



Identificación de factores asociados al ausentismo laboral en una Institución Prestadora de Servicios de salud mediante modelos de analítica predictiva

Sara Lizeth Rojas González

Artículo de investigación presentado para optar al título de
Magíster en Gerencia del Talento Humano

Asesor
Andrés Alberto Osorioro, Doctor (PhD) en Ingeniería

Universidad de Manizales
Facultad de Ciencias Sociales y Humanas
Maestría en Gerencia del Talento Humano
Manizales, Caldas, Colombia

2026

Citar/How to cite	(Rojas González, 2026)
Referencia/Reference	Rojas González, (2026). <i>Identificación de factores asociados al ausentismo laboral en una institución prestadora de servicios de salud mediante modelos de analítica predictiva</i> [Tesis de maestría]. Universidad de Manizales. RIDUM: Repositorio Institucional Universidad de Manizales.
Estilo/Style: APA 7ma ed. (2020)	



Maestría en Gerencia del Talento Humano, XXIV

Biblioteca y Centro de Recursos: biblioteca.umanizales.edu.co

Repositorio Institucional: ridum.umanizales.edu.co

Universidad de Manizales: umanizales.edu.co

Revistas: revistasum.umanizales.edu.co

El contenido de esta obra corresponde al derecho de expresión de los autores y no compromete el pensamiento institucional de la Universidad de Manizales ni desata su responsabilidad frente a terceros. Los autores asumen la responsabilidad por los derechos de autor y conexos.

Resumen

Este estudio identifica los factores asociados al ausentismo laboral en una institución hospitalaria de alta complejidad. Su carácter novedoso radica en la aplicación de analítica predictiva avanzada en la gestión del talento humano dentro del sector salud, un ámbito en el que el uso de Human Resource Analytics para anticipar comportamientos laborales aún es limitado. Este enfoque se integra con la Visión de la Firma basada en Recursos, reconociendo la capacidad analítica en gestión humana como un activo estratégico de la organización. Se desarrolló un diseño cuantitativo, no experimental y retrospectivo, a partir de datos archivísticos de SES Hospital Universitario de Caldas (Manizales, Colombia), y se emplearon modelos supervisados de aprendizaje automático de regresión basados en Gradient Boosting y LightGBM con el propósito de predecir patrones de ausentismo y estimar la relevancia de variables organizacionales y sociodemográficas asociadas. Los modelos alcanzaron buena capacidad predictiva y estabilidad entre algoritmos, identificando de forma consistente un conjunto central de variables asociadas al ausentismo, entre ellas las vacaciones del año anterior, la edad, la antigüedad, la presencia de hijos y las horas extra, mientras que los turnos nocturnos y el área de trabajo mostraron menor influencia. A partir de estos hallazgos, el estudio aporta evidencia para el diseño de prácticas preventivas de gestión del talento humano orientadas a mitigar el ausentismo laboral en organizaciones hospitalarias complejas.

Palabras claves: ausentismo laboral, analítica de datos, sector salud, machine learning.

Abstract

This study identifies the factors associated with workplace absenteeism in a high-complexity hospital institution. Its novelty lies in the application of advanced predictive analytics in human resource management within the healthcare sector, a field where the use of Human Resource Analytics to anticipate workforce behavior remains limited. This approach is integrated with the Resource-Based View of the firm, recognizing analytical capability in human resource management as a strategic organizational asset. A quantitative, non-experimental, and retrospective design was implemented using archival data from SES Hospital Universitario de Caldas (Manizales, Colombia). Supervised machine learning regression models based on Gradient Boosting and LightGBM were employed to predict absenteeism patterns and estimate the relevance of associated organizational and sociodemographic variables. The models achieved good predictive capacity and stability across algorithms,

consistently identifying a core set of variables associated with absenteeism, including previous-year vacations, age, organizational tenure, having children, and overtime work, while night shifts and work area showed lower influence. Based on these findings, the study provides evidence for the design of preventive human resource management practices aimed at mitigating workplace absenteeism in complex hospital organizations.

Keywords: Workplace absenteeism, Human Resource Analytics, healthcare sector, machine learning, predictive analytics.

INTRODUCCIÓN

En los últimos años, la gestión del talento humano ha empezado a incorporar herramientas de analítica predictiva para anticipar comportamientos laborales y apoyar decisiones estratégicas, y en instituciones hospitalarias, donde la continuidad del servicio depende directamente de la disponibilidad del personal, esta capacidad adquiere especial relevancia; una herramienta que permite anticipar comportamientos laborales y optimizar la toma de decisiones estratégicas (Espgren & Hugosson, 2025). En este sentido, la analítica de datos aplicada a los recursos humanos (HR analytics - HRA de aquí en adelante) ha demostrado ser clave para identificar patrones de ausentismo laboral, facilitando intervenciones oportunas y efectivas para reducirlo (Fernández-Solís et al., 2024).

Investigaciones recientes muestran que la incorporación de HRA permite integrar datos organizacionales, métricas de desempeño y variables laborales para generar modelos predictivos que apoyen la toma de decisiones estratégicas en talento humano y mejoren los resultados organizacionales (Margherita, 2021).

La literatura en gestión humana coincide en que el uso de modelos predictivos en talento humano no solo mejora la eficiencia operativa, sino que fortalece la toma de decisiones desde una perspectiva estratégica (Madhani, 2023). Empresas que operan sectores altamente demandantes, como el de salud, han reportado una disminución significativa del ausentismo al incorporar sistemas de HRA, pues esto les permite implementar acciones o prácticas de gestión humana preventivas para reducirlo (Deloitte, 2023; HumanSmart, 2023). En consecuencia, comprender el comportamiento del ausentismo laboral mediante analítica predictiva se convierte en una estrategia crítica para fortalecer la gestión humana y anticipar escenarios de riesgo en contextos organizacionales complejos como el hospitalario (Rao et al., 2025).

En este sentido, SES Hospital Universitario de Caldas, una institución prestadora de servicios de salud con aproximadamente 1300 trabajadores, enfrenta desafíos significativos relacionados con el ausentismo laboral, especialmente aquel que se asocia a enfermedad común. Aunque cuenta con un programa de planeación táctica que garantiza la cobertura del personal en casos de ausencia, las causas subyacentes de dichas ausencias no han sido suficientemente analizadas desde una perspectiva basada en datos. Esta limitación impide que el área de gestión humana identifique patrones predictivos que permitan anticiparse y orientar estrategias efectivas de intervención que posibiliten la prevención del ausentismo. Además de los efectos directos en la continuidad del servicio, estas ausencias generan consecuencias indirectas como el incremento del cansancio laboral en quienes deben cubrir a sus compañeros, el deterioro en la calidad de la atención y mayores costos operativos.

A pesar del avance de la analítica predictiva en gestión humana en sectores como la banca o la manufactura, su aplicación en el contexto hospitalario sigue siendo incipiente (Marler & Boudreau, 2017; Minbaeva, 2018), pero particularmente escasa en el análisis del ausentismo por enfermedad común en instituciones de salud (Martin-Ríos & Erhardt, 2023). Esta brecha evidencia la necesidad de estudios aplicados que integren herramientas de analítica predictiva con la toma de decisiones estratégicas en gestión humana, adaptadas a la complejidad y especificidad del sector salud (Rao et al., 2025). En este sector, la analítica de datos en gestión humana permitiría alinear las decisiones del talento con los objetivos estratégicos de la organización (Huselid, 2018).

Por lo tanto, este estudio tiene como objetivo identificar los factores que inciden en el ausentismo laboral de una institución prestadora de servicios de salud (SES Hospital Universitario de Caldas), con base en modelos de analítica de datos predictiva. Por un lado, para lograr este objetivo, este estudio se sustenta en la Visión de la Firma basada en Recursos, dado que la HRA se aborda aquí como un recurso estratégico que permite mejorar el desempeño de las organizaciones; por otro lado, a nivel de la implementación de la analítica predictiva en gestión humana, el estudio sigue las fases del modelo CRISP-DM, ya que es un estándar metodológico que permite estructurar de forma lógica y sistemática los proyectos de análisis de datos en las organizaciones.

Desde una perspectiva conceptual, el presente estudio se apoya en la Visión de la Firma basada en Recursos (Resource-Based View, RBV), la cual plantea que los recursos estratégicos de una organización, entre ellos el capital humano, pueden constituir una fuente de ventaja competitiva cuando son valiosos, escasos y difíciles de imitar (Barney, 1991; Wright & McMahan, 2011). En este marco, la analítica de recursos humanos (Human Resource Analytics)

se ha consolidado como una capacidad organizacional emergente que permite transformar los datos laborales en información estratégica para la toma de decisiones en la gestión del talento (Minbaeva, 2018; Madhani, 2023). Particularmente, el análisis del ausentismo laboral mediante modelos predictivos permite comprender un fenómeno organizacional complejo influenciado por factores sociodemográficos, laborales y contextuales, tal como ha sido documentado en estudios sobre comportamiento laboral y salud ocupacional (Allebeck & Mastekaasa, 2004; Qamar & Samad, 2022). En este sentido, integrar herramientas de analítica predictiva con marcos teóricos organizacionales permite avanzar hacia modelos de gestión humana más estratégicos y basados en evidencia, especialmente en contextos organizacionales complejos como el sector salud.

Teoría central del estudio: la Visión de la Firma basada en Recursos

Teóricamente, la presente investigación se sustenta en la Visión de la Firma basada en Recursos (Resource-Based View of the Firm, RBV, por sus siglas en inglés), la cual plantea que los recursos estratégicos de una organización, entre ellos el capital humano pueden constituir una fuente de ventaja competitiva cuando son valiosos, escasos y difíciles de imitar (Barney, 1991; Wright & McMahan, 2011). Desde esta perspectiva, las personas son consideradas activos estratégicos que, cuando son gestionados de manera adecuada, pueden contribuir de forma diferenciadora al desempeño organizacional y a la sostenibilidad de las instituciones en entornos altamente demandantes.

Aplicar esta teoría al contexto de SES Hospital Universitario de Caldas implica reconocer que, si bien existen mecanismos de reemplazo temporal ante las ausencias, la pérdida recurrente de capacidad instalada humana afecta directamente la sostenibilidad del servicio. Así, el uso de modelos de analítica predictiva se convierte en una herramienta que, al integrarse con los postulados de la RBV, permite identificar factores internos asociados al ausentismo, anticiparse a su ocurrencia y proponer acciones estratégicas de gestión humana basadas en evidencia (Rao et al., 2025). De este modo, no solo se contribuye a una mejora operativa, sino que se fortalece la capacidad de la organización para sostener su ventaja competitiva en un entorno altamente exigente como el de la salud (Becker & Huselid, 2006; Minbaeva, 2018).

La analítica de datos, al ser aplicada estratégicamente, se convierte en un activo crucial para la organización como un todo (Minbaeva, 2018). En el contexto de los recursos humanos, la analítica predictiva no solo ayuda a gestionar la fuerza laboral de manera eficiente, sino que también facilita la toma de decisiones informadas que pueden mejorar los resultados organizacionales (Huselid, 2018). Según Pfeffer (2018), las organizaciones que integran la

analítica en sus procesos estratégicos son capaces de aprovechar los datos de sus empleados, transformándolos en información valiosa para la mejora continua y la alineación de sus objetivos.

La analítica predictiva, en particular, permite predecir patrones y comportamientos que afectan el desempeño organizacional, como el ausentismo, el desempeño individual o incluso el compromiso del recurso humano (Huselid, 2018). Este enfoque no solo optimiza los recursos existentes, sino que también fortalece la competitividad al permitir que la organización se adapte rápidamente a los cambios y desafíos del entorno (Minbaeva, 2018). A través de la analítica, las organizaciones pueden desarrollar estrategias proactivas que, al anticiparse a las necesidades del talento humano, mejoran la eficiencia operativa y, en consecuencia, los resultados financieros y de calidad (Becker & Huselid, 2006). Así, la analítica de datos se posiciona como un recurso estratégico que potencia la sostenibilidad a largo plazo de la organización, maximizando su capacidad para anticipar y responder a las necesidades tanto del personal como del entorno externo (Wright & McMahan, 2011).

Analítica de datos aplicada a la gestión humana (Human resource analytics o Workforce analytics)

La analítica de datos aplicada a la gestión humana, conocida también como *Human Resource Analytics* (HRA) o *Workforce Analytics*, es una disciplina emergente dentro del campo de la gestión del talento que busca transformar datos en conocimiento útil para la toma de decisiones estratégicas en recursos humanos (Minbaeva, 2018). Este enfoque combina herramientas de análisis estadístico, minería de datos, inteligencia artificial y modelado predictivo con el fin de entender, anticipar y optimizar fenómenos laborales como el ausentismo, la rotación, el desempeño individual, el compromiso y el bienestar organizacional (Bahuguna et al., 2024; Qamar & Samad, 2022).

A lo largo de la última década, la HRA ha evolucionado desde su rol tradicional orientado a la generación de reportes descriptivos, hacia un enfoque más proactivo y estratégico, en el que la analítica se convierte en un activo intangible que permite vincular las decisiones de talento con los resultados del negocio (Madhani, 2023). Desde esta perspectiva, el valor de la analítica no reside únicamente en el tratamiento técnico de los datos, sino en su capacidad de facilitar procesos de toma de decisiones basados en evidencia que respondan a objetivos organizacionales concretos (Minbaeva, 2018).

Uno de los retos esenciales en la consolidación de la HRA es la necesidad de superar las barreras de legitimidad dentro de las organizaciones (Huselid, 2018). Belizón y Kieran

(2022) advierten que la aceptación de la HRA como práctica estratégica de gestión humana requiere construir una narrativa sólida de credibilidad, en la que los profesionales de recursos humanos puedan comunicar con claridad el valor generado por los hallazgos analíticos, especialmente ante las altas directivas. En este sentido, la analítica se convierte en una herramienta estratégica que debe insertarse en la cultura organizacional para generar un impacto real (Huselid, 2018).

Adicionalmente, diversos estudios identifican tendencias clave en la adopción de la analítica de talento, destacando el crecimiento exponencial en la generación de datos, la integración de sistemas ERP y la necesidad de capacidades analíticas dentro de los equipos de gestión humana. Esto implica una transformación de los perfiles profesionales del área, pues deben incorporar competencias en ciencia de datos, pensamiento crítico y dominio de herramientas tecnológicas (Minbaeva, 2018).

Madhani (2023) señala que la analítica de recursos humanos permite a las organizaciones alinear sus prácticas de talento con la estrategia empresarial, mediante indicadores clave que faciliten la toma de decisiones basada en datos. De este modo, la HRA se posiciona como una vía para maximizar el desempeño del capital humano, consolidando su papel como ventaja competitiva sostenible (Minbaeva, 2018).

A pesar de su potencial, la investigación en HRA aún presenta vacíos importantes, especialmente en sectores complejos como el de la salud, en el que los fenómenos laborales están influenciados por múltiples variables contextuales, emocionales y organizacionales (Rao et al., 2025). Según Qamar y Samad (2022), la literatura en esta área sigue siendo incipiente, con una creciente necesidad de estudios empíricos que muestren aplicaciones prácticas de la analítica para resolver problemas específicos, como el ausentismo, desde una perspectiva predictiva.

Ausentismo laboral y su impacto en las organizaciones

El ausentismo laboral es una de las principales problemáticas que enfrentan las organizaciones contemporáneas, especialmente aquellas que operan en sectores intensivos en talento humano como la salud (Guzmán & Ríos, 2021). Este fenómeno hace referencia a la ausencia del trabajador en su puesto de trabajo durante el tiempo que estaba programado para laborar, por razones justificadas o injustificadas, lo que afecta la dinámica operativa y los resultados organizacionales (Aloğlu & Güllü, 2022).

Diversas investigaciones han evidenciado que el ausentismo no solo implica la interrupción de las funciones asignadas, sino que genera una cadena de efectos negativos como

la sobrecarga del personal activo, la disminución en la productividad, el deterioro en la calidad del servicio y el aumento de los costos asociados a reemplazos o reprogramaciones (Borda et al., 2017). En el sector salud, estas consecuencias adquieren mayor relevancia, puesto que la continuidad y calidad del cuidado al paciente pueden verse directamente comprometidas (Aloğlu & Güllü, 2022).

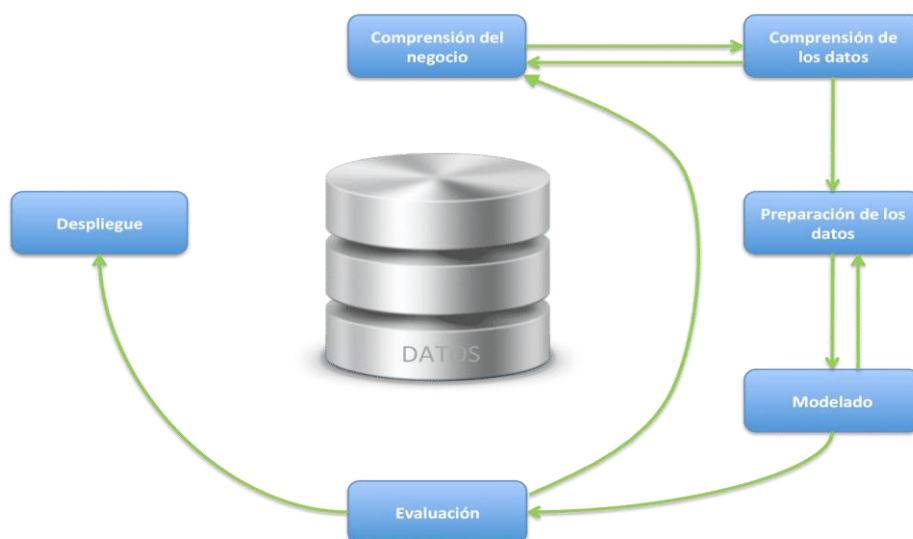
Entre los factores más comunes que explican el ausentismo se encuentran las enfermedades comunes, los accidentes laborales, el estrés crónico, el burnout, como también condiciones asociadas con la salud mental (Ardila Díaz, 2023). En efecto, la dimensión psicosocial del trabajo ha cobrado mayor protagonismo en los últimos años, y se reconoce que ambientes laborales tóxicos, exigencias excesivas o la falta de conciliación entre vida laboral y personal pueden detonar episodios de incapacidad o ausencias recurrentes (Borda et al., 2017).

La literatura señala que el ausentismo también está vinculado a la deficiente gestión de riesgos laborales y a la ausencia de estrategias efectivas de prevención. En este sentido, el ausentismo no debe ser visto únicamente como una condición individual, sino como un síntoma organizacional que revela brechas en la cultura de cuidado, el liderazgo, las condiciones de trabajo y la gestión del talento humano (Guzmán & Ríos, 2021).

Modelo CRISP-DM aplicado en el análisis del ausentismo laboral

El modelo CRISP-DM (Cross Industry Standard Process for Data Mining) es una metodología estandarizada para el desarrollo de proyectos de minería de datos que ha sido adoptada ampliamente por su estructura flexible, iterativa y adaptable a diversos sectores (Mejía Rodríguez et al., 2024). Está compuesto por seis fases: comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación y despliegue. Cada etapa permite una aproximación progresiva desde la identificación del problema hasta la implementación de soluciones basadas en datos (Moine et al., 2013).

Figura 1. Fases del modelo CRISP-DM



Fuente: Blasco (2021).

Particularmente, en el análisis del ausentismo laboral, este modelo permite abordar de forma estructurada los factores que inciden en las ausencias, facilitando la creación de modelos predictivos para anticipar comportamientos de los empleados. La fase de comprensión del negocio implica definir el problema desde una perspectiva estratégica, es decir, establecer qué variables influyen en el ausentismo. Posteriormente, en la fase de comprensión de los datos, se identifican las fuentes relevantes, como registros de ausencias, causas médicas, edad, antigüedad y otros datos demográficos y laborales.

Durante la preparación de los datos, estos se limpian, transforman y estructuran para asegurar su calidad y relevancia analítica. La fase de modelado permite aplicar algoritmos de clasificación, regresión o clustering para identificar patrones que expliquen las fuentes del ausentismo. En la etapa de evaluación, se validan los resultados obtenidos y se determina su capacidad predictiva, mientras que, en la fase de despliegue, se generan reportes y visualizaciones que facilitan la toma de decisiones en la gestión del talento humano (Barrera Castillo et al., 2023).

Por último, el CRISP-DM ha sido utilizado para desarrollar sistemas de apoyo a la toma de decisiones en entornos de salud, demostrando su potencial para reducir las tasas de ausentismo mediante intervenciones anticipadas y focalizadas (Moine et al., 2013). Su aplicación permite pasar de una gestión reactiva a una gestión proactiva del talento, aportando valor estratégico a la organización al anticipar riesgos y mejorar la asignación de recursos humanos. La literatura reciente destaca que la analítica de talento permite transformar datos

organizacionales en conocimiento estratégico para la planificación de la fuerza laboral y la toma de decisiones basada en evidencia, especialmente mediante el uso de modelos predictivos y técnicas de aprendizaje automático (Menon & Sethi, 2025; Madhuri & Kumar, 2025).

Variables predictoras del ausentismo por incapacidad en el sector salud

El ausentismo por incapacidad médica en el sector salud es un fenómeno de alta importancia organizacional debido a su impacto en la continuidad asistencial, los costos laborales y la carga operativa de las instituciones sanitarias (Aloğlu & Güllü, 2022). La literatura científica sobre ausentismo ha demostrado que la ocurrencia y duración de los días de incapacidad responden a un fenómeno multifactorial en el que convergen aspectos individuales, condiciones laborales, organización del tiempo de trabajo y antecedentes de salud (Allebeck & Mastekaasa, 2004; Merkus et al., 2012). En estudios recientes basados en registros administrativos de hospitales y sus sistemas de información organizacional, este enfoque multidimensional ha permitido modelar el ausentismo con datos objetivos, los cuales reflejan con claridad lo que sucede respecto a dicha variable en las dinámicas organizacionales del hospital; Así pues, la selección de variables predictoras debe justificarse tanto por evidencia empírica como por su disponibilidad en registros institucionales. A continuación, se sintetiza la literatura que respalda cada grupo de predictores empleados en el presente estudio.

Factores sociodemográficos

Consistentemente, la edad ha sido considerada como un determinante relevante del ausentismo por incapacidad (Aloğlu & Güllü, 2022). Estudios realizados en trabajadores del sector salud muestran que los empleados de mayor edad presentan mayor probabilidad de ausencias prolongadas, lo que se atribuye a una mayor prevalencia de condiciones crónicas y tiempos de recuperación más extensos (Sukhee et al., 2023). No obstante, otros estudios advierten que trabajadores más jóvenes tienden a presentar mayor frecuencia de ausencias cortas (Allebeck & Mastekaasa, 2004). En consecuencia, la edad se considera un predictor necesario en modelos de días totales de incapacidad.

El sexo también ha mostrado asociaciones con el ausentismo en el sector salud. Investigaciones poblacionales indican que parte de las diferencias por sexo, en materia de ausentismo, se explican por segregación ocupacional, doble carga trabajo–familia y exposición diferencial a demandas emocionales (Laaksonen et al., 2010). Por ende, el sexo también debe integrarse en modelos que buscan predecir la incapacidad laboral.

Adicionalmente, desde lo sociodemográfico, el estado civil y la presencia de hijos se incluyen habitualmente como aproximaciones a responsabilidades familiares y soporte social.

Revisiones sobre factores individuales del ausentismo muestran que estas variables presentan efectos heterogéneos y dependientes del contexto organizacional y cultural (Allebeck & Mastekaasa, 2004). En el sector salud, la rigidez de los turnos y la carga asistencial pueden amplificar la influencia de demandas familiares, aunque la evidencia empírica sugiere que estos predictores suelen presentar efectos modestos sobre los días totales de incapacidad (Aloğlu & Güllü, 2022).

Factores laborales y trayectoria organizacional

La antigüedad en la empresa suele asociarse con el ausentismo de largo plazo en organizaciones del sector salud. Estudios basados en registros hospitalarios indican que ciertos rangos de antigüedad presentan mayor riesgo de ausencias prolongadas, posiblemente debido a la acumulación de exposición a riesgos físicos y psicosociales (Sukhee et al., 2023). Sin embargo, investigaciones longitudinales también evidencian que una mayor antigüedad puede asociarse con mayor adaptación al puesto y estabilidad laboral, generando efectos no lineales. Por ello, su inclusión como predictor es ampliamente recomendada en modelos que pretenden predecir el ausentismo.

En adición, el área o unidad de trabajo constituye un proxy estructural de exposición laboral en el sector salud. Diferencias entre unidades asistenciales, administrativas o técnicas implican variaciones en carga física, riesgo biológico, intensidad emocional y presión operativa. Estudios en hospitales han documentado diferencias sustantivas en tasas de ausentismo entre servicios clínicos, quirúrgicos y administrativos. Esto respalda el uso de esta variable como predictor de tipo categórico en modelos de ausentismo institucional.

Organización del tiempo del trabajo

El trabajo por turnos nocturnos ha sido uno de los factores más consistentemente asociados al ausentismo por incapacidad en el sector salud. Revisiones sistemáticas de la literatura establecen que los turnos nocturnos y rotatorios se relacionan con alteraciones del sueño, fatiga crónica y mayor riesgo de enfermedades, incrementando la probabilidad de ausencias médicas (Merkus et al., 2012). Evidencia reciente basada en registros hospitalarios demuestra además efectos agudos, dado que la probabilidad de que un empleado del sector salud se sienta enfermo aumenta inmediatamente después de secuencias de turnos nocturnos (Larsen et al., 2023).

También dentro de la organización del tiempo de trabajo, las horas extra representan una medida de intensificación laboral. Investigaciones en personal sanitario muestran que jornadas extendidas y turnos prolongados se asocian con mayor fatiga y riesgo de ausentismo,

particularmente por causas de salud mental (Dall’Ora et al., 2018; Inoue et al., 2022). Sin embargo, otros estudios observan que el sobretiempo puede coexistir con menor ausentismo de corto plazo, generando resultados empíricos mixtos (Bernstrøm & Houkes, 2018). Esta evidencia respalda su inclusión como predictor en el modelo que se evalúa en esta investigación, cuya dirección esperada puede depender del contexto institucional.

Asimismo, el goce de vacaciones en el último año se vincula teóricamente con procesos de recuperación y restauración del bienestar y la salud. Estudios longitudinales sobre efectos de las vacaciones muestran mejoras temporales en salud y bienestar psicológico tras periodos de descanso (de Bloom et al., 2013). Sin embargo, análisis cuasi-experimentales sobre políticas de vacaciones pagadas han encontrado efectos limitados sobre indicadores de salud y ausentismo prolongado (Hofmarcher, 2021). En consecuencia, esta variable se justifica como una variable proxy organizacional de oportunidades de recuperación, con efectos potencialmente contingentes al contexto.

Seguridad y salud en el trabajo

Los accidentes laborales se relacionan directamente con días de incapacidad, particularmente cuando implican lesiones musculoesqueléticas o eventos traumáticos. Estudios en personal sanitario indican que los trabajadores que experimentan accidentes laborales presentan mayor riesgo de ausencias prolongadas posteriores, incluyendo secuelas físicas y psicológicas (Farrants et al., 2024; Laaksonen, 2013).

METODOLOGÍA

Diseño de investigación

Este estudio adoptó como diseño un enfoque cuantitativo con un diseño no experimental y retrospectivo. Se trabajó con una base de datos institucional consolidada que integra los registros de ausentismo laboral medido en número de días y sus variables predictoras correspondientes al periodo 2023–2024. Aunque algunos empleados aparecen en múltiples registros a lo largo del periodo observado, el análisis se efectuó sobre un *dataset* integrado, sin modelar trayectorias individuales ni efectos temporales, por lo que el estudio no se configura como longitudinal en sentido estricto.

Desde el punto de vista analítico, en este estudio se ejecutaron modelos de aprendizaje automático supervisado de regresión, destinados a optimizar la capacidad predictiva del ausentismo individual.

Contexto organizacional

El presente estudio se desarrolló en SES Hospital Universitario de Caldas, una Empresa Mixta ubicada en Manizales, Colombia, que opera con un modelo docente-asistencial, sin ánimo de lucro y con un nivel de complejidad alto; La institución presta servicios especializados y de alta complejidad, incluyendo hospitalización, cirugía cardiovascular, atención oncológica, unidades de cuidado intensivo y urgencias, respaldados por una infraestructura moderna y tecnología biomédica avanzada. Además, cuenta con certificaciones como la Acreditación en Salud por ICONTEC, el Galardón Nacional Hospital Seguro y la certificación como Empresa Familiarmente Responsable (EFR).

Con un equipo humano de aproximadamente 1300 trabajadores, el hospital se distingue por una cultura organizacional centrada en la humanización, el bienestar del talento humano y la innovación continua. Bajo este enfoque, ha desarrollado programas como el Gimnasio de Habilidades Esenciales, Hospital Sana con Arte, y el Mes de I+D+i, reafirmando su compromiso con la salud integral de sus colaboradores y la mejora constante en la atención en salud.

En este contexto, surge la oportunidad de explorar el potencial de la analítica predictiva como herramienta estratégica en gestión humana. Más allá de abordar un problema puntual, la investigación se propone comprender con mayor profundidad el comportamiento del ausentismo laboral, integrando modelos analíticos que permitan anticipar tendencias, identificar patrones y orientar decisiones organizacionales basadas en datos. Este estudio está alineado con los retos actuales del sector salud y con la visión institucional de alcanzar resultados superiores en salud, sostenibilidad e innovación hacia el año 2026.

Participantes en el estudio

La población del estudio estuvo constituida por la totalidad de los empleados activos de la organización durante el periodo 2023–2025, por lo que se trabajó con un censo organizacional y no con una muestra.

Fuente de los datos y estrategia de investigación adoptada

Los datos utilizados en este estudio fueron obtenidos a partir del sistema ERP institucional de la organización, que integra los módulos de gestión humana, nómina y control de ausentismo. En consecuencia, no se emplearon instrumentos de recolección primaria como encuestas o entrevistas, sino que se trabajó con datos secundarios organizacionales registrados de manera rutinaria en los sistemas de información corporativos. Este enfoque corresponde a una estrategia de investigación denominada análisis de datos archivísticos (archival data), ampliamente utilizada en estudios de HR analytics y modelamiento organizacional.

Operacionalización de las variables

Variables predictoras

Edad: se midió como variable continua en años al cierre del año. La edad es un predictor habitual en ausentismo por enfermedad en estudios generales y del sector salud, por su asociación con perfiles de salud y probabilidad de ausencias de mayor duración (Allebeck & Mastekaasa, 2004).

Sexo: se codificó según el registro administrativo del ERP (0 = hombre, 1 = mujer: categorías equivalentes del ERP). El sexo es un determinante estructural frecuentemente incluido en modelos de ausentismo por su asociación con diferencias ocupacionales, exposición y patrones de ausencia, especialmente en sectores feminizados como salud (Allebeck & Mastekaasa, 2004; Koopmans et al., 2010).

Estado civil: se registró según la clasificación del ERP (p. ej., soltero, casado, unión libre, separado, y viudo). Su inclusión es consistente con revisiones que lo consideran un factor individual asociado (de forma variable) al ausentismo, usado como covariable sociodemográfica en modelos explicativos/predictivos (Allebeck & Mastekaasa, 2004).

Hijos: Se codificó como indicador binario de parentalidad (0 = no, 1 = sí). Este proxy se utiliza ampliamente en investigación de ausentismo por representar demandas familiares y contextos de vida (Allebeck & Mastekaasa, 2004).

Antigüedad en la empresa: se operacionalizó como el número de años desde la fecha de ingreso hasta el cierre del año. Esta operacionalización es coherente con enfoques longitudinales administrativos donde la trayectoria laboral se asocia a exposición acumulada y adaptación saludable, y suele incluirse como predictor en modelos de ausentismo (Laaksonen et al., 2010).

Turno nocturno: se definió como pertenencia (durante el año) a un esquema de trabajo que incluye turnos nocturnos, de acuerdo con la codificación del ERP (dentro del modelo se incluye como: 0 = no, 1 = sí). Aunque el ideal sería capturar intensidad (número de noches), el binario es una aproximación válida cuando los registros no permiten granularidad. La evidencia muestra asociaciones consistentes entre trabajo por turnos/noches y ausentismo por enfermedad, incluyendo evidencia específica en trabajadoras del sector salud (Merkus et al., 2012; Larsen et al., 2023).

Área donde trabaja: se registró como la unidad funcional/área definida por la institución (p. ej., urgencias, hospitalización, UCI, laboratorio, administrativo), transformada en variables nominales dentro del modelo. Esta operacionalización es adecuada porque el área

funciona como proxy de exposición ocupacional (carga física, riesgo biológico y demandas emocionales), y los estudios muestran diferencias sistemáticas de ausentismo según condiciones y arreglos de trabajo vinculados a entornos laborales (Laaksonen et al., 2010).

Horas extra: Se codificó como indicador binario de realización de horas extra durante el año (0 = no, 1 = sí), conforme a la trazabilidad en el ERP. Si bien un indicador continuo (horas totales) sería más sensible, el binario es defendible cuando la medición detallada no está disponible. La literatura apoya que jornadas extendidas y sobretiempo se asocian con riesgos de salud y ausentismo (Bernstrøm & Houkes, 2018; Inoue et al., 2022).

Vacaciones en el último año: Se operacionalizó como indicador de si el trabajador registró vacaciones durante el año, según el ERP (0 = no, 1 = sí). Aunque no es un predictor clásico en todos los modelos de ausentismo, su inclusión es conceptualmente consistente con la literatura de recuperación, puesto que las vacaciones pueden favorecer la restauración de bienestar y la salud (de Bloom et al., 2013; Hofmarcher, 2021). En consecuencia, el binario se justifica como proxy organizacional de oportunidades de recuperación.

Accidentes laborales: se definió como ocurrencia de al menos un accidente laboral registrado durante el año (0 = no, 1 = sí). Esta operacionalización es consistente con literatura que asocia lesiones/accidentes con patrones posteriores de incapacidad y ausentismo, y es común en registros donde no siempre se dispone de severidad detallada (Farrants & Alexanderson, 2024; Johansson et al., 2021).

Variable dependiente: incapacidad médica

Se operacionalizó como el número total de días de incapacidad médica acumulados por cada empleado dentro del año (0 a n días), incluyendo todas las causas disponibles en el registro. Esta forma de medición es estándar en investigación basada en registros de archivo, en la que el ausentismo se analiza por duración o días acumulados y permite captar mejor la carga total para la organización que únicamente el conteo de episodios (Blomgren & Jäppinen, 2021; Laaksonen, 2013).

Preprocesamiento de datos

Previo al modelamiento estadístico y predictivo, se realizó una fase de procesamiento orientada a garantizar la calidad y consistencia de los datos extraídos del sistema ERP institucional. La base de datos se estructuró en formato empleado-año, integrando registros sociodemográficos, laborales y de incapacidad médica. Este enfoque es consistente con estudios basados en registros administrativos, donde la calidad del dato depende principalmente de la trazabilidad operativa de los sistemas de información organizacionales (Redman, 2013).

Auditoría y control de duplicados

Como primer paso, el equipo de investigación efectuó una auditoría de integridad de registros, verificando la unicidad de cada combinación empleado–año. No se identificaron registros duplicados, lo que indica consistencia en los procesos de consolidación provenientes del ERP. La detección y eliminación de duplicados es una práctica estándar en procesos de *data cleaning*, dado que su presencia puede sesgar estimaciones de parámetros y métricas predictivas (Rahm & Do, 2000). La ausencia de duplicados confirma la estabilidad estructural de la base y la confiabilidad del sistema transaccional de origen.

Valores faltantes

La revisión de completitud evidenció ausencia de valores faltantes en las variables analizadas. Esto se explica porque los datos emergen directamente de módulos operativos del ERP (gestión humana, nómina, incapacidades y programación de turnos), donde los campos analizados son obligatorios para la ejecución de procesos administrativos y contractuales. En estudios basados en registros organizacionales, es frecuente encontrar niveles bajos de valores perdidos en variables operativas, en contraste con estudios basados en encuestas o autoinforme (Little & Rubin, 2019). Por tanto, no fue necesario aplicar técnicas de imputación o exclusión de casos.

Tratamiento de valores atípicos (outliers)

A diferencia de estudios experimentales o encuestas, en bases de datos empresariales los valores extremos suelen corresponder a eventos reales de interés, como incapacidades prolongadas por enfermedad grave o accidentes laborales. La literatura metodológica en *business analytics* advierte que la eliminación automática de outliers en datos operativos puede implicar la pérdida de información crítica sobre fenómenos organizacionales relevantes (Aggarwal, 2017). En consecuencia, no se eliminaron valores atípicos, preservando la distribución real de los días de incapacidad y manteniendo la validez ecológica del análisis.

Normalización y codificación

Las variables continuas (edad, antigüedad y días de incapacidad) se conservaron en su escala original para permitir interpretación directa en los modelos de regresión. Las variables categóricas fueron transformadas mediante codificación binaria o variables indicadoras (*dummy variables*), creando columnas independientes para cada categoría con valores 0 y 1. Para evitar problemas de colinealidad perfecta o *dummy variable trap*, se eliminó una categoría de referencia en cada conjunto de variables categóricas, permitiendo que el modelo estimara los

efectos relativos respecto a dicha categoría base. Este procedimiento es consistente con prácticas estándar en modelamiento estadístico y aprendizaje supervisado (James et al., 2021).

En particular:

- El área de trabajo fue transformada en variables indicadoras tomando como categoría de referencia el área administrativa.
- El estado civil se codificó tomando como categoría base “soltero(a)”.
- El nivel educativo se codificó tomando como categoría base el nivel técnico/tecnológico.
- Las variables binarias originales (sexo, hijos, turnos nocturnos, vacaciones, accidentes, horas extra) se mantuvieron en formato 0–1 sin requerir expansión adicional.

Consideración sobre balance de clases

El presente estudio implementó modelos de aprendizaje automático supervisado de regresión, dado que la variable objetivo corresponde al número total de días de incapacidad, una medida continua. En contraste con los problemas de clasificación supervisada, donde la variable dependiente es categórica y puede presentarse desbalance entre clases, en los modelos de regresión no existen clases discretas cuya proporción de observaciones pueda sesgar el proceso de aprendizaje (James et al., 2021; Hastie et al., 2017). Por tanto, los procedimientos tradicionales de class balancing, tales como oversampling, undersampling o generación sintética de observaciones (SMOTE), no resultan aplicables en este contexto.

En modelos de regresión supervisada, la principal consideración no es el balance entre clases, sino la distribución empírica de la variable dependiente. La literatura metodológica en Machine Learning Regression señala que, cuando los valores extremos representan fenómenos reales y no errores de medición —como ocurre en registros de incapacidades prolongadas por enfermedad grave o accidente laboral—, preservar la distribución original mejora la capacidad predictiva y la interpretabilidad del modelo (Aggarwal, 2017; Hastie et al., 2017). En consecuencia, no se aplicaron técnicas de re-muestreo ni transformaciones artificiales de la distribución de los días de incapacidad, con el fin de mantener la fidelidad del modelo respecto al comportamiento real del ausentismo institucional.

Este enfoque es coherente con estudios recientes de analítica organizacional basados en datos administrativos, donde los modelos de regresión supervisada buscan capturar la variabilidad natural del fenómeno y generar predicciones útiles para la gestión operativa, más que forzar distribuciones estadísticas idealizadas (Blomgren & Jäppinen, 2021).

Procedimiento de análisis de datos

Machine Learning

El análisis de datos se desarrolló en el entorno de programación Google Colab (Python) utilizando las librerías Pandas y Numpy para la manipulación y estructuración de los datos. La base de información fue importada desde un archivo en formato CSV. La variable dependiente, correspondiente al total de días de incapacidad acumulados por cada trabajador, fue inicialmente leída como texto, por lo que se realizó una transformación de formato reemplazando separadores decimales inconsistentes y convirtiéndola a tipo numérico flotante, garantizando su adecuada lectura por los algoritmos de aprendizaje automático. Posteriormente, se definieron dos matrices de trabajo: la matriz X, compuesta por variables predictoras de naturaleza sociodemográfica, laboral y organizacional, y la columna y, correspondiente a la variable objetivo de ausentismo laboral.

Para el desarrollo de los modelos predictivos se utilizó la librería Scikit-learn, ampliamente empleada en investigación científica y proyectos de aprendizaje automático; La referencia académica de esta librería corresponde al trabajo de Pedregosa et al. (2011), quienes describen su arquitectura y funcionalidades para la implementación de algoritmos de machine learning en el lenguaje Python.

El conjunto total de datos fue dividido en particiones de entrenamiento (80%) y prueba (20%) mediante muestreo aleatorio con semilla fija, con el propósito de asegurar la reproducibilidad del proceso analítico. Con el fin de garantizar la estabilidad del flujo de procesamiento ante la eventual presencia de valores faltantes, se implementó un procedimiento de imputación simple mediante la media aritmética de cada variable predictora utilizando la clase *SimpleImputer* de *Scikit-learn*. La imputación preventiva constituye una práctica recomendada en flujos reproducibles de aprendizaje automático, puesto que asegura que el pipeline pueda ejecutarse sin intervención manual ante actualizaciones futuras de la base de datos (Kuhn & Johnson, 2013).

Posteriormente, las variables predictoras fueron sometidas a un proceso de estandarización mediante *StandardScaler*, ajustando cada variable a media cero y desviación estándar unitaria. El escalamiento se calculó exclusivamente sobre el conjunto de entrenamiento y posteriormente se aplicó al conjunto de prueba, evitando fuga de información entre particiones y garantizando coherencia estadística en el proceso de normalización. La estandarización de variables es una etapa habitual en algoritmos basados en gradiente, en el que la escala de los predictores influye en la estabilidad del entrenamiento (Hastie et al., 2009).

Formulación del problema de aprendizaje automático

Dado que la variable objetivo corresponde a una variable continua (días de incapacidad), el problema para resolver la pregunta de estudio fue formulado como una tarea de aprendizaje supervisado de regresión múltiple. Se seleccionaron algoritmos de aprendizaje basados en ensamblaje de árboles mediante *Boosting* debido a su capacidad para modelar relaciones no lineales, capturar interacciones complejas entre predictores y ofrecer mecanismos de interpretación a través de análisis de importancia de variables (Friedman, 2001).

Modelo Gradient Boosting Regressor

Como primer enfoque de modelamiento se implementó el algoritmo *Gradient Boosting Regressor* de *Scikit-learn*. Este método construye secuencialmente múltiples árboles de decisión poco profundos, donde cada nuevo árbol se entrena para corregir los errores residuales del conjunto anterior. Este algoritmo fue seleccionado por su capacidad para manejar relaciones no lineales entre variables predictoras, su robustez frente a diferentes escalas de medición y su posibilidad de extraer métricas internas de importancia de variables, aspecto relevante para la interpretación de fenómenos organizacionales complejos (Hastie et al., 2009). El modelo fue entrenado empleando las variables predictoras previamente estandarizadas en el conjunto de entrenamiento.

Optimización de hiperparámetros mediante validación cruzada

Con el propósito de evitar configuraciones arbitrarias de hiperparámetros y mejorar la capacidad de generalización del modelo, se implementó un procedimiento de búsqueda exhaustiva mediante *GridSearchCV* utilizando validación cruzada de cinco particiones. La optimización sistemática de hiperparámetros mediante validación cruzada permite reducir el riesgo de sobreajuste y proporciona estimaciones más estables del desempeño esperado del modelo (Arlot & Celisse, 2010).

Modelo LightGBM Regressor como contraste metodológico

Adicionalmente, se implementó un segundo algoritmo de *Boosting* denominado *LightGBM Regressor*, una variante optimizada de *Gradient Boosting* basada en histogramas y crecimiento de árboles por hojas. La inclusión de *LightGBM* tuvo como finalidad contrastar la estabilidad del aprendizaje con un algoritmo alternativo de *Boosting* y verificar la consistencia de la estructura de aprendizaje obtenida. *LightGBM* ha sido ampliamente adoptado por su eficiencia computacional y su capacidad para manejar conjuntos de datos pequeños o medianos con múltiples predictores (Ke et al., 2017).

Evaluación de importancia de variables

Con el fin de dotar de interpretabilidad a los modelos entrenados, se implementaron dos estrategias complementarias de análisis de importancia de variables. En primer lugar, se extrajo la importancia interna de cada modelo, basada en la frecuencia y calidad de las divisiones realizadas por cada variable dentro de los árboles que componen los ensambladores. En segundo lugar, se aplicó el método de importancia por permutación, que consiste en desordenar aleatoriamente los valores de cada variable en el conjunto de prueba y cuantificar la pérdida de capacidad predictiva del modelo. Este procedimiento proporciona una medida de importancia basada en el impacto real de cada variable sobre el aprendizaje del modelo, independiente de la arquitectura interna del algoritmo (Breiman, 2001; Fisher et al., 2019).

RESULTADOS

Desempeño de los modelos predictivos

De acuerdo con lo planteado en la metodología, con el fin de estimar la capacidad predictiva de los algoritmos seleccionados, se entrenaron y evaluaron tres modelos de regresión supervisada: (1) Gradient Boosting en configuración base, (2) Gradient Boosting optimizado mediante búsqueda en malla de hiperparámetros y (3) LightGBM en configuración base. La evaluación se realizó sobre el conjunto de prueba, empleando como métricas el error cuadrático medio (RMSE), el error absoluto medio (MAE) y el coeficiente de determinación (R^2). La Tabla 1 presenta los indicadores de desempeño obtenidos para cada modelo.

Tabla 1. Indicadores de desempeño de los modelos predictivos

Modelo	RMSE	MAE	R^2
Gradient Boosting (Base)	36.838	18.668	0.252
Gradient Boosting (Optimizado)	39.631	20.040	0.134
LightGBM (Base)	36.577	19.150	0.263

Los valores de RMSE y MAE se sitúan en rangos estrechos entre modelos, lo que indica magnitudes de error promedio similares en la predicción de los días de incapacidad. En cuanto al coeficiente de determinación, los valores de R^2 oscilan entre 0.134 y 0.263, reflejando la proporción de varianza explicada por cada algoritmo sobre el conjunto de prueba. Aunque los valores de R^2 pueden parecer moderados desde una perspectiva estrictamente estadística, en el contexto organizacional del hospital donde el ausentismo está influido por múltiples factores no registrados en el ERP, estos niveles de explicación resultan razonables y operativamente útiles, dado que el comportamiento humano y laboral está influenciado por múltiples factores no observados o difícilmente cuantificables. Cohen (1988) estableció criterios ampliamente

aceptados en ciencias del comportamiento, donde valores de R^2 alrededor de 0.13 representan efectos de magnitud moderada y valores iguales o superiores a 0.26 representan efectos sustanciales. Revisiones metodológicas posteriores han reafirmado que un R^2 en el rango aproximado de 0.10 a 0.30 son frecuentes y adecuados en estudios no experimentales sobre comportamiento social y organizacional (Ferguson, 2009; Ellis, 2010).

De manera complementaria, investigaciones recientes en predicción de ausentismo laboral mediante Machine Learning en bases organizacionales han reportado coeficientes R^2 en rangos comparables, destacando su utilidad operativa en contextos de small data empresarial (Llamas-Blázquez et al., 2025). En consecuencia, los valores obtenidos en este estudio se sitúan dentro de los rangos reportados como adecuados para investigación predictiva aplicada en ciencias sociales (Cohen, 1988; Ferguson, 2009).

Comparación del desempeño entre algoritmos

El modelo LightGBM en configuración base presentó el mayor coeficiente de determinación ($R^2 = 0.263$), seguido del modelo Gradient Boosting base ($R^2 = 0.252$). Ambos algoritmos exhiben valores de RMSE y MAE similares, lo que evidencia estabilidad en la magnitud de los errores de predicción entre enfoques de boosting distintos. Por su parte, el modelo Gradient Boosting optimizado mostró un descenso en el desempeño ($R^2 = 0.134$), acompañado de incrementos en RMSE y MAE. Este resultado indica que la optimización de hiperparámetros no produjo mejoras en la capacidad de generalización sobre el conjunto de prueba.

Importancia de variables en el modelo Gradient Boosting (configuración base)

Importancia interna de variables

La Tabla 2 presenta las diez variables con mayor importancia interna estimada por el modelo Gradient Boosting en su configuración base.

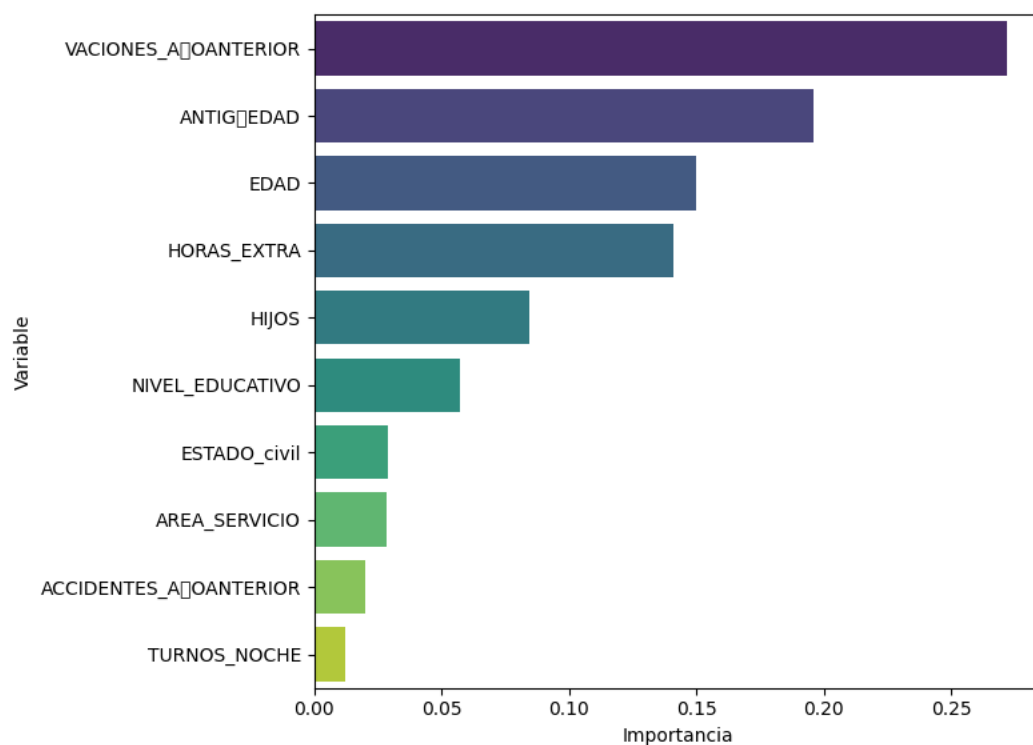
Tabla 2. Importancia interna de variables (Gradient Boosting base)

Variable	Importancia
Vacaciones del año anterior	0.272
Antigüedad	0.196
Edad	0.150
Horas extra	0.141
Hijos	0.084
Nivel educativo	0.057
Estado civil	0.029
Área o servicio	0.029
Accidentes año anterior	0.020

Turnos noche	0.012
--------------	-------

La Figura 1 presenta de forma gráfica los anteriores resultados.

Figura 1. Importancia interna de variables (Gradient Boosting base)



Importancia por permutación

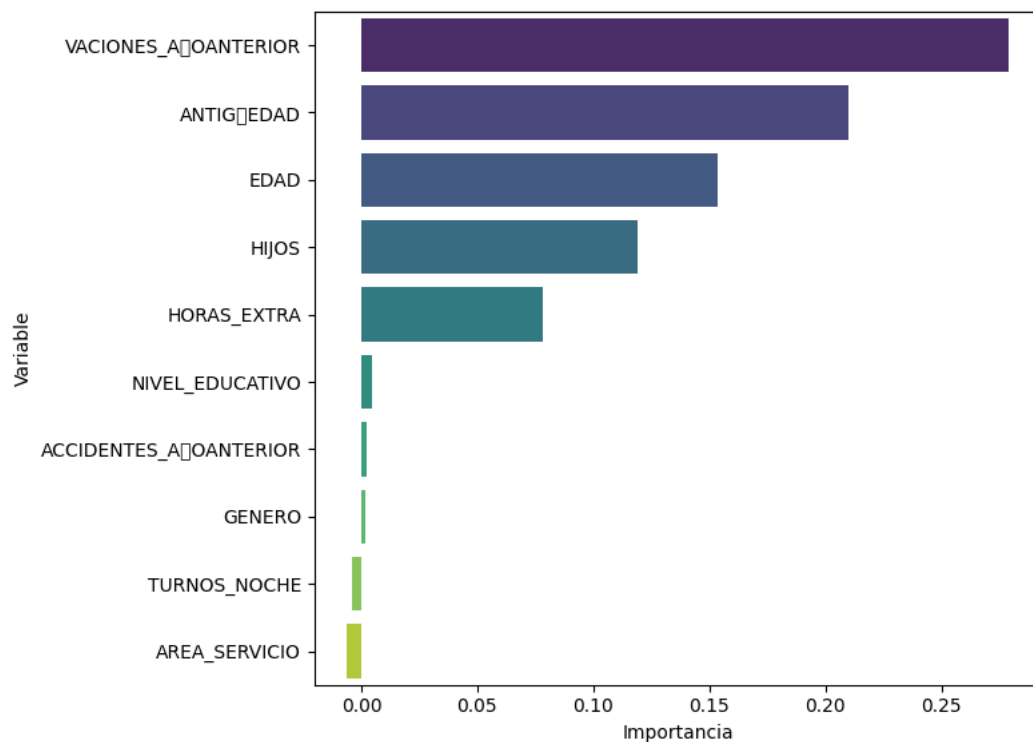
La Tabla 3 presenta la importancia de variables estimada mediante permutación para el mismo modelo.

Tabla 3. Importancia por permutación (Gradient Boosting base)

Variable	Importancia
Vacaciones del año anterior	0.279
Antigüedad	0.210
Edad	0.153
Hijos	0.119
Horas extra	0.078
Nivel educativo	0.005
Accidentes año anterior	0.002
Género	0.002
Turnos noche	-0.004
Área o servicio	-0.006

La Figura 2 presenta de forma gráfica los anteriores resultados.

Figura 2. Importancia por permutación (Gradient Boosting base)



Importancia de variables en el modelo Gradient Boosting (configuración optimizada)

Importancia interna de variables

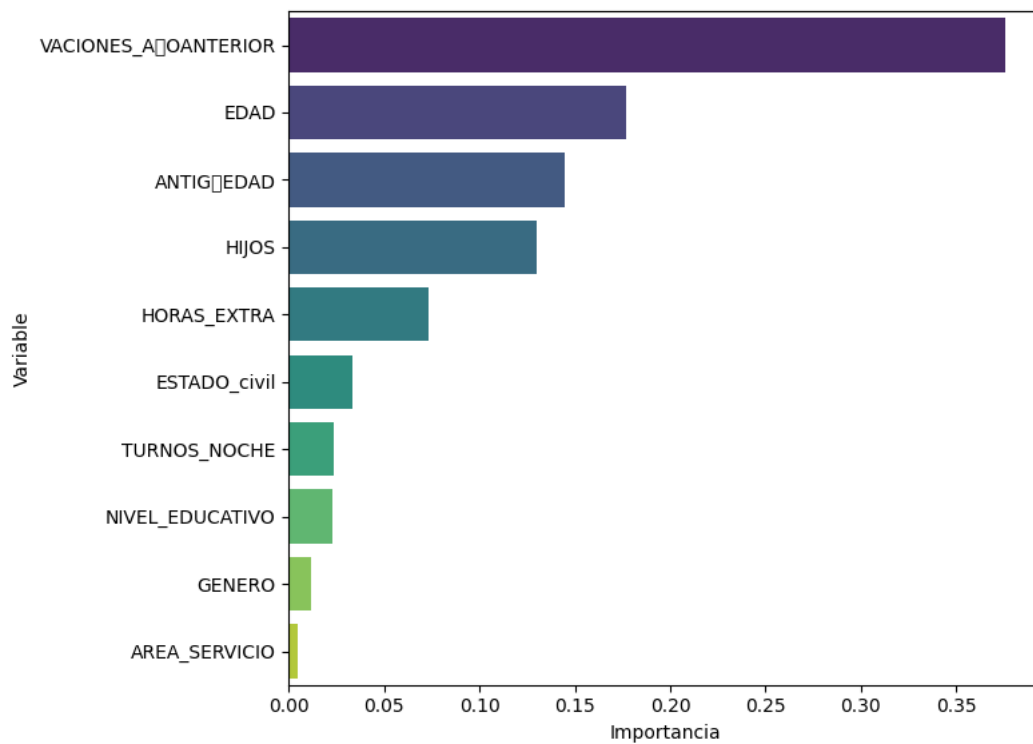
La Tabla 4 presenta las diez variables con mayor importancia interna estimada por el modelo Gradient Boosting optimizado.

Tabla 4. Importancia interna de variables (Gradient Boosting optimizado)

Variable	Importancia
Vacaciones del año anterior	0.376
Edad	0.177
Antigüedad	0.145
Hijos	0.130
Horas extra	0.073
Estado civil	0.034
Turnos noche	0.023
Nivel educativo	0.023
Género	0.012
Área o servicio	0.005

La Figura 3 presenta de forma gráfica los anteriores resultados.

Figura 3. Importancia interna (Gradient Boosting-optimizado)



Importancia por permutación

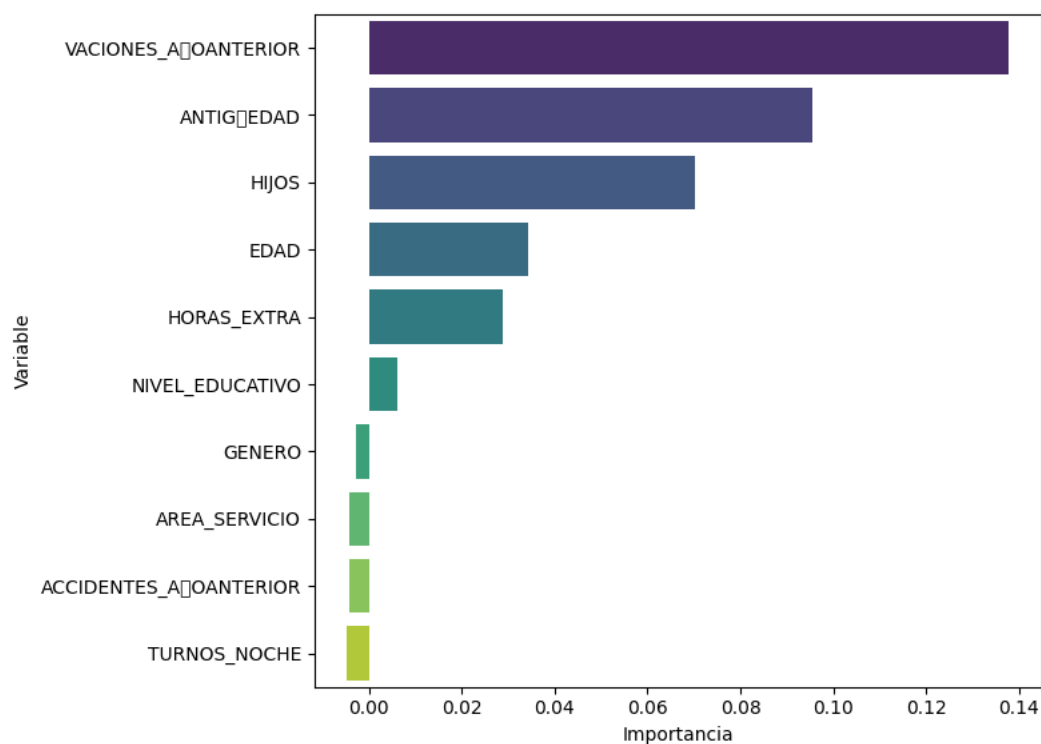
La Tabla 5 presenta la importancia por permutación de variables en el modelo optimizado.

Tabla 5. Importancia por permutación (Gradient Boosting optimizado)

Variable	Importancia
Vacaciones del año anterior	0.138
Antigüedad	0.096
Hijos	0.070
Edad	0.034
Horas extra	0.029
Nivel educativo	0.006
Género	-0.003
Área o servicio	-0.004
Accidentes año anterior	-0.004
Turnos noche	-0.005

La Figura 4 presenta de forma gráfica los anteriores resultados.

Figura 4. Importancia por permutación (Gradient Boosting-optimizado)



Importancia de variables en el modelo LightGBM (configuración base)

Importancia interna de variables

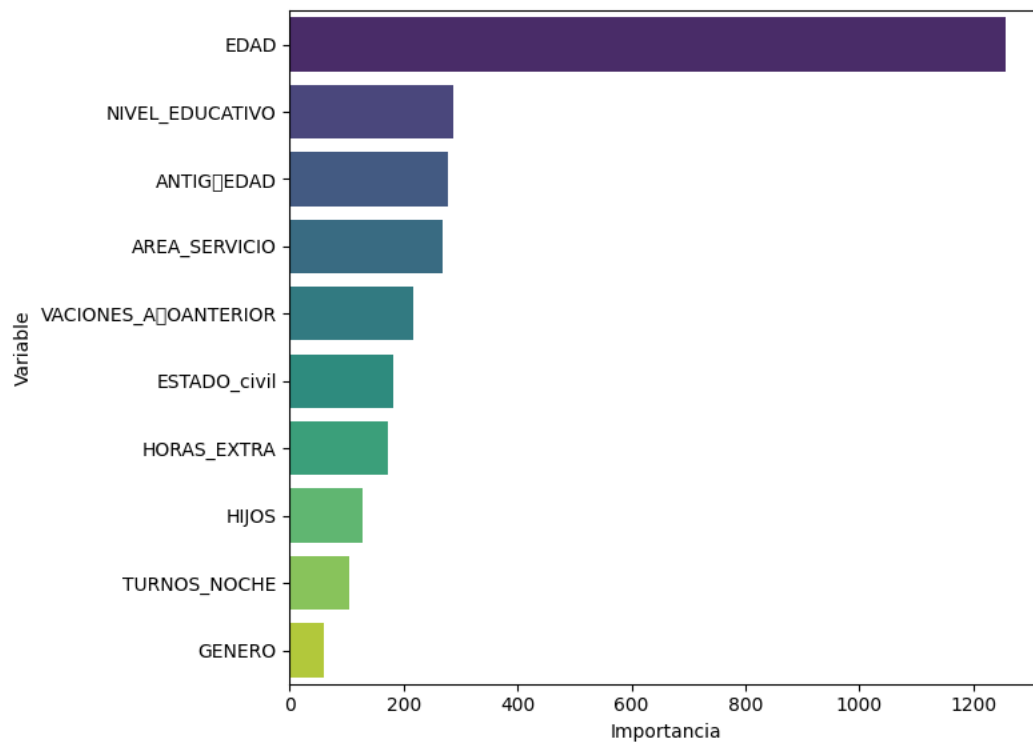
La Tabla 6 presenta las diez variables con mayor importancia interna estimada por el modelo LightGBM.

Tabla 6. Importancia interna de variables (LightGBM base)

Variable	Importancia
Edad	1257
Nivel educativo	286
Antigüedad	278
Área o servicio	268
Vacaciones del año anterior	217
Estado civil	182
Horas extra	172
Hijos	129
Turnos noche	105
Género	61

La Figura 5 presenta de forma gráfica los anteriores resultados.

Figura 5. Importancia interna (LightGBM)



Importancia por permutación

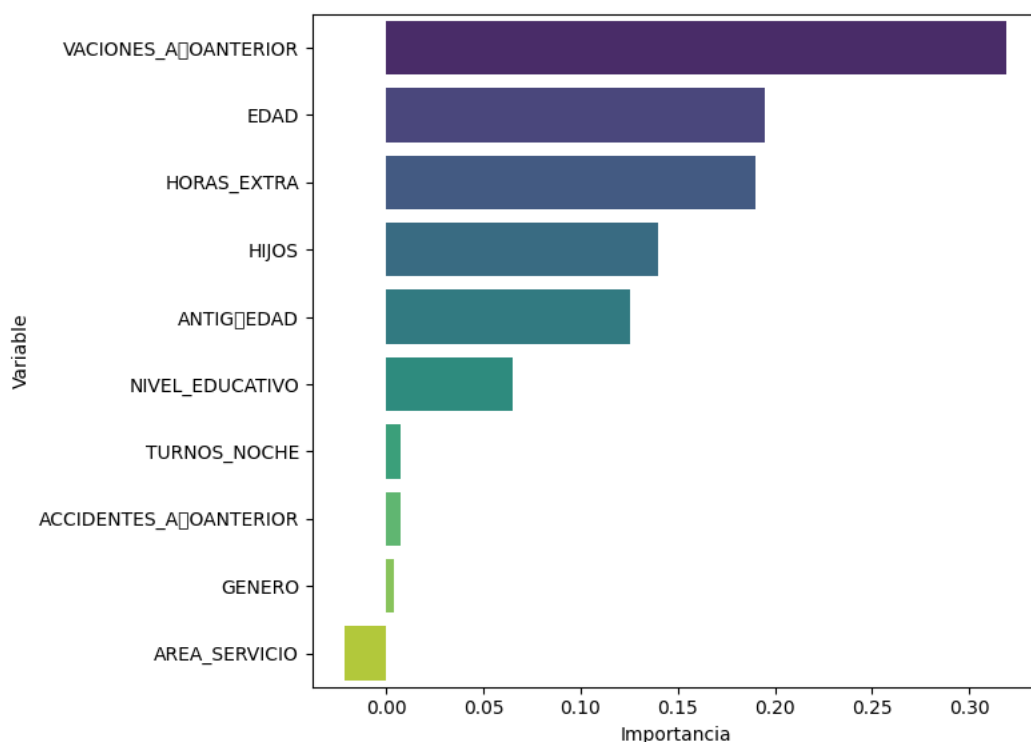
La Tabla 7 presenta la importancia por permutación para el modelo LightGBM.

Tabla 7. Importancia por permutación (LightGBM)

Variable	Importancia
Vacaciones del año anterior	0.319
Edad	0.195
Horas extra	0.190
Hijos	0.140
Antigüedad	0.125
Nivel educativo	0.065
Turnos noche	0.007
Accidentes año anterior	0.007
Género	0.004
Área o servicio	-0.021

La Figura 6 presenta de forma gráfica los anteriores resultados.

Figura 6. Importancia por permutación (LightGBM)



Comparación de variables relevantes entre algoritmos

Al contrastar los rankings de importancia interna y por permutación de los tres modelos, se identificó coincidencia recurrente en las siguientes variables dentro de los primeros lugares: vacaciones del año anterior, edad, antigüedad, hijos, horas extra. Estas variables aparecen de forma consistente tanto en Gradient Boosting (base y optimizado) como en LightGBM, y bajo ambos métodos de evaluación de importancia. Asimismo, variables como nivel educativo y estado civil presentan relevancia intermedia, mientras que género, área o servicio y turnos noche muestran relevancia baja o cercana a cero en los análisis por permutación.

Con el fin de facilitar la interpretación de los resultados para la gestión del talento humano, se presenta a continuación una síntesis aplicada del impacto de las principales variables identificadas:

- **Vacaciones del año anterior:** Su alta importancia sugiere que los procesos de descanso y recuperación influyen significativamente en la probabilidad de acumulación de días de incapacidad. Una gestión estratégica de la programación de vacaciones podría contribuir a la prevención del ausentismo.
- **Edad y antigüedad:** Reflejan posibles efectos de exposición acumulada a demandas físicas y psicosociales propias del entorno hospitalario, lo que orienta hacia intervenciones diferenciadas según trayectoria laboral.

- **Horas extra:** Indican que la carga laboral adicional puede estar asociada a mayor probabilidad de incapacidad, lo cual resalta la importancia de monitorear esquemas de extensión de jornada.
- **Hijos:** Sugiere que las responsabilidades familiares pueden incidir en la dinámica de ausentismo, destacando la relevancia de políticas de conciliación trabajo–familia.
- **Variables con baja importancia (género, área, turnos nocturnos):** en el conjunto de datos analizado no mostraron influencia significativa, lo que sugiere que el fenómeno no está concentrado en un grupo demográfico o servicio específico, sino que responde a dinámicas más estructurales.

Esta lectura permite trasladar los resultados técnicos a decisiones concretas de gestión preventiva dentro del área de talento humano.

Implicaciones de los hallazgos desde la gestión del talento humano

Desde una perspectiva de gestión del talento humano, los hallazgos permiten interpretar el ausentismo no únicamente como un evento médico individual, sino como un fenómeno influenciado por condiciones organizacionales y trayectorias laborales acumuladas. La relevancia consistente de variables como vacaciones del año anterior, antigüedad, edad, hijos y horas extra sugiere que los patrones de incapacidad no dependen exclusivamente de factores clínicos, sino también de dinámicas relacionadas con la carga laboral, los ciclos de recuperación, la conciliación trabajo–familia y la exposición prolongada a exigencias del entorno hospitalario.

En particular, la influencia de las vacaciones como predictor central puede interpretarse como un indicador de la importancia de los procesos de descanso y recuperación en la prevención del ausentismo acumulado. De igual manera, la antigüedad y la edad podrían reflejar efectos de desgaste progresivo o acumulación de exposición a demandas físicas y psicosociales. Por su parte, las horas extra y la presencia de hijos sugieren que la gestión del tiempo de trabajo y la conciliación laboral–familiar constituyen dimensiones relevantes para el diseño de estrategias preventivas.

En consecuencia, los resultados obtenidos orientan al área de talento humano hacia intervenciones focalizadas, tales como políticas de programación de vacaciones más estratégicas, monitoreo de cargas laborales en perfiles con alta exposición acumulada y diseño de acciones diferenciales para grupos con mayores demandas familiares.

DISCUSIÓN

La presente investigación modeló el ausentismo por incapacidad médica en una institución hospitalaria de alta complejidad, integrando analítica predictiva y teoría organizacional. Los hallazgos permiten discutir el fenómeno no únicamente como un problema operativo de reemplazos o costos, sino como una manifestación estratégica relacionada con la gestión de un recurso organizacional central: el capital humano.

Desde la perspectiva de la visión basada en recursos (RBV), las personas constituyen activos valiosos, difíciles de sustituir y determinantes para el desempeño organizacional cuando son gestionadas adecuadamente (Barney, 1991; Wright & McMahan, 2011). En consecuencia, la pérdida recurrente de capacidad laboral a través del ausentismo representa no solo una interrupción operativa, sino una ineficiencia en el aprovechamiento de un recurso estratégico. En este sentido, la incorporación de analítica predictiva en la gestión humana del hospital constituye un mecanismo para fortalecer la capacidad organizacional de anticipar riesgos, sostener la continuidad asistencial y preservar su ventaja competitiva en un entorno altamente exigente como el sector salud.

Con base en un enfoque metodológico, la aplicación del modelo CRISP-DM permitió estructurar de manera sistemática el proceso analítico, asegurando coherencia entre la comprensión del problema organizacional, el entendimiento y la preparación de los datos provenientes del ERP institucional, el modelamiento predictivo, la evaluación de desempeño y la interpretación de resultados. Esta integración entre metodología de minería de datos y necesidades estratégicas de gestión humana responde al llamado de la literatura de Human Resource Analytics y Workforce Analytics, que enfatiza que el valor de la analítica no reside únicamente en la técnica estadística, sino en su capacidad para traducir datos organizacionales en conocimiento accionable para la toma de decisiones (Huselid, 2018; Minbaeva, 2018; Madhani, 2023). Así, el presente estudio contribuye empíricamente a reducir la brecha señalada por Qamar y Samad (2022), quienes destacan la necesidad de aplicaciones prácticas de analítica en problemas específicos de gestión humana, particularmente en contextos complejos como el sector salud.

En términos empíricos, los modelos predictivos desarrollados exhibieron valores de R^2 entre 0.134 y 0.263, rangos consistentes con investigaciones en ciencias sociales y organizacionales donde los fenómenos laborales presentan alta variabilidad individual y contextual. Tal como lo plantea Cohen (1988), coeficientes de determinación alrededor de 0.13 corresponden a efectos moderados y valores cercanos a 0.26 representan efectos sustanciales

en ciencias del comportamiento. Esto indica que, aun sin pretender determinismo individual, los modelos construidos poseen capacidad predictiva suficiente para orientar procesos de gestión preventiva en workforce analytics. Desde la lógica de la RBV, esta capacidad de anticipación constituye en sí misma un recurso organizacional intangible, pues permite transformar datos operativos en inteligencia estratégica difícilmente replicable por otras instituciones (Minbaeva, 2018; Wright & McMahan, 2011).

Un hallazgo central del estudio fue la convergencia de los algoritmos de boosting en un núcleo común de variables relevantes: vacaciones en el último año, edad, antigüedad, hijos y horas extra. Esta convergencia se observó tanto en los modelos Gradient Boosting como en LightGBM, y bajo estrategias de importancia interna y por permutación, lo que sugiere estabilidad del patrón de aprendizaje pese al uso de arquitecturas algorítmicas distintas. La inclusión del modelo LightGBM permitió confirmar que los resultados no son dependientes de un único algoritmo, sino que reflejan relaciones estructurales persistentes en los datos organizacionales. Esta estabilidad empírica encuentra respaldo en la literatura sobre ausentismo, que ha demostrado consistentemente el carácter multifactorial del fenómeno, en el que confluyen características individuales, trayectoria organizacional y organización del tiempo de trabajo (Allebeck & Mastekaasa, 2004; Merkus et al., 2012).

La relevancia de la edad y la antigüedad se alinea con estudios que muestran asociaciones entre exposición acumulada a riesgos laborales, desgaste profesional y mayor probabilidad de ausencias prolongadas en trabajadores del sector salud. Desde la RBV, estos hallazgos resaltan la necesidad de estrategias diferenciadas de gestión del ciclo vital laboral, en tanto la experiencia y la permanencia representan simultáneamente un activo organizacional y una fuente potencial de vulnerabilidad si no se acompaña de políticas de cuidado, rediseño de tareas y adaptación progresiva de cargas laborales.

La aparición consistente de las vacaciones como predictor relevante resulta particularmente significativa. La literatura sobre recuperación indica que los periodos de descanso facilitan la restauración del bienestar y la salud, aunque sus efectos pueden ser contingentes al contexto organizacional y a la carga laboral posterior (de Bloom et al., 2013; Hofmarcher, 2021). En este sentido, la gestión de vacaciones no solo constituye un proceso administrativo, sino un mecanismo estratégico para preservar la capacidad instalada del recurso humano. Desde workforce analytics, este hallazgo ofrece evidencia concreta para que la institución fortalezca sistemas de planeación anticipada de vacaciones, evite acumulaciones excesivas de días pendientes y diseñe esquemas de rotación que aseguren descansos efectivos sin comprometer la continuidad asistencial.

De manera complementaria, las horas extra emergieron como uno de los predictores con mayor recurrencia en los modelos. Esto es coherente con investigaciones que vinculan la intensificación laboral y las jornadas extendidas con mayor fatiga y riesgo de ausencias por motivos de salud, especialmente mental, en el personal sanitario (Bernstrøm & Houkes, 2018; Dall’Ora et al., 2018; Inoue et al., 2022). Desde la perspectiva de la RBV, la sobreutilización sostenida del recurso humano puede erosionar su valor estratégico en el mediano plazo. En términos prácticos, esto significa que equipos que hoy sostienen la operación pueden convertirse mañana en focos de ausentismo si no se gestionan adecuadamente las cargas laborales generando ciclos de desgaste, disminución de desempeño y pérdida de disponibilidad laboral; además, más allá de la discusión teórica, en el contexto de SES Hospital Universitario de Caldas, la pérdida recurrente de días laborales no solo representa un costo, sino una tensión real en la operación asistencial diaria. La analítica predictiva, en este contexto, se convierte en un recurso organizacional que permite monitorear patrones de sobrecarga, anticipar unidades con riesgo elevado de fatiga colectiva y activar decisiones de redistribución de turnos, contratación temporal o redimensionamiento de dotación.

La variable hijos, utilizada como proxy de demandas familiares, también mostró relevancia consistente en los modelos. Esto coincide con revisiones que señalan efectos heterogéneos de las responsabilidades familiares sobre el ausentismo, especialmente en contextos donde la rigidez de los turnos y la carga asistencial dificultan la conciliación trabajo–familia (Allebeck & Mastekaasa, 2004; Laaksonen et al., 2010). Desde la gestión humana estratégica, estos hallazgos refuerzan la importancia de implementar políticas de flexibilidad en la programación de turnos, opciones de intercambio planificado de jornadas y programas de acompañamiento psicosocial que faciliten la estabilidad del recurso humano sin afectar la operación clínica.

Por otra parte, algunas variables teóricamente relevantes, como turno nocturno y área de trabajo, presentaron menor estabilidad en los análisis de permutación. Este resultado no contradice la literatura, que ha demostrado asociaciones sólidas entre turnos nocturnos, alteraciones del sueño y ausentismo en el sector salud (Merkus et al., 2012; Larsen et al., 2023), sino que probablemente refleja limitaciones de la operacionalización binaria utilizada en este estudio, donde no fue posible capturar intensidad ni secuencias de turnos. Este aspecto ilustra uno de los desafíos señalados en Human Resource Analytics: la calidad y granularidad de los datos disponibles condiciona la capacidad analítica y la potencia explicativa de los modelos. En consecuencia, la ampliación futura de los registros de turnicidad permitirá fortalecer la precisión predictiva y explicativa de los modelos desarrollados.

En cuanto al comportamiento del modelo Gradient Boosting optimizado, la disminución del desempeño tras la búsqueda de hiperparámetros sugiere un fenómeno de sobreajuste típico en contextos de small data organizacional. Tal como advierte la literatura metodológica, en conjuntos de datos limitados la optimización puede favorecer configuraciones que se ajustan mejor a particiones internas de entrenamiento sin mejorar la generalización en datos nuevos (Hawkins et al., 2003; Riley et al., 2019). Desde la perspectiva de CRISP-DM, este hallazgo confirma la relevancia de la fase de evaluación del modelo como instancia crítica para evitar la adopción de soluciones algorítmicas más complejas sin ganancia real en valor organizacional.

Implicaciones prácticas para la gestión humana y el liderazgo institucional

Desde el punto de vista práctico, los hallazgos permiten delinear acciones concretas de gestión humana y liderazgo administrativo orientadas a la reducción preventiva del ausentismo. En primer lugar, la identificación de vacaciones como predictor relevante sugiere la necesidad de fortalecer la planeación estratégica del descanso. Esto implica diseñar calendarios anuales de vacaciones basados en analítica, evitando acumulación de periodos sin descanso, garantizando pausas regulares y priorizando unidades con mayores cargas asistenciales. Asimismo, la institución puede incorporar indicadores de riesgo de ausencia futura en sus tableros de gestión, permitiendo a los líderes anticipar picos de incapacidad y ajustar la programación de personal.

En segundo lugar, la relevancia de las horas extra evidencia la necesidad de monitorear sistemáticamente la sobrecarga laboral. Desde gestión humana, esto se traduce en implementar sistemas de alerta temprana que identifiquen colaboradores o equipos con exceso recurrente de horas adicionales, activar revisiones de dotación, redistribuir cargas de trabajo y fortalecer procesos de contratación temporal o bancos de suplencias. Adicionalmente, los resultados respaldan la pertinencia de programas de prevención de fatiga, pausas activas y evaluación periódica de riesgos psicosociales en áreas críticas.

En tercer lugar, la importancia de edad y antigüedad sugiere la conveniencia de desarrollar estrategias de gestión diferenciada del ciclo laboral. Esto puede incluir rediseño progresivo de tareas para personal de mayor antigüedad, esquemas de mentoría que aprovechen su experiencia sin sobrecargar físicamente, y planes de movilidad interna hacia roles menos demandantes en fases avanzadas de la carrera. Tales acciones permiten preservar el valor estratégico del recurso humano experimentado, coherente con la lógica de la RBV.

En cuarto lugar, la relevancia de demandas familiares invita a consolidar políticas de conciliación trabajo-familia, tales como flexibilidad controlada en intercambios de turno,

permisos planificados, acompañamiento psicológico y servicios de apoyo social institucional. Estas medidas contribuyen a fortalecer el compromiso organizacional y reducir ausencias asociadas a tensiones externas al trabajo.

Finalmente, la convergencia de resultados entre Gradient Boosting y LightGBM demuestra que la institución cuenta con evidencia suficientemente robusta para avanzar hacia la fase de despliegue de CRISP-DM, integrando los modelos predictivos en sistemas de apoyo a la decisión. Esto implica crear tableros dinámicos de riesgo de ausentismo por unidad, perfil sociodemográfico y carga laboral, facilitando que los líderes administrativos y clínicos tomen decisiones preventivas basadas en datos y no únicamente en reacción operativa.

Limitaciones e investigaciones futuras

Este estudio presenta limitaciones que deben reconocerse. En primer lugar, el análisis se realizó en un solo caso organizacional, lo que restringe la generalización directa de los resultados a otras instituciones. En segundo lugar, el uso exclusivo de datos de ERP, aunque fortalece la validez ecológica, limita la inclusión de variables psicosociales, clínicas y contextuales que la literatura reconoce como influyentes en el ausentismo, como estrés, burnout, clima organizacional o carga asistencial real. En tercer lugar, la estructura empleado-año impidió modelar trayectorias individuales y recurrencia temporal del ausentismo, aspecto relevante señalado en estudios longitudinales. Finalmente, los modelos desarrollados fueron predictivos y descriptivos en términos de importancia de variables, pero no permiten establecer relaciones causales directas.

Estas limitaciones abren oportunidades de investigación futura en analítica de datos desde la gestión humana. Entre ellas se destacan el enriquecimiento de la base institucional con indicadores de carga asistencial y riesgos psicosociales, el modelamiento longitudinal de episodios de incapacidad, la validación externa en otras instituciones de salud y la incorporación de técnicas avanzadas de explicabilidad como SHAP para fortalecer la transparencia y legitimidad de la analítica en contextos organizacionales. De igual forma, avanzar hacia sistemas de apoyo a la decisión integrados en los tableros de gestión permitirá consolidar la analítica como recurso estratégico sostenible en la institución.

CONCLUSIÓN

La presente investigación tuvo como propósito identificar los factores asociados al ausentismo por incapacidad médica en una institución hospitalaria de alta complejidad, mediante la aplicación de modelos de analítica predictiva. Los resultados evidencian que, aun en contextos organizacionales caracterizados por alta complejidad operativa y múltiples

variables no observadas, es posible anticipar patrones relevantes de ausentismo utilizando datos administrativos institucionales.

En términos metodológicos, el estudio demuestra la viabilidad de integrar el modelo *CRISP-DM*, la *Human Resource Analytics* y la Visión basada en Recursos como marco articulador para el análisis de fenómenos organizacionales complejos. La aplicación de algoritmos de *Gradient Boosting* y *LightGBM* permitió identificar estructuras consistentes de variables asociadas al ausentismo, validando la pertinencia del enfoque predictivo en el ámbito hospitalario. Desde la perspectiva de la RBV, la capacidad analítica desarrollada puede comprenderse como una competencia organizacional que fortalece la toma de decisiones basada en evidencia.

No obstante, el aporte central de esta investigación trasciende la validación técnica de los modelos. La identificación consistente de variables como vacaciones del año anterior, edad, antigüedad, hijos y horas extra sugiere que el ausentismo acumulado no responde exclusivamente a eventos médicos individuales, sino a dinámicas organizacionales relacionadas con la trayectoria laboral, la organización del tiempo de trabajo, la conciliación familiar y los procesos de recuperación del personal.

En particular, la relevancia de las vacaciones como predictor destacado resalta la importancia de los mecanismos de descanso como estrategia preventiva. De igual manera, la influencia de la antigüedad y la edad puede interpretarse como efecto de exposición acumulada a exigencias físicas y psicosociales propias del entorno hospitalario. Las horas extra y la presencia de hijos, por su parte, evidencian la necesidad de una gestión más estratégica de la carga laboral y de las políticas de conciliación trabajo–familia.

Estos hallazgos permiten reconfigurar la comprensión del ausentismo dentro de la organización, desplazándolo de una lógica reactiva centrada en la sustitución operativa hacia una lógica preventiva basada en la anticipación de riesgos. En este sentido, la analítica predictiva no constituye un fin en sí mismo, sino un medio para fortalecer la gestión estratégica del talento humano, facilitando intervenciones focalizadas, decisiones más informadas y una administración más sostenible del capital humano.

Finalmente, si bien los modelos presentan niveles de explicación moderados propios de fenómenos organizacionales multifactoriales, su valor radica en la generación de información accionable para la gestión preventiva. Futuras investigaciones podrían incorporar variables psicosociales, clínicas o longitudinales que amplíen la comprensión integral del fenómeno. No obstante, el presente estudio demuestra que incluso a partir de datos administrativos

institucionales es posible avanzar hacia una gestión del talento humano más proactiva, basada en evidencia y orientada a la sostenibilidad organizacional en el sector salud.

Referencias

- Aggarwal, C. C. (2017). *Outlier Analysis* (2nd ed.). Springer. <https://doi.org/10.1007/978-3-319-47578-3>
- Aloğlu, N., & Güllü, A. (2022). Investigation of unplanned absenteeism of nurses working in a university hospital: cross-sectional study. *Journal of Radiology Nursing*, 41(4), 326-330. <https://doi.org/10.1016/j.jradnu.2022.06.001>
- Allebeck, P., & Mastekaasa, A. (2004). Risk factors for sick leave: General studies. *Scandinavian Journal of Public Health*, 32(63), 49–108. <https://doi.org/10.1080/14034950410021853>
- Ardila Díaz, M. (2023). *Impacto de la salud mental en el ausentismo laboral en las organizaciones*. Fundación Universitaria del Área Andina.
- Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4, 40–79. <https://doi.org/10.1214/09-SS054>
- Bahuguna, P., Sivarajah, U., & Irani, Z. (2024). Human resource analytics and organizational performance: A systematic literature review. *Information Systems Frontiers*, 26(1), 125–145.
- Barney, J. (1991). Firm resources and sustained competitive advantage. *Journal of Management*, 17(1), 99–120. <https://doi.org/10.1177/014920639101700108>
- Bernstrøm, V. H., & Houkes, I. (2018). Long working hours and sickness absence—A fixed effects design. *BMC Public Health*, 18, 578. <https://doi.org/10.1186/s12889-018-5485-2>
- Blasco, M. (2021). CRISP-DM: Towards a standard process model for data mining projects. *Data Science and Engineering*, 6(2), 132–141.

- Blomgren, J., & Jäppinen, S. (2021). Incidence and length of sickness absence among hierarchical occupational classes and non-wage-earners: A register study of 1.6 million Finns. *International Journal of Environmental Research and Public Health*, 18(2), 501. <https://doi.org/10.3390/ijerph18020501>
- Borda, M. R. (2017). *Ausentismo laboral: impacto en la productividad y estrategias de control desde los programas de salud empresarial*. Universidad del Rosario.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5–32. <https://doi.org/10.1023/A:1010933404324>
- Dall’Ora, C., Ball, J., Recio-Saucedo, A., & Griffiths, P. (2018). Characteristics of shift work and their impact on employee performance and wellbeing: A literature review. *International Journal of Nursing Studies*, 57, 12–27. <https://doi.org/10.1016/j.ijnurstu.2016.01.007>
- Deloitte. (2023). *Global Human Capital Trends 2023: New fundamentals for a boundaryless world*. Deloitte Insights.
- de Bloom, J., Geurts, S. A. E., & Kompier, M. A. J. (2013). Vacation (after-)effects on employee health and well-being, and the role of vacation activities, experiences and sleep. *Journal of Happiness Studies*, 14(2), 613–633. <https://doi.org/10.1007/s10902-012-9345-3>
- Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Lawrence Erlbaum.
- Ellis, P. D. (2010). *The Essential Guide to Effect Sizes*. Cambridge University Press.
- Espegren, Y., & Hugosson, M. (2025). *HR analytics-as-practice: a systematic literature review*. *Journal of Organizational Effectiveness: People and Performance*, 12(5), 83–111. <https://doi.org/10.1108/JOEPP-11-2022-0345>
- Farrants, K., & Alexanderson, K. (2024). Sickness absence after work accidents and post-traumatic stress among white-collar workers in the retail and wholesale industry: A

- longitudinal Swedish cohort study. *BMC Public Health*, 24, 2389.
<https://doi.org/10.1186/s12889-024-19865-0>
- Ferguson, C. J. (2009). An effect size primer: A guide for clinicians and researchers. *Professional Psychology: Research and Practice*, 40(5), 532–538.
<https://doi.org/10.1037/a0015808>
- Fisher, A., Rudin, C., & Dominici, F. (2019). All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, 20(177), 1–81.
- Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5), 1189–1232. <https://doi.org/10.1214/aos/1013203451>
- Guzmán, L. D., & Ríos, J. A. (2021). Impacto del ausentismo laboral en organizaciones del sector salud de la ciudad de Medellín. *CIES Revista de Investigación Económica, Social y Administrativa*, 12(2), 115–132.
- Hastie, T., Tibshirani, R., & Friedman, J. (2017). *The Elements of Statistical Learning* (2nd ed.). Springer. <https://doi.org/10.1007/978-0-387-84858-7>
- Hawkins, D. M., Basak, S. C., & Mills, D. (2003). Assessing model fit by cross-validation. *Journal of Chemical Information and Computer Sciences*, 43(2), 579–586.
<https://doi.org/10.1021/ci025626i>
- Hofmarcher, T. (2021). The effect of paid vacation on health: Evidence from Sweden. *Journal of Population Economics*, 34, 941–977. <https://doi.org/10.1007/s00148-020-00811-1>
- HumanSmart. (2023). *Workforce analytics and employee wellbeing report 2023*. HumanSmart Consulting.
- Huselid, M. A. (2018). The science and practice of workforce analytics: Introduction to the HRM special issue. *Human Resource Management*, 57(3), 679–684.
<https://doi.org/10.1002/hrm.21916>

- Inoue, Y., Yamamoto, S., Stickley, A., Kuwahara, K., Miyamoto, T., Nakagawa, T., Honda, T., Imai, T., Nishihara, A., Kabe, I., Mizoue, T., & Dohi, S. (2022). Overtime work and the incidence of long-term sickness absence due to mental disorders: A prospective cohort study. *Journal of Epidemiology*, 32(6), 283–289. <https://doi.org/10.2188/jea.JE20200382>
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning* (2nd ed.). Springer. <https://doi.org/10.1007/978-1-0716-1418-1>
- Johansson, M. K., Leijon, O., Falkstedt, D., & Alexanderson, K. (2021). Work and sick leave patterns following a work injury among young adults: A mixed methods study. *BMJ Open*, 11, e045143. <https://doi.org/10.1136/bmjopen-2020-045143>
- Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T. Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, 30, 3146–3154.
- Koopmans, P. C., Bültmann, U., Roelen, C. A. M., Hoedeman, R., van der Klink, J. J. L., & Groothoff, J. W. (2010). Recurrence of sickness absence due to common mental disorders. *International Archives of Occupational and Environmental Health*, 83(5), 457–467. <https://pmc.ncbi.nlm.nih.gov/articles/PMC3020308/>
- Kuhn, M., & Johnson, K. (2013). *Applied Predictive Modeling*. Springer.
- Laaksonen, M. (2013). The durations of past sickness absences predict future absence episodes. *Scandinavian Journal of Work, Environment & Health*, 39(5), 455–464. <https://doi.org/10.5271/sjweh.3369>
- Laaksonen, M., Pitkaniemi, J., Rahkonen, O., & Lahelma, E. (2010). Work arrangements, physical working conditions, and psychosocial working conditions as risk factors for sickness absence: Bayesian analysis of prospective data. *Annals of Epidemiology*, 20(5), 332–338. <https://doi.org/10.1016/j.annepidem.2010.02.004>

- Larsen, A. D., Hannerz, H., & Rugulies, R. (2023). Acute effects of night and evening shifts on risk of calling in sick. *Scandinavian Journal of Work, Environment & Health*, 49(2), 117–125. <https://doi.org/10.5271/sjweh.3972>
- Little, R. J. A., & Rubin, D. B. (2019). *Statistical analysis with missing data* (3rd ed.). Wiley.
- Llamas-Blázquez, P., Gómez-Martínez, R., & Ortega-Martín, J. (2025). Predicting workplace absenteeism using machine learning: A pilot study in occupational health. *Journal of Occupational Medicine and Toxicology*, 20(1), 38. <https://doi.org/10.1186/s12995-025-00482-5>
- Madhani, P. M. (2023). Resource-based view (RBV): Concepts and practices. *IUP Journal of Business Strategy*, 20(1), 7–32.
- Madhuri, A., & Kumar, B. R. (2025). HR analytics and decision-making: A data-driven approach to employee performance management. *Journal of Neonatal Surgery*, 14(7S). https://www.researchgate.net/publication/390477819_HR_Analytics_and_Decision-Making_A_Data-Driven_Approach_to_Employee_Performance_Management
- Margherita, A. (2021). Human resources analytics: A systematization of research topics and directions for future research. *Human Resource Management Review*, 31(2), 100795. <https://doi.org/10.1016/j.hrmr.2020.100795>
- Menon, M. S. R., & Sethi, D. (2025). The role of HR analytics in workforce planning. *Asian Journal of Research in Business and Management*, 28(5). <https://ajprui.com/index.php/ajpr/article/view/395>
- Merkus, S. L., van Drongelen, A., Holte, K. A., Lund, T., & van Mechelen, W. (2012). The association between shift work and sick leave: A systematic review. *Occupational and Environmental Medicine*, 69(10), 701–712. <https://doi.org/10.1136/oemed-2011-100488>

- Minbaeva, D. (2018). Building credible human capital analytics for organizational competitive advantage. *Human Resource Management*, 57(3), 701–713.
<https://doi.org/10.1002/hrm.21848>
- Moine, J., Haedo, A., Godoy, D., & Villagra, A. (2013). Data mining with CRISP-DM methodology: A case study. *Journal of Computer Science & Technology*, 13(2), 64–70.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.
- Qamar, A., & Samad, S. (2022). Human resource analytics and employee outcomes: A systematic review. *Personnel Review*, 51(5), 1384–1405.
- Rahm, E., & Do, H. H. (2000). Data cleaning: Problems and current approaches. *IEEE Data Engineering Bulletin*, 23(4), 3–13.
- Rasmussen, T., & Ulrich, D. (2023). Learning from practice: How HR analytics avoids being a management fad. *Organizational Dynamics*, 52(1), 100932.
<https://doi.org/10.1016/j.orgdyn.2022.100932>
- Rao, P. B., Raghavendra, R., Pawar, C., Patel, P., Sharma, P. R., & Harshitha, R. (2025). The impact of HR analytics and big data on decision making in healthcare organizations. In P. K. Pattnaik & R. Kumar (Eds.), *Cutting-edge technologies for business sectors* (pp. 335–366). IGI Global. <https://www.igi-global.com/chapter/the-impact-of-hr-analytics-and-big-data-on-decision-making-in-healthcare-organizations/359675>
- Redman, T. C. (2013). *Data driven: Profiting from your most important business asset*. Harvard Business Review Press.
- Riley, R. D., Ensor, J., Snell, K. I. E., Harrell, F. E. Jr., Martin, G. P., Reitsma, J. B., Moons, K. G. M., Collins, G. S., & Debray, T. P. A. (2019). Overfitting in prediction models —

problems, consequences and solutions. *BMJ*, 365, 1131.

<https://doi.org/10.1136/bmj.1131>

Wright, P. M., & McMahan, G. C. (2011). Exploring human capital: Putting “human” back into strategic human resource management. *Human Resource Management Journal*, 21(2), 93–104. <https://doi.org/10.1111/j.1748-8583.2010.00165.x>