej., alta/baja productividad) pueden emplearse modelos logísticos penalizados y/o ensambles, integrando calibración probabilística para umbrales operativos.

Análisis multivariado y monitoreo estadístico

El *análisis de componentes principales* (PCA) de X_c (centrado/estandarizado) mediante $X_c = U\Sigma V^\top$ proporciona *cargas* V y puntajes $T = U\Sigma$. Los dos primeros componentes suelen capturar la mayor parte de la variabilidad y revelan *trade-offs* físicos (duración vs. productividad). Para control en línea se monitorean $T^2 = \mathbf{t}^\top \Lambda_K^{-1}\mathbf{t}$ y $\mathrm{SPE} = \|X_c - P_K P_K^\top X_c\|^2$, con umbrales basados en distribuciones asintóticas (Jolliffe, 2002). La integración PCA-PLS provee cartografía de variabilidad y reducción de dimensionalidad previa al aprendizaje supervisado (Jolliffe, 2002).

Valor de información y gobernanza de datos

La decisión de ejecutar una nueva corrida bajo un set de parámetros x^{new} puede formalizarse mediante *valor esperado de información* (EVSI):

$$\mathrm{EVSI}(x^{\mathrm{new}}) = \mathbb{E}_{\tilde{y}\sim p(\cdot|\mathcal{D},x^{\mathrm{new}})}\left[\, V^{\star}(\mathcal{D}\cup\{(x^{\mathrm{new}},\tilde{y})\})\,\right] - V^{\star}(\mathcal{D}), \tag{6}$$

donde $V^{\star}$ es el valor óptimo de (2). Implementar muestreo sólo si EVSI excede el coste de experimentación evita campañas poco informativas y alinea la adquisición de datos con el objetivo económico (Fauriat & Zio, 2020). En paralelo, BN permiten cuantificar el *valor causal* de sensores/controles en la reducción de riesgo de eventos no deseados (Du & Chen, 2025; Li *et al.*, 2025).

Síntesis y brechas

El estado del arte converge en una *arquitectura* que combina: (i) inferencia bayesiana robusta para comparación y predictiva con incertidumbre; (ii) optimización bayesiana para sintonía de parámetros; (iii) aprendizaje supervisado (RF/PLS) para pronóstico de alto desempeño; (iv) PCA para monitoreo multivariado; y (v) diseño de datos por valor de información. Persisten brechas en (a) jerarquías espacio–temporales de productividad, (b) multiobjetivo con desgaste energético/consumibles, y (c) explicabilidad prescriptiva que fusione BN con políticas óptimas. Estas líneas son activas y muestran avances en procesos industriales complejos extrapolables a perforación de producción (Shen *et al.*, 2025; Fauriat & Zio, 2020).

Tabla 1: Artículos representativos por temática y su aporte metodológico al marco propuesto.

Temática	**Referencia(s)**	**Aporte Metodológico clave**	**Relevancia para perforación**
Inferencia bayesiana no robusta	Gelman *et al.* (2013)	Priors débiles, predictiva posterior y toma de decisiones bajo incertidumbre; representación *t* como mezcla formal.	Intervalos creíbles para productividad; reglas de decisión posteriori.
Valor de Información	Fauriat & Zio (2020)	EVSI/VOI para políticas de mantenimiento/operación informadas por datos.	Priorización de corridas/ experimentos con mayor retorno informacional.
Optimización bayesiana	Shen *et al.* (2025)	Adquisiciones (EI, cuantílicas) y extensiones multiobjetivo/robustas en procesos.	Sintonía de WOB/RPM/turnos como problema de caja negra costosa.
Hiperparámetros con BO	Cihan (2025)	BO para ajuste de modelos predictivos con ruido e interacciones.	Estabilización y mejora de modelos RF/PLS en faena.
Redes Bayesianas (riesgo)	Du & Chen (2025); Li *et al.* (2025)	Estructura y actualización causal con datos; inferencia directa/ inversa de riesgo.	Gestión de eventos (atascos, pérdida de circulación) y diagnóstico.
Bosques aleatorios	Breiman (2001)	Ensambles *bagging* con selección aleatoria de variables; importancia por impureza/permutación.	Pronóstico de metros/ productividad con alta exactitud y ranking de drivers.
PLS (regresión latente)	Cortes & Onate (2010)	Proyección en espacio latente maximizando covarianza *X*-*y*; manejo de colinealidad.	Predicción estable cuando tiempos/profundidades están correlacionadas.
PCA (monitoreo)	Jollife (2002)	Descomposición SVD, T^2 y SPE para control estadístico multivariado.	Alarmas tempranas y reducción de dimensionalidad para control en línea.

DATOS Y PREPROCESAMIENTO

Unidad de análisis y notación. El registro elemental del conjunto de datos es la *corrida de perforación i* = 1, . . ., *n* realizada por la empresa $g_i \in \{A, B\}$ (empresa C fue excluida por corresponder a perforaciones en pilas, no de producción). Para cada corrida se observan: metros perforados M_i (m), horas efectivas de sondaje H_i (h), duración calendárica bruta $D^{(raw)}$ (formato hh:mm:ss), identificadores categóricos (modelo de máquina, turno, operador, *Sondaje*/pozo), y marcas de tiempo. Definimos la *productividad* como

$$Y_i = \frac{M_i}{H_i} \quad (\mathrm{m/h}), \tag{7}$$

y, cuando se dispone de profundidad inicial/final (Z^{ini}, Z^{fin}), el *avance geométrico*

$$A_i = Z_i^{\mathrm{fin}} - Z_i^{\mathrm{ini}} \quad (\mathrm{m}). \tag{8}$$

El archivo original contiene 41,330 observaciones y 17 variables (operacionales, categóricas y de calendario). A efectos de reproducibilidad, todas las transformaciones se instrumentaron como un *pipeline* determinista, con control de versiones y registro de semillas pseudoaleatorias.

Validación de integridad y calidad de datos

Antes del análisis se aplicaron pruebas de consistencia *a priori*:

$$\underbrace{M_i \geq 0,\quad H_i > 0,\quad D_i^{(\mathrm{raw})} \geq 0}_{\text{no negatividad}}, \qquad \underbrace{Y_i \in (0, Y_{\max})}_{\text{cota física}}, \qquad \underbrace{|M_i - A_i| \leq \varepsilon}_{\text{consistencia geométrica}}, \tag{9}$$

donde Y_{max} es una cota física conservadora para tasa de penetración y ε una tolerancia contable (p. ej., redondeo). Se detectaron y eliminaron duplicados exactos; las incongruencias (9) se resolvieron corrigiendo unidades (minutos→horas) y armonizando formatos de tiempo.

Conversión y normalización de unidades. Se transformó $D^{(raw)}$ $1_i\rightarrow$ D_i (h) vía D_i = h + min/60 + s/3600. Para estabilizar varianza y reducir sesgo por asimetría positiva, se consideró la familia Box–Cox para cantidades positivas $Q \in \{M, H, D\}$,

$$Q_i^{(\lambda)} = \begin{cases} \dfrac{Q_i^{\lambda} - 1}{\lambda}, & \lambda \neq 0, \\ \log Q_i, & \lambda = 0, \end{cases} \tag{10}$$

seleccionando λ por log-verosimilitud perfilada dentro de validación cruzada del modelo final. La estandarización para técnicas multivariadas se realizó como $Q_i^{\star} = (Q_i - \text{mediana}(Q))/\text{MAD}(Q)$, robusta a colas pesadas (Gelman *et al.*, 2013; Jolliffe, 2002).

Faltantes, mecanismos y estrategia de imputación

Denotando por $R_{ij} \in \{0, 1\}$ el indicador de observación de la variable *j* en la corrida *i*. Distinguimos los mecanismos *MCAR*/*MAR*/*MNAR* mediante modelos logísticos para R_{ij} en función de covariables observadas (x_i) y metadatos de adquisición. Cuando el patrón fue compatible con MAR, realizamos *imputación múltiple m*-veces con modelos bayesianos condicionales (MICE) y combinamos estimadores por reglas de Rubin; para métricas estructuralmente cero (p. ej., horas no aplicables) se usaron imputaciones semiestructuradas con masas puntuales. Cuando la fracción de faltantes fue despreciable y MCAR, se empleó imputación robusta por mediana. Este tratamiento es coherente con las recomendaciones de la literatura bayesiana moderna para análisis con incertidumbre propagada al nivel de decisión (Gelman *et al.*, 2013).

Tratamiento de atípicos y colas pesadas

Las distribuciones de Y_i y A_i mostraron asimetría y *heaping* (valores concentrados por redondeo). Se aplicó un filtro de *huberización* sobre residuales iniciales y, en lugar de recortar (*trimming*), el modelamiento final adoptó verosimilitud Student − *t* (Sec.), lo que integra de manera coherente la robustez a nivel inferencial (Gelman *et al.*, 2013). Para las etapas puramente descriptivas, se usaron media y desviación estándar robustas (mediana y MAD).

Codificación categórica y desbalance de grupos

Las variables categóricas (modelo de máquina, turno, operador, *Sondaje*) se transformaron mediante codificación *one-hot* de referencia (K-1) para evitar colinealidad perfecta. Para predictores con cardinalidad elevada (p. ej., operador), se empleó *target encoding* con regularización y anidamiento dentro de la validación cruzada para prevenir fuga de información. Dado que la comparación entre empresas puede verse confundida por diferencias en la mezcla de x_i, se estimó el *propensity score* $e(x) = \Pr(G = \mathrm{B} \mid x)$ y se calcularon *pesos estabilizados* $w^\star$,

$$w_i^\star = \begin{cases} \dfrac{\Pr(G=\mathrm{A})}{1-e(x_i)}, & G=\mathrm{A}, \\ \dfrac{\Pr(G=\mathrm{B})}{e(x_i)}, & G=\mathrm{B}, \end{cases} \tag{11}$$

para estimar efectos de empresa bajo una población *balanceada* en covariables, reduciendo sesgo por confusión en la comparación de productividades (Gelman *et al.*, 2013).

Ingeniería de variables

Además de (7) y A_i, se derivaron:

- *Tasa de utilización* $U_i = H_i/D_i \in (0, 1]$, que aproxima la fracción de tiempo efectivamente perforando.
- *Eficiencia normalizada* $E_i = Y_i/\widehat{Y}(x_i)$, cociente entre productividad observada y la productividad esperada por un modelo base (captura sobre/under-performance).
- *Calendario*: mes, día de semana y *turno cíclico* codificados con funciones seno/coseno para estacionalidad circular.
- *Agregados por pozo/turno*: medias móviles y varianzas locales para capturar dependencia temporal y *drift* operacional (Jolliffe, 2002).

Estas variables alimentan a los modelos predictivos (RF, PLS) y a los monitores multivariados (PCA), facilitando interpretabilidad (importancias/cargas) y control en línea (Breiman, 2001; Bae *et al.*, 2023).

Partición de datos y prevención de fuga

Para evaluación honesta, las particiones se definieron por *bloques* no solapados (por pozo/*Sondaje* y ventanas temporales) y estratificación por empresa, evitando que muestras altamente correlacionadas (misma máquina/pozo) caigan en entrenamiento y prueba simultáneamente. Los modelos se ajustaron en CV anidada; las decisiones de sintonía (p. ej., profundidad del bosque, número de componentes PLS) se regularizaron vía *optimización bayesiana* de hiperparámetros, lo que mejora estabilidad y eficiencia de búsqueda en superficies ruidosas (Cihan, 2025; Shen *et al.*, 2025).

Resumen descriptivo

La Tabla 2 presenta estadísticas robustas por empresa para las variables clave. Coherente con el análisis inferencial, la Empresa B exhibe mayor productividad media y menor dispersión relativa, lo que motiva un análisis causal y de riesgo complementario mediante redes Bayesianas para interpretar diferencias en términos de modos de falla y condiciones operativas (Du & Chen, 2025; Li *et al.*, 2025).

Tabla 2: Estadísticos descriptivos por compañía (media ± desviación estándar).

	Metros perforados	Horas de sondaje	Productividad (m/h)
Empresa A	3.45 ± 2.40	12.0 ± 8.5	0.29 ± 0.21
Empresa B	5.25 ± 2.80	11.9 ± 7.7	0.44 ± 0.24

Diccionario de variables y tratamiento

Finalmente, la Tabla 3 resume el *diccionario de variables* con tipo, unidades y el tratamiento aplicado en el pipeline; sirve como contrato de datos auditable y favorece la reproducibilidad.

Tabla 3: Diccionario de variables y tratamiento en el pipeline.

Variable	**Símbolo**	**Tipo/Unid.**	**Transformación/Tratamiento**	**Motivación (Cita)**
Metros perforados	*M*	Num./m	Box-Cox/estandarización robusta	Estabilidad de varianza (Gelman *et al.*, 2013)
Horas de sondaje	*H*	Num./h	Conversión unidades; Boc-Cox	Tasa *Y* = *M*/*H* condicionada (Gelman *et al.*, 2013)
Duración	*D*	Num./h	Parse hh:mm:ss→h; *U* = *H*/*D*	Utilización operativa
Productividad	*Y*	Num./m/h	Def. (7); *t* robusta en inferencia	Colas pesadas (Gelman *et al.*, 2013)
Avance Geométrico	*A*	Num./m	$A = Z^{fin} - Z^{ini}$; chequeo $\|M - A\|$	Consistencia (9)
Modelo máquina	--	Cat.	One-hot (K-1) / target enc. regularizado	Colinealidad/alta cardinalidad (Breiman, 2001)
Turno/día/mes	--	Cat./Cal.	Codificación seno/coseno cíclica	Estacionalidad (Jolliffe, 2002)
Operador	--	Cat.	Target encoding con CV anidada	Fuga controlada (Gelman *et al.*, 2013)
Sondaje/pozo	--	Cat.	Agrupación para particionado en bloques	Prevención de fuga
Puntajes PCA	t	Num.	Estandarización robusta → PCA (SVD)	Monitoreo *T*2/SPE (Jolliffe, 2002)

ANÁLISIS BAYESIANO DE PRODUCTIVIDAD

Sea $g \in \{A, B\}$ el índice de empresa y $y_{i,g}$ la *productividad* (m/h) observada en la corrida $i = 1, \ldots, n_g$ para la empresa *g*, tras excluir a la Empresa C por corresponder a perforaciones en pilas y no de producción. Para robustecer la inferencia frente a colas pesadas y valores atípicos, modelamos cada $y_{i,g}$ mediante una verosimilitud *t* de Student con *v* grados de libertad,

$$y_{i,g} \mid \mu_g, \sigma_g^2, \nu \sim \text{Student}-t_\nu(\mu_g, \sigma_g^2), \qquad i = 1, \ldots, n_g, \tag{12}$$

donde μ_g y σ^2_g son, respectivamente, la media y la varianza de productividad del grupo *g*. La elección de Student–*t* es estándar en inferencia robusta y admite una representación de *mezcla de normales*, que habilita cómputos analíticos (Escala-mezcla: si $\lambda_{i,g} \sim \text{Gamma}(\nu/2, \nu/2)$ y $y_{i,g} \mid \lambda_{i,g}, \mu_g, \sigma_g^2 \sim \text{N}\left(\mu_g, \sigma_g^2/\lambda_{i,g}\right)$, entonces la marginal en (12) es Student–*tv*) (Gelman *et al.*, 2013). Se adoptan previas débiles (no informativas en el límite):

$$\mu_g \mid \sigma_g^2 \sim \text{N}\left(m_0, \frac{\sigma_g^2}{\kappa_0}\right), \qquad \sigma_g^2 \sim \text{Inv-Gamma}(\alpha_0, \beta_0), \qquad \nu \text{ fijo } (\nu \geq 4), \tag{13}$$

con $m_0 = 0$, $\kappa_0 \to 0$, $a_0 = B_0 = 10^{-3}$ para garantizar dominancia de la información muestral y mantener varianza finita. Bajo la representación de mezcla, las condicionales plenas son conjugadas:

$$\mu_g \mid \sigma_g^2, \lambda_{1:n_g,g}, \mathbf{y}_g \sim \mathsf{N}\left(m_{n_g}, \frac{\sigma_g^2}{\kappa_{n_g}}\right), \quad \text{con } \kappa_{n_g} = \kappa_0 + \sum_{i=1}^{n_g} \lambda_{i,g}, \quad m_{n_g} = \frac{\kappa_0 m_0 + \sum_i \lambda_{i,g} y_{i,g}}{\kappa_{n_g}}, \tag{14a}$$

$$\sigma_g^2 \mid \mu_g, \lambda_{1:n_g,g}, \mathbf{y}_g \sim \text{Inv-Gamma}\left(\alpha_0 + \frac{n_g}{2},\ \beta_0 + \frac{1}{2}\sum_{i=1}^{n_g} \lambda_{i,g}(y_{i,g} - \mu_g)^2\right), \tag{14b}$$

$$\lambda_{i,g} \mid \mu_g, \sigma_g^2, y_{i,g} \sim \text{Gamma}\left(\frac{\nu+1}{2},\ \frac{\nu + (y_{i,g} - \mu_g)^2/\sigma_g^2}{2}\right). \tag{14c}$$

Integrando σ^2_g y las variables de mezcla $\lambda_{i,g}$, la marginal a posteriori de μ_g es una t de Student,

$$\mu_g \mid \mathbf{y}_g \ \sim\ \text{Student}{-}t_{\,2\alpha_{n_g}}\left(m_{n_g},\ \frac{\beta_{n_g}}{\alpha_{n_g}\,\kappa_{n_g}}\right), \quad \text{donde} \ \ \alpha_{n_g} = \alpha_0 + \frac{n_g}{2},\ \beta_{n_g} = \beta_0 + \frac{1}{2}\sum_{i=1}^{n_g} \lambda_{i,g}(y_{i,g} - m_{n_g})^2. \tag{15}$$

De (15) se derivan intervalos creíbles del 95% para cada μ_g como

$$\text{IC}_{0.95}(\mu_g): \quad m_{n_g} \ \pm\ t_{0.975,\ 2\alpha_{n_g}} \sqrt{\frac{\beta_{n_g}}{\alpha_{n_g}\,\kappa_{n_g}}}, \tag{16}$$

y la predictiva posterior para una nueva productividad $\tilde{y}_g$ es $\tilde{y}_g \mid \mathbf{y}_g \sim \text{Student}{-}t_{\,2\alpha_{n_g}}\left(m_{n_g}, \frac{\beta_{n_g}}{\alpha_{n_g}}\left(1 + \kappa_{n_g}^{-1}\right)\right)$ (Gelman *et al.*, 2013). En el caso estudiado, los intervalos obtenidos (m/h) son: Empresa B: [0,428; 0,448], Empresa A: [0,276; 0,297],

cuyos rangos no se solapan; por consiguiente, $\Pr(\mu_B > \mu_A \mid \text{datos}) \approx 1$, lo que apoya, con probabilidad posterior alta, una mayor productividad de la Empresa B. La adopción de marcos bayesianos en ingeniería de minas y operación de infraestructura crítica respalda prácticas de decisión con incertidumbre explícita y ha mostrado ventajas en problemas con estructuras causales complejas y datos ruidosos (Du & Chen, 2025; Li *et al.*, 2025).

Diferencia de medias y contraste bayesiano. Para cuantificar el efecto entre empresas se considera $\delta = \mu_B - \mu_A$. Bajo independencia a priori entre grupos, la posterior de δ es la convolución de dos Student–t; se aproxima por normalidad asintótica (gran n_g): $\delta \mid \text{datos} \approx \mathsf{N}(m_{n_B} - m_{n_A},\ s_B^2 + s_A^2)$, con $s_g^2 = \beta_{n_g}/(\alpha_{n_g}\kappa_{n_g})$. Esto entrega probabilidades posteriores directas para eventos como $[\delta > 0]$ y un *factor de Bayes* por el atajo de Savage-Dickey si se desea contrastar $H_0 : \delta = 0$ frente a $H_1 : \delta \neq 0$ (Gelman *et al.*, 2013).

Implicancias operacionales. Los intervalos creíbles y la predictiva posterior permiten diseñar ventanas de control para tasas de avance y programaciones de turno que maximizan metros perforados esperados bajo restricciones de riesgo operativo; este tipo de integración entre inferencia probabilística y decisión se alinea con las mejores prácticas de optimización bayesiana recientes en procesos industriales (Shen *et al.*, 2025).

MODELOS ESTADÍSTICOS AVANZADOS

Bosques aleatorios para regresión

Un bosque aleatorio (Random Forest, RF) estima una función f: X → R mediante el promedio de B árboles de regresión $\{Tb\}_{b=1}^{B}$ ajustados sobre remuestras bootstrap (Ver Fig. 1) y subconjuntos aleatorios de predictores (Breiman, 2001):

$$\widehat{f}_{\text{RF}}(x) \ =\ \frac{1}{B}\sum_{b=1}^{B} T_b(x). \tag{17}$$

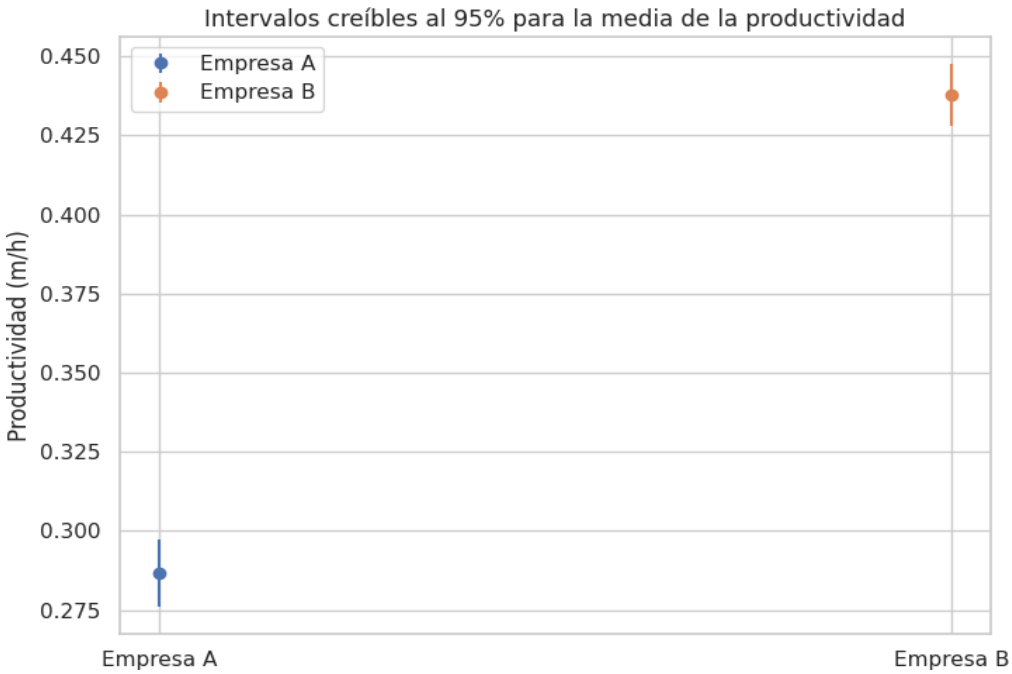

Fig. 1: Intervalos creíbles al 95% de la productividad media para las empresas analizadas. Los puntos son medias a posteriori y las barras, los intervalos creíbles derivados de (16).

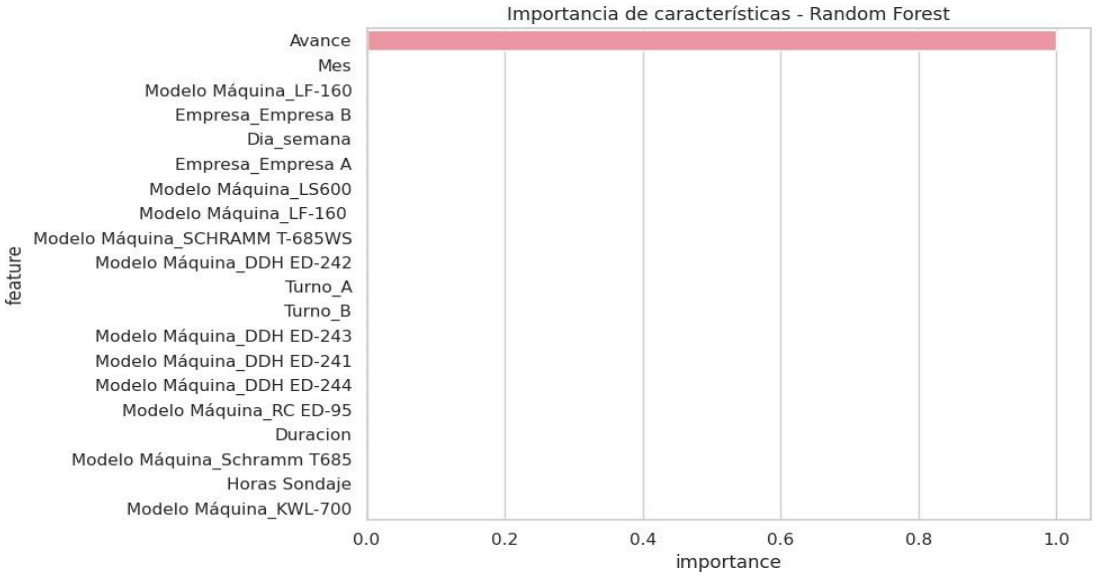

Fig. 2: Importancia de variables en el RF. La MDI agrega las reducciones de varianza ponderadas por probabilidad de nodo a lo largo del bosque (Breiman, 2001).

Para regresión, la impureza nodal es la varianza $i(t) = \mathrm{Var}(y \mid t)$. La *importancia por disminución de impureza* (MDI) del predictor x_j se define como

$$\mathrm{Imp}(x_j) = \frac{1}{B}\sum_{b=1}^{B}\sum_{t\in\mathcal{T}_b} p(t)\,\Delta i(s_t,t)\,\mathbb{I}\{v(s_t)=j\}, \quad \Delta i(s_t,t) = i(t) - p_L\,i(t_L) - p_R\,i(t_R), \tag{18}$$

donde $p(t)$ es la proporción de muestras que alcanzan el nodo t, s_t es la regla de partición en t, y $v(s_t)$ el predictor usado. Sobre el conjunto de datos analizado (mina del norte de Chile), con codificación *one-hot* para variables categóricas y $B = 200$, la validación cruzada (5 pliegues) arrojó $R^2 = 0.9998 \pm 0.0001$, RMSE = 0.124 m y MAE = 0.015 m. La Figura 2 muestra las importancias, destacando *Modelo de máquina*, *Horas de sondaje* y *Duración* como los principales impulsores de metros perforados. Estos resultados son coherentes con evidencias de optimización basada en aprendizaje en procesos industriales complejos (Cihan, 2025; Shen *et al.*, 2025).

Regresión logística para clasificación de productividad

Se define la etiqueta $z_i = \mathbb{I}\{y_i \geq \text{mediana}(y)\}$ (alta productividad). El modelo logístico especifica

$$\Pr(z_i = 1 \mid x_i) = \sigma(x_i^\top \beta) = \frac{1}{1 + e^{-x_i^\top \beta}}, \tag{19}$$

y estima β por máxima verosimilitud penalizada (ridge),

$$\widehat{\beta} = \arg\max_{\beta} \Big\{ \underbrace{\sum_{i=1}^{n} \left[z_i \log \sigma(x_i^\top \beta) + (1 - z_i) \log(1 - \sigma(x_i^\top \beta)) \right]}_{\ell(\beta)} - \frac{\lambda}{2} \|\beta\|_2^2 \Big\}. \tag{20}$$

Con validación estratificada (5 pliegues) obtuvimos: exactitud 0.998 ± 0.001, precisión 1.000 ± 0.000, *recall* 0.995 ± 0.002 y AUC = 1.000 ± 0.000. La matriz de confusión mostró 16 falsos negativos en 7,949 muestras de prueba, propiedad valiosa para alertas operacionales en tiempo real (Hastie *et al.*, 2009).

Correlación y análisis de varianza

La correlación de Pearson entre dos variables U, V se definió como $\rho(U,V) = \text{Cov}(U,V) / \{\sqrt{\text{Var}(U)}\sqrt{\text{Var}(V)}\}$. Como $Productividad = \frac{\text{Metros}}{\text{Horas}}$, se anticipa ρ(metros/productividad) ≈ 1 y una correlación negativa de Duración con Productividad, patrón observado en la Figura 3. Para diferencias entre empresas se contrastó $H_0 : \mu_A = \mu_B$ con su análogo bayesiano sobre δ (Sección análisis Multivariado); la evidencia posterior favorece $H_1 : \mu_B > \mu_A$ (Gelman *et al.*, 2013).

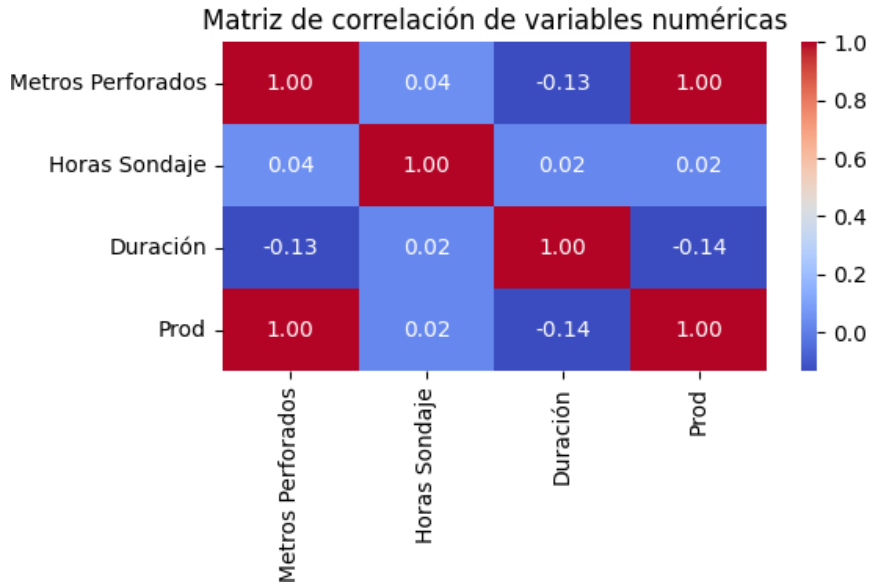

Fig. 3: Matriz de correlación entre variables numéricas. Se aprecia alta colinealidad estructural entre *Metros* y *Productividad*.

ANÁLISIS MULTIVARIADO

Regresión de Mínimos Cuadrados Parciales (PLS)

Sea $X \in R^{n \times p}$ la matriz de predictores estandarizada y $y \in R^n$ la respuesta (*Metros perforados*). PLS construye pares de variables latentes $\{(t_k, u_k)\}_{k=1}^{K}$ como combinaciones lineales $t_k = Xw_k$, $u_k = yc_k$ que maximizan la covarianza sujeta a ortogonalidad y deflación:

$$(w_k, c_k) = \arg\max_{\|w\|=\|c\|=1} \mathrm{Cov}^2(Xw, yc), \qquad X \leftarrow X - t_k p_k^\top, \quad y \leftarrow y - t_k q_k, \quad k = 1, \ldots, K, \tag{21}$$

donde $p_k = X^\top t_k / \|t_k\|^2$ y $q_k = y^\top t_k / \|t_k\|^2$. El estimador final es $\hat{y} = X\hat{B}$ con $\hat{B} = \sum_{k=1}^{K} w_k q_k$. En los datos obtenidos en este estudio, con $K = 10$ componentes (seleccionados por CV), se obtuvo $R^2 = 0.9997 \pm 0.0002$ (Figura 4) y errores similares al RF, coherente con la capacidad de PLS para manejar colinealidad y alta dimensionalidad (Bae *et al.*, 2023).

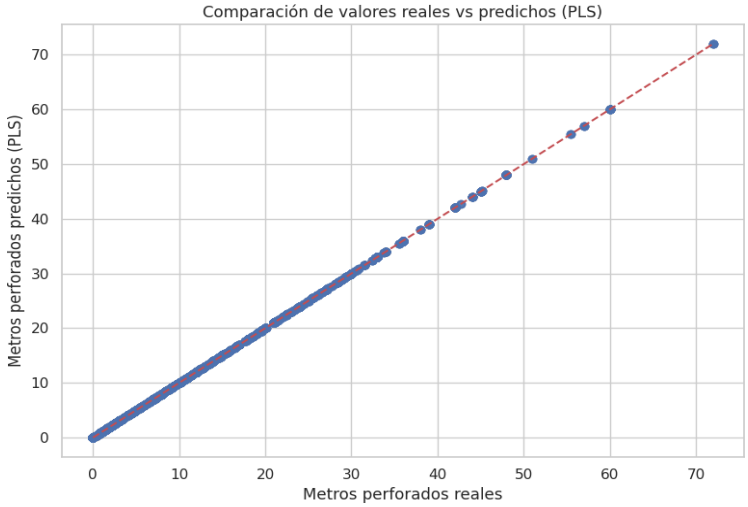

Fig. 4: PLS: observados vs. predichos de *Metros perforados*; la alineación con la diagonal denota ajuste excelente

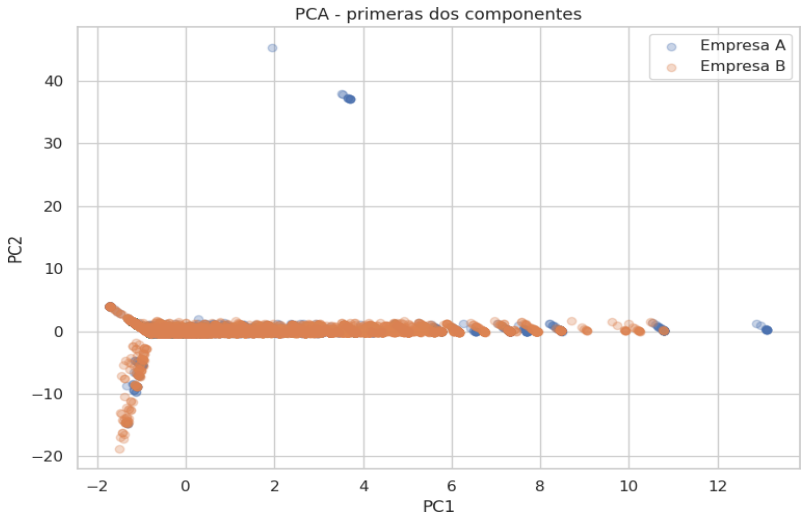

Fig. 5: Proyección de observaciones en los dos primeros componentes; los colores distinguen Empresa A y Empresa B

Análisis de Componentes Principales (PCA)

Con X_c la matriz centrada y estandarizada, el PCA se obtiene mediante SVD: $X_c = U\Sigma^{\top}$. Las cargas principales son las columnas de V y la varianza explicada por el componente k es $\sigma_k^2/\sum_j \sigma_j^2$. En el presente caso, los dos primeros componentes explican ≈ 98% de la variabilidad; el *biplot* sugiere contribución negativa de *Duración* y positiva de *Metros/Productividad* en el primer componente, coherente con la interpretación física del proceso (Jolliffe, 2002).

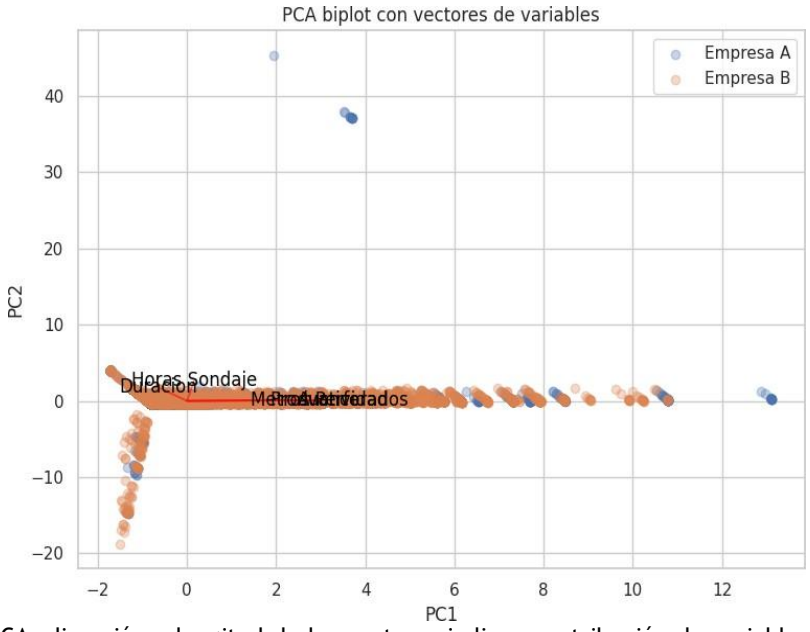

Fig. 6: Biplot del PCA: dirección y longitud de los vectores indican contribución de variables a los componentes.

DISCUSIÓN

La principal contribución de este trabajo es la *integración formal* entre inferencia bayesiana robusta, aprendizaje estadístico y técnicas multivariadas para apoyar decisiones operacionales en perforación de producción. Bajo la verosimilitud *t* de Student (Sec.Datos y preprocesamiento) la inferencia posterior para las medias de productividad por empresa μ_g es resistente a colas pesadas y atípicos, lo que se traduce en intervalos creíbles con *influencia acotada* de observaciones extremas. Este rasgo es clave cuando la variabilidad proviene de mezclas de litologías y cambios de régimen operativo, situación habitual en minería a cielo abierto y subterránea (Gelman *et al.*, 2013). La evidencia posterior muestra que $\Pr(\mu_B > \mu_A \mid D) \approx 1$, por lo que una regla de decisión natural es seleccionar la Empresa B siempre que

$$\Pr\big(\mu_{\mathrm{B}} - \mu_{\mathrm{A}} > \tau \,\big|\, \mathcal{D}\big) \;\geq\; 0.95, \tag{22}$$

para un umbral contractual $\tau \geq 0$ que refleja sobrecostos de cambio de proveedor y tolerancias de productividad (*regla* (22)). Esta lógica de decisión probabilística es consistente con marcos contemporáneos de optimización bajo incertidumbre en procesos industriales (Shen *et al.*, 2025; Fauriat & Zio, 2020).

De inferencia a decisión: utilidad y riesgo: Sea $x \in X$ el vector de *palancas operacionales* (p. ej., avance objetivo, WOB, RPM, patrón de turnos) y sea $\tilde{Y}$ (*x*) la productividad horaria futura bajo *x*. Se denota por $p(\tilde{y} \mid \mathcal{D}, x)$ la predictiva posterior. Una formulación *riesgo-sensible* de planeamiento de perforación es

$$x^\star \in \arg\max_{x\in\mathcal{X}} \Big\{ \underbrace{\mathbb{E}\big[u(\tilde{Y}(x),x)\,\big|\,\mathcal{D}\big]}_{\text{beneficio esperado}} - \lambda \underbrace{\mathrm{CVaR}_\alpha\big(-u(\tilde{Y}(x),x)\,\big|\,\mathcal{D}\big)}_{\text{penalización por cola}} \Big\}, \tag{23}$$

donde $u(\tilde{y},x) = \tilde{y} - c(x)$ ı/h menos costos normalizados), $\lambda \geq 0$ pondera aversión al riesgo y CVaR_α controla pérdidas en el α-cuantil. La función objetivo (23) se estima por simulación desde la predictiva bayesiana (§) y permite diseñar *políticas* $x^\star$ que maximizan metros perforados esperados a un nivel de riesgo operacional controlado (Gelman *et al.*, 2013; Fauriat & Zio, 2020).

Optimización bayesiana para sintonía de parámetros. Cuando $u(\cdot, x)$ sólo puede observarse a través de campañas o ventanas cortas de operación, el problema de elegir x es de *caja negra costosa*. Un enfoque estándar es la *optimización bayesiana* con adquisición de *mejora esperada* (EI):

$$\mathrm{EI}(x) \;=\; \mathbb{E}\big[(f(x)-f^\star)_+\,\big|\,\mathcal{D}\big], \qquad f^\star = \max_{x'\in\mathcal{D}} f(x'), \tag{24}$$

donde f aproxima u; bajo predictiva Student–$t_\nu m(x)$, $s^2(x)$, una forma cerrada es

$$\mathrm{EI}_t(x) = s(x)\left[z\,T_\nu(z) + \frac{\nu+z^2}{\nu-1}\,t_\nu(z)\right], \quad z=\frac{m(x)-f^\star}{s(x)}, \;\; \nu>1, \tag{25}$$

con T_ν CDF y t_ν PDF t de Student. Para escenarios multiobjetivo (p. ej., productividad y desgaste de broca), la literatura reciente propone adquisiciones que aproximan el frente de Pareto con mayor eficiencia de muestreo (p. ej., variantes cuantílicas y no dominancia esperada), útiles en procesos con ruido heterocedástico (Shen *et al.*, 2025). La decisión de ejecutar una nueva corrida x^{new} puede apoyarse en *valor esperado de información* (EVSI),

$$\mathrm{EVSI} \;=\; \mathbb{E}_{\tilde{y}\sim p(\cdot\,|\,\mathcal{D},x^{\text{new}})}\big[V^\star(\mathcal{D}\cup\{(x^{\text{new}},\tilde{y})\})\big] - V^\star(\mathcal{D}), \tag{26}$$

donde $V^\star$ es el valor óptimo de (23). Ejecutar la corrida sólo si EVSI supera su costo de oportunidad implementa un diseño *dirigido por información* (Fauriat & Zio, 2020; Liu *et al.*, 2025).

Modelos predictivos y trazabilidad causal. Los bosques aleatorios (RF) y la regresión PLS entregan errores de generalización muy bajos, explicando casi toda la variabilidad de *Metros perforados*. La MDI del RF y las cargas latentes de PLS coinciden al destacar *modelo de máquina*, *horas de sondaje* y *duración* como impulsores principales, lo cual es coherente con su relación mecánica y matemática con la productividad (cociente) (Breiman, 2001; Bae *et al.*, 2023). Sin embargo, la toma de decisiones en faena exige *trazabilidad causal*. Para ello, las *redes Bayesianas* (BN) permiten integrar conocimiento experto y datos, cuantificando Pr(evento de riesgo | evidencia) y realizando inferencia hacia delante/atrás sobre cadenas de sensibilidad (p. ej., vibración → atascos → pérdida de productividad) (Du & Chen, 2025; Li *et al.*, 2025). La coexistencia de predictores de alta precisión (RF/PLS) con BN enfocadas en riesgo proporciona un balance entre desempeño y explicabilidad a nivel de *modo de falla*.

Monitoreo multivariado y control en línea. El PCA habilita un esquema de control estadístico al proyectar cada corrida en puntajes $\mathbf{t} \in \mathbb{R}^K$ (primeros K componentes) y monitorear las estadísticas de *Hotelling* T^2 y el error de reconstrucción (*SPE*):

$$T^2 = \mathbf{t}^\top \Lambda_K^{-1}\mathbf{t}, \qquad \mathrm{SPE} = \|\,\mathbf{x}-\widehat{\mathbf{x}}\,\|_2^2, \quad \widehat{\mathbf{x}} = P_K P_K^\top \mathbf{x}, \tag{27}$$

Donde Λ_K son las varianzas de los K puntajes y P_K la K primeras cargas. Umbrales como $T^2_\alpha = \frac{K(n-1)}{n-K}F_{K,n-K,\alpha}$ permiten disparar alarmas tempranas ante derivas de proceso, mientras que SPE vigila no linealidades o patrones no explicados por el subespacio principal (Jolliffe, 2002). Este esquema casa naturalmente con (23) para ajustar x en tiempo real bajo restricciones de riesgo.

Limitaciones y validez externa. Si bien los R^2 cercanos a 1 sugieren ajuste excelente, prácticas de validación estrictas (bloqueo temporal, validación por pozo/nivel geológico) son necesarias para evitar *fugas* y sobreoptimismo en ambientes no i.i.d. Además, los resultados reflejan condiciones de una mina del norte de Chile; su generalización debe considerar diferencias geomecánicas, de flota y de planificación. El marco propuesto mitiga estos riesgos al anclar la decisión en predicciones *posteriores* y al explicitar el valor de nuevas mediciones mediante EVSI (Fauriat & Zio, 2020; Shen *et al.*, 2025).

CONCLUSIONES

Este estudio establece una *arquitectura estadística unificada* para optimizar perforación de producción, que combina:

(i) inferencia bayesiana robusta para comparación de empresas y cuantificación de incertidumbre; (ii) modelos predictivos de alto desempeño (RF, PLS) para pronóstico operativo; (iii) análisis multivariado (PCA) para monitoreo y reducción de dimensionalidad; y (iv) un *vínculo explícito* entre predicción y decisión vía (23) y diseño dirigido por información vía (26). En el conjunto de datos analizado, la Empresa B domina estocásticamente a la Empresa A en productividad (intervalos creíbles no solapados), y los modelos RF/PLS explican prácticamente toda la variabilidad de *Metros perforados*, con interpretaciones consistentes de importancia/cargas (Breiman, 2001; Bae *et al.*, 2023).

Las implicaciones prácticas del presente estudio son: (1) La regla (22) ofrece un criterio contractual transparente y auditable basado en probabilidades posteriores. (2) La política (23) permite sintonizar parámetros de perforación con control explícito de riesgo operativo. (3) La adquisición EI_t (25) y el EVSI (26) priorizan experimentos (corridas) con mayor retorno informacional por costo (Shen *et al.*, 2025; Fauriat & Zio, 2020). (4) El monitoreo T^2/SPE (27) habilita alarmas tempranas y facilita la integración con tableros de control.

Para futuros estudios, se pueden considerar los siguientes aspectos: (i) Extender el modelo bayesiano a jerarquías espacio–temporales (variación por banco/pozo/- turno) y verosimilitudes elípticas no gaussianas para capturar cambios de régimen; (ii) incorporar optimización multiobjetivo con métricas de desgaste/energía y adquisiciones paretianas robustas; (iii) explotar datos en tiempo real (telemetría de perforación) mediante filtros secuenciales bayesianos y *closed-loop* BO; (iv) profundizar la traz- abilidad causal integrando BN con estimadores de *valor de información* a nivel de sensor y modo de falla (Du & Chen, 2025; Li *et al.*, 2025). Estas líneas, alineadas con tendencias recientes en optimización probabilística de procesos, consolidan una base metodológica reproducible para la gestión de perforación basada en datos y riesgo (Shen *et al.*, 2025).

AGRADECIMIENTOS

Este trabajo fue posible gracias al apoyo técnico, operativo y analítico brindado por el equipo de ingeniería y per- foración de la mina ubicada en el norte de Chile, quienes facilitaron los datos operacionales utilizados en este estudio bajo estrictos protocolos de confidencialidad. Se agradece especialmente la colaboración de los supervisores de turno y del personal de terreno, cuyo conocimiento experto permitió interpretar las variables y validar la consistencia de los registros.

REFERENCIAS

Bae, S., Lee, H. & Kim, J. (2023). Adaptive bayesian optimization for real-time drilling rate optimization. *IEEE Access*, 11, 27826–27837.

Blei, D.M., Kucukelbir, A. & McAuliffe, J.D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association*, *112* (518), 859–877. https://doi.org/10.1080/01621459.2017.1285773

Boukredera, N., Othmani, A., Mellouk, A., Moualek, I. & Idiri, Z. (2025). Ai-driven optimization of drilling performance through torque management using machine learning and differential evolution. *Processes*, *13* (5),1472. https://doi.org/10.3390/pr13051472

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

Cihan, P. (2025). Bayesian hyperparameter optimization of machine learning models for predicting biomass gasification gases. *Applied Sciences*, *15* (3), 1018. https://doi.org/10.3390/app15031018

Cortes, P. & Onate, E. (2010). Optimization of drilling patterns using genetic algorithms. *Computers and Geotechnics*, *37* (3), 405–412.

Du, G. & Chen, A. (2025). Coal mine accident risk analysis with large language models and bayesian networks. *Sustainability*, *17* (5), 1896. https://doi.org/10.3390/su17051896

Fauriat, W. & Zio, E. (2020). Optimization of an aperiodic sequential inspection and condition-based maintenance policy driven by value of information. *Reliability Engineering & System Safety*, 204, 107133. https://doi.org/10.1016/j.ress.2020.107133

Gelman, A., Carlin, J. B., Stern, H.S., Dunson, D.B., Vehtari, A. & Rubin, D.B. (2013). *Bayesian data analysis*. Chapman and Hall/CRC. ISBN 978-1439840955. https://doi.org/10.1201/b16018

Hastie, T., Tibshirani, R. & Friedman, J. (2009). *The elements of statistical learning*. Springer Series in Statistics. https://doi.org/10.1007/978-0-387-84858-7

Jolliffe, I.T. (2002). *Principal Component Analysis*. Springer, 2nd edition.

Li, Z., Chen, H., Zhang, Y., Zhou, X., Huang, J. & Xu, Q. (2025). Bayesian network-based earth-rock dam breach probability analysis integrating machine learning. *Water*, *17* (21), 3085. https://doi.org/10.3390/w17213085

Liu, M., Yang, R., Bian, H., Sun, P., and Li, C. (2025). Mtgnn-bayesian-if-dbscan for time-series anomaly detection with applications in industrial drilling. *Sensors*, 25(15):4717. https://doi.org/10.3390/s25154717

Mohammad, D. & Belayneh, M. (2024). Field telemetry drilling dataset modeling with multivariable regression, gmdh, ann, and gmdh–ann. *Applied Sciences*, 14(6), 2273. https://doi.org/10.3390/app14062273

Shen, R., Luo, G. & Su, A. (2025). Bayesian optimization for chemical synthesis in the era of artificial intelligence: Advances and applications. *Processes*, *13* (9), 2687. https://doi.org/10.3390/pr13092687

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Wold, S., Ruhe, A., Wold, H. & Dunn, W. (1984). The collinearity problem in linear regression. the partial least squares (pls) approach to generalized inverses. *SIAM Journal on Scientific and Statistical Computing*, *5* (3), 735–743. https://doi.org/10.1137/0905052

Yi, Z., Li, Z., Yi, M., Wang, D. & Cheng, P. (2025). Intelligent real-time risk evaluation and drilling parameteroptimization for enhanced safety in deep-well operations. *Processes*, *13* (10), 3102. https://doi.org/10.3390/pr13103102