Optimización en el proceso de consolidación y reporte de indicadores institucionales mediante análisis de datos y Machine Learning

Sandra Milena Bonza Sanchez

Asesor

Julio Eduardo Mejia Manzano

Universidad Nacional Abierta y a Distancia UNAD

Escuela de Ciencias Básicas, Tecnología e Ingeniería ECBTI

Especialización en Ciencia de Datos y Analitica

2026

Resumen

La consolidación manual de indicadores hospitalarios genera reprocesos, inconsistencias y limitaciones para el análisis oportuno de la información institucional. Este trabajo presenta el diseño e implementación de un sistema automatizado orientado a optimizar el procesamiento y reporte de indicadores de hospitalización en la Clínica de Marly Jorge Cavelier Gaviria, integrando técnicas de analítica de datos y aprendizaje automático bajo la metodología CRISP-DM.

El sistema fue desarrollado sobre la información correspondiente al período 2021–2024, procesando 22.723 registros de ingresos hospitalarios y una base de indicadores institucionales consolidada que comprende 48 períodos mensuales. Como fuente principal para el modelado se utilizó la base oficial de indicadores de hospitalización, que incluye variables de ocupación, flujo de pacientes y días de estancia, consolidada por el área de estadística de la institución a partir del sistema SERVINTE. Se implementó un pipeline en Python que automatiza la extracción, limpieza, estandarización y consolidación de la información, generando archivos estructurados compatibles con herramientas de Business Intelligence y eliminando la dependencia de procedimientos manuales propensos a error.

En la fase de modelado se aplicaron tres enfoques complementarios sobre la tabla mensual consolidada, que registra un promedio de 417 ingresos y 421 egresos mensuales, una estancia promedio de 4,15 días y una ocupación promedio del 82,1% para el período analizado. El algoritmo K-Means segmentó los períodos mensuales en tres perfiles operativos diferenciados —alto volumen, volumen medio y bajo volumen con alta complejidad—, con un índice Silhouette de 0,4804, identificando como hallazgo relevante que los primeros ocho meses de 2021 presentaron la estancia promedio más alta del período (5,68 días) con el menor volumen de

pacientes, configurando un perfil cualitativamente distinto al de los períodos posteriores. El modelo de regresión lineal para la predicción de egresos mensuales demostró una relación casi lineal con los ingresos, explicando el 91,1% de la variabilidad del indicador con un error relativo del 4,3%, lo que lo posiciona como una herramienta de planificación operativa de alta utilidad práctica. En contraste, el modelo de regresión para días de estancia totales presentó un R^2=0,143, evidenciando que este indicador depende principalmente de la complejidad clínica individual de cada hospitalización y no del volumen de pacientes, resultado que delimita con precisión qué información adicional se requiere para modelarlo de forma confiable. El árbol de decisión complementó el análisis alcanzando una exactitud del 93,3% y un F1-Score ponderado de 0,9325 en la clasificación de los tres perfiles operativos, generando reglas explícitas basadas en umbrales de egresos directamente interpretables por los equipos de gestión sin mediación técnica.

Los resultados demuestran que la automatización del procesamiento es condición necesaria para garantizar la confiabilidad del análisis, y que la combinación de técnicas de segmentación, predicción y clasificación permite comprender el comportamiento hospitalario desde perspectivas complementarias. El sistema desarrollado es escalable, documentado y replicable, y constituye una base concreta para avanzar hacia una gestión hospitalaria sostenida en evidencia cuantitativa.

Palabras clave: automatización de datos, indicadores hospitalarios, machine learning, CRISP-DM, gestión hospitalaria, analítica de datos en salud.

Abstract

The manual consolidation of hospital indicators generates rework, inconsistencies, and limitations for the timely analysis of institutional information. This paper presents the design and implementation of an automated system aimed at optimizing the processing and reporting of hospitalization indicators at Clínica de Marly Jorge Cavelier Gaviria, integrating data analytics techniques and machine learning under the CRISP-DM methodology.

The system was developed using data from the 2021–2024 period, processing 22,723 hospital admission records and an official institutional indicators dataset comprising 48 monthly periods. The primary source for modeling was the official hospitalization indicators base, which includes occupancy, patient flow, and length-of-stay variables, consolidated by the institution's statistics department from the SERVINTE system. A Python pipeline was implemented to automate the extraction, cleaning, standardization, and consolidation of information, generating structured files compatible with Business Intelligence tools and eliminating dependence on error-prone manual procedures.

In the modeling phase, three complementary approaches were applied to the consolidated monthly table, which records an average of 417 admissions and 421 discharges per month, an average length of stay of 4.15 days, and an average occupancy rate of 82.1% for the analyzed period. The K-Means algorithm segmented monthly periods into three distinct operational profiles —high volume, medium volume, and low volume with high complexity—, with a Silhouette index of 0.4804, identifying as a key finding that the first eight months of 2021 presented the highest average length of stay (5.68 days) with the lowest patient volume, constituting a qualitatively distinct profile from subsequent periods. The linear regression model for predicting monthly discharges demonstrated a near-linear relationship with admissions,

explaining 91.1% of the indicator's variability with a relative error of 4.3%, positioning it as a highly practical operational planning tool. In contrast, the regression model for total length of stay yielded R^2=0.143, evidencing that this indicator depends primarily on the individual clinical complexity of each hospitalization rather than patient volume, a result that precisely identifies what additional information is needed to model it reliably. The decision tree complemented the analysis by achieving 93.3% accuracy and a weighted F1-Score of 0.9325 in classifying the three operational profiles, generating explicit rules based on discharge thresholds directly interpretable by management teams without technical mediation.

The results demonstrate that processing automation is a necessary condition for ensuring analytical reliability, and that combining segmentation, prediction, and classification techniques allows hospital behavior to be understood from complementary perspectives. The system developed is scalable, documented, and replicable, and provides a concrete foundation for advancing toward hospital management sustained by quantitative evidence.

Keywords: data automation, hospital indicators, machine learning, CRISP-DM, hospital management, health data analytics.

Tabla de Contenido

Lista de Tablas

Lista de Figuras

Introducción

La gestión de información en el sector salud ha adquirido una importancia estratégica creciente, dado su impacto directo sobre la calidad asistencial, la eficiencia operativa y la sostenibilidad institucional. En entornos hospitalarios, los sistemas de información generan continuamente grandes volúmenes de datos relacionados con el flujo de pacientes, la utilización de recursos y el desempeño de los servicios. Sin embargo, la disponibilidad de estos datos no garantiza por sí sola su aprovechamiento efectivo: cuando su consolidación depende de procesos manuales, se incrementan los riesgos de error, reproceso e inconsistencia, limitando la capacidad de las instituciones para analizar su desempeño de forma oportuna y confiable.

Este es el contexto que motivó el presente estudio. La Clínica de Marly Jorge Cavelier Gaviria, institución de mediana complejidad ubicada en Bogotá, realizaba la consolidación mensual de sus indicadores de hospitalización mediante procedimientos manuales apoyados en hojas de cálculo, lo que generaba reprocesos recurrentes, inconsistencias entre fuentes y limitaciones para el análisis sistemático de indicadores como ingresos, egresos, días de estancia y porcentaje de ocupación. Esta situación representaba una barrera concreta para la toma de decisiones basada en evidencia y para la planificación anticipada de recursos.

Frente a este problema, el presente trabajo tuvo como objetivo diseñar e implementar un sistema automatizado para el procesamiento y análisis mensual de indicadores de hospitalización, integrando técnicas de analítica de datos y aprendizaje automático. La solución desarrollada no solo buscó eliminar las ineficiencias del proceso manual, sino también generar conocimiento analítico a partir de los datos históricos disponibles, mediante la identificación de patrones de comportamiento, la estimación de indicadores operativos y la clasificación de períodos según su nivel de carga asistencial.

El estudio se desarrolló bajo la metodología CRISP-DM, que proporcionó un marco estructurado e iterativo para organizar el proceso desde el entendimiento del problema institucional hasta el despliegue de la solución, garantizando coherencia entre los objetivos planteados, las técnicas analíticas seleccionadas y los resultados obtenidos. La fuente principal para el análisis fue la base oficial de indicadores institucionales de hospitalización, consolidada por el área de estadística de la clínica a partir del sistema SERVINTE, que comprende 48 períodos mensuales correspondientes al período 2021–2024. Esta base registra un promedio mensual de 417 ingresos y 421 egresos, una estancia promedio de 4,15 días y una ocupación promedio del 82,1%, valores coherentes con el perfil operativo de una institución de mediana complejidad. Adicionalmente, se procesaron 22.723 registros individuales de ingresos hospitalarios para el análisis transaccional complementario.

Entre los hallazgos más relevantes del estudio se destacan cuatro resultados. Primero, la segmentación mediante K-Means identificó tres perfiles operativos diferenciados, revelando que los primeros ocho meses de 2021 constituyeron un período de bajo volumen con alta complejidad clínica, caracterizado por la estancia promedio más alta del período analizado (5,68 días), patrón que no era visible en el análisis descriptivo convencional. Segundo, el modelo de regresión lineal para egresos demostró una relación casi lineal con los ingresos mensuales, con un R^2=0,911 y un error relativo del 4,3%, convirtiéndose en una herramienta de planificación operativa de alta utilidad práctica. Tercero, el modelo de regresión para días de estancia totales obtuvo un R^2=0,143, evidenciando que este indicador depende de la complejidad clínica individual y no del volumen de pacientes, resultado que delimita con precisión qué información adicional se requiere para abordarlo en fases futuras del sistema. Cuarto, el árbol de decisión

alcanzó una exactitud del 93,3% en la clasificación de los tres perfiles operativos, generando reglas interpretables directamente aplicables por los equipos de gestión sin mediación técnica.

El documento se organiza de la siguiente manera: el capítulo dos presenta el marco teórico que sustenta el enfoque analítico adoptado; el capítulo tres describe el planteamiento del problema, los objetivos y la justificación del estudio; el capítulo cuatro desarrolla la metodología aplicada siguiendo las fases de CRISP-DM; el capítulo cinco presenta los resultados obtenidos; y los capítulos finales recogen las conclusiones, recomendaciones y referencias bibliográficas del trabajo.

Justificación

La gestión eficiente de la información se ha convertido en un factor crítico para la toma de decisiones en las organizaciones de salud, especialmente en contextos caracterizados por la alta complejidad operativa y el crecimiento sostenido del volumen de datos. La adecuada integración de las tecnologías de la información y la comunicación en los procesos sanitarios permite mejorar la calidad de los datos, reducir errores y optimizar el desempeño institucional, impactando de manera directa la eficiencia operativa y la calidad de la atención en salud (Berg, 2004).

En el ámbito hospitalario, los indicadores de gestión constituyen una herramienta fundamental para el seguimiento del desempeño operativo, la planificación de recursos y la evaluación de la eficiencia de los servicios. Sin embargo, cuando los procesos de consolidación y reporte de indicadores se realizan de forma manual o con bajo nivel de estandarización, se incrementa el reproceso, se generan inconsistencias en la información y se limita el aprovechamiento del potencial analítico de los datos disponibles. Sedkaoui (2018) señala que la analítica de datos y el big data permiten transformar grandes volúmenes de información en conocimiento útil para la gestión organizacional, siempre que existan procesos estructurados de automatización, limpieza y control de calidad de los datos.

De manera particular, Santos et al. (2020), en el capítulo dedicado a la aplicación del big data en sanidad, destacan que el análisis avanzado de datos sanitarios facilita la optimización de procesos hospitalarios, la mejora del seguimiento de indicadores clínicos y administrativos, y el fortalecimiento de la toma de decisiones basada en evidencia. Los autores subrayan que la automatización del procesamiento de datos y la incorporación de modelos analíticos contribuyen

a una gestión más eficiente de los recursos hospitalarios y a una mejor comprensión del comportamiento de los servicios de salud.

Adicionalmente, Rajkomar et al. (2019) señalan que los modelos de machine learning aplicados a datos clínicos y operativos permiten identificar patrones ocultos, predecir comportamientos futuros y generar alertas tempranas que contribuyen a una gestión más proactiva de los servicios hospitalarios. Los autores destacan que, si bien el machine learning ha mostrado resultados prometedores en el diagnóstico médico, su aplicación en la optimización de procesos operativos y en la gestión de indicadores institucionales representa una oportunidad significativa para mejorar la eficiencia y la calidad de la atención.

En este contexto, el presente trabajo se justifica por la necesidad de optimizar los procesos de consolidación y reporte de indicadores institucionales en la Clínica de Marly Jorge Cavelier Gaviria, mediante el desarrollo de un sistema automatizado que integre técnicas de análisis de datos y machine learning. Actualmente, los métodos manuales utilizados en la gestión de datos generan ineficiencias, limitan la precisión de los informes y restringen la capacidad de la institución para responder de manera ágil a las demandas del entorno. Estas deficiencias no solo afectan los procesos internos, sino que también comprometen la calidad del servicio al paciente y la competitividad de la institución. La solución propuesta busca reducir el reproceso manual, mejorar la trazabilidad de la información, garantizar la consistencia de los indicadores y apoyar la toma de decisiones institucionales basadas en evidencia, alineándose con los planteamientos teóricos sobre el uso del big data y machine learning en sanidad, y con las tendencias actuales de transformación digital en el sector salud.

Objetivos

Objetivo General

Diseñar y desarrollar un sistema automatizado de consolidación y análisis de indicadores institucionales en la Clínica de Marly Jorge Cavelier Gaviria, mediante técnicas de análisis de datos y machine learning, con el fin de apoyar la toma de decisiones basada en evidencia

Objetivos Específicos

Diseñar y desarrollar un sistema automatizado de extracción, limpieza y consolidación de datos institucionales, orientado a reducir la dependencia de procesos manuales en la generación de indicadores.

Aplicar modelos de machine learning para analizar patrones operativos y evaluar el comportamiento de indicadores hospitalarios.

Generar archivos estructurados y resultados analíticos que permitan su integración en herramientas de visualización y Business Intelligence.

Marco Teórico

Gestión de la Información en el Sector Salud

La gestión de la información en el sector salud constituye un componente estratégico para garantizar la calidad asistencial, la eficiencia operativa y la sostenibilidad institucional. En entornos hospitalarios, los sistemas de información clínica y administrativa generan grandes volúmenes de datos relacionados con ingresos, egresos, diagnósticos, tiempos de estancia y utilización de recursos. Sin embargo, la mera disponibilidad de datos no garantiza su aprovechamiento efectivo. Berg (2003) señala que la integración de tecnologías de la información en el trabajo sanitario no solo implica la digitalización de procesos, sino también la reorganización del flujo de información para apoyar la toma de decisiones clínicas y administrativas. En este sentido, la adecuada estructuración, estandarización y análisis de los datos hospitalarios resulta fundamental para transformar la información en conocimiento útil que oriente la gestión institucional.En este marco, la gestión de indicadores institucionales adquiere relevancia estratégica, pues permite operacionalizar el análisis del desempeño hospitalario y monitorear variables críticas como la ocupación de camas, la rotación de pacientes y la eficiencia de los servicios. Sin embargo, cuando esta consolidación se realiza mediante procesos manuales, se incrementa el riesgo de errores, reprocesos e inconsistencias, lo que limita la capacidad analítica de la institución y obstaculiza la toma de decisiones oportuna. Esta problemática justifica la exploración de alternativas automatizadas para el procesamiento y análisis de indicadores, tal como se aborda en el presente estudio

Big Data y Analítica de Datos en Sanidad

El concepto de Big Data en salud se refiere al manejo y análisis de grandes volúmenes de datos clínicos y administrativos con el fin de generar valor para la gestión hospitalaria y la

atención al paciente. Sedkaoui (2018) destaca que la analítica de datos permite enfrentar los desafíos de los sistemas de salud modernos, siempre que se cuente con procesos estructurados de limpieza, integración y control de calidad de la información; precisamente los elementos que este trabajo busca sistematizar en el contexto del procesamiento de indicadores hospitalarios.

En esta línea, Santos y Costa (2020) señalan que la aplicación del Big Data en sanidad contribuye a la optimización del flujo de pacientes, la planificación de recursos y el seguimiento de indicadores operativos, y enfatizan que la automatización del procesamiento de datos es un paso fundamental para garantizar la consistencia y confiabilidad de los análisis. Estos planteamientos son coherentes con el enfoque del presente estudio, en el que la automatización del procesamiento constituye la base sobre la cual se construyen los modelos analíticos.

En el contexto hospitalario, la analítica de datos trasciende el diagnóstico clínico para abarcar la gestión operativa: la evaluación de la ocupación, el análisis de la duración de la estancia y la identificación de patrones en el uso de los servicios. Estos elementos fortalecen la toma de decisiones basada en evidencia y contribuyen a mejorar la eficiencia institucional, conectando así con la necesidad de implementar modelos de aprendizaje automático que permitan ir más allá de la descripción hacia la predicción.

Machine Learning Aplicado a la Gestión Hospitalaria

El Machine Learning (ML) ha demostrado ser una herramienta poderosa en el análisis de datos sanitarios, permitiendo identificar patrones complejos y generar modelos predictivos a partir de información histórica. Rajkomar, Dean y Kohane (2019) señalan que el aprendizaje automático puede aplicarse no solo al diagnóstico médico, sino también a la optimización de procesos operativos en instituciones de salud, abriendo un campo amplio para su uso en la gestión de indicadores hospitalarios.

Los modelos supervisados, como la regresión lineal y los árboles de decisión, permiten analizar relaciones entre variables y generar predicciones interpretables a partir de datos etiquetados. Por su parte, los modelos no supervisados, como el clustering K-Means, facilitan la identificación de patrones de comportamiento sin necesidad de clasificaciones previas. En el presente estudio, la selección de modelos consideró tanto el desempeño estadístico como la naturaleza y el volumen de los datos disponibles, priorizando enfoques que ofrecieran resultados interpretables por los equipos clínicos y administrativos de la institución, sin requerir formación estadística avanzada.

La interpretabilidad resulta especialmente relevante en entornos hospitalarios, donde las decisiones deben poder sustentarse y comunicarse con claridad. Por ello, la elección de modelos no respondió únicamente a criterios de precisión, sino también a su aplicabilidad práctica dentro del contexto organizacional analizado.

Metodología CRISP-DM en Proyectos de Analítica de Datos

La metodología CRISP-DM (Cross Industry Standard Process for Data Mining) fue propuesta como un marco estructurado para el desarrollo de proyectos de minería de datos y análisis predictivo (Chapman et al., 2000). Este modelo organiza el proceso en seis fases interrelacionadas: entendimiento del negocio, entendimiento de los datos, preparación de los datos, modelado, evaluación y despliegue. Una de sus ventajas más destacadas frente a marcos lineales es su carácter iterativo, que permite revisar y ajustar decisiones tomadas en fases anteriores a medida que avanza el análisis.

CRISP-DM es ampliamente adoptada en proyectos de ciencia de datos por su enfoque práctico y su capacidad para garantizar coherencia entre los objetivos organizacionales y las técnicas analíticas empleadas. Su aplicación en el sector salud facilita la estructuración

sistemática del análisis, desde la identificación del problema hasta la implementación de soluciones basadas en datos, asegurando que cada decisión técnica esté respaldada por una comprensión clara del contexto institucional.

En el presente estudio, CRISP-DM se adoptó como marco metodológico rector para organizar el proceso de automatización y modelado de indicadores hospitalarios. Su estructura permitió establecer una trazabilidad clara entre las fases de preparación de datos, construcción de modelos y análisis de resultados, garantizando que las decisiones técnicas respondieran consistentemente a los objetivos institucionales planteados desde el inicio del proyecto.

Metodología CRISP-DM

La metodología CRISP-DM fue adoptada como marco estructurador del presente proyecto por su capacidad para alinear de forma sistemática los objetivos organizacionales con las técnicas analíticas seleccionadas, y por su carácter iterativo, que permite revisar y ajustar decisiones a medida que avanza el proceso. A diferencia de enfoques lineales, CRISP-DM concibe el desarrollo de proyectos de analítica como un ciclo continuo en el que cada fase alimenta a las siguientes y puede ser revisada a la luz de los hallazgos posteriores. En las secciones que siguen se describe la aplicación de cada una de sus seis fases al contexto específico del estudio.

Entendimiento del Negocio

La primera fase tuvo como propósito comprender el contexto operativo de la Clínica de Marly Jorge Cavelier Gaviria e identificar con precisión el problema que el proyecto buscaba resolver. Esta institución gestiona mensualmente un conjunto de indicadores hospitalarios relacionados con ingresos, egresos y días de estancia, los cuales constituyen insumos fundamentales para la evaluación del desempeño asistencial y la planificación de recursos.

El diagnóstico inicial reveló que el proceso de consolidación y reporte de estos indicadores se realizaba mediante procedimientos manuales, apoyados en hojas de cálculo sin automatización. Esta forma de trabajo generaba reprocesos recurrentes en la integración de información proveniente de distintas fuentes, incrementaba el riesgo de inconsistencias entre bases de datos, limitaba la capacidad de análisis oportuno y creaba una dependencia de tareas operativas repetitivas que reducían el tiempo disponible para el análisis estratégico. En conjunto, estas condiciones configuraban un problema analítico con impacto directo sobre la calidad de la información disponible para la toma de decisiones institucionales.

A partir de este diagnóstico, se definió como objetivo central del proyecto el diseño e implementación de un sistema automatizado para el procesamiento y análisis mensual de indicadores hospitalarios, con énfasis en las variables de hospitalización. Este objetivo se operacionalizó a través de tres lineamientos metodológicos: adoptar el periodo mensual como unidad de análisis, en coherencia con la periodicidad con que la institución gestiona sus indicadores; priorizar modelos interpretables y aplicables en el entorno hospitalario, donde la transparencia de los resultados es condición para su adopción; y desarrollar un sistema escalable, replicable y compatible con las herramientas de Business Intelligence disponibles en la institución.

Los principales casos de uso identificados para el análisis de datos en este contexto incluyeron la evaluación del flujo de pacientes, el análisis de la carga asistencial por período y unidad, la identificación de patrones de ocupación hospitalaria y el apoyo a la toma de decisiones administrativas. Estos casos de uso orientaron tanto la selección de variables como el diseño de los modelos implementados en fases posteriores, asegurando que el proyecto respondiera a necesidades concretas y verificables de la institución.

Entendimiento de los Datos

Con el problema analítico claramente definido, la segunda fase se orientó a caracterizar las fuentes de información disponibles, evaluar su calidad estructural y determinar su potencial para responder a los objetivos del proyecto. Se identificaron tres bases de datos principales, todas descargadas desde el sistema SERVINTE en formato Excel: la base de ingresos hospitalarios, la base de egresos hospitalarios y la base de indicadores institucionales consolidados.

La revisión estructural de estas fuentes permitió identificar diferencias en la nomenclatura de las variables entre bases, variaciones en los formatos de encabezado y ausencia

de estandarización en los tipos de datos asignados a campos equivalentes. Adicionalmente, se detectaron diferencias en la frecuencia de registro: mientras las bases de ingresos y egresos contenían registros a nivel diario e individual, la base de indicadores institucionales presentaba información con periodicidad mensual, lo que implicaba la necesidad de una agregación temporal para homogeneizar las fuentes antes de su integración.

El análisis exploratorio de datos (EDA) realizado en esta fase permitió obtener una caracterización inicial del conjunto de información disponible. Se examinaron las distribuciones de las variables numéricas principales, identificando el rango y la dispersión de los días de estancia y los volúmenes mensuales de ingresos y egresos. Se detectaron valores atípicos en la variable días de estancia, asociados a estancias de duración inusualmente prolongada, cuyo tratamiento fue contemplado en la fase de preparación. Asimismo, el análisis de tendencias temporales mostró variaciones estacionales en los volúmenes de ingresos que resultarían relevantes para la interpretación de los resultados del modelado.

La exploración de las relaciones entre variables, particularmente entre el tipo de paciente, la unidad de hospitalización y los días de estancia, confirmó la existencia de diferencias sistemáticas entre segmentos que justificaban el uso de técnicas de agrupamiento para identificar perfiles de comportamiento. En conjunto, los hallazgos de esta fase confirmaron que los datos disponibles eran suficientes en calidad y volumen para el propósito del estudio, aunque requerían procesos de estandarización, consolidación temporal y tratamiento de valores atípicos antes de proceder al modelado. Estas necesidades orientaron directamente el diseño de la fase de preparación de datos descrita en el apartado siguiente.

Preparación de los Datos

La fase de preparación de los datos tuvo como objetivo transformar la información proveniente de las bases institucionales en una estructura consistente, íntegra y adecuada para el análisis estadístico y el modelado predictivo. Esta etapa es determinante en cualquier proyecto de ciencia de datos, puesto que la calidad de los resultados depende directamente de la confiabilidad y coherencia de los datos sobre los que se construyen los modelos.

Estandarización y Limpieza Inicial

El proceso comenzó con la carga de las bases de datos de ingresos, egresos e indicadores institucionales, disponibles en formato Excel. Dado que estas fuentes fueron generadas por distintos sistemas y en diferentes períodos, presentaban variaciones en la nomenclatura de columnas, formatos de encabezado y criterios de registro que requerían ser unificados antes de cualquier análisis.

Para garantizar consistencia en el tratamiento de la información, se implementaron procedimientos de estandarización que incluyeron la eliminación de espacios en blanco en los encabezados, la conversión de nombres de columnas a mayúsculas y la homogeneización de denominaciones a través de un mecanismo de identificación flexible de variables equivalentes entre bases. Este último procedimiento resultó especialmente relevante, dado que una misma variable podía estar nombrada de forma distinta según la fuente de origen.

Adicionalmente, se verificó la coherencia de los tipos de datos asignados a cada campo y se realizó una inspección sistemática para identificar valores faltantes, registros duplicados o entradas inconsistentes que pudieran comprometer la integridad del análisis posterior.

Selección y Transformación de Variables

Una vez garantizada la consistencia estructural de las bases, se seleccionaron las variables relevantes para el análisis del comportamiento operativo hospitalario. Entre ellas se consideraron el tipo de paciente, el mes de ingreso y egreso, la unidad de hospitalización y los días de estancia, por ser las dimensiones que permiten caracterizar la dinámica asistencial y construir los indicadores de interés.

La variable días de estancia, central en el análisis, fue convertida a formato numérico con el fin de permitir su uso en cálculos agregados y modelos predictivos. Los registros que contenían valores no válidos o no convertibles fueron eliminados de manera controlada, dado que su conservación habría introducido sesgos en el cálculo de indicadores como la estancia promedio. Esta decisión se tomó tras verificar que el volumen de registros afectados no comprometía la representatividad del conjunto de datos.

Durante la inspección de la variable días de estancia en la base de indicadores institucionales, se identificó una inconsistencia de formato en un subconjunto de registros: los valores correspondientes a determinados períodos estaban expresados en unidades reducidas por un factor de mil, lo que producía estancias promedio inferiores a un día, incompatibles con el comportamiento operativo real de un servicio de hospitalización convencional. Este hallazgo fue corregido de forma automatizada mediante una regla de transformación que identifica valores inferiores a 100 y aplica el factor de corrección correspondiente, garantizando que todos los registros expresen días de estancia en la misma unidad antes de proceder al cálculo de indicadores agregados.

Las variables categóricas que se incorporaron en los modelos de aprendizaje automático fueron codificadas a formato numérico mediante esquemas de codificación apropiados según la

naturaleza de cada variable, con el propósito de hacerlas compatibles con los algoritmos implementados.

Consolidación Mensual de la Información

Dado que los indicadores institucionales se gestionan y reportan con periodicidad mensual, la información fue agregada temporalmente mediante operaciones de agrupación por mes y unidad de análisis. Esta decisión metodológica responde directamente a la lógica de seguimiento operativo de la institución, asegurando que la estructura de los datos sea coherente con los procesos reales de monitoreo y toma de decisiones.

Como resultado de esta consolidación se obtuvieron, para cada mes, el total de ingresos, el total de egresos, la suma de días de estancia y la estancia promedio. Estas variables agregadas constituyeron la base analítica sobre la que se construyeron los modelos predictivos y de agrupamiento, y permitieron reducir la dimensionalidad de los datos sin pérdida de información relevante para los objetivos del estudio.

Preparación para Modelado

Con los datos consolidados y estructurados cronológicamente, se realizaron los ajustes finales requeridos por cada tipo de modelo. Para los algoritmos basados en distancia, como K-Means, se aplicó estandarización de variables mediante escalado, con el fin de evitar que diferencias en la magnitud de las variables distorsionaran los resultados del agrupamiento.

Para los modelos supervisados, el conjunto de datos fue dividido en subconjuntos de entrenamiento y prueba, lo que permitió evaluar la capacidad de generalización de los algoritmos sobre datos no vistos durante el ajuste del modelo. La proporción de división utilizada se definió en función del volumen total de observaciones disponibles, buscando un equilibrio entre la representatividad del conjunto de entrenamiento y la robustez de la evaluación.

En conjunto, las decisiones adoptadas en esta fase garantizan que los datos utilizados en el modelado sean representativos, íntegros y coherentes con los objetivos institucionales planteados, constituyendo una base sólida para la interpretación de los resultados obtenidos en las fases siguientes.

Modelado

La fase de modelado tuvo como propósito aplicar técnicas de aprendizaje automático que permitieran, de manera complementaria, describir patrones históricos, predecir indicadores operativos y clasificar el comportamiento mensual de la institución en categorías de ocupación. La selección de los modelos respondió a tres criterios articulados: la naturaleza agregada y de tamaño moderado de los datos disponibles, la necesidad de obtener resultados interpretables por equipos clínicos y administrativos sin formación estadística avanzada, y la complementariedad entre enfoques no supervisados y supervisados para abordar distintas dimensiones del problema.

Con base en estos criterios, se implementaron tres modelos: clustering K-Means para la identificación de patrones de comportamiento mensual, regresión lineal múltiple para la estimación de indicadores operativos, y árbol de decisión para la clasificación de niveles de ocupación. A continuación se describe cada uno.

Modelo de Clustering K-Means

El algoritmo K-Means fue seleccionado con el propósito de identificar grupos de periodos mensuales con comportamientos operativos similares, sin partir de categorías predefinidas. Este enfoque no supervisado resulta adecuado cuando el objetivo es explorar la estructura latente de los datos y detectar patrones que no son evidentes a partir del análisis descriptivo convencional (Rajkomar et al., 2019).

Las variables de agrupamiento incluyeron el total mensual de ingresos, el total mensual de egresos, los días de estancia totales y la estancia promedio mensual, por ser las dimensiones que caracterizan de forma más completa el comportamiento operativo en cada período.

Previamente al entrenamiento, se aplicó estandarización mediante StandardScaler, procedimiento necesario para evitar que las diferencias en la magnitud de las variables sesgaran el cálculo de distancias sobre el que se basa el algoritmo.

Para determinar el número óptimo de clústeres se empleó el método del codo, que evalúa la reducción de la inercia interna a medida que aumenta el valor de K, identificando el punto a partir del cual incrementar el número de grupos no aporta una mejora sustancial. Como validación complementaria, se calculó el índice Silhouette, que mide la cohesión interna de cada

clúster y su separación respecto a los demás, ofreciendo una valoración cuantitativa de la calidad del agrupamiento obtenido.

Modelo de Regresión Lineal Múltiple

Se implementó un modelo de regresión lineal múltiple con el objetivo de estimar el comportamiento de indicadores operativos clave a partir de variables de entrada disponibles en las bases institucionales. La regresión lineal fue seleccionada por su interpretabilidad, su adecuación a muestras de tamaño moderado y porque las relaciones entre las variables operativas analizadas presentaban un comportamiento aproximadamente lineal, lo que hace coherente su aplicación en este contexto (Santos y Costa, 2020).

Se implementaron dos modelos de regresión lineal. El primero utilizó el total de ingresos y el total de egresos mensuales como variables independientes para predecir los días de estancia totales acumulados en cada período. El segundo utilizó el total de ingresos como única variable independiente para predecir el total de egresos mensuales. En ambos casos, el conjunto de datos

fue dividido en subconjuntos de entrenamiento (70%) y prueba (30%), y el desempeño se evaluó mediante el coeficiente de determinación R^2, que indica la proporción de varianza explicada por el modelo, y el error absoluto medio (MAE), que cuantifica en unidades originales la magnitud promedio de las desviaciones entre los valores observados y los estimados. No se incorporaron variables de tendencia temporal ni rezagos en ninguno de los dos modelos, dado que el análisis se orientó a evaluar las relaciones operativas contemporáneas entre variables del mismo período mensual.

Modelo Supervisado Árbol de Decisión

Como complemento a los enfoques anteriores, se implementó un árbol de decisión para clasificar los periodos mensuales en categorías de ocupación hospitalaria, definidas a partir de umbrales derivados del comportamiento histórico de la estancia promedio mensual. Este modelo supervisado permite generar reglas explícitas de clasificación expresadas en lenguaje condicional, lo que facilita su comprensión e interpretación por parte de los equipos administrativos y clínicos.

Las variables de entrada incluyeron el total mensual de ingresos y el total mensual de egresos, mientras que la variable objetivo correspondió al nivel de ocupación categorizado. El modelo fue entrenado y evaluado bajo el mismo esquema de división en conjuntos de entrenamiento y prueba aplicado en la regresión lineal, garantizando consistencia metodológica entre los modelos supervisados.

El desempeño del clasificador se evaluó mediante exactitud global (accuracy) y F1-score, esta última especialmente relevante en contextos donde las categorías pueden presentar distribución desbalanceada. La elección del árbol de decisión frente a algoritmos de mayor complejidad, como bosques aleatorios o redes neuronales, responde al principio de parsimonia y

a la prioridad de la interpretabilidad sobre la sofisticación técnica, criterio especialmente pertinente en entornos institucionales donde los resultados deben poder explicarse y sustentarse ante audiencias no especializadas.

Los tres modelos implementados abordan dimensiones distintas pero articuladas del mismo problema: K-Means describe la estructura histórica de los datos identificando tipologías de comportamiento mensual; la regresión lineal proyecta el valor futuro de indicadores operativos clave; y el árbol de decisión traduce esa información en categorías de ocupación accionables para la gestión institucional. Esta complementariedad garantiza que los resultados del modelado respondan de manera integral a los objetivos del estudio.

Evaluación de los Modelos

Una vez construidos los modelos de aprendizaje automático, se procedió a su evaluación mediante criterios cuantitativos y cualitativos, con el propósito de verificar su desempeño estadístico, su estabilidad ante datos no vistos y su coherencia con los objetivos operativos del estudio. Esta fase es determinante dentro del marco CRISP-DM, pues permite determinar si los modelos desarrollados son suficientemente robustos para respaldar decisiones institucionales o si requieren ajustes antes de su implementación.

Evaluación Cuantitativa

Cada modelo fue evaluado con métricas acordes a su naturaleza y propósito. Para el modelo de clustering K-Means, el índice Silhouette constituyó la medida principal de validación interna, complementando el criterio del codo utilizado durante la selección del número de clústeres. Un valor de Silhouette cercano a 1 indica que los periodos agrupados presentan alta cohesión interna y están bien diferenciados de los demás grupos, lo que permite juzgar si la segmentación obtenida es estadísticamente válida y operativamente significativa.

Para los modelos de regresión lineal, el coeficiente de determinación (R^2) permitió estimar la proporción de variabilidad de la variable dependiente explicada por el modelo, mientras que el Error Absoluto Medio (MAE) cuantificó la magnitud promedio de las desviaciones entre los valores observados y los estimados, expresada en las unidades originales de cada indicador. Esta combinación de métricas ofrece una visión complementaria del desempeño: R^2 informa sobre el ajuste global del modelo y MAE sobre la precisión práctica de sus predicciones.

En el árbol de decisión, la exactitud global (accuracy) indicó la proporción de periodos correctamente clasificados, mientras que el F1-score permitió evaluar el equilibrio entre precisión y exhaustividad en cada categoría de ocupación. Esta última métrica resultó especialmente relevante dado que las categorías definidas podían presentar distribución desigual en el conjunto de datos disponible.

En todos los modelos supervisados, la división de los datos en conjuntos de entrenamiento y prueba permitió estimar la capacidad de generalización de cada algoritmo y detectar posibles señales de sobreajuste, garantizando que los resultados reportados reflejen el comportamiento esperado del modelo ante datos reales no utilizados durante su construcción.

Evaluación Cualitativa

Más allá de las métricas estadísticas, se realizó una evaluación cualitativa orientada a determinar si los modelos implementados son útiles, comprensibles y pertinentes en el contexto institucional en que se aplicarán. Este tipo de valoración es especialmente relevante en entornos hospitalarios, donde la aceptación de un modelo por parte de los equipos administrativos y clínicos depende en gran medida de su transparencia y de la posibilidad de vincular sus resultados con el conocimiento experto de los profesionales.

En este sentido, se analizó si los clústeres identificados por K-Means correspondían a tipologías de comportamiento operativo reconocibles y diferenciables desde una perspectiva institucional, verificando que los grupos no fueran artefactos estadísticos sino patrones con sentido práctico. Del mismo modo, se examinó si los coeficientes de la regresión lineal reflejaban relaciones coherentes con la lógica operativa esperada, y si las reglas generadas por el árbol de decisión podían ser interpretadas y utilizadas por los equipos de gestión sin requerir formación técnica especializada.

La prioridad otorgada a la interpretabilidad sobre la complejidad algorítmica responde a un criterio metodológico deliberado: en contextos donde los modelos deben integrarse en procesos institucionales reales, un modelo comprensible con desempeño aceptable es preferible a un modelo de alta precisión cuyo funcionamiento resulte opaco para sus usuarios.

Alineación con los Objetivos del Estudio

Como cierre de la fase de evaluación, se verificó de manera explícita la contribución de cada modelo al cumplimiento del objetivo general del estudio. El modelo K-Means permitió identificar patrones de comportamiento mensual que no eran visibles a través del análisis descriptivo convencional, aportando una base para la segmentación operativa institucional. Los modelos de regresión lineal ofrecieron estimaciones cuantificables de indicadores clave, apoyando la planificación anticipada de recursos. El árbol de decisión, por su parte, tradujo la información operativa en categorías de ocupación accionables, facilitando la comunicación de resultados a audiencias no especializadas.

Esta evaluación integral permitió determinar que los modelos seleccionados son pertinentes para el propósito del estudio y que sus resultados son coherentes con la realidad operativa de la institución. Al mismo tiempo, el proceso evidenció oportunidades de mejora que

podrán ser abordadas en iteraciones futuras, en consonancia con el carácter cíclico de la metodología CRISP-DM, que concibe la evaluación no como un punto de cierre definitivo sino como un insumo para el refinamiento continuo del sistema analítico.

Despliegue

La fase de despliegue tuvo como propósito materializar los resultados del procesamiento y modelado en una solución funcional, reproducible y preparada para su integración progresiva con los procesos de análisis institucional. En el marco de este estudio, el despliegue no implicó la implementación formal del sistema dentro de la infraestructura tecnológica de la institución, decisión coherente con el alcance del proyecto y con las condiciones de acceso a los entornos productivos. No obstante, el sistema desarrollado fue diseñado desde su concepción para facilitar dicha integración en una etapa posterior, priorizando la modularidad, la documentación y la compatibilidad con las herramientas de análisis ya disponibles en la institución.

Desarrollo del Sistema Automatizado

Se desarrolló un script en Python estructurado como un pipeline de procesamiento de datos, capaz de ejecutar de forma secuencial y reproducible las etapas de carga, limpieza, consolidación, modelado y exportación de resultados. La decisión de implementar el sistema como un pipeline unificado responde a la necesidad de garantizar que el proceso completo pueda ejecutarse con mínima intervención manual, reduciendo el riesgo de errores operativos y asegurando la trazabilidad entre los datos de entrada y los resultados generados.

El diseño del script siguió un principio de modularidad, organizando cada etapa del proceso en funciones independientes que pueden ser ejecutadas, modificadas o reemplazadas sin afectar la arquitectura general del sistema. Esta característica resulta especialmente relevante para la sostenibilidad del proyecto, pues permite incorporar nuevas variables, actualizar los modelos o adaptar el sistema a cambios en la estructura de las fuentes de datos sin necesidad de reescribir el código desde cero.

Generación de Archivos Estructurados

Como resultado del procesamiento automatizado, el sistema genera archivos de salida en formato Excel organizados según el tipo de información producida. Se generan bases depuradas de ingresos y egresos, resúmenes mensuales consolidados, una tabla mensual integrada para análisis gerencial y los resultados derivados de cada modelo analítico implementado.

La elección del formato Excel como medio de exportación responde a criterios pragmáticos: es el formato de mayor uso y familiaridad en los equipos administrativos de la institución, no requiere instalación de software especializado para su consulta y es directamente compatible con las principales herramientas de visualización y Business Intelligence disponibles

en el mercado. De este modo, los archivos generados pueden ser utilizados como insumo directo para la construcción de tableros de control dinámicos, sin requerir transformaciones adicionales.

Integración con Herramientas de Visualización

La estructura y nomenclatura de los archivos de salida fueron diseñadas para facilitar su conexión con plataformas de visualización de datos como Power BI, permitiendo que los indicadores consolidados y los resultados de los modelos puedan ser monitoreados de forma continua a través de tableros interactivos. Esta integración posibilita el seguimiento mensual de indicadores operativos, la visualización de tendencias históricas, la identificación de patrones de comportamiento y el apoyo a la toma de decisiones basada en evidencia, sin que los usuarios finales necesiten interactuar directamente con el código o los modelos subyacentes.

La orientación hacia herramientas de Business Intelligence responde a una decisión metodológica deliberada: separar la capa de procesamiento y modelado, gestionada por el pipeline en Python, de la capa de visualización y consulta, accesible para los equipos de gestión. Esta separación favorece la adopción institucional del sistema, al presentar los resultados en un formato familiar y accionable para los tomadores de decisiones.

Documentación y Escalabilidad

Con el propósito de garantizar la sostenibilidad y replicabilidad del sistema más allá del alcance del presente estudio, se elaboró documentación técnica que describe las funciones principales del script, las variables utilizadas en cada etapa, el flujo de procesamiento y las métricas de evaluación aplicadas a cada modelo. Esta documentación no constituye un elemento accesorio, sino una condición necesaria para que el sistema pueda ser mantenido, auditado o ampliado por otros profesionales de la institución en iteraciones futuras.

La escalabilidad del sistema fue considerada desde el diseño. La arquitectura modular del pipeline permite incorporar nuevas fuentes de datos, añadir modelos adicionales o extender el análisis a otras unidades o servicios sin comprometer el funcionamiento del núcleo del sistema. Esta capacidad de evolución es coherente con el carácter iterativo de la metodología CRISP-DM, que concibe el despliegue no como el cierre definitivo del proyecto sino como el punto de partida para ciclos sucesivos de refinamiento y mejora.

En conjunto, la fase de despliegue cierra el ciclo metodológico establecido por CRISP-DM, demostrando que los objetivos planteados en la fase de entendimiento del negocio se tradujeron en una solución concreta, documentada y orientada a la generación de valor operativo para la institución. El sistema desarrollado constituye una base funcional sobre la cual es posible avanzar hacia una implementación formal e integrada en los procesos de gestión hospitalaria.

Resultados

Implementación del Sistema Automatizado

Como resultado de la aplicación de la metodología CRISP-DM, se desarrolló un sistema automatizado orientado al procesamiento, consolidación y análisis mensual de los indicadores de hospitalización de la Clínica de Marly Jorge Cavelier Gaviria. Este sistema constituye el producto técnico central del proyecto y materializa la respuesta al problema identificado en la fase de entendimiento del negocio: la dependencia de procesos manuales para la consolidación de indicadores hospitalarios, con los riesgos de error, reproceso e inconsistencia que ello implicaba.

El sistema fue aplicado sobre la información correspondiente al período 2021–2024, procesando 22.723 registros individuales de ingresos hospitalarios y una base oficial de indicadores institucionales que comprende 48 períodos mensuales consolidados por el área de estadística de la clínica a partir del sistema SERVINTE. Esta base oficial constituyó la fuente principal para el modelado analítico, dado que representa los indicadores con los que la institución gestiona operativamente su desempeño y garantiza consistencia y trazabilidad a lo largo del período analizado.

La integración y procesamiento de estas fuentes requirió resolver diferencias estructurales entre bases, incluyendo variaciones en la nomenclatura de variables, inconsistencias en los tipos de datos y problemas de formato en algunas columnas numéricas. Un hallazgo relevante del proceso de preparación fue que la columna de sumatoria de días de estancia presentaba valores expresados en distintas unidades según el período, situación que fue identificada y corregida de forma automatizada mediante el pipeline desarrollado. El sistema resolvió estas diferencias de

forma reproducible, garantizando que los resultados sean consistentes independientemente del operador que ejecute el proceso y del período analizado.

Como producto central, el sistema construye una tabla mensual consolidada que integra en una única estructura analítica los principales indicadores operativos de hospitalización: total mensual de ingresos, total mensual de egresos, días de estancia acumulados, estancia promedio mensual y porcentaje de ocupación hospitalaria. Para el período 2021–2024, esta tabla refleja un promedio mensual de 417 ingresos y 421 egresos, una estancia promedio de 4,15 días y una ocupación promedio del 82,1%, valores coherentes con el perfil operativo de una institución de mediana complejidad en funcionamiento continuo.

En conjunto, la implementación del sistema automatizado demostró que es posible transformar un proceso manual, fragmentado y propenso a errores en un flujo de procesamiento estructurado, reproducible y escalable, sentando las bases para el análisis de patrones y la construcción de modelos predictivos que se describen en las secciones siguientes.

Productos Generados por el Sistema Automatizado

Como resultado del procesamiento automatizado de la información institucional, el sistema desarrollado generó un conjunto de archivos estructurados en formato Excel, organizados según el nivel de transformación de los datos y su propósito analítico. Estos productos constituyen los entregables concretos del pipeline implementado y representan la materialización del proceso de transformación que va desde los registros individuales en bruto hasta los insumos listos para el análisis gerencial y el modelado predictivo.

El producto de mayor relevancia analítica es la tabla mensual consolidada, que integra en una única estructura los principales indicadores operativos de hospitalización para cada uno de los 48 períodos mensuales del período 2021–2024: total de ingresos, total de egresos, días de

estancia acumulados, estancia promedio mensual y porcentaje de ocupación. Este archivo constituye el insumo central para el análisis gerencial y la base sobre la cual se construyeron los modelos de aprendizaje automático implementados en el estudio. Su valor radica no solo en la información que contiene, sino en que esa información es el resultado de un proceso estandarizado y reproducible, que incluye la corrección automática de inconsistencias de formato presentes en las fuentes originales, garantizando su consistencia independientemente del período o del operador que ejecute el sistema.

Como productos intermedios, el sistema genera la base depurada de ingresos hospitalarios, con los registros previamente estandarizados y filtrados según las variables definidas para el análisis, eliminando inconsistencias estructurales presentes en la fuente original. Adicionalmente, se genera una base de egresos filtrada exclusivamente por el servicio de hospitalización, condición necesaria para aislar los registros clínicamente relevantes del conjunto total de egresos institucionales, que incluye servicios ambulatorios y de otra naturaleza.

Finalmente, el sistema genera un archivo de resultados de los modelos analíticos, que almacena las segmentaciones por clúster obtenidas mediante K-Means, las predicciones de días de estancia y egresos producidas por los modelos de regresión lineal, y las clasificaciones de niveles de ocupación hospitalaria derivadas del árbol de decisión. Este archivo permite comparar directamente los resultados modelados con el comportamiento histórico real, facilitando la validación práctica de los modelos y su comunicación a los equipos de gestión institucional.

En conjunto, estos productos garantizan trazabilidad completa entre las bases fuente, los procesos de transformación aplicados y los resultados obtenidos, condición fundamental para auditar el proceso, replicarlo en períodos futuros y sostener la confianza institucional en los análisis generados.

Construcción y Análisis de la Tabla Mensual Consolidada

Como resultado del procesamiento automatizado de la base oficial de indicadores institucionales, se construyó una tabla mensual consolidada que integra los principales indicadores operativos de hospitalización para cada uno de los 48 períodos del período 2021–2024. Esta estructura constituye la unidad de análisis central del estudio y el insumo sobre el cual se aplicaron los modelos de aprendizaje automático.

Para cada mes, la tabla registra el total de ingresos, el total de egresos, los días de estancia acumulados, la estancia promedio mensual y el porcentaje de ocupación hospitalaria. La Tabla 1 presenta una selección de períodos representativos que ilustra la variabilidad observada a lo largo del período analizado.

Tabla 1

Ejemplo de Tabla Mensual Consolidada de Hospitalización

Mes	Año	Ingresos	Egresos	Días estancia totales	Estancia promedio	% Ocupación
Enero	2021	256	248	1.526	6,15	77,40%
Junio	2021	265	247	1.684	6,82	95,80%
Marzo	2022	482	498	2.078	4,17	104,50%
Julio	2022	441	460	1.029	2,24	73,70%
Marzo	2023	554	541	1.659	3,07	82,00%
Mayo	2024	570	552	2.431	4,4	92,80%

El análisis de la tabla consolidada permite identificar patrones de comportamiento operativo con relevancia para la gestión institucional. Las estadísticas generales del período muestran un promedio mensual de 417 ingresos y 421 egresos, una estancia promedio general de

4,15 días y una ocupación promedio del 82,1%, valores coherentes con el perfil operativo de una institución de mediana complejidad en funcionamiento continuo.

En cuanto a la variabilidad de la estancia promedio, se observa un rango entre 2,24 días (julio 2022) y 6,82 días (junio 2021), lo que evidencia diferencias sustanciales en la complejidad o duración de las hospitalizaciones según el período. Los valores más altos de estancia promedio se concentran en el primer semestre de 2021, con una tendencia a estabilizarse entre 3,5 y 4,5 días a partir de 2022, comportamiento que puede asociarse a cambios en el perfil de los pacientes atendidos o en los protocolos de gestión del alta médica durante ese período.

Respecto al porcentaje de ocupación, se registran meses con valores superiores al 100%, como marzo y mayo de 2022, lo que indica que en esos períodos la demanda superó la capacidad nominal de camas disponibles, configurando escenarios de alta presión asistencial. En contraste, meses como enero de 2023 (64,7%) y septiembre de 2023 (66,8%) presentaron ocupaciones significativamente menores, sugiriendo variabilidad estacional en la demanda hospitalaria.

La relación entre ingresos y egresos es consistente a lo largo del período, con diferencias mensuales menores a 30 pacientes en la mayoría de los casos, lo que refleja una dinámica de rotación estable. Las mayores discrepancias se observan en meses de transición entre períodos de alta y baja demanda, donde los egresos pueden superar a los ingresos como resultado de la resolución de casos ingresados en el período anterior.

En conjunto, la tabla mensual consolidada transforma información operativa dispersa en una estructura analítica coherente que evidencia variaciones estacionales, diferencias en la presión asistencial entre períodos y relaciones entre variables que no son identificables a partir del análisis manual de registros individuales. Estos hallazgos preliminares fundamentan y

orientan la aplicación de las técnicas de segmentación y modelado predictivo presentadas en las secciones siguientes.

Segmentación de Periodos Mediante K-Means

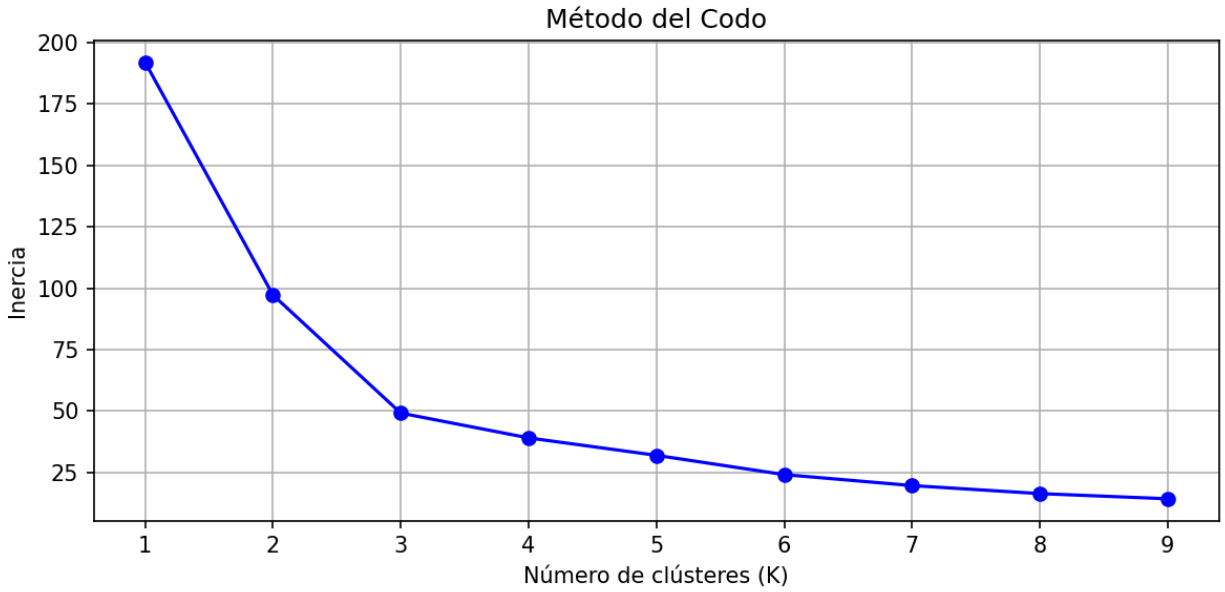
Con el propósito de identificar patrones de comportamiento en la ocupación hospitalaria mensual más allá del análisis descriptivo convencional, se aplicó el algoritmo K-Means sobre la tabla mensual consolidada de 48 períodos. Las variables utilizadas para el agrupamiento fueron el total de ingresos, el total de egresos, la sumatoria de días de estancia y la estancia promedio mensual, por ser las dimensiones que caracterizan de forma más completa la carga operativa en cada período. Previamente al entrenamiento, las variables fueron estandarizadas mediante StandardScaler para garantizar que las diferencias en magnitud no distorsionaran el cálculo de distancias sobre el que se basa el algoritmo.

Determinación del Número Óptimo de Clústeres

Para seleccionar el número adecuado de grupos se aplicó el método del codo, evaluando la variación de la inercia interna para valores de K entre 1 y 9. Como se observa en la Figura 1, la curva presenta una inflexión clara en K=3, punto a partir del cual la reducción de inercia se estabiliza progresivamente y el incremento en el número de grupos no aporta una mejora sustancial en la cohesión interna del agrupamiento. En consecuencia, se seleccionó K=3 como número óptimo de clústeres, decisión que además favorece la interpretabilidad operativa de los resultados.

Figura 1

Método del Codo para la Selección del Número Óptimo de Clústeres



Evaluación de la Calidad del Agrupamiento

La calidad de la segmentación fue evaluada mediante el índice Silhouette, que mide simultáneamente la cohesión interna de cada clúster y su separación respecto a los demás. El valor obtenido fue de 0,4804, correspondiente a una segmentación de calidad moderada-buena según los criterios establecidos en la literatura especializada. Este resultado indica que los tres grupos presentan diferenciación suficiente para sostener interpretaciones operativas, con un nivel de solapamiento limitado entre clústeres adyacentes.

Interpretación de los Clústeres

La Figura 2 y el análisis de los valores promedio de cada grupo permitieron identificar tres perfiles operativos diferenciados, cuyos atributos se resumen en la Tabla 2.

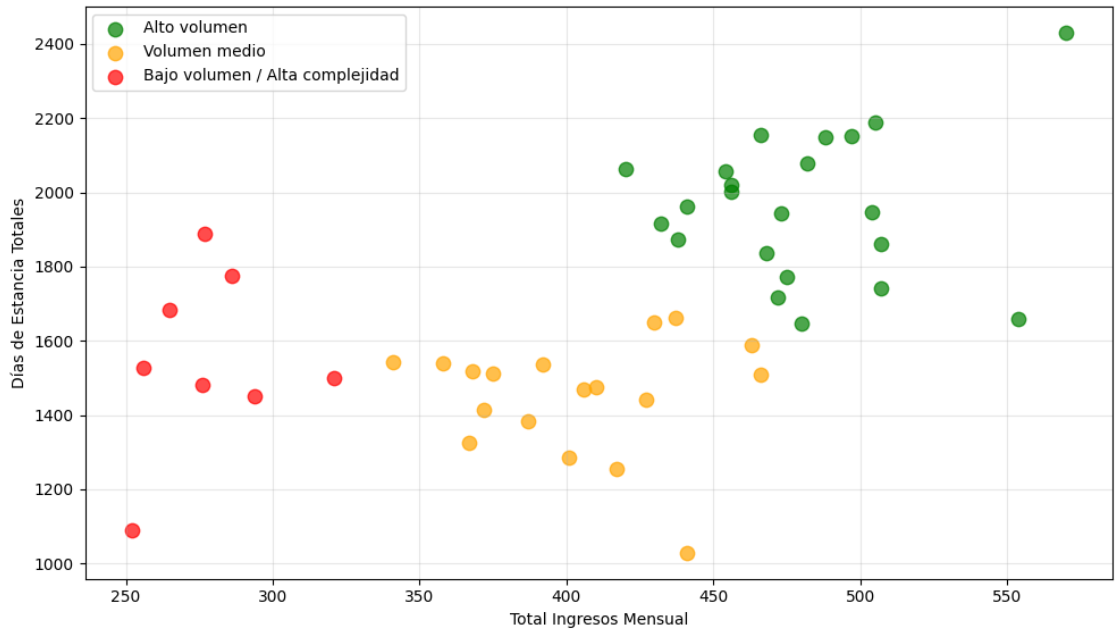
Tabla 2

Caracterización de los Clústeres Obtenidos Mediante K-Means

Clúster	Perfil	Períodos	Ingresos prom	Días estancia prom	Estancia prom	Ocupación prom
0	Alto volumen	22	479	1.962	4,01 días	85%
1	Volumen medio	18	403	1.452	3,65 días	77%
2	Bajo volumen, alta complejidad	8	278	1.550	5,68 días	85%

Figura 2

Segmentación de Períodos Mensuales K-Means K=3



Nota. Segmentación de períodos mensuales según niveles de carga asistencial – K-Means K=3

El Clúster 0 (alto volumen) agrupa 22 períodos con los mayores volúmenes de ingresos y egresos y la mayor acumulación de días de estancia totales, concentrados principalmente entre marzo de 2022 y noviembre de 2024. Este perfil representa la operación de mayor escala de la

institución durante el período analizado, con una ocupación promedio del 85% y una estancia promedio de 4,01 días.

El Clúster 1 (volumen medio) reúne 18 períodos con valores intermedios en todas las variables, distribuidos principalmente en los meses de inicio y cierre de año a lo largo del período completo. Su ocupación promedio del 77% sugiere condiciones de menor presión asistencial relativa, típicas de los meses de transición entre temporadas de alta y baja demanda.

El Clúster 2 constituye el hallazgo más relevante del análisis. Agrupa exclusivamente los ocho primeros meses de 2021, período caracterizado por el menor volumen de ingresos y egresos del conjunto pero con la estancia promedio más alta (5,68 días, frente a 4,01 y 3,65 días de los otros grupos) y una ocupación equivalente a la del clúster de alto volumen (85%). Este patrón evidencia que durante ese período la institución atendía menos pacientes simultáneamente, pero cada hospitalización era de mayor duración, configurando un perfil operativo cualitativamente distinto que no puede interpretarse simplemente como "baja carga" sino como una combinación de bajo volumen con alta complejidad o mayor duración de los procesos de atención.

Valor Institucional del Modelo

La segmentación obtenida aporta una clasificación objetiva y reproducible de los períodos mensuales que revela dimensiones del comportamiento operativo no visibles en el análisis descriptivo convencional. En particular, la identificación del perfil de bajo volumen con alta complejidad en 2021 tiene implicaciones directas para la planificación de recursos: la presión sobre el sistema no depende únicamente del número de pacientes sino también de la duración y complejidad de las hospitalizaciones, variable que debe considerarse en la programación de personal y en la evaluación de la capacidad instalada. La posibilidad de

reclasificar automáticamente nuevos períodos a medida que se incorporen datos futuros dota al modelo de un valor sostenido más allá del período analizado en este estudio.

Predicción de Indicadores Operativos Mediante Regresión Lineal

Se implementaron dos modelos de regresión lineal múltiple con el propósito de evaluar la capacidad predictiva de las variables operativas mensuales sobre indicadores clave de hospitalización. El primer modelo estimó los días de estancia totales mensuales y el segundo el total de egresos mensuales. El análisis comparativo de sus resultados constituye uno de los hallazgos más relevantes del estudio, al revelar diferencias sustanciales en la predictibilidad de dos indicadores que operan sobre la misma base de datos.

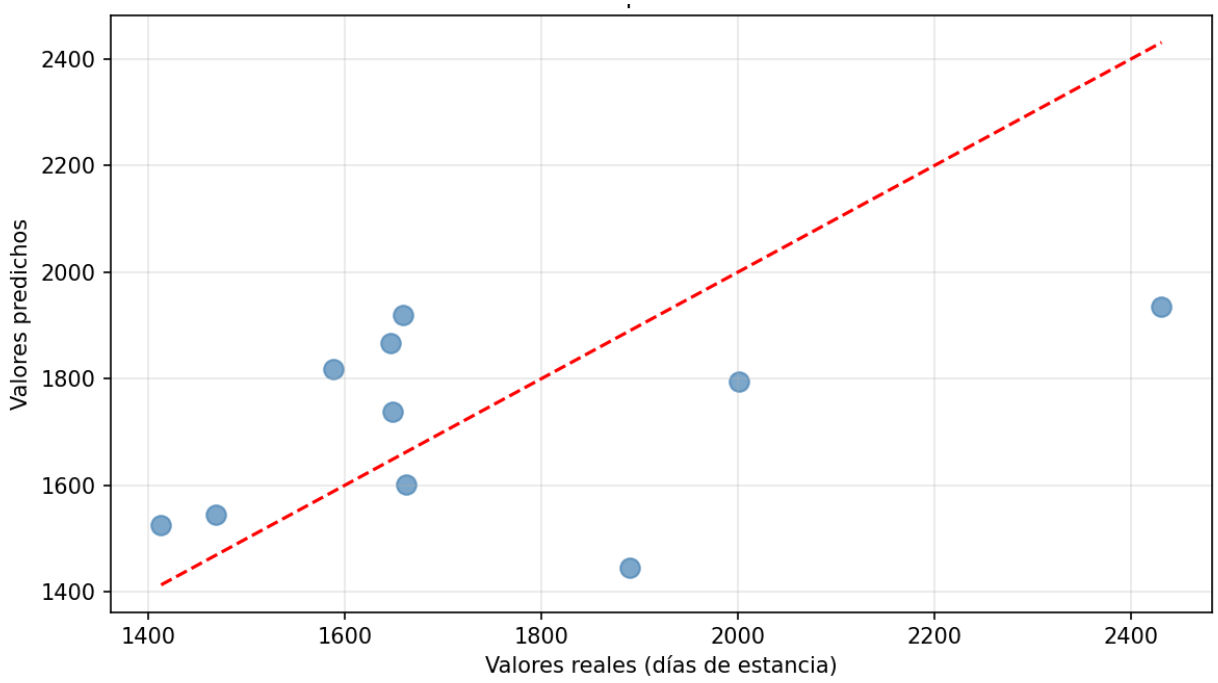
Modelo 1 Predicción de Días de Estancia Totales

El modelo de predicción de días de estancia totales utilizó como variables independientes el total de ingresos y el total de egresos mensuales, obteniendo un coeficiente de determinación R^2=0,143 y un error absoluto medio de 219,64 días, equivalente a un error relativo del 12,9% respecto al promedio mensual del indicador.

Este resultado indica que el volumen mensual de ingresos y egresos explica únicamente el 14,3% de la variabilidad de los días de estancia acumulados, lo que evidencia que este indicador depende en mayor medida de la duración individual de cada hospitalización que del número de pacientes atendidos. El gráfico de valores reales versus predichos (Figura 3) confirma esta limitación: los puntos presentan dispersión considerable respecto a la línea ideal, con desviaciones sistemáticas que el modelo no logra capturar.

Figura 3

Regresión Lineal Días de Estancia



Nota. Regresión lineal – Valores reales versus predichos para días de estancia totales

La interpretación operativa de este resultado es relevante: la acumulación mensual de días de estancia está determinada principalmente por la complejidad clínica de los casos atendidos y la duración de los procesos de alta médica, variables que no están disponibles en la base de indicadores agregados utilizada. Este hallazgo señala con precisión qué información adicional sería necesaria para construir un modelo predictivo confiable de este indicador, orientando el desarrollo de futuras fases del sistema analítico.

Modelo 2 Predicción de Egresos Mensuales

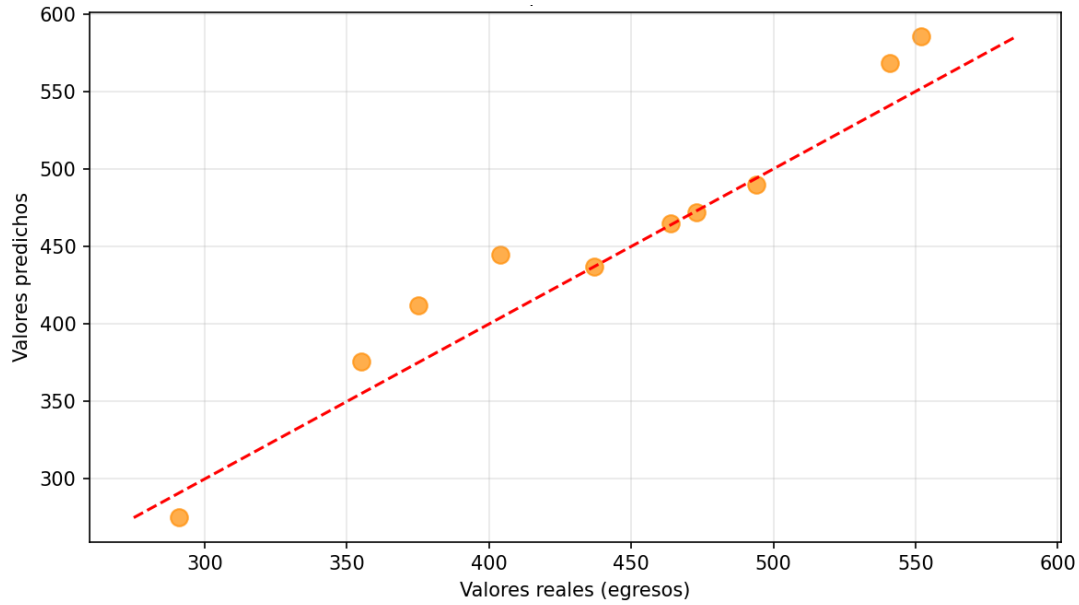
El modelo de predicción de egresos mensuales utilizó el total de ingresos como variable independiente, obteniendo un coeficiente de determinación R^2=0,911 y un error absoluto medio de 18,06 egresos, equivalente a un error relativo del 4,3% respecto al promedio mensual del

indicador. El coeficiente de regresión obtenido fue de 1,0597, lo que indica que por cada ingreso adicional en un mes se espera aproximadamente 1,06 egresos, relación coherente con la dinámica operativa de hospitalización donde los pacientes que ingresan en un período egresan en ese mismo período o en el inmediatamente siguiente.

El gráfico de valores reales versus predichos (Figura 4) muestra una distribución de puntos con alta proximidad a la línea diagonal ideal a lo largo de todo el rango de valores, confirmando un ajuste consistente y sin sesgos sistemáticos. Este resultado posiciona al modelo de egresos como una herramienta de planificación operativa de alta utilidad práctica, capaz de estimar el volumen mensual de salidas hospitalarias con una precisión del 95,7% a partir únicamente del registro de ingresos.

Figura 4

Regresión Lineal – Egresos Mensuales



Nota. Regresión lineal – Valores reales versus predichos para egresos mensuales

Análisis Comparativo y Valor Metodológico

El contraste entre R^2=0,143 para días de estancia y R^2=0,911 para egresos, sobre la misma base de datos y en el mismo período, constituye el hallazgo metodológico más relevante de esta fase del análisis. Demuestra que los indicadores operativos hospitalarios tienen naturalezas predictivas fundamentalmente distintas: los egresos responden de forma casi lineal al flujo de ingresos, mientras que los días de estancia acumulados están condicionados por la complejidad clínica individual de cada caso, dimensión que escapa al alcance de las variables operativas agregadas disponibles.

Este resultado tiene dos implicaciones prácticas directas. Primera, el modelo de egresos puede incorporarse de forma inmediata en los ciclos de planificación mensual de la institución, con un nivel de confianza alto y un margen de error operativamente aceptable. Segunda, la mejora del modelo de días de estancia requiere acceso a variables clínicas desagregadas, como el diagnóstico principal, el nivel de complejidad del caso o el servicio tratante, lo que establece una agenda concreta para el desarrollo de fases futuras del sistema analítico.

Clasificación de Niveles de Ocupación Mediante Árbol de Decisión

Como complemento a los modelos de segmentación y predicción desarrollados, se implementó un modelo supervisado de árbol de decisión con el objetivo de clasificar los períodos mensuales en los tres perfiles operativos identificados mediante K-Means: alto volumen, volumen medio y bajo volumen con alta complejidad. La variable objetivo fue construida directamente a partir de las etiquetas de agrupamiento obtenidas en la fase anterior, garantizando coherencia entre ambos modelos y permitiendo evaluar si las reglas generadas por el árbol son consistentes con los patrones identificados mediante el agrupamiento no supervisado.

Las variables explicativas utilizadas fueron el total de ingresos y el total de egresos mensuales, por ser las dimensiones que reflejan de forma más directa el flujo de pacientes en cada período. El conjunto de datos fue dividido en subconjuntos de entrenamiento (33 períodos, 70%) y prueba (15 períodos, 30%), aplicando estratificación para garantizar representación proporcional de cada clase en ambos conjuntos.

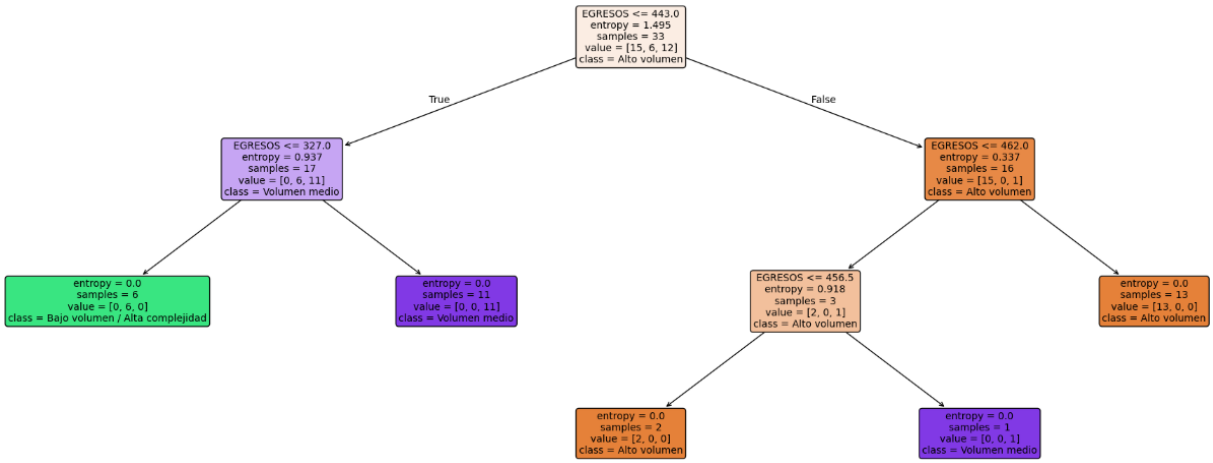
Estructura y Reglas del Modelo

Como se observa en la Figura 5, el árbol generó una estructura de clasificación con profundidad máxima de tres niveles, expresada en reglas condicionales directamente interpretables por los equipos de gestión. La variable con mayor capacidad discriminante en el nodo raíz fue el total de egresos mensuales, estableciendo un umbral de 443 egresos como primera división que separa los períodos de alto volumen de los restantes. Esta jerarquía es coherente con los resultados del modelo de regresión, donde los egresos demostraron ser el indicador con mayor predictibilidad y relación lineal con el flujo operativo mensual.

En la rama izquierda (egresos ≤ 443), un segundo umbral de 327 egresos permite distinguir los períodos de bajo volumen con alta complejidad (correspondientes a 2021) de los períodos de volumen medio. En la rama derecha (egresos > 443), un umbral adicional de 462 egresos refina la clasificación dentro del perfil de alto volumen, con un nodo de baja impureza que concentra 13 de los 15 períodos de ese perfil.

Figura 5

Árbol de Decisión



*Nota***.** Árbol de decisión para clasificación de niveles de ocupación hospitalaria

Evaluación del Desempeño

El modelo alcanzó una exactitud global (accuracy) de 93,3%, clasificando correctamente 14 de los 15 períodos del conjunto de prueba. El único error correspondió a un período de alto volumen clasificado como volumen medio, lo que representa una confusión entre categorías adyacentes operativamente próximas. El F1-Score ponderado fue de 0,9325, indicando un equilibrio sólido entre precisión y exhaustividad en las tres categorías.

Tabla 3

Desempeño del Árbol de Decisión por Categoría

Categoría	Precisión	Exhaustividad	F1-Score	Soporte
Alto volumen	0,88	1,00	0,93	7
Bajo volumen / Alta complejidad	1,00	1,00	1,00	2
Volumen medio	1,00	0,83	0,91	6
Promedio ponderado	0,94	0,93	0,93	15

El perfil de bajo volumen con alta complejidad fue clasificado con precisión y exhaustividad perfectas (F1=1,00), resultado especialmente relevante dado que este perfil representa el comportamiento operativo más diferenciado del conjunto y el de mayor valor diagnóstico para la gestión institucional.

Consideraciones sobre el Tamaño de la Muestra

Los resultados deben interpretarse considerando que el conjunto de prueba comprende únicamente 15 períodos, lo que limita la robustez estadística de las métricas obtenidas. No obstante, el alto desempeño del modelo es consistente con la clara separación entre perfiles identificada por K-Means (Silhouette=0,4804) y con la fuerte relación lineal entre ingresos y

egresos demostrada por el modelo de regresión (R^2=0,911). Esta coherencia entre los tres modelos refuerza la validez de los resultados y sugiere que las reglas generadas por el árbol capturan patrones genuinos del comportamiento operativo institucional.

Valor Institucional del Modelo

Las reglas de clasificación generadas por el árbol permiten identificar con transparencia el perfil operativo de cualquier período mensual a partir de dos variables fácilmente disponibles: el total de egresos supera o no los umbrales de 443 y 327 pacientes. Esta simplicidad es precisamente su mayor fortaleza en el contexto hospitalario: los equipos de gestión pueden aplicar estas reglas directamente en sus procesos de monitoreo mensual sin requerir herramientas estadísticas especializadas, facilitando la integración del modelo en los flujos de trabajo institucionales existentes.

Síntesis de los Resultados del Modelado

Los tres modelos de aprendizaje automático implementados no constituyen análisis independientes sino componentes articulados de una estrategia analítica diseñada para comprender el comportamiento de los indicadores hospitalarios desde perspectivas complementarias. Examinados en conjunto, sus resultados permiten extraer conclusiones que trascienden lo que cada modelo revela por separado, y que en varios casos resultaron contraintuitivas respecto a las hipótesis iniciales del análisis.

El hallazgo metodológico más relevante del proceso de modelado emerge del contraste entre los dos modelos de regresión. La diferencia entre R^2=0,911 para egresos y R^2=0,143 para días de estancia, sobre la misma base de datos y con variables operativas equivalentes, demuestra que estos dos indicadores tienen naturalezas predictivas fundamentalmente distintas. Los egresos mensuales responden de forma casi lineal al flujo de ingresos, con un coeficiente de 1,06 y un error relativo del 4,3%, lo que los convierte en un indicador altamente predecible y útil para la planificación operativa de corto plazo. Los días de estancia acumulados, en cambio, están condicionados principalmente por la complejidad clínica individual de cada hospitalización, dimensión que no es capturada por las variables operativas agregadas disponibles. Este resultado no representa una limitación del estudio sino un hallazgo con valor propio: delimita con precisión qué puede y qué no puede analizarse de forma confiable con la información actualmente disponible, orientando decisiones concretas sobre qué variables incorporar en desarrollos futuros.

El modelo K-Means aportó la dimensión exploratoria más valiosa del conjunto, identificando tres perfiles operativos diferenciados con un índice Silhouette de 0,4804. El hallazgo más relevante de esta fase fue la identificación del perfil de bajo volumen con alta

complejidad, correspondiente exclusivamente a los primeros ocho meses de 2021, caracterizado por el menor número de pacientes pero la estancia promedio más alta del período analizado (5,68 días). Este patrón no era visible en el análisis descriptivo convencional y tiene implicaciones directas para la planificación de recursos: la presión sobre el sistema hospitalario no depende únicamente del volumen de pacientes sino también de la duración y complejidad de las hospitalizaciones, variable que debe considerarse de forma independiente en la programación de personal y capacidad instalada.

El árbol de decisión cerró el ciclo analítico con el desempeño más sólido del conjunto, alcanzando una accuracy del 93,3% y un F1-Score ponderado de 0,9325 en la clasificación de los tres perfiles operativos. Su estructura de reglas basada en umbrales de egresos (443 y 327 pacientes mensuales) es directamente aplicable por los equipos de gestión sin mediación técnica, y su coherencia con los resultados de K-Means y regresión refuerza la validez del enfoque analítico en su conjunto.

Tabla 4

Resumen Comparativo de los Modelos de Aprendizaje Automático

Modelo	Objetivo	Métrica principal	Resultado	Valor institucional
K-Means (K=3)	Segmentación de períodos	Silhouette	0,4804	Identificación de 3 perfiles operativos diferenciados
Regresión lineal – Egresos	Predicción de egresos	R^2 / MAE	0,911 / 18 egresos (4,3%)	Herramienta de planificación mensual de alta precisión

Modelo	Objetivo	Métrica principal	Resultado	Valor institucional
Regresión lineal – Días estancia	Predicción días de estancia	R^2 / MAE	0,143 / 220 días (12,9%)	Delimita necesidad de variables clínicas adicionales
Árbol de decisión	Clasificación operativa	Accuracy / F1	93,3% / 0,933	Reglas interpretables de clasificación mensual

En conjunto, estos resultados demuestran que el análisis basado en variables operativas agregadas de periodicidad mensual tiene un alcance definido pero genuinamente útil: es suficiente para identificar perfiles de comportamiento, predecir con alta precisión el volumen de egresos y generar reglas interpretables de clasificación operativa. Al mismo tiempo, los resultados señalan con precisión los límites de este alcance, identificando los indicadores y las dimensiones del problema que requieren fuentes de información más ricas para ser abordados con mayor profundidad analítica. Esta combinación de capacidades demostradas y limitaciones explícitas constituye, en sí misma, una contribución valiosa para la institución y para el campo de la analítica aplicada a la gestión hospitalaria.

Conclusiones

El presente estudio demostró que es posible transformar un proceso institucional dependiente de consolidación manual en un sistema automatizado, reproducible y orientado al análisis, a partir de fuentes de datos operativos disponibles en la institución. Esta transformación no es únicamente técnica: implica un cambio en la forma en que la información hospitalaria se gestiona y se pone al servicio de la toma de decisiones, pasando de registros dispersos a una estructura analítica coherente con la periodicidad y los objetivos institucionales de seguimiento.

La automatización del procesamiento constituyó el fundamento sobre el cual fue posible aplicar técnicas de aprendizaje automático con resultados interpretables y operativamente relevantes. Un hallazgo importante de esta fase fue la identificación de inconsistencias en las fuentes de datos originales, incluyendo problemas de formato en la columna de días de estancia y la presencia de registros de servicios no hospitalarios en la base de egresos, que habrían comprometido la validez de cualquier análisis posterior de no haber sido detectados y corregidos. Este resultado confirma lo señalado por Santos y Costa (2020) respecto a que la automatización del procesamiento, con sus procedimientos de validación y corrección incorporados, es condición necesaria para garantizar la confiabilidad de los análisis en entornos hospitalarios.

En cuanto al modelado, los resultados evidencian que los indicadores operativos hospitalarios no son igualmente predecibles a partir de variables agregadas mensuales, y que esta diferencia tiene valor explicativo propio. Los egresos mensuales demostraron una relación casi lineal con los ingresos, con un R^2=0,911 y un error relativo del 4,3%, convirtiéndose en el indicador más predecible del conjunto y en una herramienta concreta de planificación operativa de corto plazo. Los días de estancia acumulados, en cambio, presentaron un R^2=0,143, evidenciando que su comportamiento está determinado principalmente por la complejidad clínica

individual de cada hospitalización y no por el volumen de pacientes. Este contraste no representa un éxito y un fracaso, sino la demostración de que distintos indicadores requieren distintas fuentes de información para ser modelados de forma confiable.

La segmentación mediante K-Means identificó tres perfiles operativos diferenciados con un índice Silhouette de 0,4804, entre los cuales el hallazgo más relevante fue la caracterización del período enero-agosto 2021 como un perfil de bajo volumen con alta complejidad, con una estancia promedio de 5,68 días frente a 4,01 y 3,65 días de los otros perfiles. Este patrón, no visible en el análisis descriptivo convencional, evidencia que la presión sobre el sistema hospitalario no depende únicamente del número de pacientes sino también de la duración y complejidad de las hospitalizaciones, dimensión que debe considerarse de forma independiente en la planificación de recursos.

El árbol de decisión alcanzó una accuracy del 93,3% y un F1-Score ponderado de 0,9325, clasificando correctamente 14 de los 15 períodos del conjunto de prueba mediante reglas basadas en umbrales de egresos directamente interpretables por los equipos de gestión. La coherencia entre los resultados de los tres modelos, donde los egresos emergen como el indicador operativo más predecible y discriminante en todos los enfoques, refuerza la validez del enfoque analítico en su conjunto.

Desde una perspectiva más amplia, este trabajo contribuye a demostrar la viabilidad de implementar soluciones de analítica de datos en instituciones de salud de mediana complejidad, utilizando exclusivamente información operativa disponible en los sistemas de registro institucional, sin requerir infraestructura tecnológica adicional ni acceso a datos clínicos sensibles. Esta condición reduce las barreras de adopción y hace replicable el enfoque en otras instituciones con características similares.

El estudio tiene limitaciones que deben reconocerse con claridad. El volumen de 48 períodos mensuales, aunque suficiente para un análisis exploratorio, restringe la robustez estadística de los modelos supervisados y limita su capacidad de generalización. La ausencia de variables clínicas desagregadas impide modelar los días de estancia con mayor precisión. Y el sistema desarrollado, aunque funcional y documentado, no fue integrado formalmente en los procesos institucionales durante el alcance de este estudio, lo que constituye el paso siguiente necesario para validar su utilidad en condiciones reales de operación.

Con base en estos resultados y limitaciones, se identifican tres líneas prioritarias para el desarrollo futuro. En primer lugar, la incorporación de variables clínicas como el diagnóstico principal, el nivel de complejidad del caso y el servicio tratante, que permitirían mejorar sustancialmente la capacidad predictiva del modelo de días de estancia. En segundo lugar, la integración formal del sistema en las plataformas de Business Intelligence institucionales, que permitiría que los resultados del análisis sean consultados de forma continua por los equipos de gestión. En tercer lugar, la ampliación del período de análisis a medida que se acumulen nuevos datos, lo que fortalecería la robustez de los modelos y permitiría evaluar su estabilidad en el tiempo. En conjunto, este trabajo evidencia que la ciencia de datos aplicada a la gestión hospitalaria no requiere grandes volúmenes de datos ni modelos de alta complejidad para generar valor institucional. Requiere, en cambio, rigor metodológico en el procesamiento de los datos, claridad sobre el alcance del análisis y orientación consistente hacia las necesidades reales de quienes toman decisiones en la institución.

Recomendaciones

Las recomendaciones se organizan en tres niveles según su horizonte de implementación.

Implementación Inmediata

La primera acción prioritaria es la institucionalización del pipeline desarrollado en Python como herramienta formal de consolidación mensual de indicadores. Su adopción en los procesos regulares de la institución permitirá eliminar el reproceso manual, garantizar la consistencia de los datos y liberar tiempo operativo para el análisis estratégico. Para que esta adopción sea sostenible, se recomienda designar un responsable técnico de su mantenimiento y establecer un protocolo documentado de ejecución mensual que incluya los procedimientos de validación y corrección de formato incorporados en el sistema.

Se recomienda conectar los archivos de salida generados por el sistema con una plataforma de Business Intelligence, preferiblemente Power BI, para que los indicadores consolidados y los resultados de los modelos sean consultados de forma continua por los equipos de gestión a través de tableros interactivos, sin requerir intervención técnica en cada ciclo de análisis.

El modelo de regresión lineal para egresos, con R^2=0,911 y un error relativo del 4,3%, está en condiciones de ser utilizado de inmediato como herramienta de planificación mensual. Por cada ingreso registrado en el período, el modelo estima aproximadamente 1,06 egresos, relación que puede incorporarse directamente en los ciclos de programación de recursos, asignación de personal y gestión de la capacidad instalada.

La segmentación en tres perfiles operativos y las reglas del árbol de decisión pueden utilizarse desde ya para clasificar cada nuevo período mensual según su nivel de carga, identificar meses históricamente críticos y diseñar estrategias de contingencia diferenciadas

según el perfil esperado. Los umbrales de 443 y 327 egresos mensuales constituyen referencias operativas concretas y verificables.

Ajustes de Mediano Plazo

El bajo poder explicativo del modelo de días de estancia (R^2=0,143) señala con precisión la necesidad de enriquecer las fuentes de datos. Se recomienda incorporar variables clínicas como el diagnóstico principal, el nivel de complejidad del caso y el tipo de servicio, así como variables administrativas relacionadas con los tiempos de gestión del alta médica. Esta ampliación requiere coordinación entre los equipos de sistemas, estadística y gestión clínica, y constituye la condición más importante para mejorar la capacidad predictiva del sistema en una segunda fase.

Se recomienda establecer un protocolo de recalibración periódica de los modelos con frecuencia mínima semestral, incorporando los nuevos datos disponibles y evaluando si el desempeño se mantiene estable. El hallazgo de que 2021 constituye un perfil operativo diferenciado sugiere que el comportamiento institucional puede cambiar de forma significativa en el tiempo, y la vigencia de los modelos depende de su actualización sistemática.

Líneas de Desarrollo Futuro

A medida que se amplíe el volumen de datos y se enriquezcan las fuentes de información, se recomienda explorar modelos de series temporales como ARIMA o Prophet para la predicción de indicadores con componentes estacionales, y modelos de gradient boosting para mejorar la clasificación operativa manteniendo un nivel razonable de interpretabilidad. La exploración de estas alternativas debe realizarse de forma comparativa, evaluando no solo el desempeño estadístico sino también la viabilidad de adopción institucional de cada enfoque.

Finalmente, se recomienda explorar la integración del sistema con otras fuentes de información institucional, como los registros de urgencias, consulta externa y cirugía, que permitirían construir una visión analítica más completa del funcionamiento de la institución y ampliar el alcance de los modelos hacia una gestión hospitalaria verdaderamente basada en datos.

Referencias Bibliográficas

Berg, M. (2003). Health information management: Integrating information technology in health care work. Routledge.

Castellanos, M., Sánchez, E., Ríos, J., Barrera, Á., Erazo, L., Marroquín, E., . . . Suárez, M. (2020). Instructivo para la usabilidad de Normas APA. Universidad Nacional Abierta y a Distancia UNAD.

Correa Restrepo, J. S. & Murillo Ospina, J. H. (2015). *Escritura e investigación académica: una guía para la elaboración del trabajo de grado:* (1 ed.). Colegio de Estudios Superiores de Administración - CESA. https://elibro-net.bibliotecavirtual.unad.edu.co/es/ereader/unad/222461?page=48

Distancia, U. N. (2013). Opciones de grado especialización. Obtenido de https://estudios.unad.edu.co/opcion-de-grado-especializacion

Función pública. (2006). Obtenido de DECRETO 3851 DE 2006: https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=22088

García, a RJL (2003). *Cómo elaborar un proyecto de investigación (3ª edición)* . Digitalia.

Ministerio de salud y protección social. (2016). Obtenido de Resoluciòn 256 de 2016: https://www.minsalud.gov.co/Normatividad_Nuevo/Resoluci%C3%B3n%200256%20de%202016.pdf

Peset, F., & Millan, L. (2017). Ciencia abierta y gestión de datos de investigación (RDM). Ediciones trea.

Pozzo, M. I. (2020). *Escritura de tesis de posgrado: desde el proyecto hasta la defensa:* (ed.). Editorial Biblos.

Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. New England Journal of Medicine, 380(14), 1347-1358.

Reglamento Estudiantil Universidad Nacional Abierta y a Distancia UNAD. (2013). UNAD.

Ríos Insua, D., & Gómez-Ullate Oteiza, D. (2019). Big data: Conceptos, tecnologías y aplicaciones. Consejo Superior de Investigaciones Científicas (CSIC).

Sampieri, R. H., & Mendoza Torres, C. P. (2018). Metodología de la Investigación Las Rutas Cuantitativa, Cualitativa y Mixta. McGraw-Hill Interamericana.

Sedkaoui, S. (2018). How data analytics and big data can help to handle the challenges of healthcare systems. En Big data analytics for entrepreneurial success (pp. 1-27).