Escuela de Administración

Escuela de Ingeniería Ciencia y Tecnología

Maestría en Business Analytics

Solución basada en datos para el fortalecimiento del proceso de quimioterapia en la ruta Oncológica del Hospital Universitario Méderi

Presentado por:

Alejandra Hernández Rodríguez, Juan Camilo Parra Torres y Paula Andrea Guiot Martínez

Bogotá, D.C. ,25 de mayo 2025

Escuela de Administración

Escuela de Ingeniería Ciencia y Tecnología

Maestría en Business Analytics

Solución basada en datos para el fortalecimiento del proceso de quimioterapia en la ruta Oncológica del Hospital Universitario Méderi

Presentado por:

Alejandra Hernández Rodríguez, Juan Camilo Parra Torres y Paula Andrea Guiot Martínez

Bajo la dirección de:

Erika Salazar

Bogotá, D.C. ,25 de mayo 2025

Contenido

Declaración de originalidad y autonomía

Declaramos bajo la gravedad del juramento, que hemos escrito el presente Reto Empresarial, en la modalidad de proyecto de emprendimiento por nuestra propia cuenta y que, por lo tanto, su contenido es original.

Declaramos que hemos indicado clara y precisamente todas las fuentes directas e indirectas de información y que este Reto Estratégico no ha sido entregado a ninguna otra institución con fines de calificación o publicación.

Alejandra Hernández Rodríguez

Paula Andrea Guiot Martínez

Juan Camilo Parra Torres

Firmado en Bogotá, D.C. el 25 de mayo de 2025

Declaración de exoneración de responsabilidad

Declaramos que la responsabilidad intelectual del presente trabajo es exclusivamente de sus autores. La Universidad del Rosario no se hace responsable de contenidos, opiniones o ideologías expresadas total o parcialmente en él.

Alejandra Hernández Rodríguez

Paula Andrea Guiot Martínez

Juan Camilo Parra Torres

Firmado en Bogotá, D.C. el 25 de mayo de 2025

Lista de figuras

Índice de tablas

Glosario

Administración de medicamentos: “Secuencia de actividades mediante las cuales un fármaco, es proporcionado por el personal de salud idóneo (enfermera o auxiliar de enfermería) a las personas, por diferentes vías de administración, según indicación médica escrita, dicha actividad queda debidamente registrada en el formato de notas de enfermería” (Hospital Universitario Méderi, 2022)

Ciclo de quimioterapia: “Hace referencia al número de veces en un periodo determinado durante el cual el paciente recibirá un protocolo de quimioterapia (8 días, 15 días, 21 o 28 días)” (Hospital Universitario Méderi, 2022)

Medicamento Citotóxico: “Medicamento utilizado para la destrucción de las células tumorales” (Hospital Universitario Méderi, 2022)

Preparación (como acción): “Toda operación que permite alistar un medicamento a las necesidades específicas de un paciente y/o adaptarlo para su administración o utilización. Por ejemplo, personalizar las dosis o reconstituir un medicamento para que esté listo para su administración. También se incluyen las operaciones asociadas a estas actividades, como la adquisición de los materiales de partida, los controles de calidad, la aprobación de la preparación final y su almacenamiento” (Hospital Universitario Méderi, 2022)

Protocolo de quimioterapia: “Es el nombre que se designa a la combinación de medicamentos citotóxicos que hacen parte del tratamiento específico para las diferentes neoplasias sólidas o hemato-linfoides en un determinado periodo de tiempo” (Hospital Universitario Méderi, 2022)

Quimioterapia: “Tratamiento médico, basado en la administración de sustancias químicas (fármacos), de manera conjunta, con el fin de producir sinergismo de las sustancias y efectos de

destrucción y no reproducción de células neoplásicas en el organismo del paciente" (Hospital Universitario Méderi, 2022)

Sala de quimioterapia: "Área intra-hospitalaria, que cuenta con la adecuada estructura, equipos médicos, medicamentos y personal entrenado para la atención y manejo de pacientes que requieren la administración de quimioterapia ambulatoria" (Hospital Universitario Méderi, 2022).

Giro silla: Indicador que mide la ocupación de pacientes en las salas de quimioterapia.

Metodología CRISP-DM: Metodología utilizada para el desarrollo de proyectos analíticos, la cual cuenta con 6 fases: comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación e implementación.

HUM: "Méderi es un Hospital Universitario que tiene convenio con la Universidad del Rosario y otras instituciones universitarias del país, por lo que basa su actividad en el rigor científico y en la gestión y producción del conocimiento" (Méderi)

ACHC: "La Asociación Colombiana de Hospitales y Clínicas es una entidad sin ánimo de lucro de carácter gremial, con Personería Jurídica reconocida por el Ministerio de Justicia mediante la Resolución No.1258 del 29 de mayo de 1956. La Junta Directiva Nacional de la Asociación está compuesta por 9 Miembros principales y 9 suplentes procedentes de diferentes zonas del país, elegidos por la Asamblea General conformada por todos los afiliados a la entidad." (Asociación Colombiana de Hospitales y Clínicas, 2025)

Índice de Anexos

Resumen Ejecutivo

Solución basada en datos para el fortalecimiento del proceso de quimioterapia en la ruta Oncológica del Hospital Universitario Méderi

En 2022, el Fondo Colombiano de Enfermedades de Alto Costo reportó 509,727 casos de cáncer, de los cuales el 62% corresponden a mujeres y el 38% a hombres, con una mediana de edad de 63 años. Según el Instituto Nacional de Cancerología, los servicios oncológicos en Colombia han experimentado un crecimiento notable, siendo la consulta externa el servicio más ofrecido. Según Bibiana Meneses, coordinadora del servicio de oncología en comunicación personal el 18 de noviembre del 2024, El Hospital Universitario Méderi, sujeto de este proyecto empresarial, también registró un incremento del 14.5% en sus servicios oncológicos entre 2020 y 2021. Para 2023, la ruta oncológica en Méderi se consolidó como una de las principales fuentes de ingreso para el hospital, con la quimioterapia como eje central en su impacto financiero. Este proyecto tiene como objetivo desarrollar una solución analítica, basada en la metodología CRISP-DM, para mejorar el desempeño del servicio de quimioterapia, el cual actualmente carece de una herramienta que permita gestionar de manera eficiente el uso de las salas de administración de medicamentos. Como parte de la solución, se diseñó un tablero interactivo que describe las características de los pacientes atendidos con quimioterapia, junto con un modelo predictivo de inasistencia y modelo prescriptivo de programación de citas en los turnos. El propósito es proporcionar al hospital herramientas que faciliten la toma de decisiones orientadas a mejorar el indicador de giro de silla. Mediante la implementación de modelos predictivos, como XGBoost y técnicas de balanceo de clases, se estimó la inasistencia de los pacientes a sus citas, alcanzando un desempeño del 69 % en la predicción de ausencias. Adicionalmente, se desarrolló un modelo prescriptivo, basado en

programación lineal entera mixta, que permite asignar de manera óptima a los pacientes en las unidades disponibles (sillas o camas), maximizando el uso del recurso físico.

Los resultados evidencian que variables como el ciclo de tratamiento, el turno asignado y la programación en fines de semana o días festivos tienen una influencia significativa en la asistencia de los pacientes. El proyecto sienta las bases para una futura implementación en tiempo real por parte del hospital, lo cual permitiría ajustar los modelos, incorporar técnicas de ensamblado y evaluar su desempeño con el apoyo de expertos clínicos.

Palabras clave

Indicador Giro Silla, Oncología, quimioterapia, CRISP-DM, solución analítica, optimización, XGBoost

Abstract

Data-Driven Solution for Strengthening the Chemotherapy Process in the Oncology Pathway at Méderi University Hospital

In 2022, the Colombian High-Cost Diseases Fund reported 509,727 cancer cases, of which 62% were women and 38% men, with a median age of 63 years. According to the National Cancer Institute, oncology services in Colombia have shown significant growth, with outpatient consultation being the most offered service. According to Bibiana Meneses (personal communication, November 18, 2024), Méderi University Hospital—subject of this business project—also recorded a 14.5% increase in its oncology services between 2020 and 2021. By 2023, the oncological care pathway at Méderi accounted for 45% of hospital revenue, with chemotherapy treatments contributing 42% of that income. This project aims to develop an analytical solution, based on the CRISP-DM methodology, to optimize the chemotherapy service, which currently lacks a tool for efficiently managing the use of drug administration rooms. As part of the solution, an interactive dashboard was designed to describe the characteristics of patients receiving chemotherapy, along with a predictive model for appointment absenteeism. The goal is to provide the hospital with tools that support data-driven decision-making to improve the "chair turnover" indicator. By implementing predictive models such as XGBoost and class balancing techniques, the probability of patient absenteeism was estimated, achieving 69% performance in predicting missed appointments. Additionally, a prescriptive model based on mixed-integer linear programming was developed to optimally assign patients to available units (chairs or beds), maximizing the use of physical resources.

The results show that variables such as treatment cycle, assigned shift, and scheduling on weekends or holidays significantly influence patient attendance. This project lays the groundwork for future real-time implementation by the hospital, which would allow for model refinement, integration of ensemble methods, and performance evaluation with the support of clinical experts.

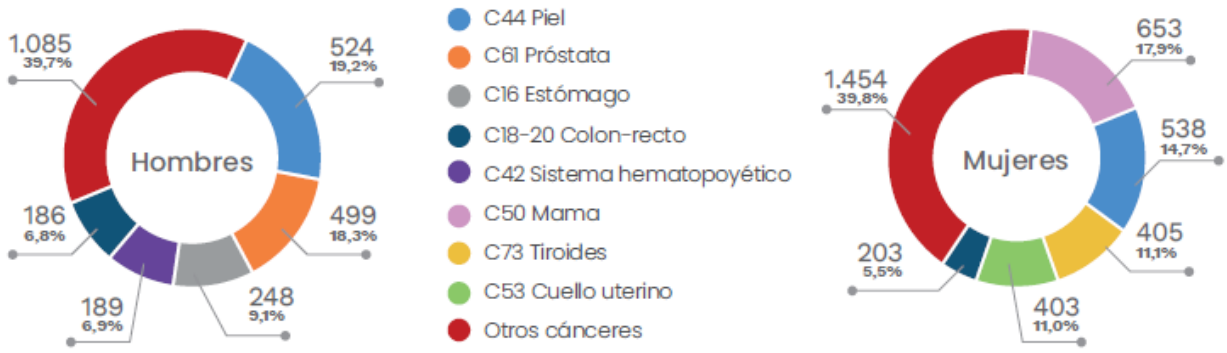
Keywords

Chair Turnover Indicator, Oncology, Chemotherapy, CRISP-DM, Analytical Solution, Optimization, XGBoost

1. Introducción

El Fondo Colombiano de Enfermedades de Alto Costo registró 509,727 casos de pacientes con cáncer (46,870 casos nuevos y 462,857 casos prevalentes), de los cuales el 62% son mujeres y el 38% hombres, con una mediana de edad de 63 años (Fondo Colombiano de Enfermedades de Alto Costo, 2022) .De acuerdo con el Instituto Nacional de Cancerología (Instituto Nacional de Cancerología,2023) los tipos de cáncer más frecuentes de nuevos casos en hombres son: piel(19,2%), próstata (18,3%) y estómago (9,1%); y en mujeres: mama (17,9%), piel(14,7%) y tiroides (11,1%), como se observa en la Figura 1.

Figura 1. *Distribución porcentual de nuevos casos reportados de acuerdo con tipología de cáncer y sexo*



Fuente: (Instituto Nacional de Cancerología,2023)

Además, según el Instituto Nacional de Cancerología (Instituto Nacional de Cancerología, 2022, p. 26), los servicios prestados a pacientes oncológicos han aumentado en Colombia a lo largo de los años, siendo la consulta externa el servicio con mayor oferta, como se muestra en Figura 2

Figura 2. *Oferta de servicios oncológicos durante el tiempo*

Fuente: (Instituto Nacional de Cancerología, 2022, p. 26)

En línea con lo anterior, Hospital Universitario Méderi (HUM), el lugar de estudio de este proyecto también ha experimentado un crecimiento en las consultas oncológicas, con un aumento del 14.5% entre 2020 y 2021 (Hospital Universitario Mederi, 2021, p. 47) en diferentes tipos de servicios prestados, como se observa en la figura 3 en 2021, el hospital atendió a 9,072 pacientes oncológicos (tanto ambulatorios como hospitalizados).

Figura 3.*Crecimiento de servicios oncología Mederi*

Fuente: (Hospital Universitario Mederi, 2021, p. 47)

Este proyecto se enfocó en fortalecer la ruta oncológica del HUM, específicamente, en los tratamientos de quimioterapia, debido a que esta unidad es crítica para el hospital a nivel de costos y facturación. Actualmente, la unidad oncológica del HUM cuenta con un indicador llamado giro silla. Este es un indicador mensual, calculado como parte del Sistema de Gestión de Calidad del hospital, que consiste en totalizar el número de pacientes agendados sobre la disponibilidad total de 31 ubicaciones. Durante los últimos tres meses del año, se alcanzó un cumplimiento del 67%, con una tasa de inasistencia del 19%.

Este indicador busca medir la eficiencia en el uso de los recursos: a mayor porcentaje, mayor rotación de pacientes a través de las ubicaciones disponibles, lo que refleja un mejor aprovechamiento de la capacidad instalada. Esto se traduce en una atención más oportuna para un mayor número de pacientes y en un incremento en la facturación del área de oncología.

En este contexto, el proyecto tiene como objetivo crear una herramienta analítica que permita a la unidad oncológica tomar decisiones basadas en datos respecto a la programación de pacientes a las sesiones de quimioterapia, mediante el desarrollo de una solución construida bajo la metodología CRISP-DM orientada a mejorar el desempeño del servicio.

Como resultado de esta investigación, se quiere obtener una herramienta analítica con el fin de que la unidad oncológica pueda tomar decisiones basadas en datos respecto a la programación de pacientes a su quimioterapia.

Para la ejecución de esta metodología, es importante conocer el negocio, realizar la limpieza de la data de acuerdo con los puntos recogidos en el entendimiento de negocio, creación y evaluación de cada una de las herramientas propuestas, el modelo predictivo para determinar la asistencia de los pacientes, el modelo prescriptivo para su correcta ubicación en sala. Adicional a esto, un dashboard que permite visualizar información relevante sobre los

pacientes. Por último, los resultados junto con las conclusiones del proyecto y los próximos pasos sugeridos para el hospital.

2. Objetivos

2.1. Objetivo general

Diseñar una solución analítica basada en datos, siguiendo la metodología CRISP-DM, que apoye la toma de decisiones en el servicio de quimioterapia mediante la identificación de oportunidades de mejora en el uso de sillas y camas en sala, así como en la programación de los pacientes.

2.2. Objetivos específicos

- Depurar, transformar y estructurar los datos proporcionados por el área oncológica, aplicando las reglas del negocio y los principios de calidad de datos definidos por la metodología DAMA, garantizando su integridad, coherencia y utilidad para análisis posteriores.

- Diseñar un tablero de visualización interactivo que permita al área oncológica explorar características relevantes de los pacientes y monitorear indicadores clave relacionados con la asistencia a las sesiones de quimioterapia, apoyando tanto el análisis descriptivo como la toma de decisiones informadas.

- Desarrollar un modelo predictivo supervisado que estime la inasistencia de los pacientes a las sesiones de quimioterapia, utilizando técnicas de aprendizaje automático y validación estadística para garantizar la precisión y la capacidad de generalización del modelo empleando algoritmos de machine learning y métodos de validación estadística para asegurar su precisión

- Diseñar un modelo prescriptivo que optimice el proceso de asignación de pacientes a las unidades disponibles en sala, con el objetivo de maximizar el uso eficiente de sillas y camas y contribuir a la mejora del indicador de ocupación.

3. Alcance

El presente proyecto tiene como propósito diseñar una solución analítica basada en datos que optimice la programación de tratamientos de quimioterapia en la ruta oncológica del Hospital Universitario Méderi, con el propósito de optimizar el uso de recursos físicos en sala (indicador de giro silla). La solución fue estructurada bajo el enfoque de la metodología CRISP-DM, cubriendo desde el entendimiento del negocio hasta el modelado y evaluación de herramientas predictivas y prescriptivas.

En alineación con los objetivos específicos, el alcance del proyecto contempla las siguientes acciones:

Depuración, transformación y estructuración de datos: se realiza la integración y limpieza de las bases de datos de programación, asistencia y características demográficas de los pacientes oncológicos, aplicando principios de calidad de datos (validez, precisión, completitud y unicidad) definidos por la metodología DAMA. Esta fase busca garantizar la integridad de la información para su posterior análisis y modelado.
Diseño de un tablero de visualización interactivo: se desarrolla una herramienta de business intelligence (Power BI) que permita a los usuarios explorar variables relevantes como edad, EPS, protocolo, turno y estado de asistencia, y que facilite el monitoreo continuo de indicadores clave asociados al servicio de quimioterapia.

Desarrollo de un modelo predictivo de inasistencia: se construye un modelo de aprendizaje supervisado que estime la probabilidad de que un paciente no asista a su cita programada. El modelo es validado estadísticamente para asegurar su precisión, sensibilidad y capacidad de generalización, cuyo propósito es que el hospital pueda generar una lista de espera, en la cual se pueda tener un listado de pacientes, con la posibilidad de asistir en caso tal que se confirme la inasistencia de un paciente, esto generaría una mayor atención de pacientes logrando así aumentar el indicador de gestión del hospital.

Diseño de un modelo prescriptivo de asignación óptima: utilizando los resultados del modelo predictivo, se diseñó una herramienta que sugiera asignaciones eficientes de pacientes a turnos y sillas disponibles, con el fin de maximizar el uso de los recursos físicos y mejorar el indicador de giro silla.

Entrega de documentación técnica y operativa: como parte de los entregables, se incluirá un conjunto de manuales de usuario para cada una de las soluciones desarrolladas (dashboard, modelo predictivo y modelo prescriptivo), diseñados para asegurar su comprensión y correcta utilización por el equipo hospitalario. Asimismo, se presentan recomendaciones concretas sobre los pasos siguientes que el hospital podría considerar para continuar con el fortalecimiento del proceso, como el ajuste del sistema de captura de datos, la integración con fuentes en tiempo real, y la gobernanza de datos para asegurar la sostenibilidad de los modelos.

Los principales usuarios de esta solución son la coordinación del área de oncología, los jefes de programación y el personal clínico, responsables de la planificación operativa del servicio. La frecuencia de uso de las herramientas dependerá de la periodicidad con la que el hospital actualice sus datos, estimándose su uso diario o semanal en el caso de la programación de pacientes.

Este proyecto no contempla la implementación operativa de los modelos ni su integración en tiempo real con los sistemas del hospital. No obstante, se entregaron los insumos necesarios —incluyendo modelos analíticos, dashboards funcionales, manuales de usuario y recomendaciones de continuidad— que permiten una futura la implementación por parte del equipo de tecnología y datos del hospital.

Como riesgos potenciales se identifican: la falta de estandarización en variables clave (como protocolos, duración y turnos), posibles vacíos en las bases de datos demográficas, y la dependencia del conocimiento experto para completar o interpretar ciertas variables. Además, la viabilidad de una futura implementación dependerá del fortalecimiento de la gobernanza de datos y de las capacidades tecnológicas de la institución.

4. Cronograma

Se diseñó e implementó una solución analítica para mejorar la programación de pacientes en quimioterapia en el Hospital Universitario Méderi, aplicando la metodología CRISP-DM y un enfoque ágil. El proyecto avanzó desde la comprensión del proceso hasta la entrega de modelos predictivo y prescriptivo, herramientas visuales y documentación final.

Primer semestre 2024 (febrero – junio)

Se inició el proyecto con reuniones de levantamiento de información, análisis del proceso y diseño preliminar de la solución analítica.

- Febrero: Inicio formal del proyecto y reuniones con oncología.
- Marzo: Visitas al hospital y observación directa de procesos.
- Abril: Identificación de variables clave e indicadores prioritarios.

- Mayo: Validación con oncología y cierre de la fase de negocio.
- Junio: Diseño preliminar del tablero y definición del ecosistema analítico.

Segundo semestre 2024 (septiembre – diciembre)

Se revisaron y prepararon los datos, se construyó el dataset y se desarrollaron los primeros modelos predictivos.

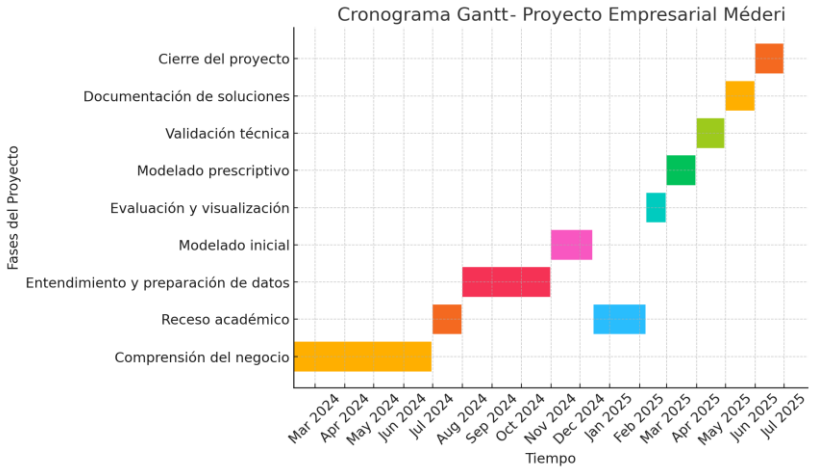
- Julio - Agosto: Receso académico, avance en visuales preliminares.
- Septiembre: Revisión de bases y análisis de calidad de datos.
- Octubre: Limpieza, transformación y consolidación en el dataset. Noviembre: Entrenamiento del modelo predictivo de inasistencia.
- Diciembre (hasta el 14): Ajustes al modelo y retroalimentación clínica.

Primer semestre 2025 (febrero – junio)

Se completaron los ajustes finales, se desarrolló el modelo prescriptivo y se realizó la entrega oficial del proyecto.

- **Enero - 7 febrero:** Receso académico y planificación de entregables.
- **Febrero:** Ajuste del tablero visual y validación con Méderi.
- **Marzo:** Desarrollo del modelo prescriptivo de asignación de turnos.
- **Abril:** Validación conjunta de modelos con simulaciones.
- **Mayo:** Elaboración de manuales de usuario y documentación técnica.
- **Junio:** Presentación final y cierre del proyecto.

Figura 4. Cronograma.



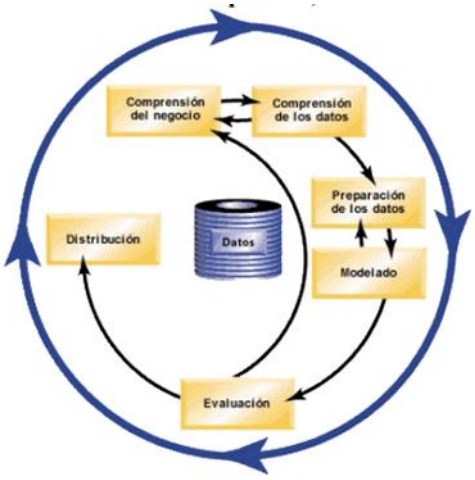
Fuente: Elaboración propia

5. Metodología

Para el cumplimiento de los objetivos propuestos en el proyecto fue necesario acceder a fuentes internas de información del Hospital que facilitaron la validación de hipótesis y la definición de criterios clave, específicamente sobre el histórico de asignación de pacientes en sala de quimioterapia, así como una metodología basada en minería de datos y lograr un mejor indicador giro silla, para ello se hizo uso de una combinación de CRISP-DM y metodología ágil SCRUM.

CRISP-DM sigla de Cross-Industry Standard Process for Data Mining es una **metodología** "incluye descripciones de las fases normales de un proyecto, las tareas necesarias en cada fase y una explicación de las relaciones entre las tareas" (IBM Corporation, 1994), así como un **modelo de procesos** "ofrece un resumen del ciclo vital de minería de datos" (IBM Corporation, 1994) como se observa en la figura 5.

Figura 5. Modelo CRISP-DM



Fuente (IBM Corporation, 1994)

Consiste en una secuencia de 6 fases que permiten tener una estructura de trabajo bien definida que va desde entender la lógica del negocio, sus fuentes de información, modelado, hasta una solución e implementación. Por otra parte, aunque es estructurada, no es rígida, permite regresar a etapas anteriores para realizar ajustes y es aplicable a todo tipo de industria. Así mismo, se enfoca en los objetivos de negocio y su relevancia, que para el caso de Méderi es una mayor ocupación en sala, entendiendo que es el tratamiento que mayor ingreso genera (tener sillas o jornadas ociosas implica menor facturación). Un beneficio importante de esta metodología y de SCRUM es la reducción de riesgos, gracias a la validación continua de los resultados obtenidos de cara a la pertinencia para el negocio, dado que el proyecto se realiza en un sector sensible que impacta vidas, personas con diagnóstico confirmado de cáncer, es necesaria dicha validación con el área oncología recurrentemente para evitar imprecisiones en la estimación de asistencia de pacientes, para ello se organiza el flujo de trabajo en sprints que se detallan en el cronograma del proyecto, los cuales siempre reciben retroalimentación de la coordinadora de oncología y los jefes de programación.

A continuación, se explica cómo se abordó el proyecto cada una de las fases de CRISP DM.

6. Entendimiento del negocio

Se evaluó el contexto de negocio, sistema de salud en Colombia, las IPS y como se encuentra el hospital Méderi frente a otros hospitales y clínicas el país a nivel general y en oncología, por último, qué indicadores o métricas tiene el hospital actualmente para evaluar su gestión.

El presente proyecto empresarial se enmarca en el sector salud, uno de los más relevantes en la economía colombiana. Según el Presupuesto General de la Nación para 2023, los sectores de salud y protección social representan el 16% del gasto público, por encima de defensa (14%), trabajo (12%) y minas y energía (10%) a sectores como defensa con el 14%, trabajo 12%, minas y energía 10%, entre otros (Ministerio de Hacienda y crédito público, 2024). Este sector también tuvo un crecimiento en términos del PIB, pasando del 5.83% en 2022 al 5.9% en 2023 (*Colombia - Gasto Público Salud*, s.f.)

Dentro del sistema de salud colombiano, existen distintos tipos de entidades prestadoras de servicios "Con corte al 1 de Noviembre de 2016 se encontraron un total de 45.563 prestadores de servicios de salud inscritos en el Registro Especial de Prestadores de Servicios de Salud (REPS); de estos, el 72.7% (33.130) eran profesionales independientes, el 22.7% (10.366) Instituciones prestadoras de servicios de salud, el 3.7% (1.728) entidades con objeto social diferente y el 0.7 % (339) servicios de Transporte especial de pacientes" (Gaviria Uribe et al., 2016), a diciembre del 2022 la Asociación Colombiana de Hospitales y Clínica (ACHC) indicó que se contaba con 11.382 IPS habilitadas, de las cuales 2.193 son clínicas y hospitales (Quiceno, 2024). Para el corte de noviembre del 2016 de estas IPS el 90% corresponde a entidades privadas y el mayor número se encuentra ubicadas en la ciudad de Bogotá con 1.650 IPS. (Gaviria Uribe et al., 2016)

Este proyecto se desarrollará específicamente en el Hospital Universitario Méderi, una de las IPS más importantes del país, ubicada en Bogotá. De acuerdo con el ranking 2024 de la firma de investigación IntelLat, Méderi ocupa el puesto 33 entre las 61 mejores instituciones hospitalarias de América Latina, y el puesto 15 entre las 28 entidades colombianas presentes en dicho listado. (Intellat, 2024)

Méderi cuenta con dos sedes; sin embargo, el proyecto se concentrará exclusivamente en la operación del Hospital Universitario Mayor (HUM) En este hospital "se prestan principalmente servicios asistenciales médicos y quirúrgicos de alta complejidad (Urgencias, hospitalización, cuidado intermedio e intensivo adulto y neonatal, cirugía hospitalaria y ambulatoria y consulta externa" (Méderi, 2023), en el 2023 HUM atendió 344.951 servicios de salud, cerca del 50% de sus pacientes son personas con más de 60 años y el 55% de esta población son mujeres. (Méderi, 2024).

El presidente ejecutivo de Méderi, Mauricio Rubio Buitrago, ha destacado que uno de los enfoques prioritarios de la institución es el fortalecimiento de la atención oncológica. Según sus declaraciones: "La gestión de la red Hospitalaria continuó alineada con nuestra misión de generar bienestar a través de una atención integral de calidad y humanizada, que provee a los pacientes el tratamiento, la recuperación y el mejor desenlace posible, presentando un crecimiento en las actividades de los diferentes servicios, con énfasis en **oncología**" (Méderi, 2024). En línea con lo anterior, el hospital reportó un crecimiento del 32% en la atención de pacientes oncológicos con diagnóstico confirmado (Méderi, 2023), y ha catalogado la atención ambulatoria en oncología como uno de sus procesos misionales.

Según Bibiana Meneses, coordinadora del servicio de oncología en comunicación personal, 18 de noviembre de 2024, **la ruta oncológica representa el 45% de los ingresos totales del hospital, y dentro de ésta, el tratamiento con quimioterapia equivale al 42%**. En 2023, las

áreas de oncología y hemato-oncología registraron 24.520 consultas presenciales y 7.944 por telemedicina. La aplicación de quimioterapia tuvo un crecimiento significativo, pasando de 15.975 sesiones en 2021 a 24.610 en 2022 (54% de incremento). La central de mezclas preparó 29.068 combinaciones de quimioterapia en 2021 y 33.636 en 2022 (Méderi, 2023).

Dada la relevancia del tratamiento con quimioterapia dentro del modelo operativo y financiero del hospital, el proyecto busca generar una solución analítica que optimice procesos críticos de esta ruta, particularmente en lo relacionado con la programación de sesiones y la utilización eficiente de los recursos disponibles.

El flujo operativo actual implica, tras la confirmación diagnóstica, la autorización de la EPS (Ej. Nueva EPS mediante web, Compensar vía correo electrónico). Una vez aprobada, el auxiliar administrativo entrega la documentación al jefe de programación (JP), quien ingresa al sistema Servintec la fecha tentativa de tratamiento según protocolo y ciclo. Posteriormente, el auxiliar contacta al paciente para agendar la cita y sus exámenes de laboratorio, los cuales deben realizarse dos días antes. La realización efectiva de la quimioterapia depende del resultado de dichos exámenes. Bibiana Meneses destaca que es frecuente la reprogramación por causas como anemia o neutropenia, y que las mezclas de medicamentos pueden alcanzar valores cercanos a 20 millones de pesos. En caso de inasistencia, el hospital no puede facturar a la EPS, y se pierde la mezcla. Por esta razón, el auxiliar realiza una segunda confirmación el día anterior o el mismo día en la mañana.

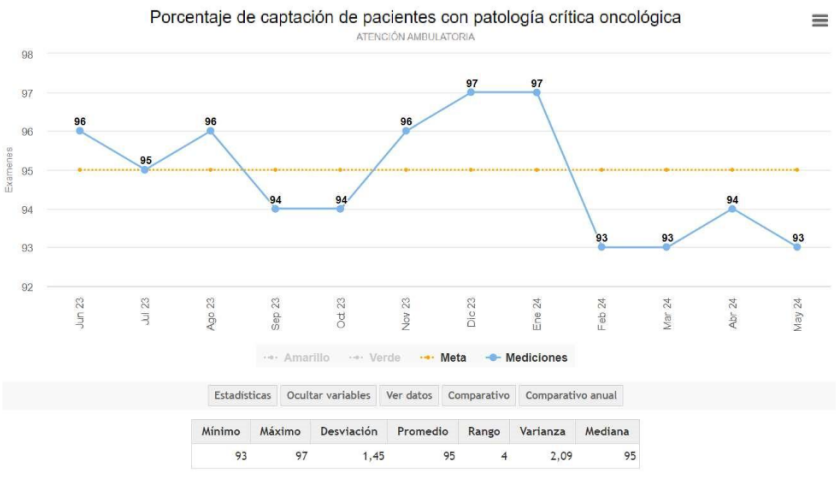
En cuanto a la infraestructura, el HUM dispone de dos salas para quimioterapia ambulatoria (tratamientos entre 30 minutos y 10 horas). La sala principal tiene 22 sillas y 2 camas; la secundaria, 7 sillas. Cada sala opera con 2 jefes y 3 auxiliares por turno, distribuidos en cuatro bloques horarios: mañana (7:00), mediodía (12:00), tarde (13:30) y tarde-tarde (15:30), la hora máxima de salida de un paciente es a las 7:30 pm.

El día anterior, la enfermera jefa recibe en físico el listado de pacientes confirmados, y asigna a cada uno su silla o cama para el día siguiente. También se generan rótulos con información personal y protocolaria (medicación incluida), que el personal clínico usa para tachar lo administrado durante la sesión.

Finalmente, se presentan indicadores clave del desempeño de la ruta oncológica entre junio de 2023 y junio de 2024. El hospital utiliza un sistema tipo semáforo: rojo (<83%), amarillo (83–98%) y verde (>98%).

- **Porcentaje de captación de pacientes con patología crónica oncológica (número de pacientes captados / total de pacientes reportados por patología)**

Figura 6. Indicador porcentaje de captación de pacientes con patología crítica oncológica



Fuente oncológica (Hospital Universitario Méderi, 2024)

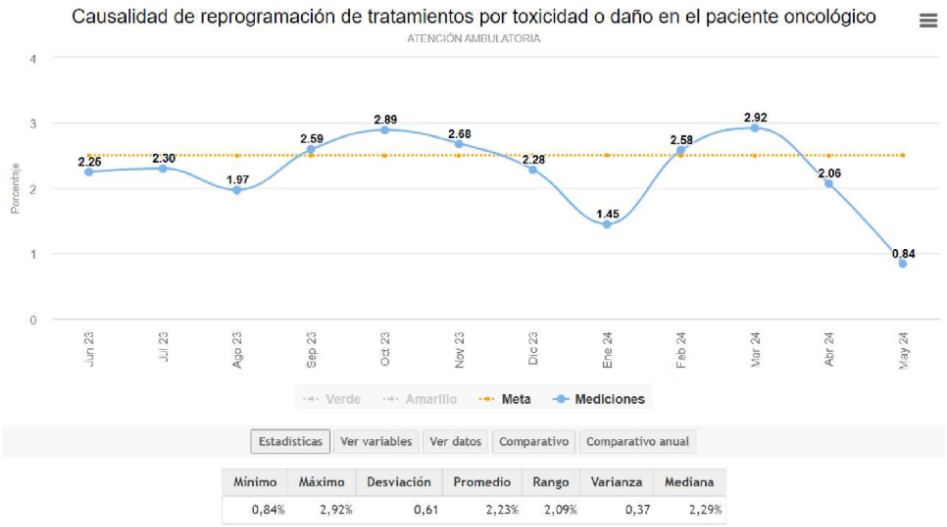
En mayo 2024 el indicador, como se muestra en la figura 6, estaba en el 93%, de los 418 pacientes reportados, se logran captar e ingresar a la ruta 389, resultado semáforo amarillo. La

mayor captación fue en enero con el 97 %, se evidencia una menor ejecución en lo corrido del 2024.

La **c**aptación de pacientes hace referencia al proceso estratégico y operativo mediante el cual el hospital busca atraer, canalizar y garantizar la atención oportuna de los pacientes dentro de su red o ruta asistencial. Es decir, implica asegurar que los pacientes diagnosticados o remitidos por las EPS principalmente, efectivamente ingresen a los servicios de tratamiento y seguimiento ofrecidos por Méderi. Por tanto, un paciente captado es aquel usuario del sistema de salud que ha sido identificado, contactado, y efectivamente vinculado a la ruta oncológica.

- **Causalidad de reprogramación de tratamientos por toxicidad o daño en el paciente oncológico (número de tratamientos abandonados por esta causal / total de tratamientos abandonados en el mismo periodo)**

Figura 7. Indicador causalidad de reprogramación de tratamientos por toxicidad o daño en el paciente oncológico



Mínimo	Máximo	Desviación	Promedio	Rango	Varianza	Mediana
0,84%	2,92%	0,61	2,23%	2,09%	0,37	2,29%

Fuente (Hospital Universitario Méderi, 2024)

En mayo, como se observa en la figura 7, de los 1.192 pacientes programados para tratamiento con quimioterapia, 10 fueron reprogramados debido a efectos adversos derivados de la toxicidad de los medicamentos. Esta toxicidad hace referencia a las reacciones físicas o clínicas que los fármacos pueden generar en el organismo del paciente, impidiendo su adecuada condición para recibir el tratamiento en la fecha prevista. Esto deja el indicador en 0,84%, por debajo de la meta establecida del 2,50% como valor máximo, ubicándose en semáforo verde.

- **Causalidad de reprogramación de tratamientos por inconvenientes de autorizaciones del paciente**

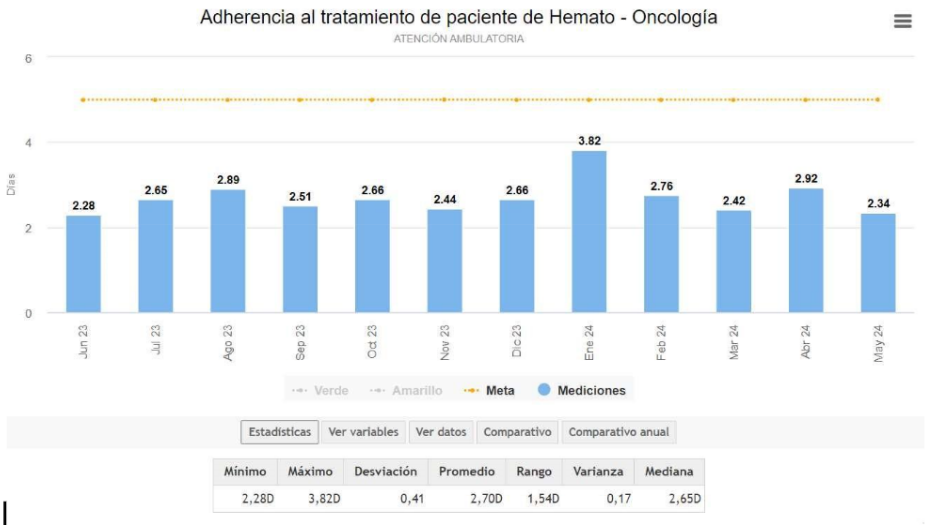
Figura 8 Indicador causalidad de reprogramación de tratamientos por inconvenientes de autorizaciones del pagador

Fuente (Hospital Universitario Méderi, 2024)

En mayo de los 1192 pacientes, como se observa en la figura 8, se reprogramaron 28 por esta causa, lo cual deja el indicador al 2,35% de los casos, de una meta del 6% como máximo. Se encuentra en semáforo verde.

- **Adherencia al tratamiento de pacientes de Hemato - Oncología (Sumatoria de días de atraso del protocolo de todos los pacientes / Número de pacientes programados y efectivamente atendidos en quimioterapia)**

Figura 9 .Indicadores adherencia al tratamiento de pacientes de Hemato-Oncología

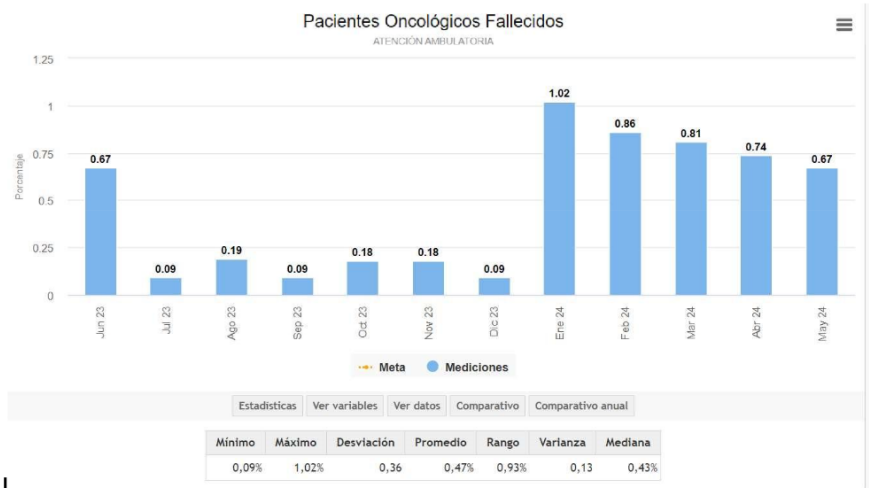


Fuente (Hospital Universitario Méderi, 2024)

En la figura 9, se muestra que la meta actual es una desviación máxima de 5 días para recibir el tratamiento de quimioterapia, para el mes de mayo se encuentra una medición de 2.3 días, el indicador se encuentra en semáforo verde. En este mes se atendieron 1192 pacientes, de los cuales 977 están de 0 a 5 días, 142 de 6 a 10 días y 73 con más de 10 días.

- **Pacientes fallecidos**

Figura 10 Indicadores pacientes Oncológicos Fallecidos



Fuente (Hospital Universitario Méderi, 2024)

En mayo de los 1192 pacientes programados para quimioterapia, fallecieron 8 pacientes, dejando indicador del 0,67%, como se evidencia en la figura 10.

6.1. Problemática

El área de Oncología del Hospital Universitario Méderi representa no solo un componente clave en la atención integral de pacientes con enfermedades de alta complejidad, sino también una fuente fundamental de sostenibilidad financiera para la institución. Actualmente, esta área representa una fuente clave de sostenibilidad financiera para el hospital, con la quimioterapia como uno de los servicios más determinantes en su estructura de ingresos. Esta relevancia clínica y económica convierte al servicio de oncología en un proceso misional prioritario, que requiere

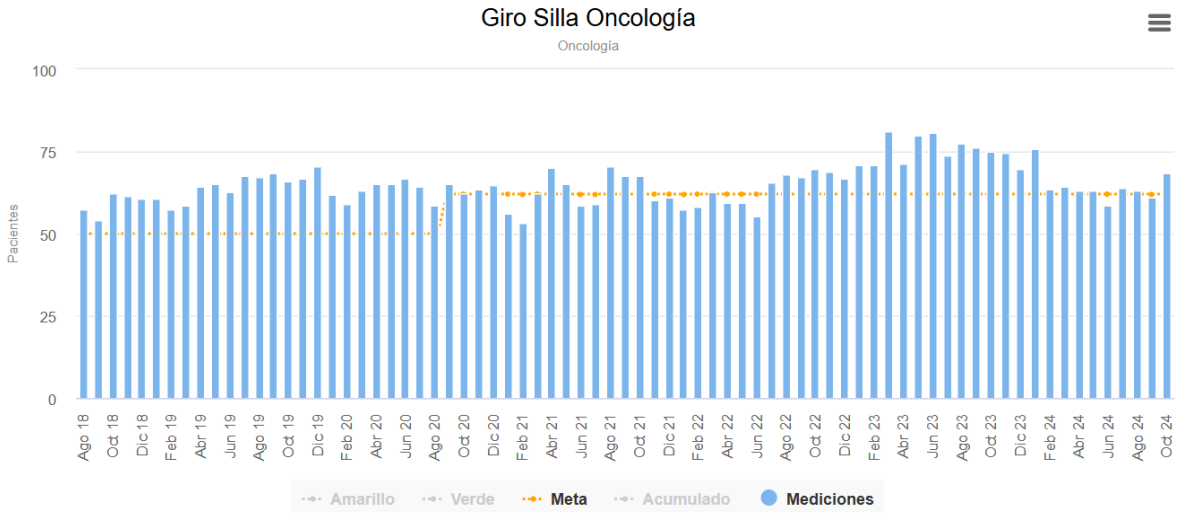
atención estratégica en términos de eficiencia operativa, calidad del servicio y uso inteligente de los datos.

Durante la fase de entendimiento del negocio, conforme a la metodología CRISP-DM, se evidenció una oportunidad significativa: el uso de herramientas analíticas para mejorar la eficiencia del área aún es limitado. En la actualidad, el análisis de datos en oncología se reduce a la generación de algunos indicadores básicos de gestión, los cuales son reportados al sistema de calidad institucional (ALMERA). Adicionalmente, gran parte del proceso operativo, desde la programación hasta la asignación de pacientes, se realiza de forma manual, lo que limita la trazabilidad, la capacidad de anticipación y el aprovechamiento de información para la toma de decisiones. A través de reuniones con la coordinadora del servicio de oncología y el levantamiento detallado de procesos, se identificó que una de las tareas más críticas es la programación de los tratamientos de quimioterapia. Esta programación debe lograr una distribución eficiente de los pacientes en dos salas disponibles, cada una con múltiples turnos diarios, de modo que se garantice tanto la continuidad del tratamiento clínico como la ocupación óptima de los recursos físicos y humanos. Sin embargo, actualmente esta tarea es realizada por tres jefes de programación en el sistema SIIFAM, quienes asignan la fecha de aplicación una vez reciben la autorización de la EPS y la confirmación del químico farmacéutico respecto a la disponibilidad del medicamento. El sistema no proporciona información consolidada o en tiempo real sobre la disponibilidad de sillas o la ocupación por turnos, lo que genera escenarios de sobre agendamiento o, por el contrario, espacios disponibles no utilizados, especialmente en los turnos de la tarde.

Esta situación impacta directamente el desempeño de uno de los principales indicadores del servicio: el Indicador de Giro de Silla en Oncología, calculado como el número total de aplicaciones de quimioterapia dividido por el número promedio de sillas disponibles en el periodo.

La meta institucional es de 62 pacientes por silla al mes. Sin embargo, como se evidencia en la figura 11, este indicador ha mostrado una tendencia decreciente durante el último año, reflejando un subuso de la capacidad instalada.

Figura 11 Indicador Giro Silla



Fuente (Hospital Universitario Méderi, 2024)

Ante este panorama, se plantea como objetivo del proyecto el diseño e implementación de una solución analítica que apoye la programación eficiente de pacientes a tratamiento, promoviendo una ocupación más equilibrada de los turnos y de las unidades físicas. Para ello, se propone el desarrollo de un modelo predictivo que permita anticipar la inasistencia de los pacientes a la quimioterapia, incorporando diversas etapas del ciclo CRISP-DM como la transformación y evaluación de variables, la selección y entrenamiento de modelos, y la validación de resultados. Con esta solución, se espera no solo reducir pérdidas operativas y mejorar la facturación hospitalaria, sino también favorecer la oportunidad y continuidad del tratamiento, generando beneficios para los usuarios internos (personal asistencial y administrativo) y externos (pacientes y cuidadores).

7. Entendimiento de los datos

En esta fase se contó con el apoyo del área de TI y oncología para validar las fuentes de información disponibles para el alcance del objetivo del proyecto, así como los aplicativos en los cuales reposa la información. Se contempla un análisis con las 4 dimensiones principales de calidad de datos (DAMA), por último, un análisis estadístico que contribuya al entendimiento del estado actual de la asignación de pacientes, así como la reprogramación de estos.

7.1. Fuentes de información

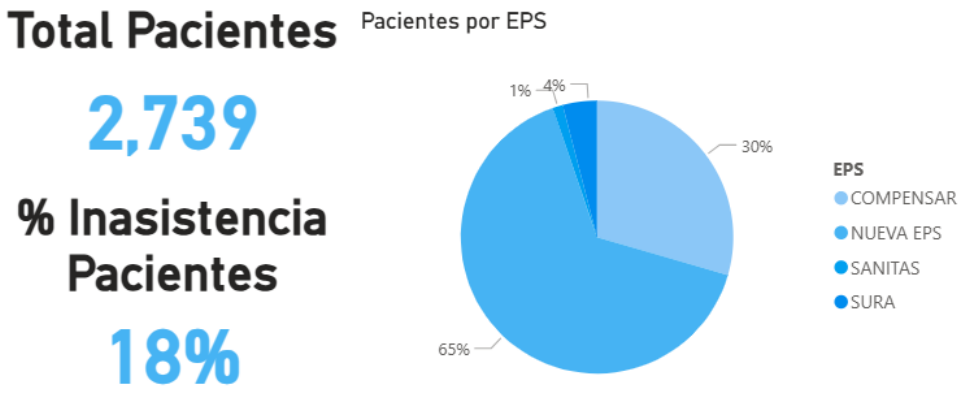
Actualmente se cuenta con una base de 2.739 pacientes del servicio oncológico que han sido programados al tratamiento de quimioterapia de enero a septiembre del 2024, esta base contiene información relacionada con la asistencia del paciente a la sala, su ubicación y medicamentos administrados.

Adicionalmente se cuenta con una base demográfica que tiene 4636 filas con la información de 3256 pacientes. En el anexo.1 se encuentran los diccionarios de datos de cada una de las bases mencionadas previamente.

7.2. Análisis descriptivo-Dashboard.

El análisis de los datos oncológicos de enero a septiembre de 2024 destaca la concentración de pacientes en ciertas EPS, la alta proporción de adultos mayores y una significativa tasa de inasistencia, lo que podría impactar la eficiencia del sistema. También se observa una concentración geográfica en zonas urbanas densamente pobladas, lo que subraya la necesidad de estrategias de atención diferenciadas. A continuación, se realiza un análisis de las principales variables, como la distribución por EPS, edad, género, estado civil, inasistencia, ubicación y diagnósticos, con el fin de identificar oportunidades de mejora en la atención y los resultados de los pacientes.

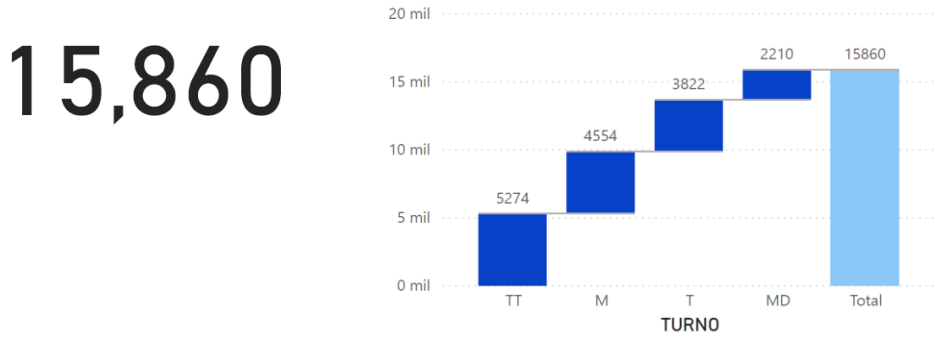
Figura 12. Distribución EPS



Fuente: Elaboración propia

El HUM cuenta con un total de 2.739 pacientes, de los cuales el 18 % presenta inasistencia a las citas y el 65 % pertenece a la NUEVA EPS, como se puede observar en la figura 12. El porcentaje de inasistencia evidencia un posible desbalance en la variable objetivo, aspecto importante a tener en cuenta en el desarrollo del modelo.

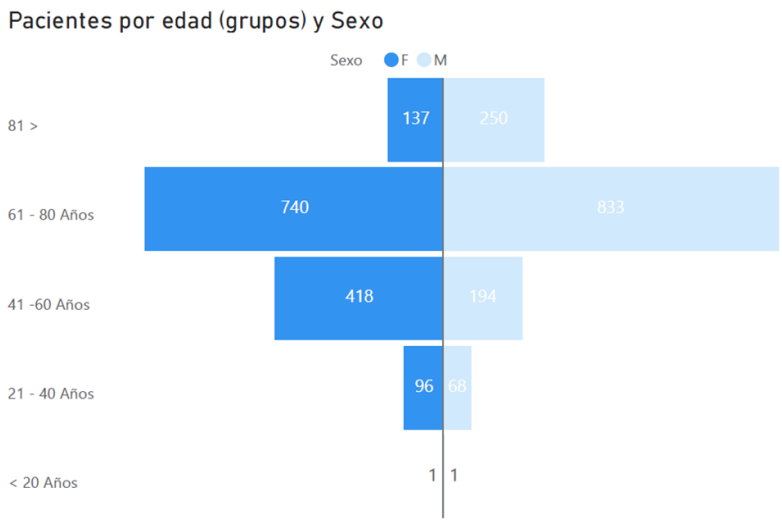
Figura 13 Número de Citas Totales y Distribución por Turno



Fuente: Elaboración propia

En la figura 13, periodo de datos evaluados, se registraron un total de 15.860 Citas de las cuales hay una mayor concentración en el turno Tarde Tarde (TT) con un total de 5.274 Citas, el siguiente turno más frecuente es el turno de la mañana (M) donde hubo 4.554 citas. finalmente, el Turno tarde (T) registro 3.822 Citas y el Turno media mañana (MD) 2.210 Citas.

Figura 14 Distribución edad y sexo de pacientes



Fuente: Elaboración propia

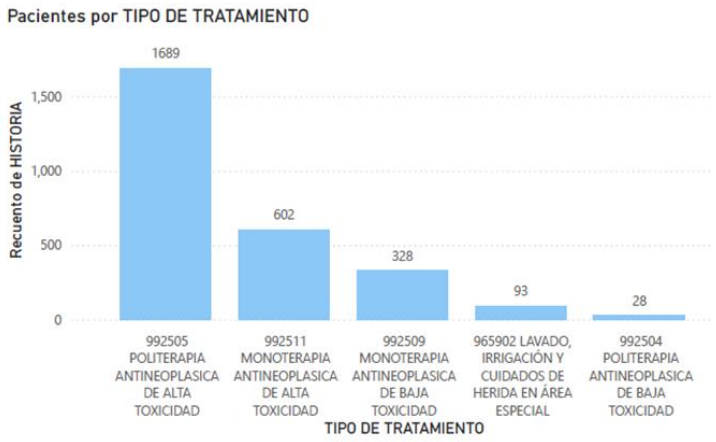
Análisis de la edad (figura 14) la mayor proporción en edades avanzadas, La mayoría de los pacientes se concentran en el rango de 61-80 años, seguido por el grupo >81 años. Esto sugiere que la población atendida tiene una tendencia hacia personas mayores.

Distribución por género: Hay más mujeres (F) 50,84% con un total de 1392 de mujeres, por otra parte, los hombres (M) representan el 49,16% con un total de 1.346 hombres, cuando revisamos por grupos de edad. en los grupos de más de 61 años identificamos que hay más pacientes hombres que mujeres, caso contrario en el grupo de 41 a 60 años presenta una mayor concentración de mujeres.

También identificamos una baja cantidad de jóvenes: Hay muy pocos pacientes menores de 40 años, lo que podría indicar un enfoque o necesidad de atención médica más relevante para adultos mayores.

El análisis indica que el promedio de edad de la población total es aproximadamente **65 años**, se debería plantear un enfoque hacia adultos mayores, lo que generaría una mejor experiencia para los pacientes de quimioterapia.

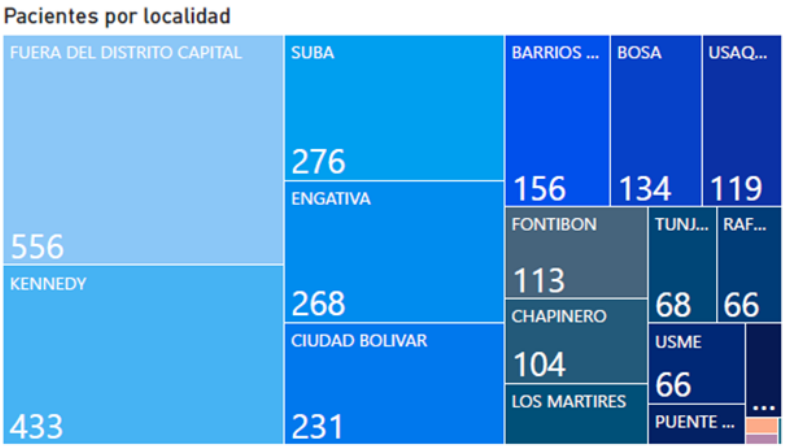
Figura 15 Tipo de tratamiento



Fuente: Elaboración propia

En cuanto a la figura 15 nos muestra que se genera una alta concentración de pacientes que necesitan tratamientos más fuertes o agresivos. Esto podría significar que hay muchos casos de cáncer avanzado o que requieren terapias intensivas. Sería importante enfocarse en estrategias para manejar los efectos secundarios, brindar apoyo paliativo y ofrecer una atención más completa a estos pacientes.

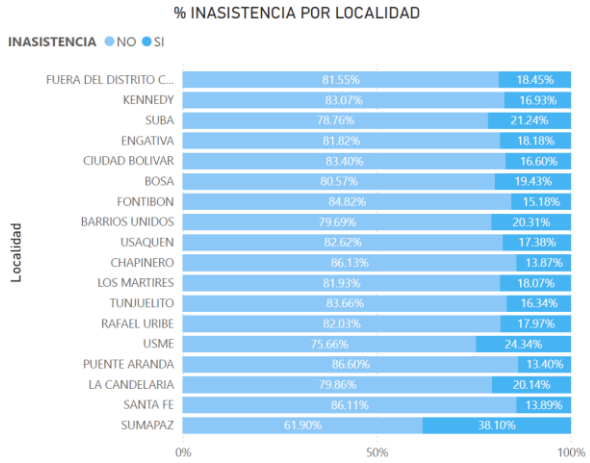
Figura 16 .Localidad pacientes



Fuente: Elaboración propia

Respecto a la ubicación geográfica de los pacientes figura 16: La mayor cantidad de pacientes se encuentra en localidades con alta densidad poblacional, como Kennedy, Suba, y Engativá. Cobertura geográfica: Fuera del Distrito Capital también muestra un número considerable de pacientes, lo que sugiere una distribución significativa en áreas externas.

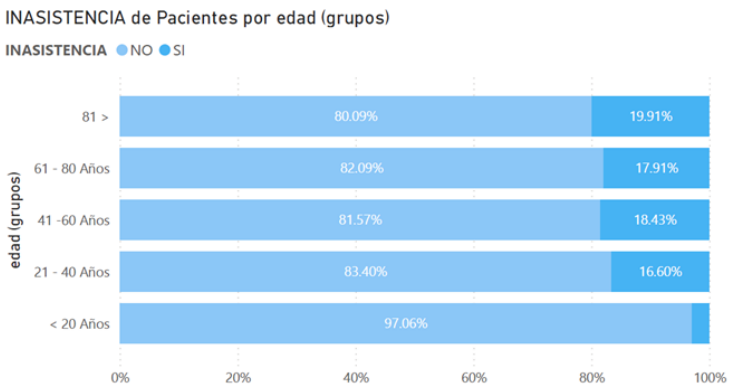
Figura 17 Porcentaje de inasistencia por Localidad



Fuente: Elaboración propia

En términos generales, la mayoría de las localidades presentan tasas de inasistencia por debajo del 20%, lo cual es una señal positiva de adherencia a las citas programadas. La figura 17 muestra el porcentaje de inasistencia por localidad, destacando que la mayoría de las zonas presentan niveles bajos de ausentismo, con porcentajes inferiores al 20%. Localidades como Puente Aranda (13.40%), Chapinero (13.87%) y Santa Fe (13.89%) registran las tasas de inasistencia más bajas, lo que sugiere un buen comportamiento de asistencia. En contraste, Sumapaz (38.10%) representa el caso más crítico, con una inasistencia considerablemente superior al promedio. Otras localidades como Usme (24.34%) y Suba (21.24%) también presentan niveles preocupantes. En general, la distribución evidencia una buena adherencia a las citas en la mayoría de las localidades.

Figura 18 Distribución inasistencia por edad



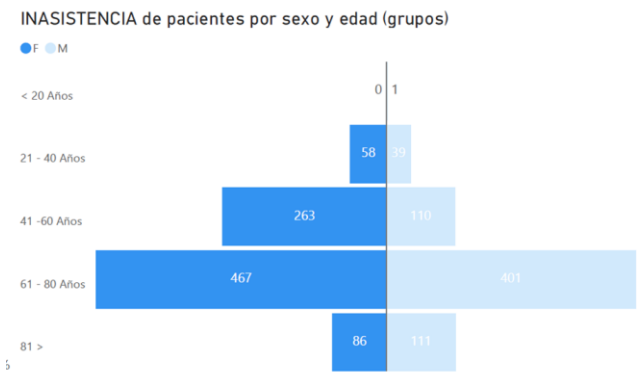
Fuente: Elaboración propia

En la figura 18, se observan las inasistencias por edad se conforman de la siguiente manera: 81 años y más: 19.91% de inasistencia, el porcentaje más alto entre todos los grupos.

- 41 a 60 años: 18.43% de inasistencia.
- 61 a 80 años: 17.91% de inasistencia.

Grupos con menor inasistencia: Menores de 20 años: Solo el 2.94% no asiste, lo que representa el menor índice de inasistencia en comparación con otros grupos. Se observa que la inasistencia aumenta con la edad, siendo más significativa a partir de los 41 años, alcanzando su pico en los mayores de 81 años.

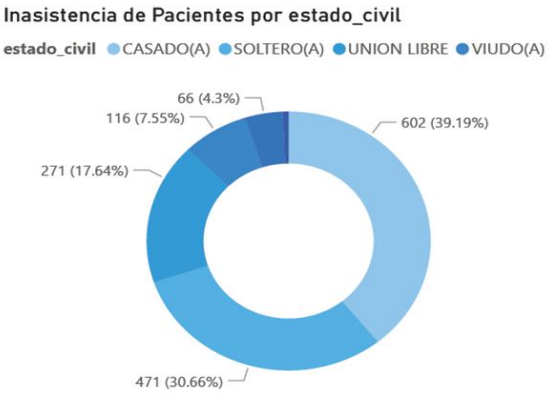
Figura 19 Inasistencia por grupo de edad y sexo



Fuente: Elaboración propia

En la figura 19, se observa los grupos de edad, las **mujeres (barras azules oscuras)** tienen mayor número de inasistencias que los hombres (barras azules claras). El grupo de **61-80 años** concentra la mayor cantidad de inasistencias en ambos sexos, con 467 mujeres y 401 hombres.

Figura 20 Inasistencia por estado civil

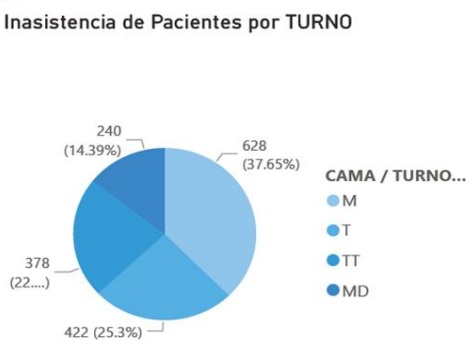


Fuente: Elaboración propia

Al validar el estado civil de los pacientes, como se muestra en la figura 20, identificamos la siguiente información. Los CASADO(A) representan el porcentaje más alto de inasistencia

(39.19%). SOLTERO(A) es el segundo grupo más frecuente con un 30.66% de inasistencias. UNIÓN LIBRE y VIUDO(A) tienen porcentajes significativamente menores (17.64% y 7.55%, respectivamente). La categoría menos común es "VIUDO(A)" con solo un 4.3% de inasistencias. Podemos concluir que el estado civil "CASADO(A)" presenta la mayor proporción de inasistencia.

Figura 21 Inasistencia por turno

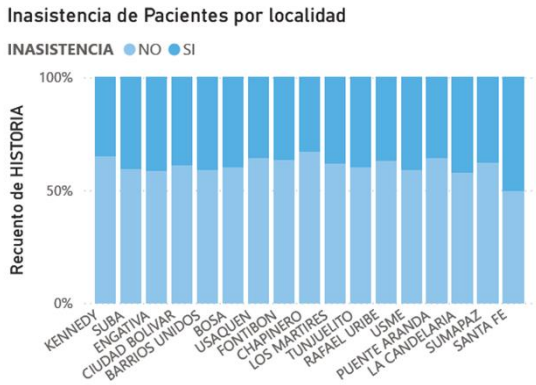


Fuente: Elaboración propia

En la figura 21, se observa que los turnos generan una información relevante de la cual se destaca: El turno **M** (mañana) tiene la mayor proporción de inasistencias con un 37.65%. El turno **MD** (medio día) sigue con un 25.3%. Los turnos **TT** (tarde) y **T** (total) tienen porcentajes menores, con 22% y 14.39%, respectivamente.

Podemos concluir que el turno de la mañana muestra la mayor cantidad de inasistencias, posiblemente por conflictos con horarios laborales o personales en las primeras horas del día.

Figura 22 Inasistencia por localidad



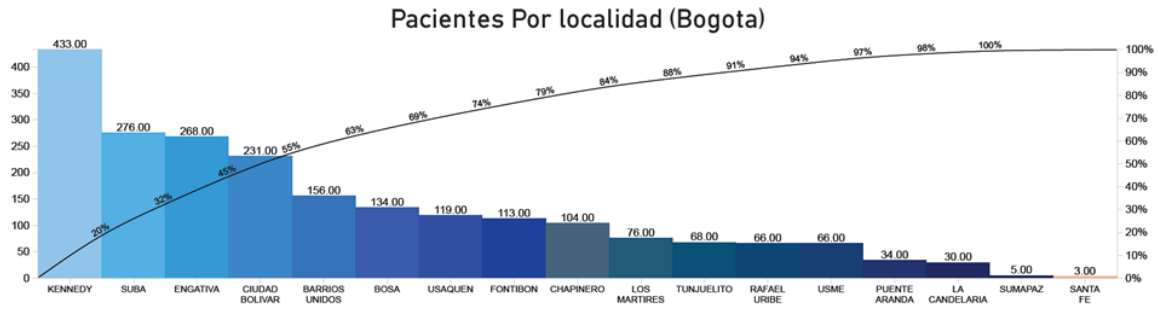
Fuente: Elaboración propia

En la figura 22, se identifica un análisis entre la inasistencia y la localidad de residencia de los pacientes destacamos que: hay un patrón de inasistencia significativo en algunas localidades específicas.

- Localidades como KENNEDY y SUBA presentan niveles más altos de inasistencia en comparación con otras.

La inasistencia varía según la localidad del paciente, probablemente influenciada por factores como la accesibilidad a la clínica, distancia o problemas de transporte.

Figura 23 Pareto pacientes por localidad



Fuente: Elaboración propia

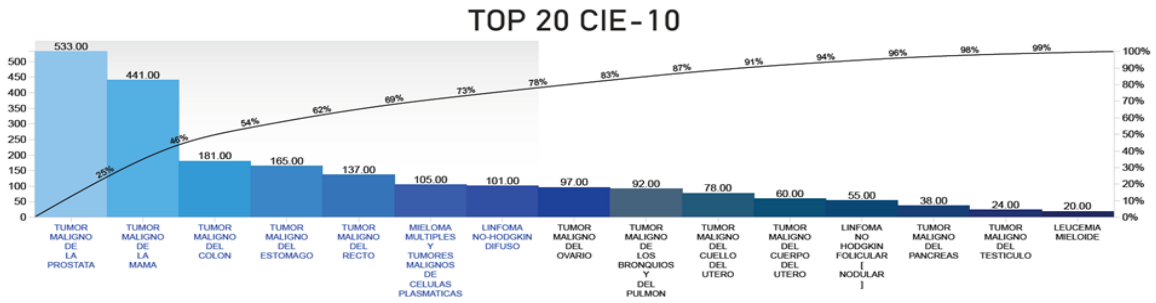
En la figura 23. se observa un Pareto, para determinar de qué localidades provienen la mayor cantidad de pacientes generando la siguiente información:

- La localidad **Kennedy** tiene el mayor número de pacientes (433), representando el 20% del total.
- **Suba** (276) y **Engativá** (268) también tienen una alta proporción de pacientes, sumando el 32% y 26% respectivamente.
- Localidades como **Santa Fe** (3) y **Sumapaz** (5) tienen la menor cantidad de pacientes, con un porcentaje insignificante en comparación con otras localidades.
- El gráfico sigue una distribución acumulativa donde las primeras localidades representan un porcentaje significativo del total de pacientes.

La atención en salud está más concentrada en localidades densamente pobladas como Kennedy, Suba y Engativá. Las localidades con menos pacientes, como Santa Fe y Sumapaz,

podrían requerir estrategias diferentes de cobertura, como telemedicina o brigadas móviles, para mejorar la accesibilidad.

Figura 24 Top 20 CIE-10

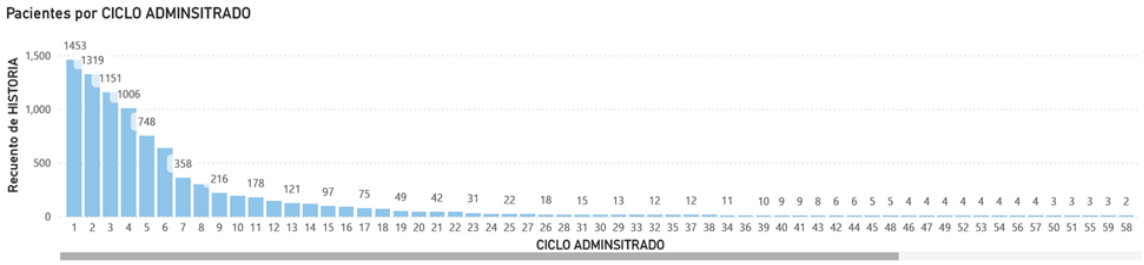


Fuente: Elaboración propia

Así mismo, en la figura 24, identificamos el Top 20 de los Diagnósticos más comunes en los pacientes del HUM generando la siguiente información:

- **Tumor maligno de la próstata** es el diagnóstico más frecuente (533 casos), representando el 25% del total.
- Le sigue el **tumor maligno de mama** (441 casos), con un 20% del total.
- Otros diagnósticos relevantes incluyen tumores malignos del colon (181), estómago (165) y recto (137), cubriendo entre el 46% y el 62% de los casos acumulados.
- Diagnósticos menos comunes en el top 20 son **leucemia mieloide** (20 casos) y **tumor maligno del testículo** (24 casos), representando menos del 5% del total acumulado.

Figura 25 Pacientes por ciclo administrado



Fuente: Elaboración propia

Por otra parte, en la figura 25 se muestra el comportamiento de los pacientes a través del número de ciclos, el mayor número de pacientes se concentra en los primeros ciclos. Los primeros cinco ciclos tienen los valores más altos, destacándose especialmente el primer ciclo con 1,453 pacientes, seguido por el segundo ciclo con 1,319 pacientes. nos muestra una tendencia decreciente a medida que aumenta el número de ciclos administrados, el número de pacientes disminuye de forma significativa.

Para el ciclo 10 en adelante, los valores son mucho menores, con menos de 300 pacientes por ciclo. creando una larga cola de valores pequeños, Más allá del ciclo 20, el número de pacientes por ciclo es muy bajo y permanece casi constante con pequeñas variaciones.

7.3. Calidad de Datos

Se realizó el cálculo de los indicadores de calidad de datos según la metodología DAMA, cuyos resultados se presentan en la Tabla 1. Se identificaron problemas de duplicidad asociados a errores en el ingreso de la fecha final del tratamiento, así como casos en los que un mismo paciente presenta múltiples motivos de reprogramación, diferentes fechas de formulación, días de retraso y ausencia de información en el campo de ciclo administrado.

En cuanto a la completitud, se observaron vacíos en campos clave como historia clínica, código CIE-10, ciclo administrado, número de cédula, sexo, fecha de nacimiento, lugar de residencia, localidad del paciente, ocupación, protocolo y duración del tratamiento. En el indicador de precisión, se detectaron inconsistencias relacionadas con ciclos negativos o atípicos. Por su parte, la validez se vio afectada principalmente en los campos de ciclo administrado y especialidad médica.

Tabla 1 Indicadores Calidad

Nombre del indicador	Valor del indicador
Validez	72,05%
Precisión	99,98%
Completitud	97,39%
Unicidad	99,92%
Índice total de calidad	92,34%

Fuente: Elaboración propia

En conclusión, se puede afirmar que la base de datos presenta una calidad general alta, con un índice total del 99.33%. Sin embargo, se evidencia una debilidad importante en el indicador de validez, especialmente en el manejo de los protocolos. Si esta situación no se corrige o no se implementa una solución adecuada, se corre el riesgo de perder hasta un 28% de la base de datos, lo que afectaría significativamente la generación de entregables y la confiabilidad de los análisis del proyecto.

Para mitigar la pérdida de información y evitar la introducción de sesgos en el desarrollo del trabajo de grado, se han implementado acciones como la creación de diccionarios de datos, la estandarización de variables (eliminación de espacios innecesarios, símbolos como "-", nombres duplicados o mal escritos), la corrección de espacios dobles y la validación de datos atípicos en conjunto con el área de negocio.

8. Preparación de los datos

Se buscó corregir los errores en calidad de datos identificados en la sección anterior, estandarización de grupos, eliminar vacíos, unificar las bases existentes con los datos útiles para el análisis, así como nuevos campos que enriquezcan el tablero de indicadores de desempeño y demás información descriptiva.

El proceso de transformación de los datos se llevó a cabo en Google Colab y comenzó con la unificación de las bases de datos que contenían la historia clínica de los pacientes y la fecha de inicio de sus ciclos. Se realizó una reclasificación de los diagnósticos CIE-10 según los capítulos establecidos por la Organización Panamericana de la Salud, reduciendo las categorías de 349 a 24. Asimismo, las ocupaciones fueron agrupadas en cinco categorías principales: empleado, desempleado, civil, estudiante y otros, simplificando 44 ocupaciones originales. Se eliminaron registros con vacíos en columnas clave como historia clínica, protocolos, ciclos administrados, CIE-10, EPS, oportunidad de inicio de QMT y especialidad, lo que representó 404 registros de un total de 16.212.

En cuanto a la variable de duración del tratamiento, se identificó que los protocolos registrados en la base de datos no contaban con un formato estandarizado ni con información clara sobre su duración, lo que dificultaba su análisis y uso en los modelos. Para resolver esta limitación, se trabajó en conjunto con el equipo de oncología del Hospital Universitario Méderi, quienes

facilitaron un diccionario en formato Excel con tres columnas: nombre del protocolo tal como aparecía en la base, su equivalencia estandarizada y su duración estimada en minutos. Esta información fue construida con base en la experiencia de los jefes de programación y permitió realizar una fusión con los datos originales para incorporar la duración como una nueva variable. La estandarización fue necesaria debido a la alta variabilidad en la forma de registrar los protocolos (abreviaciones, errores tipográficos, múltiples formas de escritura), lo que impedía su análisis confiable. Esta variable fue clave tanto para el modelo predictivo como para el modelo prescriptivo, ya que influye directamente en la asignación de recursos físicos y en la probabilidad de asistencia del paciente.

Adicionalmente, se crearon variables derivadas con el objetivo de capturar patrones de comportamiento individual que no estaban explícitamente representados en los datos originales, pero que resultan fundamentales para anticipar la inasistencia. Estas variables se fundamentaron en tres pilares. En primer lugar, el conocimiento experto del equipo clínico y de programación del Hospital Universitario Méderi, quienes identificaron que el turno de la mañana y los fines de semana o días festivos presentaban mayores tasas de inasistencia, lo que motivó la creación de variables como “inasistencia en fin de semana”, “día de la cita” y “semana del año”. En segundo lugar, se consideraron hallazgos de la literatura especializada en predicción de inasistencias en contextos hospitalarios, donde estudios como los de AlMuhaideb et al. (2019) y Alaeddini et al. (2011) destacan la relevancia de variables como “asistencia a la última cita”, “inasistencias anteriores” y “porcentaje de asistencia histórica” como predictores claves. En tercer lugar, se incorporaron variables derivadas propuestas a partir del análisis exploratorio de los datos, que permitieron identificar atributos adicionales con potencial explicativo, como “días desde la última cita” o “inasistencias anteriores”. Aunque estas variables no fueron sugeridas directamente por el equipo clínico ni extraídas de literatura previa, su inclusión se validó empíricamente durante el

entrenamiento del modelo XGBoost, donde varias de ellas se posicionaron entre las más importantes según el análisis de importancia de variables como se observar en la figura 33.

9. Modelado

Se hizo uso de diferentes modelos supervisados con los datos estandarizados en la fase anterior, que buscan predecir la inasistencia de los pacientes oncológicos al tratamiento de quimioterapia. Los modelos utilizados son regresión logística, máquinas vectoriales, Random Forest, XG Boost.

Por otra parte, un modelo prescriptivo de optimización para lograr la ubicación más eficiente de pacientes en sala.

9.1. Modelo predictivo

En esta sección se describe el proceso de selección del modelo predictivo más adecuado, mediante la aplicación de técnicas de aprendizaje supervisado en el contexto de problemas con desbalance de clases. Se evaluaron distintas estrategias de balanceo de datos (SMOTE, SMOTE-Tomek, ADASYN y SMOTE-ENN), así como procesos de transformación y generación de variables. Posteriormente, se entrenaron y compararon cuatro algoritmos de clasificación: Random Forest, XGBoost, Regresión Logística y Máquina de Soporte Vectorial (SVM).

En las secciones previas de entendimiento de los datos (sección 7) y preparación de los datos (sección 8), se identificó que el conjunto de datos presenta un desbalance en la variable objetivo (asistencia vs. inasistencia), lo cual representa un desafío para la construcción de modelos robustos. Además, se evidenció la presencia de variables tanto categóricas como numéricas, con múltiples categorías que requieren codificación adecuada.

En la sección de entendimiento del negocio (sección 6), se estableció que la inasistencia de los

pacientes impacta negativamente el indicador *giro silla*, afectando la eficiencia del servicio de quimioterapia. Esta situación constituye un problema operativo, ya que compromete tanto la oportunidad en la atención como el aprovechamiento de los recursos físicos disponibles. Asimismo, en la introducción se detalla cómo se calcula este indicador y se destaca su relevancia institucional, al ser un parámetro clave del Sistema de Gestión de Calidad del hospital. La selección de estos modelos responde a su capacidad para ajustarse a las características del conjunto de datos (variables mixtas, desbalance de clases y presencia de ruido), así como a su uso extendido en aplicaciones de predicción en salud. Cada uno fue evaluado comparativamente en términos de desempeño predictivo, interpretabilidad, escalabilidad y adecuación al contexto clínico-operativo del hospital. A continuación, se presenta una descripción de cada modelo, incluyendo sus principales fortalezas, posibles desventajas y consideraciones a tener en cuenta para su aplicación en este contexto.

Random Forest es un modelo de ensamble que resulta útil en contextos con desbalance de clases, como el del presente proyecto, en el que solo el 18 % de los pacientes registrados presentan inasistencia. Su capacidad para ponderar clases lo hace apto para mejorar la predicción de eventos poco frecuentes. Además, puede trabajar con variables categóricas y numéricas, como las utilizadas en el HUM, identificar relaciones no lineales y proporcionar medidas de importancia de variables, lo cual facilita la interpretación por parte del equipo clínico. Sin embargo, su alto costo computacional, la dificultad de interpretación por su naturaleza no paramétrica y la sensibilidad a variables con muchas categorías pueden representar limitaciones en un entorno operativo donde se requiere agilidad y claridad en la toma de decisiones.

XGBoost es particularmente adecuado para este proyecto debido a su robustez en tareas de clasificación con clases desbalanceadas, como es el caso de la inasistencia en quimioterapia. Su parámetro scale_pos_weight permite compensar la baja frecuencia de la clase minoritaria, y su

capacidad para manejar patrones complejos lo hace ideal en un entorno clínico donde múltiples variables interaccionan. También incorpora mecanismos de regularización que ayudan a prevenir el sobreajuste, lo cual es crucial al trabajar con historiales de pacientes. No obstante, su interpretación puede ser limitada para usuarios no técnicos y, al igual que Random Forest, puede ser sensible a variables con muchas categorías.

SVM es un modelo eficaz en contextos de alta dimensionalidad como el del conjunto de datos del HUM, que incluye múltiples variables clínicas, demográficas y operativas. Puede clasificar correctamente patrones complejos, especialmente mediante el uso de kernels no lineales. Esto lo convierte en una opción válida para capturar la lógica detrás de la inasistencia a tratamientos. Sin embargo, su sensibilidad al sobreajuste, la necesidad de normalización de datos, su baja escalabilidad con grandes volúmenes y su limitada interpretabilidad son factores para considerar, especialmente en entornos hospitalarios que requieren aplicabilidad.

La regresión logística es un modelo clásico ampliamente utilizado en contextos de salud, valorado por su facilidad de interpretación y bajo costo computacional. Es útil como punto de partida para establecer una línea base de predicción, y puede adaptarse al desbalance de clases mediante el parámetro class_weight. Aunque es menos flexible frente a patrones no lineales o estructuras complejas, su uso en este proyecto es pertinente como alternativa interpretable para los equipos asistenciales del HUM. Sin embargo, su rendimiento puede verse afectado por la presencia de outliers y multicolinealidad entre variables.

Para el análisis preliminar de los modelos descritos previamente, se emplearon las variables “naturales”, entendidas como los datos originales proporcionados por el hospital, los cuales se encuentran detallados en el numeral 8. El conjunto de datos incluye el historial desagregado de citas de cada paciente entre enero y agosto de 2024.

En una primera iteración, se optó por consolidar todo el historial de cada paciente en un único registro con las variables que relacionan en la tabla 2, con el objetivo de generar nuevas variables derivadas a partir de la información histórica. Sin embargo, esta estrategia presentó una limitación crítica: al resumir el historial completo en un solo registro, se introdujo una **fuga de datos** (data leakage). En consecuencia, el modelo accedió a información futura al momento de predecir eventos pasados, lo cual distorsionó su desempeño real, generando **un sobreajuste artificial** y métricas de evaluación no representativas.

Tabla 2 Variables modelo inicial

VARIABLES CREADAS
TURNO_MAS_FRECUENTE
TRATAMIENTO_PROLONGADO
PORC_CITAS_PACIENTE_TURNO_M
PORC_CITAS_PACIENTE_TURNO_MD
PORC_CITAS_PACIENTE_TURNO_T
PORC_CITAS_PACIENTE_TURNO_TT
CONCENTRACION_ INASISTENCIA_EN_UN_TURNO
NO_ASISTIO_ ULTIMA_CITA
NO_ASISTIO_ ULTIMAS_2CITAS
NO_ASISTIO_ ULTIMAS_3CITAS

Fuente: Elaboración propia

Como segunda estrategia, se optó por mantener un único registro por paciente, asegurando que la variable objetivo (inasistencia) no fuera utilizada en la construcción de las variables predictoras, con el fin de evitar cualquier tipo de fuga de datos, las variables utilizadas en esta opción se pueden observar en la tabla 3.

Tabla 3 Variables modelo 2

VARIABLES CREADAS
TURNO_MAS_FRECUENTE
TRATAMIENTO_PROLONGADO
CITAS_PACIENTE_TURNO_M
CITAS_PACIENTE_TURNO_MD
CITAS_PACIENTE_TURNO_T
CITAS_PACIENTE_TURNO_TT
CONCENTRACION_EN_UN_TURNO
CITA_FIN_DE_SEMANA
ES_PRIMERA_CITA
SEMANA_DEL_AÑO

Fuente: Elaboración propia

A partir de esta configuración, se evaluaron distintos modelos empleando tanto las variables originales proporcionadas por la institución (variables naturales), como las variables derivadas permitidas. Los resultados obtenidos se presentan a continuación en la figura 26.

Figura 26 Comparación métricas modelos

	Modelo	F1 Train	Recall Train	Precision Train	AUC-ROC Train	AUC-PR Train	F1 Test	Recall Test	Precision Test	AUC-ROC Test	AUC-PR Test
0	XGBoost	0.914019	0.983903	0.853403	0.994034	0.983609	0.813043	0.877934	0.757085	0.964343	0.914637
1	SVM	0.837910	0.951710	0.748418	0.971506	0.879514	0.791489	0.873239	0.723735	0.947973	0.839382
2	Regresión Logística	0.814815	0.907445	0.739344	0.943079	0.797740	0.789474	0.845070	0.740741	0.918725	0.753903
3	Random Forest	1.000000	1.000000	1.000000	1.000000	1.000000	0.639535	0.516432	0.839695	0.915309	0.803889

Fuente: Elaboración propia

En la figura 26, se observa el principal inconveniente es el sobreajuste, especialmente en los modelos XGBoost y Random Forest, ya que se observan amplias brechas entre las métricas de

F1, recall y precisión en los conjuntos de entrenamiento y prueba. Por otro lado, también se evidencian indicios de sobreajuste en los modelos de Regresión Logística y Máquina de Soporte Vectorial (SVM), particularmente en la diferencia del indicador recall entre los datos de entrenamiento y los de prueba, lo que sugiere una posible pérdida de capacidad de generalización.

Como tercera iteración, se optó por conservar la estructura inicial del dataset con las variables observadas en la tabla 4, en la que cada registro corresponde a una cita específica de quimioterapia asociada a un paciente y una fecha determinada.

Tabla 4 Variables modelo 3

VARIABLES CREADAS
CANT_CITAS
CITAS_ANTERIORES
INASISTENCIAS_ANTERIORES
PORC_ASISTENCIA_ANTERIOR
ASISTIO_ULTIMA_CITA
INASISTENCIAS_ULTIMAS_3
DIAS_DESDE_ULTIMA_CITA
DIA_SEMANA
ES_FIN_SEMANA
SEMANA_DEL_ANO
ASISTENCIA_FIN_SEMANA
INASISTENCIA_FIN_SEMANA
INASISTENCIAS_MISMO_TURNO

Fuente: Elaboración propia

Esta configuración permite aprovechar lo mejor de las dos aproximaciones anteriores: por un lado, posibilita la creación de variables enriquecidas a partir del comportamiento histórico de inasistencia, preservando el hábito de asistencia individual; y por otro, permite incorporar variables adicionales derivadas de los datos originales o “naturales”.

Es importante resaltar que esta estrategia requiere, para su implementación en producción, un proceso continuo de actualización del dataset, alimentándose con las nuevas citas que se generen

en el tiempo. Esto no solo garantiza un control actualizado y realista del patrón de asistencia de cada paciente, sino que también mitiga el riesgo de fuga de datos al respetar la secuencia temporal de los eventos.

El ajuste metodológico para evitar la fuga de datos fue adecuado y logró mejorar la validez del modelo. Aunque el desempeño general bajó, como se observa en la figura 27, los resultados reflejan **un escenario mucho más realista**.

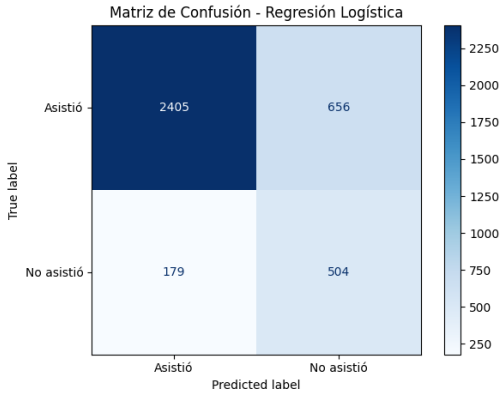
Figura 27 Comparación métricas modelos train - test

```
  warnings.warn(smsg, UserWarning)
                Modelo  F1_1_Train  Recall_1_Train  Precision_1_Train  \
0       XGBoost (Reg)      0.6648          0.8870             0.5316
1                 SVM      0.6851          0.9021             0.5523
2  Random Forest (Reg)      0.5563          0.7922             0.4287
3  Regresión Logística      0.6134          0.8236             0.4886

   F1_0_Train  Recall_0_Train  Precision_0_Train  AUC-ROC_Train  AUC-PR_Train  \
0      0.8922          0.8256             0.9704         0.9360        0.7970
1      0.9005          0.8368             0.9746         0.9474        0.8005
2      0.8443          0.7644             0.9428         0.8657        0.6317
3      0.8746          0.8077             0.9535         0.9030        0.7133

   F1_1_Test  Recall_1_Test  Precision_1_Test  F1_0_Test  Recall_0_Test  \
0     0.6637         0.8858            0.5307     0.8918         0.8252
1     0.6003         0.7862            0.4855     0.8745         0.8141
2     0.5570         0.8082            0.4249     0.8405         0.7560
3     0.5469         0.7379            0.4345     0.8521         0.7857

   Precision_0_Test  AUC-ROC_Test  AUC-PR_Test
0            0.9700        0.9322       0.7971
1            0.9447        0.8846       0.6600
2            0.9464        0.8748       0.6596
3            0.9307        0.8562       0.6432
```

Fuente: Elaboración propia

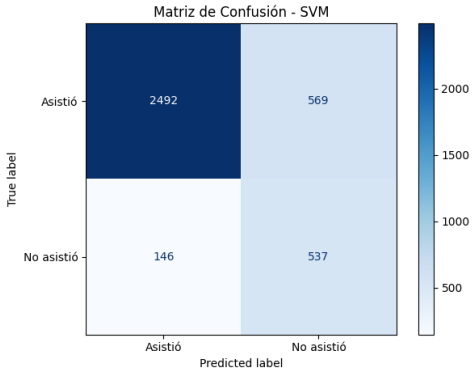
Como se observa en la figura 27, XGBoost sigue siendo el mejor modelo en métricas junto con la revisión de los falsos negativos y falsos positivos, como se muestra en las figuras 28 al 31, seguido por la máquina de soporte vectorial.

Figura 288 Matriz de confusión SVM



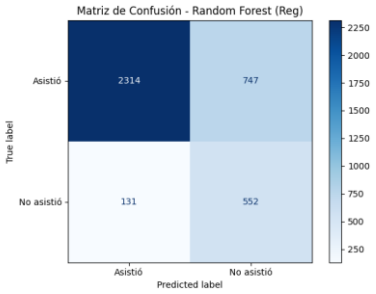
Fuente: Elaboración propia

Figura 29 Matriz de confusión regresión logística



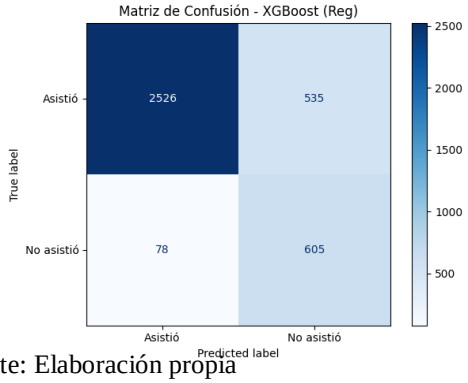
Fuente: Elaboración propia

Figura 30 Matriz de confusión Random Fs



Fuente: Elaboración propia

Figura 30 Matriz de confusión XG

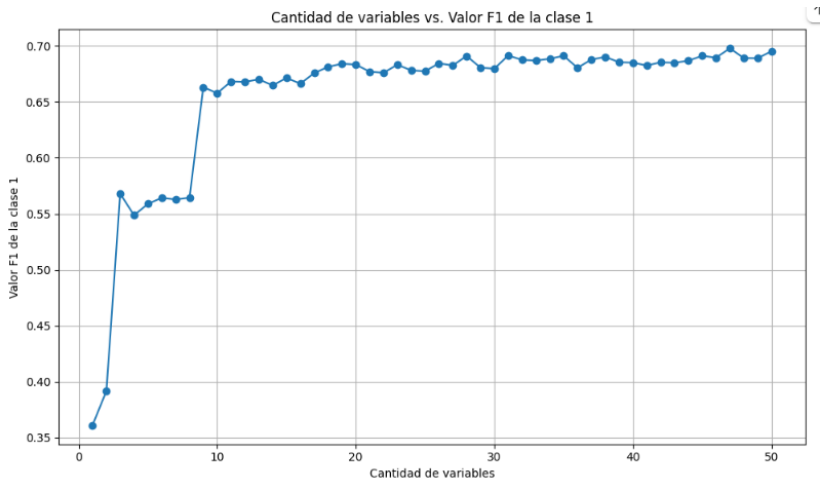


Fuente: Elaboración propia

Una vez definidas las variables predictoras, seleccionada la estructura de registros por paciente y elegido el modelo (XGBoost), el siguiente paso consistió en optimizar su desempeño a través de tres estrategias principales.

En primer lugar, se identificaron cuantas variables tomar de acuerdo con el valor del F1 como se ve en la figura 32.

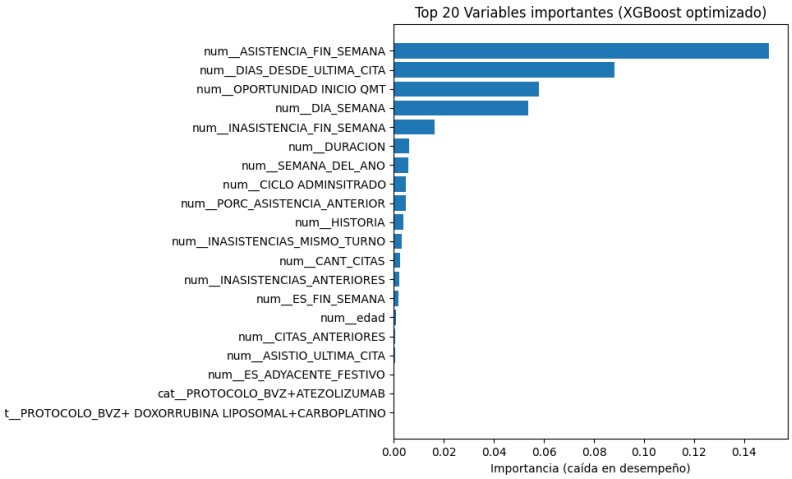
Figura 31 .Estimación número variables



Fuente: Elaboración propia

Se redujo la carga computacional del modelo al incorporar únicamente las variables con mayor relevancia según su importancia relativa teniendo en cuenta el valor del **F1-Score** vs cantidad de variables como se observa en la figura 33.

Figura 32 Top variables

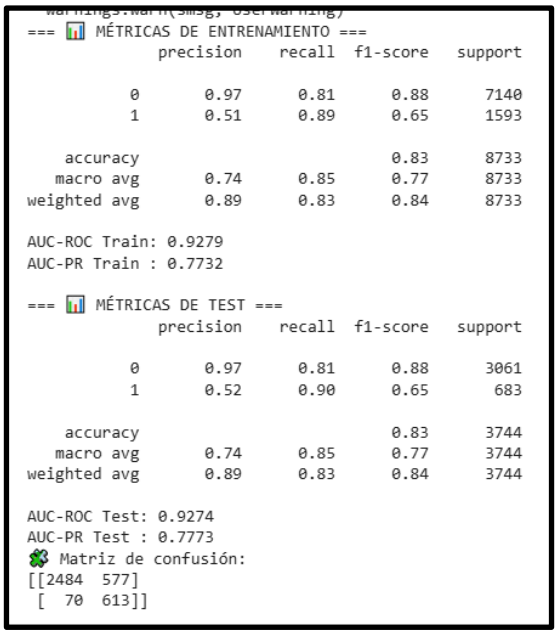


Fuente: Elaboración propia

En segundo lugar, se aplicaron distintas técnicas de balanceo para corregir el desbalance de clases presente en la variable objetivo. Finalmente, se procedió a ajustar el umbral de clasificación, identificando su punto óptimo con base en la métrica **F1-score**, la cual resulta especialmente adecuada para mejorar el desempeño en la predicción de la clase minoritaria de interés: la inasistencia.

Se volvió a ejecutar el modelo, esta vez incorporando el top 15 de variables más relevantes. Si bien los resultados no presentan mejoras significativas en las métricas generales, como se observa en la figura 34, el ahorro en carga computacional y la mayor facilidad de interpretación para el negocio justifican la adopción de esta estrategia.

Figura 33 Métricas desempeño top 15 variables.



```
=== 📊 MÉTRICAS DE ENTRENAMIENTO ===
              precision    recall  f1-score   support

           0       0.97      0.81      0.88      7140
           1       0.51      0.89      0.65      1593

    accuracy                           0.83      8733
   macro avg       0.74      0.85      0.77      8733
weighted avg       0.89      0.83      0.84      8733

AUC-ROC Train: 0.9279
AUC-PR Train : 0.7732

=== 📊 MÉTRICAS DE TEST ===
              precision    recall  f1-score   support

           0       0.97      0.81      0.88      3061
           1       0.52      0.90      0.65       683

    accuracy                           0.83      3744
   macro avg       0.74      0.85      0.77      3744
weighted avg       0.89      0.83      0.84      3744

AUC-ROC Test: 0.9274
AUC-PR Test : 0.7773
🧩 Matriz de confusión:
[[2484  577]
 [  70  613]]
```

Fuente: Elaboración propia

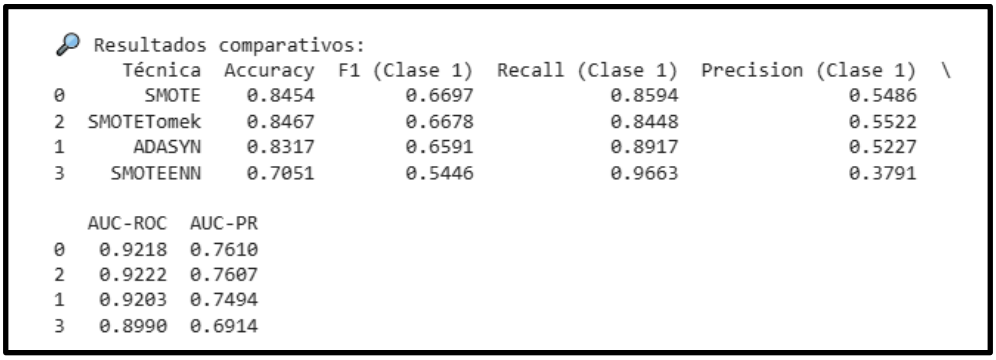
En cuanto a la matriz de confusión, al comparar con el modelo que utiliza todas las variables, se observa que este último logra identificar correctamente la inasistencia de **8 pacientes adicionales**. Sin embargo, este beneficio viene acompañado de una disminución en la precisión, al incrementar en **42 los falsos positivos** respecto al modelo basado en las 15 variables seleccionadas.

La segunda estrategia consistió en implementar técnicas de balanceo, dado que el hospital tiene un porcentaje de inasistencia entre el 18% y 20%.

En un contexto de salud y logística hospitalaria, la inasistencia es el evento que se quiere anticipar para tomar decisiones: reprogramar, optimizar recursos, llenar vacantes con lista de espera, etc. Por tanto, se necesita un modelo que tenga alta sensibilidad (recall) para esta clase minoritaria.

Un modelo que no detecta inasistencias, aunque sea preciso en promedio, no es útil para intervención, en este caso la clase minoritaria de inasistencia no está bien representada, lo que dificulta a los diferentes modelos aprender patrones de esta clase. Para ello se compararon técnicas de balanceo: SMOTE: Sobremuestreo sintético de la clase minoritaria. Undersampling: Submuestreo de la clase mayoritaria. SMOTEENN: Combinación de SMOTE + eliminación de ruido con ENN. ADASYN: Variante de SMOTE que genera muestras difíciles. Los resultados de los modelos con las técnicas mencionadas previamente se observan en la figura 35 y los resultados del modelo sin balanceo se encuentran en la figura 34.

Figura 34 Métricas desempeño técnicas de balanceo.



🔎 Resultados comparativos:

	Técnica	Accuracy	F1 (Clase 1)	Recall (Clase 1)	Precision (Clase 1)	\
0	SMOTE	0.8454	0.6697	0.8594	0.5486	
2	SMOTETomek	0.8467	0.6678	0.8448	0.5522	
1	ADASYN	0.8317	0.6591	0.8917	0.5227	
3	SMOTEENN	0.7051	0.5446	0.9663	0.3791	

	AUC-ROC	AUC-PR
0	0.9218	0.7610
2	0.9222	0.7607
1	0.9203	0.7494
3	0.8990	0.6914

Fuente: Elaboración propia

En la figura 35, se puede ver que SMOTE obtiene el mayor F1-score (0.6697), lo que refleja un buen equilibrio entre precisión y recall para la detección de inasistencias. Por su parte, SMOTETomek presenta la mayor precisión (0.5522) y un F1-score prácticamente equivalente al de SMOTE. Además, alcanza los valores más altos en AUC-ROC y AUC-PR, lo que evidencia una muy buena capacidad discriminativa entre las clases.

Si bien SMOTEENN logra identificar la mayor cantidad de pacientes inasistentes (recall de 0.9663), su precisión es considerablemente baja (0.3791), lo cual genera un número elevado de falsos positivos. Desde una perspectiva operativa, esto podría traducirse en una sobreocupación innecesaria de los recursos clínicos, restando confiabilidad a su aplicación.

En conjunto, SMOTETomek y SMOTE se posicionan como las técnicas más equilibradas y robustas, especialmente dado que el objetivo es alcanzar un balance adecuado entre precisión y sensibilidad.

No obstante, en lugar de balancear la base de entrenamiento en una proporción 50/50, se optó por una ratio de **30% de casos de inasistencia y 70% de asistencia**, en concordancia con la distribución real observada en los datos históricos. Esta decisión busca mejorar la capacidad de generalización del modelo en escenarios reales, evitando una sobreestimación de la clase minoritaria que podría derivar en un aumento excesivo de falsos positivos. Además, este enfoque permite preservar una mayor variabilidad de la clase mayoritaria, optimizando la precisión sin sacrificar significativamente el recall como se observa en la figura 34.

A continuación, en la tabla 5, se comparan los resultados de las diferentes técnicas de balanceo entendiendo que la prioridad en este caso es mantener el mejor F1 posible y precisión, la mejor alternativa es mantener un balance del 30% con la técnica de Smotetomek.

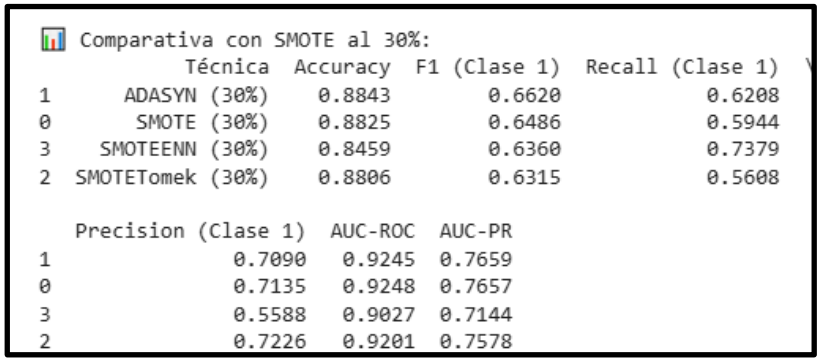
Tabla 5 Métricas desempeño por tipo de balanceo

Técnica	F1 Clase 1 (50%)	F1 Clase 1 (30%)	Recall (50%)	Recall (30%)	Precisión (50%)	Precisión (30%)
SMOTE	**0.6697**	0.6486	**0.8594**	0.5944	0.5486	**0.7135**
SMOTETomek	0.6678	0.6315	0.8448	0.5608	0.5522	**0.7226**
ADASYN	0.6591	**0.6620**	0.8917	0.6208	0.5227	0.7090
SMOTEENN	0.5446	0.6360	**0.9663**	**0.7379**	0.3791	0.5588

Fuente: Elaboración propia

Continuando con la evaluación del modelo usando la técnica de balanceo Smotetomek,en la figura 36, se muestra que las métricas de AUC-ROC y AUC-PR de este modelo son acordes a lo esperado.

Figura 35 Métricas desempeño técnicas de balanceo al 30%



📊 Comparativa con SMOTE al 30%:

	Técnica	Accuracy	F1 (Clase 1)	Recall (Clase 1)	Precision (Clase 1)	AUC-ROC	AUC-PR
1	ADASYN (30%)	0.8843	0.6620	0.6208	0.7090	0.9245	0.7659
0	SMOTE (30%)	0.8825	0.6486	0.5944	0.7135	0.9248	0.7657
3	SMOTEENN (30%)	0.8459	0.6360	0.7379	0.5588	0.9027	0.7144
2	SMOTETomek (30%)	0.8806	0.6315	0.5608	0.7226	0.9201	0.7578

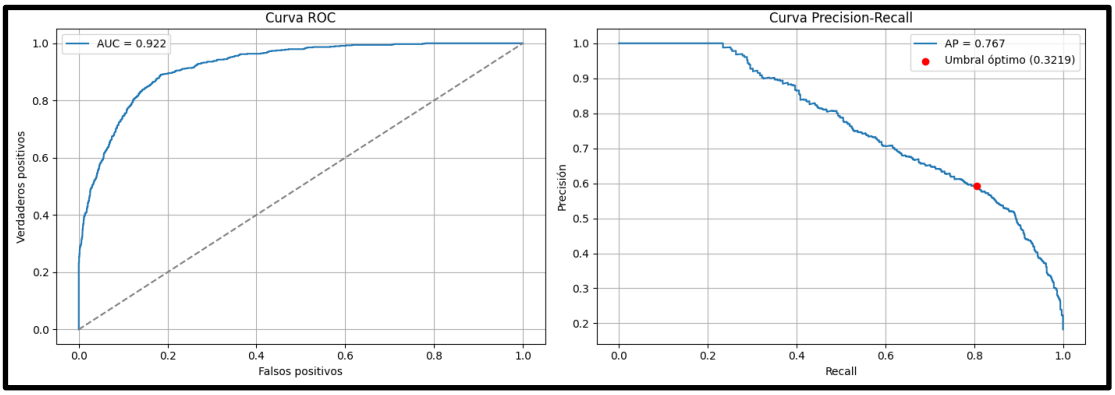
Fuente: Elaboración propia

Como tercera estrategia se ajustaron los umbrales de decisión del modelo XGBoost con técnica de balanceo Smotetomek, Smote y sin balancear, buscando un equilibrio entre la mejor precisión y recall. Los resultados de esta estrategia se observan en las figuras 37 -44.

En la figura 37. se observa la curva ROC y la curva Precision-Recall para el modelo con balanceo Smotetomek y umbral 0.322, lo cual permite evaluar la capacidad del modelo para distinguir entre clases y su rendimiento en la predicción de la clase minoritaria.

XGBoost con técnica de balanceo Smotetomek - Umbral 0.322

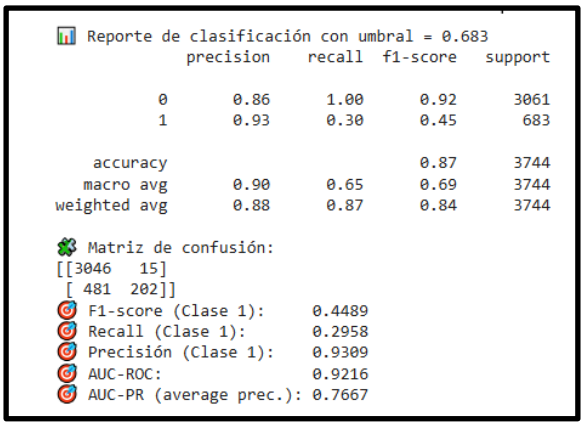
Figura 36 Curva ROC-CPR



Fuente: Elaboración propia

La Figura 38 complementa este análisis con las métricas de desempeño (precisión, recall, F1-score y matriz de confusión), que permiten observar el equilibrio alcanzado entre sensibilidad y especificidad.

Figura 37 Métricas desempeño con balanceo Smotetomek + umbral optimizando el F1

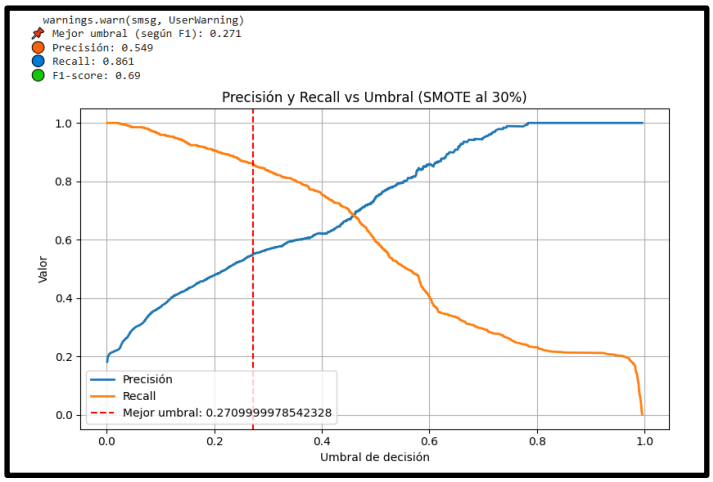


Fuente: Elaboración propia

La Figura 39 presenta la relación entre precisión y recall a medida que varía el umbral de decisión para el modelo con balanceo SMOTE y umbral 0,27, lo que facilita la identificación del punto óptimo de corte.

XGBoost con técnica de balanceo Smote - Umbral 0.2771

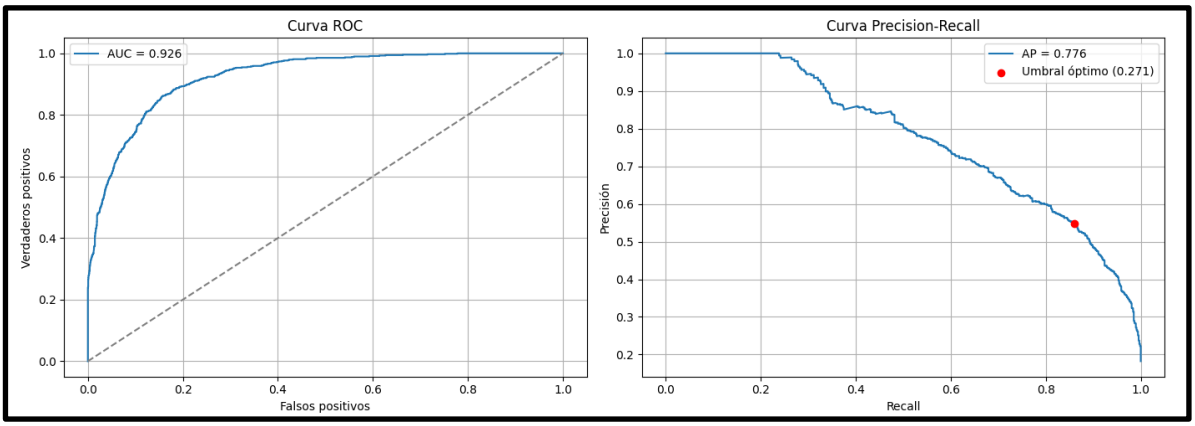
Figura 38 Precisión vs Recall



Fuente: Elaboración propia

La Figura 40 muestra las curvas ROC y Precision-Recall lo cual permite demostrar un buen rendimiento del modelo nuevamente las curvas ROC y Precision-Recall para esta configuración.

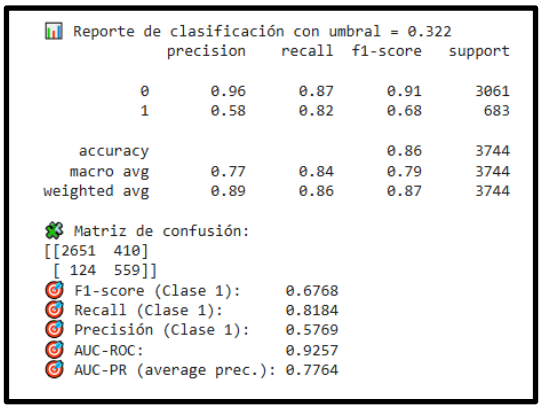
Figura 39 Curva ROC y CPR con balanceo SMOTE + umbral optimizando el F1



Fuente: Elaboración propia

La Figura 41 detalla las métricas de desempeño obtenidas con el umbral optimizado. Estas visualizaciones permiten comparar el impacto del tipo de balanceo y del umbral sobre el rendimiento del modelo.

Figura 40 .Métricas desempeño con balanceo SMOTE + umbral optimizando el F1

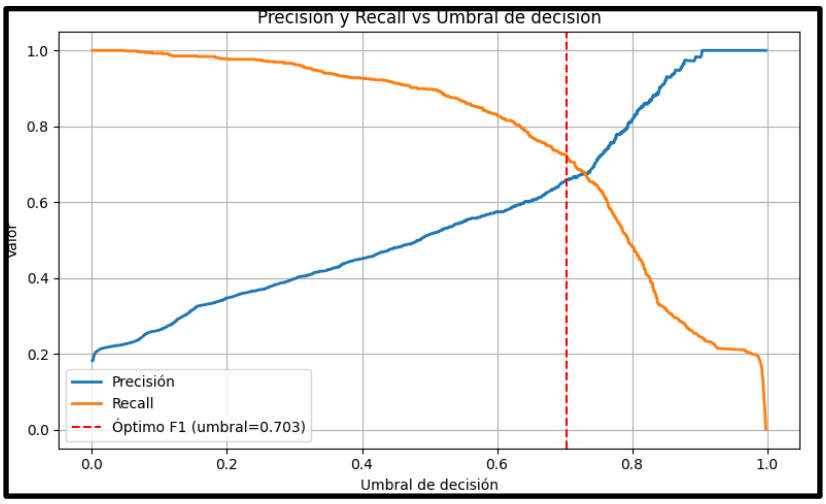


Fuente: Elaboración propia

La Figura 42 presenta la relación entre precisión y recall a medida que se ajusta el umbral de decisión para el modelo XGBoost sin aplicar técnicas de balanceo. Esta visualización permite identificar el punto óptimo de corte, que en este caso se encuentra en un umbral de 0.703. Este valor representa el mejor compromiso entre capturar la mayor cantidad de inasistencias (recall) y mantener una tasa aceptable de falsos positivos (precisión).

XGBoost sin balanceo - Umbral 0.703

Figura 41



Fuente: Elaboración propia

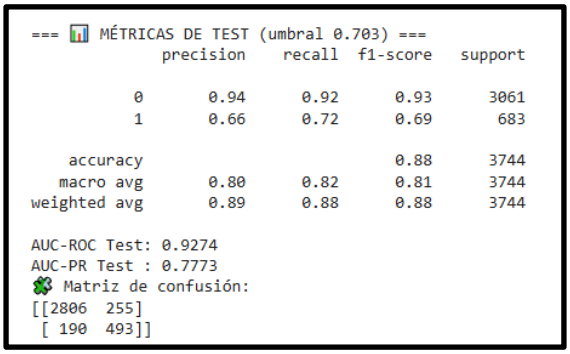
La Figura 43 muestra las métricas de desempeño obtenidas con este umbral optimizado. Se observa un F1-score competitivo, acompañado de una precisión del 66% y un recall del 72% para la clase de inasistencia. Estos resultados reflejan un modelo que, aunque no utiliza técnicas de balanceo, logra un rendimiento razonable gracias al ajuste fino del umbral, lo que lo convierte en una alternativa viable en escenarios donde se prefiera evitar la generación de datos sintéticos.

Es relevante señalar que se identificó previamente que la base de datos presentaba un desbalance en la proporción de clases (80/20), lo cual podía representar un inconveniente para el análisis y desempeño de los modelos. Por este motivo, se abordó el problema desde diferentes perspectivas: se evaluaron diversas técnicas de balanceo y, adicionalmente, en el

código del modelo XGBoost se incorporó la técnica *scale_pos_weight* para compensar el desbalance durante el entrenamiento.

A pesar de estas estrategias, se observó que el modelo XGBoost con optimización del umbral basado en la métrica F1 logró mitigar en gran medida los efectos del desbalance, proporcionando una herramienta robusta para controlar su impacto. En consecuencia, y considerando que las técnicas de balanceo implican la generación de datos sintéticos que podrían afectar la representatividad de la muestra, se tomó la decisión de trabajar con la base original para preservar la integridad de los datos.

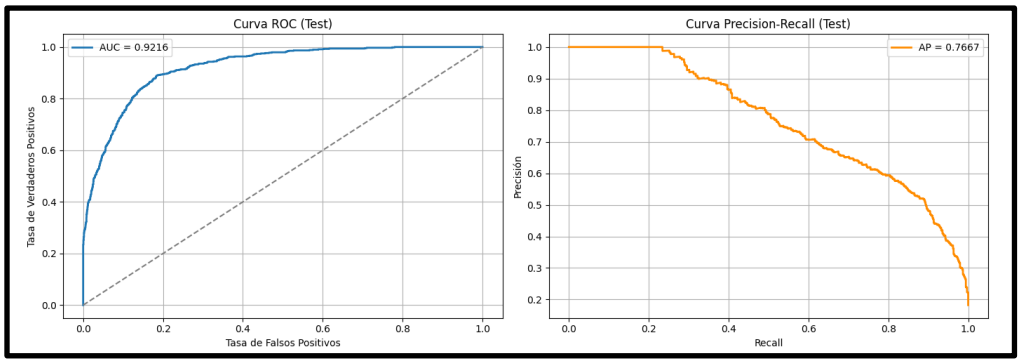
Figura 42



Fuente: Elaboración propia

La Figura 44 presenta las curvas ROC y Precision-Recall para esta configuración. Con un AUC-ROC de 0.9216 y un AUC-PR de 0.7767, el modelo demuestra una alta capacidad discriminativa y un buen equilibrio entre precisión y sensibilidad. Estos valores confirman que, incluso sin balanceo, el modelo puede adaptarse eficazmente al desbalance de clases mediante el ajuste del umbral, manteniendo un rendimiento competitivo frente a las versiones balanceadas.

Figura 43 *Métricas desempeño sin balanceo + umbral optimizando el F1*



Fuente: Elaboración propia

En conclusión, el proceso de modelado predictivo permitió construir una solución analítica robusta y adaptada al contexto clínico-operativo del Hospital Universitario Méderi, mediante la aplicación de técnicas de aprendizaje supervisado, transformación de variables, estrategias de balanceo y ajuste de umbrales. A lo largo de este capítulo se documentaron las decisiones metodológicas clave que guiaron la selección del modelo XGBoost como la alternativa más adecuada para anticipar la inasistencia de pacientes a sus sesiones de quimioterapia, abordando desafíos como el desbalance de clases, la fuga de datos y la necesidad de interpretabilidad operativa. La siguiente sección (10.1) presentará la evaluación integral del modelo seleccionado, incluyendo su desempeño en condiciones simuladas de producción y su comparación con referentes internacionales, con el fin de validar su utilidad como herramienta de apoyo en la gestión eficiente del servicio.

9.2. Modelo prescriptivo

Además del componente predictivo, el proyecto incorpora un modelo prescriptivo de optimización enfocado exclusivamente en una etapa específica de la ruta oncológica: la asignación

y ubicación de los pacientes en sala para sus sesiones de quimioterapia. Este modelo está diseñado para apoyar el proceso de programación, maximizando la eficiencia en el uso de los recursos disponibles (sillas, camas y turnos) dentro del Hospital Universitario Méderi.

Este modelo fue desarrollado utilizando la técnica de programación lineal entera mixta (MILP) mediante la librería PuLP en Python. La formulación matemática permite simular un entorno real de programación hospitalaria, considerando múltiples restricciones clínicas, logísticas y de negocio.

El modelo prescriptivo parte de la misma base de datos estructurada que alimentó al modelo predictivo, y utiliza variables relevantes tanto clínicas como operativas. Entre ellas se incluyen:

- **ID del paciente (HISTORIA)** Identificador único para garantizar asignación exclusiva a una unidad y turno.
- **Duración del tratamiento (DURACIÓN)** Expresada en minutos. Fundamental para definir si el tratamiento es ambulatorio y qué combinaciones de turnos pueden cubrirlo.
- **Protocolo estandarizado** Clasificación del tratamiento para agrupar combinaciones de medicamentos similares.
- **Tipo de unidad requerida (silla o cama)** Determinado en función de la duración del tratamiento. Por ejemplo, tratamientos de más de 240 minutos se asignan preferiblemente a camas.
- **Turnos disponibles** Cuatro turnos regulares definidos por el hospital: mañana (M), medio día (MD), tarde (T) y tarde-tarde (TT), más un turno extendido (EXT) utilizado solo como último recurso, se utiliza para protocolos de 600 minutos.

- **Horario clínico** Se permite que un tratamiento inicie hasta las 4:30pm y finalice a más tardar a las 7:30pm, (no es el escenario ideal y lo permite solo para casos puntuales).
- **Unidad física** Identifica cada silla o cama en las salas disponibles (sala 1: 22 sillas + 2 camas; sala 2: 7 sillas).

El objetivo del modelo es maximizar el uso eficiente del recurso tiempo, lo cual se traduce en:

Maximizar la cantidad total de minutos asignados a pacientes, favoreciendo tratamientos más largos (que son los turnos más complejos de asignar), pero manteniendo equidad en la asignación y cumplimiento de restricciones clínicas.

Formulación Matemática del Modelo de Asignación

Conjuntos

I: Conjunto de pacientes

U: conjunto de unidades disponibles (sillas o camas),

B: Conjunto de bloques de turnos

Parámetros

T_b: {conjunto de turnos que componen el bloque $b \in B$}

D_i: {Duración del tratamiento del paciente $i \in B$, en minutos}

H_t: {Duración del turno $t \in B$, en minutos}

S_t: {Hora de inicio del turno t ∈ T, en minutos}

$Tipo_u$: {Tipo de unidad u ∈ U}

Variable de decisión

$$Z_{iub}=\begin{cases}1 \text{ } si\ el\ paciente\ i \in I\ es\ asignado\ a\ la\ unidad\ u \in U\ en\ el\ bloque\ b \in B \\ 0\ de\ lo\ contrario\end{cases}$$

Función objetivo

Maximizar el uso del tiempo total asignado a pacientes:

$$\max Z = \sum\{i \in I\} \sum\_\{u \in U\} \sum\_\{b \in B\} D_i * Z_{iub}$$

Restricciones

1. Asignación única por paciente

$$\sum\_\{u \in U\} \sum\_\{b \in B\} Z_{iub} \leq 1 \quad \forall\ i \in I$$

2. No solapamiento de pacientes en una misma unidad

$$\sum\_\{i \in I\} \sum\_\{b \in B: t \in T_b\} Z_{iub} \leq 1 \quad \forall\ u \in U, \forall\ t \in T_b$$

3. Compatibilidad entre duración y tipo de unidad

$Z_{iub} = 0$ si $D_i \leq 240$ y $tipo(u) = cama \quad \forall\ i \in I, u \in U, b \in B$

4. Bloque suficientemente largo para cubrir tratamiento

$$\sum\_\{t \in T_b\} H_t * Z_{iub} \geq D_i * Z_{iub} \quad \forall\ i \in I, \forall\ u \in U, \forall\ b \in B$$

5. Restricción de horario clínico

$\sum_{t \in T_b} H_t * Z_{iub} + S_{\min(T_b)} * Z_{iub} \leq 1170 \quad \forall\ i \in I, u \in U, b \in B$

$S\{\min(T_b)\} * Z_{iub} \leq 990 \quad \forall\ i \in I, \forall\ u \in U, \forall\ b \in B$

6. Uso limitado del turno EXT

Esto significa que una unidad solo puede ser usada en EXT si estuvo completamente desocupada durante los turnos normales.

Formalmente:

Si una unidad u fue utilizada en algún turno M, MD, T o TT, entonces no puede ser usada en EXT:

$$Si \sum Z_{i,u,b} \geq 1 \text{ con } t \in b \text{ y } t \in \{M, MD, T, TT\} \Rightarrow \sum Z_{i,u,b} = 0 \text{ con EXT} \in b$$

En cuanto a las salidas del modelo, produce una serie de entregables que permiten su implementación práctica:

- **Archivo de programación diaria:** Listado con ID del paciente, unidad asignada, turno(s), horario y duración.
- **Validación de asignación única:** Indicador de pacientes asignados correctamente a una única unidad física.
- **Visualización tipo Gantt:** Gráfico generado automáticamente que muestra la ocupación por unidad y turno, facilitando la interpretación para jefes de programación y personal clínico.

Respecto a la validación del modelo, es importante mencionar que se aplicaron criterios tanto estructurales como empíricos para asegurar su correcto funcionamiento. En primer lugar, se verificó el cumplimiento del 100 % de las restricciones previamente definidas, incluyendo: asignación única por paciente, no solapamiento de unidades, compatibilidad entre duración del tratamiento y tipo de unidad (camas reservadas para tratamientos largos), continuidad de turnos y cumplimiento del horario clínico (inicio antes de las 16:30 y finalización antes de las 19:30). Estas validaciones se realizaron mediante inspección directa del archivo de resultados, el cálculo de métricas automatizadas y la generación de una visualización tipo Gantt que permitió corroborar gráficamente la no superposición entre pacientes en una misma unidad.

Desde el punto de vista empírico, se evaluó la eficiencia del modelo con base en la cantidad de pacientes asignados, la necesidad de utilizar el turno EXT y la capacidad del modelo para ubicar el 100 % de los pacientes en días de baja demanda. Se priorizó la asignación completa dentro de los turnos regulares (M, MD, T, TT), utilizando el turno EXT únicamente como último recurso en unidades no ocupadas previamente. Adicionalmente, se cuantificaron los pacientes recuperados en la etapa de reasignación extendida y se verificó la cobertura horaria y el uso eficiente del recurso tiempo.

Es importante aclarar que, debido a que la base de datos utilizada corresponde a una programación histórica previamente realizada por el hospital, los resultados generados por el modelo no son directamente comparables con dicha programación original, ya que el modelo parte de un escenario de asignación desde cero, sin tener en cuenta decisiones previas, sobre agendamientos manuales u otras condiciones no modeladas.

Si bien el modelo desarrollado demuestra ser una herramienta robusta y útil para apoyar el proceso de programación de pacientes en quimioterapia, existen varias **limitaciones operativas y técnicas** que deben tenerse en cuenta:

Horizonte de planificación diario:
El modelo está diseñado para optimizar la programación en un único día. No contempla la planificación multidiaria o longitudinal, lo cual sería relevante para pacientes con ciclos de tratamiento recurrentes o sesiones distribuidas a lo largo de la semana.

Asignación indivisible por bloques completos:
El modelo solo permite asignar a los pacientes a bloques **consecutivos completos** de turnos (por ejemplo, M+MD+T), siempre y cuando la suma de sus duraciones cubra la duración del tratamiento. **No se permite fraccionar un tratamiento dentro de un turno** (ej. usar solo 60 minutos del turno MD) ni repartirlo entre turnos no consecutivos (ej. M + T).

Sin prioridad clínica o de urgencia:
Todos los pacientes son tratados por igual en la función objetivo. El modelo no incorpora ponderaciones por gravedad clínica, prioridad médica, fecha del último tratamiento o tipo de cáncer, lo cual podría ser necesario en la práctica para casos urgentes o pacientes oncológicos con condiciones críticas.

Asignación en turnos fijos, no por hora exacta:
Aunque la solución calcula la hora exacta de inicio y fin dentro del bloque, las asignaciones están limitadas por los bloques predefinidos del hospital. No se permite ajustar dinámicamente los horarios fuera de los turnos M, MD, T, TT o EXT.

Uso del turno EXT solo como contingencia: El turno extendido solo se activa en una segunda etapa, lo cual limita su uso incluso cuando podría ayudar a maximizar la eficiencia general en escenarios de alta demanda. Además, solo se permite usar EXT en unidades que no hayan sido utilizadas previamente.

No integración directa con la probabilidad de inasistencia: Aunque el proyecto incluye un modelo predictivo de inasistencia, esta probabilidad no se incorpora aún dentro del modelo prescriptivo por solicitud del hospital. No se prioriza la asignación de pacientes con mayor probabilidad de asistir ni se dejan cupos disponibles para sobre agendamiento inteligente.

Sin reubicación dinámica ante cancelaciones o fallos: El modelo opera bajo el supuesto de información completa y estable. No contempla escenarios dinámicos como cancelaciones de último minuto, reubicaciones durante el día, o eventos inesperados (por ejemplo, retrasos en administración de medicamentos).

No considera restricciones de personal ni equipos médicos: El modelo asume que la disponibilidad de personal clínico y recursos adicionales (bombas de infusión, monitores, etc.) es constante y suficiente, lo cual no siempre se ajusta a la realidad hospitalaria.

Se espera que la implementación de esta herramienta le permita al hospital **aumentar la eficiencia operativa** del servicio, al reducir la cantidad de sillas vacías en turnos críticos, así como **incrementar la facturación hospitalaria**, mediante una mejor utilización del recurso físico sin necesidad de inversiones adicionales en infraestructura.

Actualmente, los modelos predictivos de inasistencia y prescriptivo de asignación fueron desarrollados de manera independiente, conforme a la solicitud del Hospital Universitario Méderi. Esta decisión responde a una consideración clave del entorno clínico: aunque el modelo predictivo estima la probabilidad de inasistencia, no se puede tomar como una certeza absoluta en un contexto médico, donde está en juego la salud de las personas. Por lo tanto, no es posible condicionar la asignación operativa directamente a dicha predicción sin la validación correspondiente. Esta precaución refuerza la necesidad de mantener ambos modelos separados en esta fase del proyecto.

El modelo predictivo de inasistencia tiene como propósito anticipar la ausencia de pacientes a las sesiones de quimioterapia. Sus resultados permiten generar una lista de espera priorizada, de manera que, ante una ausencia confirmada, se pueda contactar a otro paciente y aprovechar mejor el recurso disponible.

El modelo prescriptivo de asignación está orientado a optimizar el uso de las sillas y camas disponibles, distribuyendo de forma eficiente a los pacientes durante los diferentes turnos. Este modelo busca mitigar la sobreocupación en días de alta demanda y mejorar el uso de la sala cuando hay menor asistencia, garantizando un balance operativo.

Si bien en la actualidad ambos modelos operan por separado, como recomendación futura, se sugiere evaluar su integración bajo la supervisión del equipo médico. Esto permitiría construir un sistema más robusto de apoyo a decisiones, en el que las predicciones se usen como insumo complementario, siempre respetando los criterios clínicos y las condiciones particulares de cada paciente.

Tal como se identificó en la fase de entendimiento del negocio, el proceso actual presenta desafíos importantes como el sobreagendamiento, la existencia de turnos parcialmente

desaprovechados y una ejecución fragmentada del flujo operativo, dado que la programación y la ubicación física de los pacientes son gestionadas por equipos distintos. En este contexto, la implementación de soluciones analíticas, aunque inicialmente independientes por solicitud del hospital, permite abordar de manera específica estas problemáticas. La solución prescriptiva, por ejemplo, automatiza simultáneamente la asignación y ubicación de pacientes, lo que facilita aprovechar en tiempo real los espacios disponibles y aumentar el número de pacientes atendidos por jornada. En conjunto, estas iniciativas contribuyen al cumplimiento de las metas institucionales y fortalecen la eficiencia del servicio oncológico, abriendo la puerta a una futura integración que potencie aún más su impacto operativo y clínico.

10. Evaluación

Una vez completado el proceso de modelado, la efectividad de los modelos desarrollados será evaluada utilizando los datos históricos del hospital. Específicamente, se comparará el **indicador de giro de silla** en el proceso actual frente al obtenido mediante la herramienta analítica propuesta. Además del análisis cuantitativo, se incluirá una **validación cualitativa con el personal del área de oncología**, para determinar si el modelo resulta útil en la práctica clínica y apoya efectivamente la toma de decisiones.

10.1. Modelo predictivo

En esta sección se evaluarán todos los modelos presentados previamente (ver sección 9.1),el objetivo del modelo predictivo es anticipar la inasistencia de los pacientes a sus sesiones de quimioterapia y, con ello, que el hospital pueda habilitar acciones que mejoren la eficiencia y continuidad del tratamiento. En este sentido, se proponen al hospital **dos líneas de intervención**:

1. **Gestión proactiva**: los jefes de programación podrían reforzar el contacto con aquellos pacientes que el modelo identifica con mayor probabilidad de inasistencia, con el fin de asegurar su asistencia o reprogramar de manera oportuna, evitando pérdida de medicamentos y recursos clínicos.
2. **Listas de espera dinámicas**: se podría establecer un sistema de lista de espera sobre un porcentaje de pacientes con alto riesgo de inasistencia, permitiendo asignar a otros pacientes en caso de confirmarse una cancelación o ausencia, optimizando así la ocupación de las unidades.

Dado este enfoque, se seleccionaron como métricas de evaluación del modelo el F1-Score, la precisión, el recall, el AUC y el AUC-PR, con el fin de medir su capacidad predictiva en un contexto con desbalance de clases. El F1-Score permite valorar el equilibrio entre precisión y recall, lo que resulta útil para entender el desempeño general del modelo frente a los casos positivos. La precisión ayuda a evaluar cuántas de las alertas generadas por el modelo realmente corresponden a pacientes que no asistirán, minimizando así la generación de falsos positivos. Por su parte, el recall indica qué proporción de los pacientes que efectivamente no asistirán fueron correctamente identificados, lo cual es relevante para anticiparse a escenarios de baja ocupación. El AUC mide la capacidad general del modelo para diferenciar entre clases (asistencia e inasistencia), mientras que el AUC-PR ofrece una evaluación más detallada del equilibrio entre precisión y recall, especialmente útil cuando se trabaja con clases desbalanceadas. Estas métricas permiten comparar el desempeño entre modelos y seleccionar aquel con mejor capacidad para apoyar decisiones operativas en el servicio.

Para la evaluación de los modelos en este proyecto, se establecieron dos grupos. El primero incluye diversos tipos de modelos utilizando todas las variables recolectadas, mientras que el

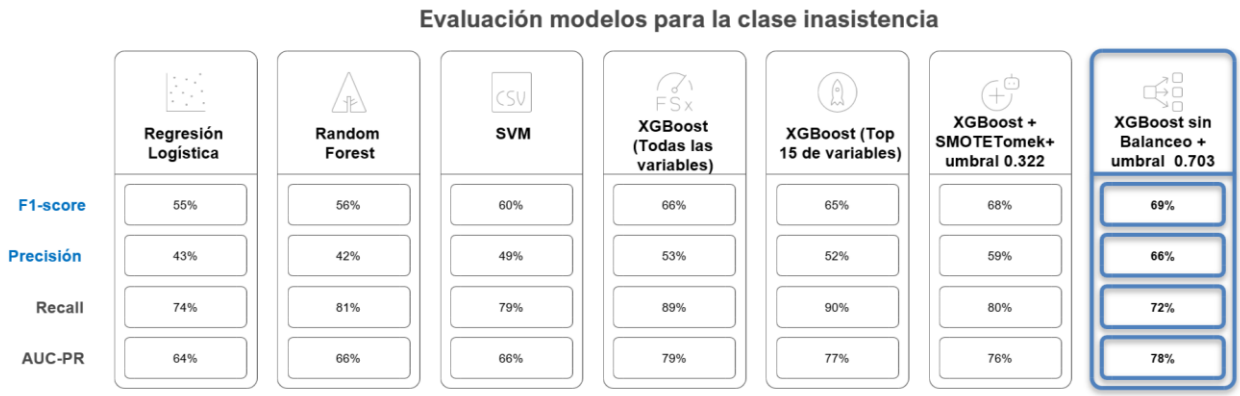
segundo se enfoca en el top 15 de variables, aplicando diferentes variaciones como balanceo de clases, optimización del umbral de decisión y una versión sin técnicas adicionales.

Los modelos del primer grupo resultaron poco eficientes debido a los tiempos prolongados requeridos para su implementación. Sin embargo, el análisis de este grupo permitió identificar que el modelo XGBoost obtuvo el mejor desempeño, al presentar la menor cantidad de falsos positivos y falsos negativos, así como los porcentajes más altos en todas las métricas evaluadas para ambas clases (clase 0 y clase 1). Esto indica que XGBoost es el modelo que mejor maneja la distribución de clases del problema y ofrece una predicción más equilibrada. En contraste, el modelo con peor desempeño fue Random Forest, ya que mostró una menor precisión en la clase 1 y un mayor número de falsos positivos, lo cual es especialmente crítico para el indicador de "giro silla", donde la correcta identificación de la clase 1 es fundamental.

El segundo grupo de modelos se caracteriza por su buena rapidez de ejecución, facilidad de implementación y un desempeño sólido en la predicción de la clase 0. Sin embargo, aunque la predicción de la clase 1 presenta un buen F1-Score, el indicador más crítico para el caso del giro silla la precisión se mantiene en niveles bajos. Dentro de este grupo de modelos, el de peor desempeño fue XGBoost con balanceo SMOTETomek y un umbral de 0.32. Aunque las métricas son similares entre modelos, esta versión introduce mayor incertidumbre al generar datos sintéticos, sin aportar mejoras significativas en los resultados. Esto evidencia que, en este caso, el balanceo no solo es innecesario, sino que puede afectar negativamente el rendimiento, especialmente considerando que XGBoost ya cuenta con un manejo interno eficiente de clases. Por otra parte, el mejor modelo del grupo y del proyecto en general, fue XGBoost con la selección de las 15 variables más representativas, sin técnicas de balanceo y con un umbral óptimo de 0.32. Este modelo cuenta con una buena predicción de la clase 0 con un F1-Score de 93%, Auc 92%, un bajo porcentaje de

falsos positivos 8%; sin embargo, cuenta con un rendimiento moderado en la predicción de la clase de inasistencia.

Figura 44 Evaluación de desempeño modelos



Fuente: Elaboración propia

La figura 45 muestra el modelo seleccionado para predecir la inasistencia de los pacientes fue XGBoost, utilizando las 15 variables más representativas, sin técnicas de balanceo y con un umbral óptimo de 0.32, por las razones expuestas en el párrafo anterior. Entre las variables más relevantes se encuentran:

Asistencia en fin de semana: cantidad de veces que el paciente ha asistido a citas programadas en fines de semana.

Días desde la última cita: número de días transcurridos entre citas programadas.

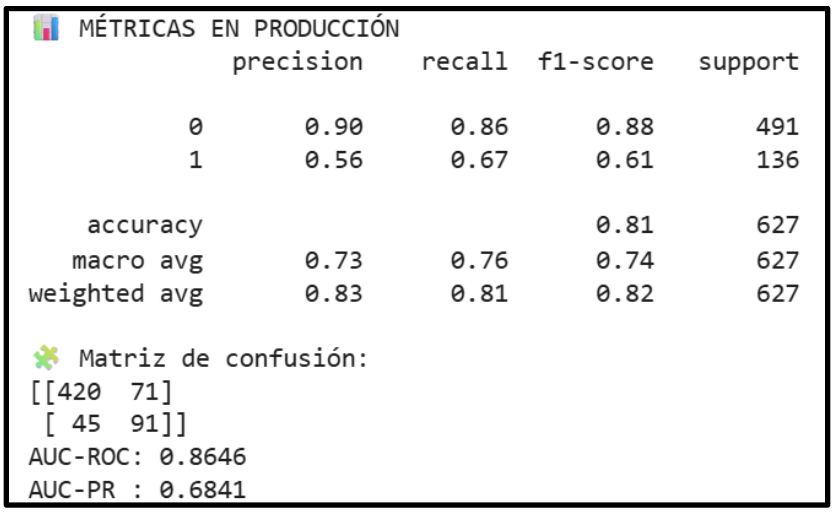
Oportunidad de inicio de quimioterapia (QMT): tiempo entre ciclos de quimioterapia.

Día de la semana: día específico en el que se programa la cita (lunes a domingo).

Duración del tratamiento: duración en horas de la aplicación del tratamiento según el protocolo.

Estas variables son clave, ya que orientan al hospital en la formulación de estrategias para mejorar el indicador de inasistencia. Este modelo fue probado en un entorno de producción, obteniendo los resultados que se muestran en la figura 46.

Figura 45 Resultados puesta en producción 1 al 15 de septiembre 2024



📊 MÉTRICAS EN PRODUCCIÓN

	precision	recall	f1-score	support
0	0.90	0.86	0.88	491
1	0.56	0.67	0.61	136
accuracy			0.81	627
macro avg	0.73	0.76	0.74	627
weighted avg	0.83	0.81	0.82	627

🧩 Matriz de confusión:

```
[[420  71]
 [ 45  91]]
AUC-ROC: 0.8646
AUC-PR : 0.6841
```

Fuente: Elaboración propia

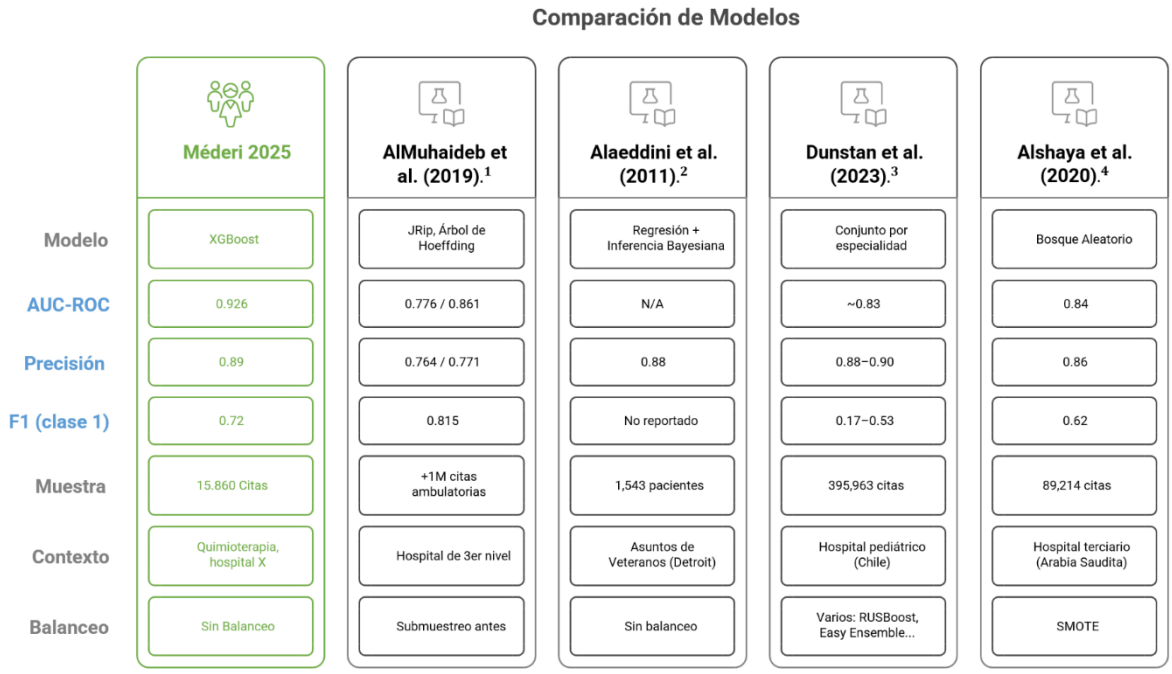
En la figura 46 se observa que el modelo, al enfrentarse a nuevos datos, mantiene una buena capacidad de predicción para la clase de asistencia, así como una adecuada clasificación general. Se redujeron los falsos positivos y, aunque los indicadores de la clase de inasistencia (precisión, recall y F1-score) disminuyeron ligeramente, el F1-score sigue siendo aceptable. Esto sugiere que el modelo continúa identificando correctamente a una proporción significativa de pacientes que no asistirán, aunque con una leve tendencia a clasificar como inasistentes a pacientes que sí asistirán (mayor recall que precisión).

Este modelo representa un valor significativo para Méderi, ya que permite anticipar qué pacientes podrían faltar a sus sesiones, facilitando acciones preventivas como el seguimiento personalizado o la creación de listas de espera dinámicas. No obstante, se recomienda considerar estas estrategias como parte de una siguiente fase del proyecto, especialmente al implementar

modelos de ensamble que puedan mejorar el F1-score y, en particular, la precisión de la clase 1, que es crítica para la toma de decisiones operativas.

Inicialmente, se realizó una comparación, como se observa en la figura 47, de distintos modelos predictivos dentro del desarrollo del proyecto del Hospital Universitario Méderi, identificando que el algoritmo **XGBoost** ofrece el mejor desempeño en términos de equilibrio entre precisión y capacidad de generalización. A partir de esta selección, se construyó el modelo final aplicado a la ruta oncológica del hospital universitario. Para darle mayor validez a los resultados obtenidos y contextualizar su desempeño, se decidió contrastar las métricas del modelo del Hospital Universitario Méderi con otros estudios internacionales que abordan la predicción de inasistencias en entornos hospitalarios. Como resultado, se elaboró una matriz comparativa que incluye investigaciones realizadas en distintos países, con muestras que van desde miles hasta millones de citas.

Figura 46 Comparativo proyectos sector salud



Fuente: Elaboración propia

Las métricas del modelo del Hospital Universitario Méderi se ubican dentro del promedio general observado, **destacándose particularmente en su precisión (0.89) y en el F1 Score para la clase de Inasistencia (0.69)**, lo cual refleja una capacidad sólida para identificar pacientes que probablemente se ausenten, sin comprometer la confiabilidad de las predicciones. Aunque estudios como el de AlMuhaideb et al. (2019) reportan un F1 Score más alto para la clase de inasistencia (0.81), el modelo de Méderi muestra un desempeño más balanceado, con el **AUC-ROC más alto (0.926)** entre todos los modelos comparados. En conclusión, la evaluación cruzada con estudios previos respalda la solidez del modelo desarrollado en el Hospital Universitario Méderi, posicionándose como una solución eficaz, robusta y alineada con las mejores prácticas internacionales para la predicción de inasistencias médicas.

10.2. Modelo prescriptivo

En cuanto a la decisión de utilizar un modelo de optimización basado en programación lineal entera mixta (MILP) y no otra alternativa (como heurísticas, simulación o modelos de aprendizaje automático):

La prioridad del hospital es cumplir reglas clínicas y operativas estrictas, lo cual requiere que las soluciones respeten todas las restricciones (no sólo aproximaciones). Así mismo, este modelo permite tomar decisiones precisas sobre la asignación o no de un paciente, por otra parte, el volumen de datos es manejable y la estructura del problema se ajusta perfectamente a un modelo de asignación binaria, lo que hace innecesario utilizar modelos más complejos que requieren un mayor uso computacional.

El modelo se puede adaptar fácilmente a cambios como nuevas restricciones (por ejemplo, prioridad por severidad clínica), ajustes en horarios, o disponibilidad variable de unidades. Esto permite que la herramienta sea sostenible y escalable a futuro. Una de sus mayores ventajas es que este modelo puede integrarse gradualmente con los sistemas actuales del hospital (como SIIFAM o Servinte), y ser comprendido por perfiles no técnicos como coordinadores clínicos o jefes de programación.

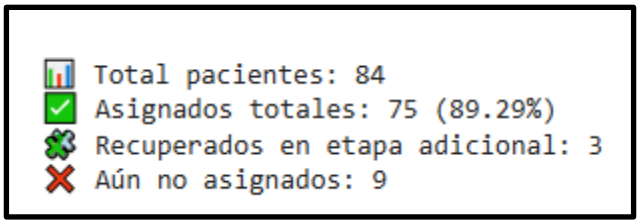
En cuanto a la evaluación del modelo, se realizaron ajustes progresivos a las restricciones con el objetivo de maximizar la función objetivo, utilizando como principal criterio el porcentaje de ocupación de las unidades disponibles. Para validar su desempeño, se seleccionaron datos históricos correspondientes a días con alta demanda de pacientes, así como días con menor volumen, con el fin de evaluar la capacidad del modelo para asignar el 100% de los pacientes en distintos escenarios operativos.

En promedio, el hospital atiende aproximadamente 62 pacientes diarios en el servicio de quimioterapia. A continuación, se presentan los resultados del modelo para dos escenarios contrastantes: el 1 de agosto de 2024, correspondiente a un día de alta demanda con 84 pacientes agendados, y el 10 de junio de 2024, día con baja demanda, con 26 pacientes agendados.

Resultados 1 de agosto (sobre agendamiento)

Los resultados del modelo se observan en la figura 48:

Figura 47 Resultados asignación pacientes 1 de agosto



Fuente: Elaboración propia

Genera la lista de los 9 pacientes que no han podido ser asignados, en este caso puntual no han sido asignados porque el 100% de los espacios se encuentran asignados como se observa en la tabla 6:

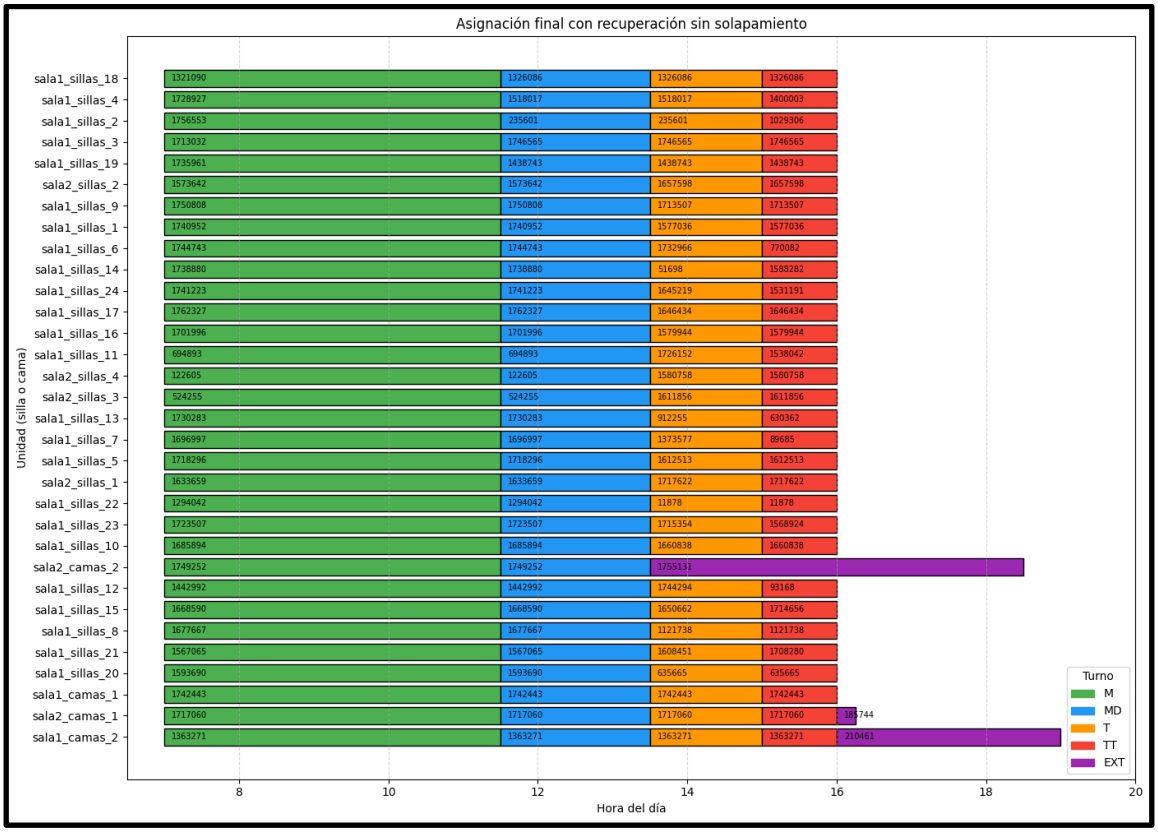
Tabla 6 Resultados pacientes sin asignación modelo optimización

aciente_ID	DURACION	Motivo
1392547	660	No se pudo asignar en ninguna etapa
1670239	300	No se pudo asignar en ninguna etapa
1662340	300	No se pudo asignar en ninguna etapa
1680869	300	No se pudo asignar en ninguna etapa
1581613	300	No se pudo asignar en ninguna etapa
1748658	300	No se pudo asignar en ninguna etapa
1756172	300	No se pudo asignar en ninguna etapa
1714237	300	No se pudo asignar en ninguna etapa
1760599	300	No se pudo asignar en ninguna etapa

Fuente: Elaboración propia

En la figura 49 se ilustra el diagrama de Gantt con la ubicación por historia clínica y por turno. Solo tres pacientes fueron asignados al turno EXT con salida 7pm.

Figura 48 Diagrama Gantt, resultados asignación óptima

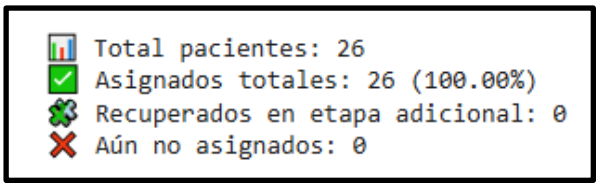


Fuente: Elaboración propia

Resultados 10 de junio (baja asignación)

Los resultados del modelo se muestran en la figura 50:

Figura 49 Resultados asignación pacientes 1 de agosto

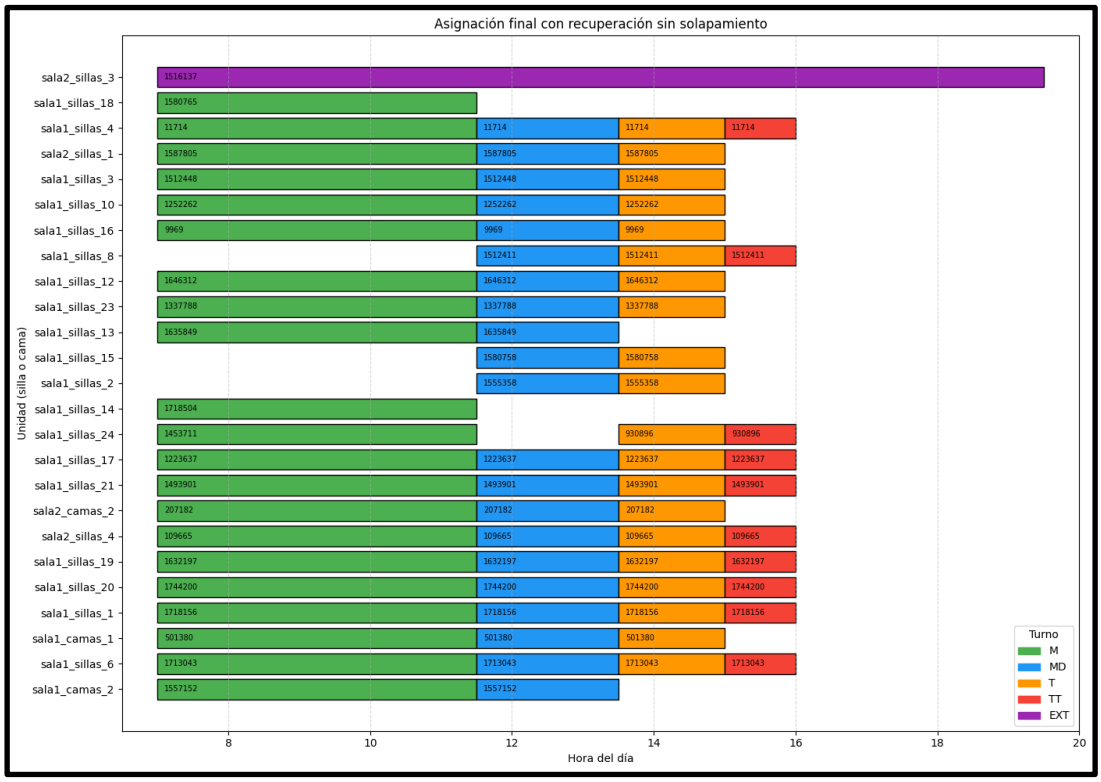


Fuente: Elaboración propia

No genera la lista de los pacientes no asignados, dado que se ubica el 100%

Diagrama de Gantt, de la figura 51, con la ubicación por historia clínica y por turno. Solo un paciente fue asignado al turno EXT con salida 7pm.

Figura 50 Diagrama Gantt, resultados asignación óptima



Fuente: Elaboración propia

Para estos casos, se espera que los jefes de programación puedan consultar el modelo en distintos momentos del día, con el fin de identificar el porcentaje de asignación alcanzado y los

espacios aún disponibles. De esta manera, podrán utilizar las listas de espera generadas por el modelo predictivo para ocupar de forma eficiente los cupos restantes.

Al evaluar el modelo con datos históricos de diferentes días, se observa que en jornadas con una demanda por debajo del promedio se logra una asignación del 100% de los pacientes. En contraste, en días con sobreocupación, el modelo permite alcanzar niveles de asignación entre el 92% y el 96% en promedio.

Si bien no es posible realizar una comparación estrictamente cuantitativa entre los resultados del modelo prescriptivo y la programación histórica realizada manualmente por el hospital —ya que el modelo parte de una programación ya establecida —, sí es posible identificar algunas diferencias cualitativas y operativas relevantes.

Criterio de asignación: El proceso manual depende en gran medida de la experiencia del personal de programación y puede incorporar decisiones subjetivas o no estandarizadas. Tal como se identificó en la fase de entendimiento del negocio, los jefes de programación no cuentan con visibilidad en tiempo real sobre las asignaciones que están realizando sus pares, lo que puede generar sobreocupación o uso ineficiente del recurso. En contraste, el modelo prescriptivo aplica un criterio uniforme, automatizado y transparente, maximizando el uso del recurso tiempo dentro de los límites clínicos establecidos. Esto permite a los jefes de programación concentrarse únicamente en los espacios aún disponibles o en los pacientes que no pudieron ser asignados inicialmente, ya sea para reagendarlos o para revisar los motivos de exclusión, como superar el tope de asignaciones por unidad.

Eficiencia operativa: En días de alta demanda, el modelo representa una herramienta de gran valor, ya que ejecuta en segundos la mejor combinación posible de asignaciones, tarea que actualmente puede tomar entre tres y cuatro horas de trabajo manual. Además de reducir

significativamente el tiempo operativo, minimiza el riesgo de errores humanos y permite al equipo clínico tomar decisiones informadas de manera oportuna.

11. Entregables

Durante el desarrollo del proyecto se construyó inicialmente un tablero descriptivo en Power BI que permitía visualizar el comportamiento general de la inasistencia a citas médicas en pacientes del servicio de quimioterapia. Este panel sirvió como punto de partida para analizar tendencias y patrones relevantes. Sin embargo, a partir de varias sesiones de trabajo conjunto con el equipo del Hospital Universitario Mayor Méderi, surgieron nuevas necesidades de información que requerían una herramienta más enfocada, práctica y alineada con la realidad operativa del hospital.

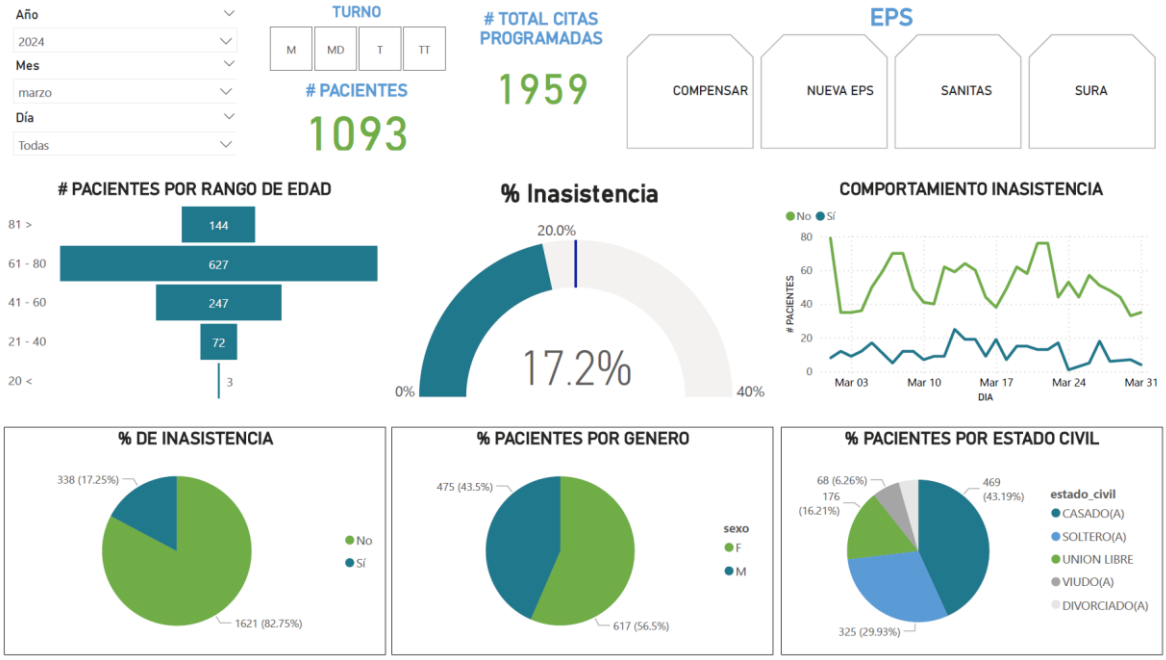
Como resultado de ese proceso colaborativo, se diseñó una nueva versión del dashboard que responde de manera más precisa a los requerimientos de los equipos clínico-asistenciales, incorporando indicadores clave y segmentaciones que permiten una lectura más clara de la situación.

A continuación, se presenta en detalle cada una de las hojas del dashboard final, destacando su estructura, funcionalidad y los principales hallazgos que aporta al análisis y monitoreo de la inasistencia en el contexto oncológico. Este recorrido busca no solo mostrar la herramienta, sino también evidenciar cómo la analítica puede convertirse en un insumo estratégico para la mejora continua de los servicios de salud.

En la figura 52 se muestra la primera hoja del dashboard presenta una vista general compuesta por indicadores clave como el total de pacientes, citas programadas, filtros por turno, EPS y fecha, así como gráficos que permiten analizar la distribución por edad, género, estado civil y el comportamiento diario de la asistencia. Su estructura está pensada para ofrecer una

comprensión rápida y flexible del panorama de inasistencia, permitiendo al usuario explorar la información de manera dinámica y detectar posibles patrones o segmentos críticos desde una perspectiva demográfica y temporal.

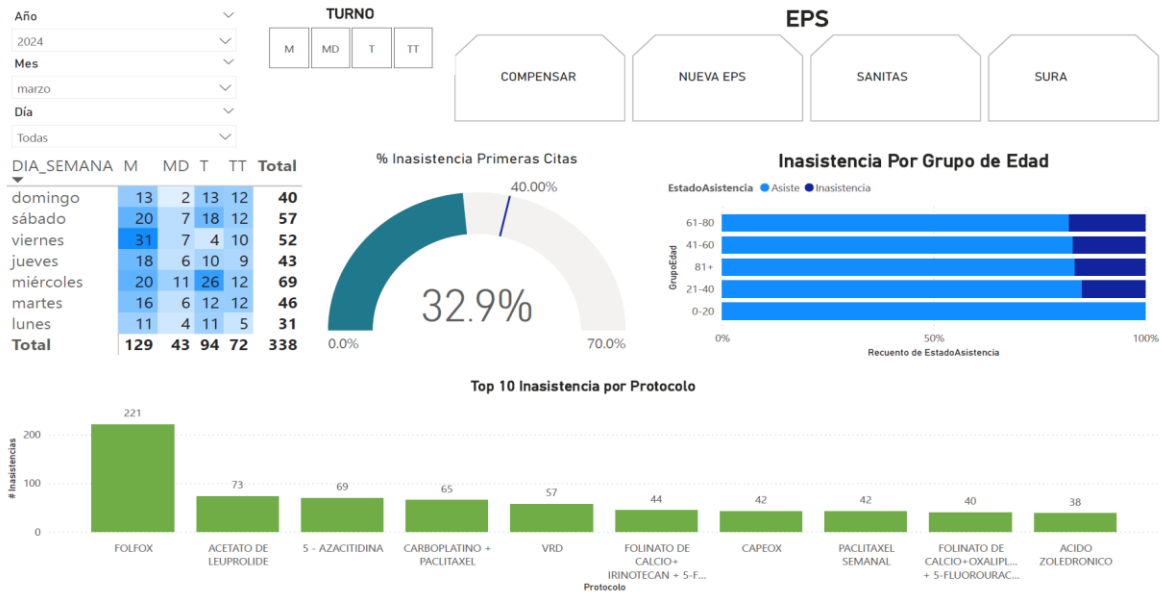
Figura 51 Visual Hoja 1 Tablero Méderi



Fuente: Elaboración propia

En la figura 53 se observa la segunda hoja del dashboard profundiza en el análisis de la inasistencia al incorporar una vista detallada por día de la semana, tipo de turno y grupo etario, junto con un indicador específico para primeras citas.

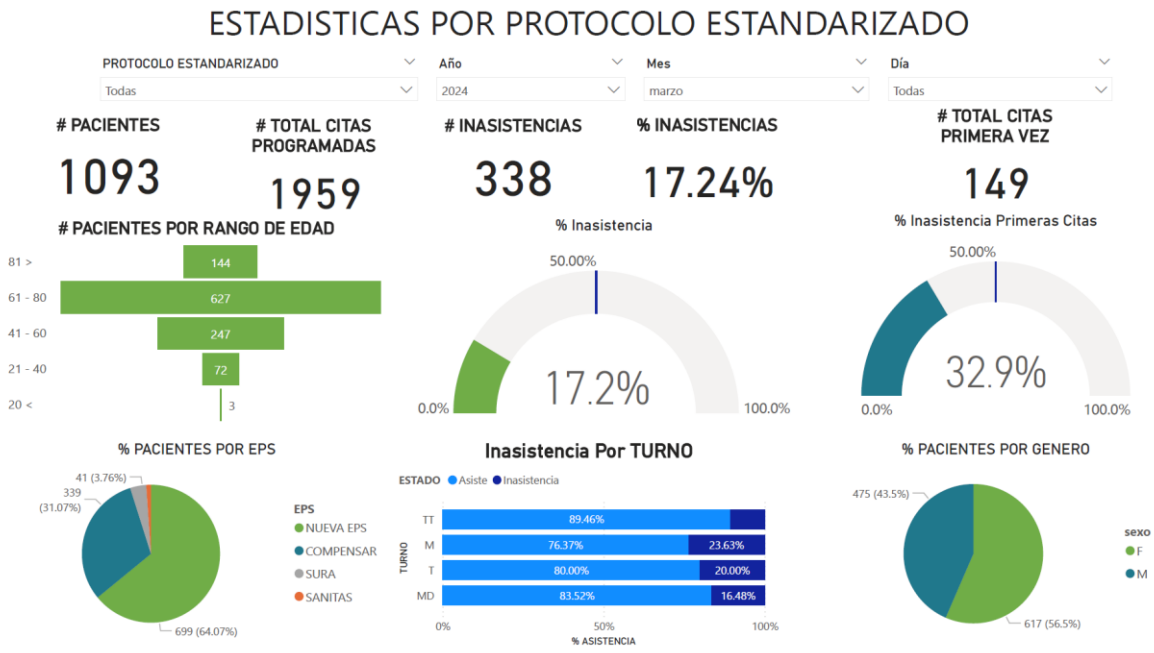
Figura 52 Visual Hoja 2 Tablero Méderi



Fuente: Elaboración propia

Además, permite identificar los protocolos clínicos con mayor cantidad de inasistencias, facilitando así el enfoque hacia tratamientos críticos. La composición de esta visual está diseñada para ayudar al equipo asistencial a reconocer patrones operativos y clínicos que pueden influir en el ausentismo, permitiendo priorizar intervenciones focalizadas según día, turno o tipo de protocolo.

Figura 53 Visual Hoja 3 Tablero Méderi



Fuente: Elaboración propia

En la figura 46 se ilustra la tercera hoja del dashboard presenta una vista consolidada de estadísticas por protocolo estandarizado, permitiendo segmentar la información por periodo de tiempo y cruzarla con variables clave como grupo etario, EPS, turno y género. Su composición integra indicadores globales (pacientes, citas programadas, inasistencias), junto con visualizaciones detalladas de comportamiento por turno y primeras citas. Esta estructura facilita una comprensión integral del desempeño operativo según protocolo clínico, apoyando al equipo de oncología en la identificación de desviaciones, priorización de protocolos con mayor impacto y evaluación de brechas de asistencia según características poblacionales.

12. Próximos pasos

Una vez se cuente con el indicador de giro silla y se haya validado la utilidad de los modelos propuestos en este proyecto, se recomienda al siguiente grupo de trabajo proceder con el despliegue de las soluciones generadas. Para facilitar esta transición, se entregará un manual de uso del código y una guía detallada para la interpretación de los tableros analíticos, lo que garantizará la continuidad técnica y operativa del proyecto. Como parte de los próximos pasos, se sugiere continuar con la maduración del gobierno de datos iniciada en esta fase, realizar un estudio de tiempos y movimientos que permita comprender con mayor precisión la dinámica operativa de las citas, llevar a cabo una calibración periódica del modelo predictivo con datos actualizados, explorar la implementación de modelos en ensamble para mejorar la predicción de inasistencias, aplicar estrategias de gestión del cambio que promuevan la adopción de las herramientas desarrolladas, y finalmente, se observa que el área de oncología cuenta con diferentes oportunidades, fuera del alcance de este proyecto, que desde la analítica se pueden abordar como por ejemplo caracterización de causales de inasistencia.

Dentro de la maduración del gobierno de datos, se contempla la implementación de los diccionarios de datos generados, la asignación de listas estandarizadas de campos (por ejemplo, EPS, motivos de reprogramación), la incorporación de identificadores únicos para cada protocolo y código CIE-10, y la definición de tipos de datos obligatorios en el sistema Almera, especialmente en campos clave como la fecha de inicio del tratamiento. También se recomienda establecer reglas de negocio que impidan inconsistencias, como evitar que la fecha de finalización del tratamiento sea anterior a la fecha de inicio. Estas acciones representan un primer paso hacia la estandarización y mejora de la calidad de la información, así como hacia la garantía de su completitud. Asimismo, se sugiere implementar los notebooks entregados en este proyecto como parte del sistema Almera,

ya que contienen toda la transformación de datos necesaria para el uso de los modelos y del tablero analítico entregado.

Se recomienda también realizar un estudio de tiempos y movimientos que permita medir con mayor precisión la operación de las citas, de modo que los modelos de asistencia y programación se ajusten cada vez más a la realidad operativa y faciliten la toma de decisiones basadas en datos. Esta precisión permitirá al área de oncología detectar fallos en la operación, como cuellos de botella o espacios vacíos, y realizar una planeación más eficiente de las tareas administrativas y del personal involucrado en la programación de citas, contribuyendo así a una mejor organización del área.

Adicionalmente, se propone realizar una calibración periódica del modelo, enriqueciendo la muestra con nuevos datos, y explorar la implementación de modelos en ensamble, lo cual podría mejorar significativamente las métricas de desempeño y generar alertas cuando sea necesario reajustar los hiperparametros del modelo, especialmente en la predicción de pacientes inasistentes a sus citas de quimioterapia. Aunque este grupo no llevará a cabo la implementación directa del modelo, se recomienda dejar documentadas las buenas prácticas para su mantenimiento, incluyendo sugerencias sobre la frecuencia de actualización, monitoreo de desempeño y reentrenamiento, para que puedan ser consideradas por los equipos responsables en el futuro.

Por otra parte, para la implementación de este proyecto en el área de oncología de Méderi, será clave gestionar adecuadamente el cambio con el personal administrativo encargado del agendamiento de citas de quimioterapia. Es fundamental que conozcan las buenas prácticas y comprendan cómo estas generan valor a través de la medición del impacto de las iniciativas, promoviendo así el uso de estas nuevas herramientas en su trabajo diario.

Finalmente, se recomienda continuar desarrollando proyectos analíticos enfocados en la caracterización de las causas de inasistencia, lo que permitirá diseñar estrategias de negocio más efectivas y mejorar la atención al paciente.

13. Conclusiones

Acorde a la problemática de cumplimiento de la meta de atención de pacientes en sala, giro silla superior a 62%, y teniendo en cuenta las fuentes de información disponibles hasta el momento, se obtienen las siguientes conclusiones:

Inicialmente, el análisis descriptivo genera valor a los responsables de la ruta oncológica al identificar criterios relevantes de sus pacientes, como una alta concentración de pacientes adultos mayores, especialmente en el rango de 61-80 años, con una predominancia de mujeres y una mayor inasistencia en el turno de la mañana y entre los mayores de 81 años. Los diagnósticos más comunes son los tumores malignos de próstata y mama, y la mayor concentración de pacientes se encuentra en localidades densamente pobladas como Kennedy, Suba y Engativá. La inasistencia es un desafío significativo, especialmente en pacientes en las primeras citas, en los turnos de la tarde, pacientes con un diagnóstico de cáncer de próstata (CIE-10 C61) y con tratamientos de baja toxicidad (992509), lo que podría afectar la continuidad y efectividad del tratamiento. Además, se observa una disminución progresiva de pacientes a medida que avanzan los ciclos de quimioterapia.

Para mejorar la eficiencia del sistema, se debe enfocar en reducir las inasistencias. En este proyecto se propone disminuir las inasistencias desde dos panoramas distintos, cuando el hospital cuenta con baja y alta demanda de pacientes para recibir su tratamiento de quimioterapia. Como primera medida del modelo predictivo, se realizó una validación utilizando datos reales del periodo comprendido entre el 1 y el 15 de septiembre de 2024. Durante este intervalo, se identificaron 136

casos de inasistencia, y el modelo logró predecir 160, de los cuales se decidió considerar únicamente el 61 %, en concordancia con el F1-score obtenido en producción (0.61). Esto se traduce en 97 inasistencias correctamente anticipadas.

Tabla 7 Resultados de predicción del 1 al 15 de septiembre del 2024.

FECHA	# PACIENTES PROGRAMADOS	INASISTENCIA REAL	PREDICCION INASISTENCIA	# MEJORA (F1 61%)	% INASISTENCIA REAL	% INASISTENCIA MODELO
1/09/2024	26	2	2	1	7,7%	3,8%
2/09/2024	52	12	12	7	23,1%	9,6%
3/09/2024	39	10	9	5	25,6%	12,8%
4/09/2024	36	6	7	4	16,7%	5,6%
5/09/2024	29	5	7	4	17,2%	3,4%
6/09/2024	44	6	12	7	13,6%	-2,3%
7/09/2024	24	6	6	3	25,0%	12,5%
8/09/2024	22	3	3	1	13,6%	9,1%
9/09/2024	59	20	22	13	33,9%	11,9%
10/09/2024	58	19	23	14	32,8%	8,6%
11/09/2024	50	15	15	9	30,0%	12,0%
12/09/2024	48	10	13	7	20,8%	6,3%
13/09/2024	66	13	20	12	19,7%	1,5%
14/09/2024	40	7	7	4	17,5%	7,5%
15/09/2024	34	2	2	1	5,9%	2,9%
Total general	**627**	**136**	**160**	**97**	**22%**	**6%**

Fuente: Elaboración propia

Bajo el supuesto de que cada una de estas inasistencias pudo ser reemplazada de forma efectiva con un paciente proveniente de una lista de espera priorizada, el porcentaje de inasistencia global se reduciría del 22 % al 6 %. Esta mejora se refleja directamente en el indicador “giro silla”, el cual mide la eficiencia en la ocupación de las sillas de quimioterapia.

Tabla 8 Resultados estimados del uso del modelo en el indicador giro silla.

GIRO SILLA ACTUAL	GIRO SILLA CON MODELO APLICADO ± 15%
62%	71%

Fuente: Elaboración propia

Considerando que este indicador actualmente promedia un 62 % de ocupación, se estima que, con la implementación efectiva del modelo de predicción y el uso operativo de una lista de espera, se podría incrementar hasta un 71 %, lo que representa un avance significativo en términos de eficiencia operativa y aprovechamiento de la capacidad instalada.

En cuanto al modelo prescriptivo demostró ser una herramienta efectiva para optimizar la asignación de pacientes a unidades de quimioterapia, logrando coberturas del 100 % en días de baja demanda y entre el 92 % y 96 % en jornadas de alta ocupación. Aunque no es posible una comparación cuantitativa directa con la programación manual realizada por el hospital, se identifican ventajas operativas claras.

A diferencia del proceso actual, que depende de la experiencia individual y carece de control en tiempo real, el modelo aplica criterios estandarizados y automatizados que permiten una mejor utilización del recurso tiempo y reducen errores. Su implementación también disminuye la carga operativa del personal, al generar en segundos soluciones que hoy pueden tardar horas.

Además, al favorecer asignaciones completas, sin solapamientos y con mayor continuidad por unidad, el modelo contribuye directamente a mejorar el indicador de giro silla, aumentando la eficiencia y el volumen de pacientes atendidos sin necesidad de ampliar la infraestructura disponible.

Así mismo, en cuanto a la calidad de datos, se recomienda al Hospital Universitario Méderi establecer identificadores numéricos o ID en los protocolos, EPS y motivo de reprogramación, junto con la adición de tipos de campos obligatorios y su tipo de dato en Almera (software utilizado en área de oncología en el hospital Méderi para almacenar la información de programación de sus pacientes), con el fin de poder realizar análisis posteriores con mayor exactitud. Por último, es

fundamental contar con el tiempo exacto de permanencia del paciente en la cama o silla asignada, dado que las fuentes de información actuales únicamente indican la fecha de administración y el turno asignado, y no se puede medir con exactitud los espacios vacíos.

En cuanto a la implementación del proyecto propuesto, se entregará al Hospital Universitario Méderi un manual de usuario detallado, que incluye los diccionarios creados y los notebooks utilizados para el desarrollo de la limpieza de datos y el modelo de optimización de asignación de pacientes en sala. Además, se proporcionarán los archivos PKL de los modelos para su puesta en producción, junto con un dashboard interactivo, permitiendo así la implementación efectiva del proyecto en el hospital y la recepción de feedback continuo del modelo con los nuevos datos que se vayan generando.

Finalmente, este proyecto abre la puerta a futuros desarrollos analíticos, tales como la estructuración y aseguramiento de la calidad de los datos, el entendimiento más profundo del fenómeno de la inasistencia de los pacientes, teniendo en cuenta la historia del paciente y factores externos como el clima o infecciones virales respiratorias. Estos proyectos adicionales tienen el potencial de aumentar significativamente la cantidad de pacientes atendidos en las salas de quimioterapia del Hospital Universitario Méderi y optimizar los procesos de programación de citas, mejorando así la eficiencia y la calidad del servicio prestado.

14. Referencias

Alaeddini, A., Yang, K., Reddy, C., & Yu, S. (2011). A probabilistic model for predicting the probability of no-show in hospital appointments. *Health Care Management Science, 14*(2), 146–157. https://doi.org/10.1007/s10729-011-9148-9

AlMuhaideb, S., Alswailem, O., Alsubaie, N., Ferwana, I., & Alnajem, A. (2019). Prediction of hospital no-show appointments through artificial intelligence algorithms. *Annals of Saudi Medicine, 39*(6), 373–381. https://doi.org/10.5144/0256-4947.2019.373

Alshaya, S., McCarren, A., & Al-Rasheed, A. (2019). Predicting no-show medical appointments using machine learning. In A. Alfaries, H. Mengash, A. Yasar, & E. Shakshuki (Eds.), *Advances in Data Science, Cyber Security and IT Applications* (pp. 211–223). Springer. https://doi.org/10.1007/978-3-030-36365-9_16

Asociación Colombiana de Hospitales y Clínicas. (2025). ¿Quiénes somos?. https://achc.org.co/

Datosmacro.com. (n.d.). *Colombia - Gasto público Salud*. https://datosmacro.expansion.com/estado/gasto/salud/colombia

Dunstan, J., Villena, F., Hoyos, J. P., Riquelme, V., Royer, M., Ramírez, H., & Peypouquet, J. (2023). Predicting no-show appointments in a pediatric hospital in Chile using machine learning. *Health Care Management Science, 26*, 313–329. https://doi.org/10.1007/s10729-022-09626-z

Fondo Colombiano de Enfermedades de Alto Costo. (2022). *Libro cáncer 2022*. https://cuentadealtocosto.org/wp-content/uploads/2023/11/librocancer-2022.pdf

Instituto Nacional de Cancerología. (2023). Anuario estadístico 2022 (Vol. 20). Instituto Nacional de Cancerología. https://www.cancer.gov.co

Gaviria Uribe, A., Correa, L. F., Dávila Guerrero, C. E., Burgos Bernal, G., & Cruz Vargas, M. F. (2016). *Caracterización Registro Especial de Prestadores de Servicios de Salud (REPS) - IPS*.

Hospital Universitario Méderi. (2022, junio 29). *Ingreso, atención y egreso del paciente en el servicio de hemato-oncología*. Almera - Sistema de Gestión Integral. https://sgi.almeraim.com/sgi/?conid=sgimederi

Hospital Universitario Méderi. (2022). *Informe Méderi 2022* (p. 47). https://www.mederi.com.co/sites/default/files/2022/informe2/InformeMederi.html

Hospital Universitario Méderi. (2024, mayo). *Indicador porcentaje de captación de pacientes con patología crítica oncológica*. Almera - Sistema de Gestión Integral. https://sgi.almeraim.com/sgi/seguimiento/?nosgim&c=sgimederi

IBM Corporation. (1994). *Manual CRISP-DM de IBM SPSS Modeler*.

IE-10. (1992). *Clasificación Estadística Internacional de Enfermedades y Problemas Relacionados con la Salud*. PAHO/WHO. https://ais.paho.org/classifications/chapters/pdf/volume1.pdf

Instituto Nacional de Cancerología. (2022). *Boletín de servicios oncológicos* (p. 26). https://www.cancer.gov.co/recursos_user/files/libros/archivos/Boletin_de_servicios_oncol%C3%B3gicos(2).pdf

Intellat. (2024). *Los mejores hospitales de América Latina*.

La República. (2024, mayo 3). *Las EPS adeudan más de $12 billones a 221 hospitales y clínicas del sistema de salud*. https://www.larepublica.co/empresas/las-eps-adeudan-mas-de-12-billones-a-los-hospitales-y-clinicas-de-colombia-3853323

Méderi. (2023). *Informe de gestión 2022*.

Méderi. (2024). *Informe de gestión 2022-2023*.

Méderi. (2024). *Informe de gestión institucional 2023*.

Méderi. (2024). *Almera - Sistema de Gestión Integral*. Bogotá, Colombia. https://sgi.almeraim.com/sgi/seguimiento/?nosgim#

Méderi. (s.f.). *En Méderi somos*. https://www.mederi.com.co/sobre-nosotros/en-mederi-somos

Ministerio de Hacienda y Crédito Público. (2024, febrero). *Presupuesto General de la Nación 2023. Informe de ejecución del presupuesto General de la Nación 2023*.

Ministerio de Salud y Protección Social. (2014). *Resolución número 000247 de 2014*.

https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=39368

15. Anexos

1. Diccionario de datos Base:

Anexo 1: BASE INDICADORES

BD INDICADORES		
VARIABLE	DESCRIPCION	TIPO
historia	ID DEL PACIENTE	NUMERICO
sexo	GENERO	TEXTO
fecha_nacimiento	FECHA DE NACIMIENTO	FECHA
fecha_actual	FECHA DEL REGISTRO	FECHA
edad	EDAD DEL PACIENTE	NUMERICO
direccion_paciente	DIRECCION DE RESIDENCIA	TEXTO
lugar_residencia	CIUDAD DE RESIDENCIA	TEXTO
localidad_paciente	LOCALIDAD DEL PACIENTE	TEXTO
barrio_paciente	BARRIO DEL PACIENTE	TEXTO
ocupacion	ACTIVIDAD DEL PACIENTE	TEXTO
zona_residencial	SI ES URBANO O RURAL	TEXTO
estado_civil	ESTADO CIVIL DEL PACIENTE	TEXTO
telefono_paciente	TELEFONO DE CONTACTO #1	NUMERICO
celular_paciente	TELEFONO DE CONTACTO #2	NUMERICO

Fuente: Elaboración propia

Anexo 2: BD PROGRAMACIÓN QUIMIOTERAPIA

BASE PROGRAMACION QUIMIO TERAPIA		
VARIABLE	DESCRIPCION	TIPO
Registros	CONTEO TOTAL REGISTRO	NUMERICO
Fec_Inicial	CORTE INICIAL DE LA INFORMACION	FECHA
Fec_Final	CORTE FINAL DE LA INFORMACION	FECHA
Id_Dia	CELDA VACIA	N/A
ID	CODIGO DEL PACIENTE	NUMERICO
CAMA / TURNO	TURNO ASIGNADO	TEXTO
MEDICAMENTO	MEDICAMENTO APLICADO	TEXTO
DOSIS	CANTIDAD SUMINISTRADA	TEXTO
UNIDAD	UNIDAD DE MEDIDA DEL MEDICAM	TEXTO
FECHA DE INICIO	FECHA DE PROGRAMACION DEL TRA	FECHA
PROTOCOLO	PROTOCOLO APLICADO AL PACIENTE	TEXTO
CICLO ADMINSITRADO	# DE CICLO ADMINISTRADO	NUMERICO
DIAS	# DE DIA DEL CICLO	NUMERICO
DOSIS DIA	# DE DOSIS DEL TRATAMIENTO	NUMERICO
PRESENTACION FARMACIA	CELDA VACIA	N/A
OBSERVACION	EN CASO DE INASITENCIA O REPROGRAMACION SE GENERA COMENTARIO	TEXTO
TELEFONO 1	# DE CONTACTO 1	NUMERICO
TELEFONO 2	# DE CONTACTO 2	NUMERICO
TELEFONO 3	# DE CONTACTO 3	NUMERICO

Fuente: Elaboración propia

Anexo 3: BASE INDICADORES

BASE DESCRIPCIÓN INDICADORES		
VARIABLE	**DESCRIPCIÓN**	**TIPO**
HISTORIA	ID PACIENTE	NUMERICO
CICLOS PROGRAMADOS	# DE CICLOS ADMINISTRADOS	NUMERICO
CODIGO CIE 10	ASIFICACIÓN DE PATOLOG	NUMERICO
EPS	ENTIDAD PRESTADORA DE SALUD	TEXTO
DIAS	# DE DIA DEL CICLO	NUMERICO
DIAS DE RETRASO POR CICLO DE ADMINISTRACION	DÍAS POSTERIORES A LA FECHA ESPERADA DE ADMINISTRACION DE LA QUIMIO	NUMERICO
FECHA DE FORMULACION	FECHA EN QUE REMITE EL ONCÓLOGO EL TRATAMIENTO	FECHA
FECHA DE INICIO	FECHA DE ADMINISTRACIÓN DE LA QUIMIOTERAPIA	FECHA
OPORTUNIDAD INICIO QMT	FECHA DE INICIO - FECHA DE FORMULACIÓN	NUMERICO
FECHA DE TERMINACION	FECHA TERMINACIÓN ADMINISTRACIÓN DEL CICLO	FECHA
CICLO ADMINSITRADO	# DE CICLO ADMINISTRADO	NUMERICO
CAMA / TURNO	TURNO ASIGNADO	TEXTO
MOTIVO REP	CAUSAL DE INASISTENCIA	TEXTO
ABANDONO	CAUSAL DE ABANDONO	TEXTO
TIPO DE TRATAMIENTO	CLASIFICACIÓN TRATAMIENTO	TEXTO
TIPO DE TRATAMIENTO2	CURATIVO O PALEATIVO	TEXTO
ALTA ONCOLOGICA Estado	ESTADO EN EL TRATAMIENTO	TEXTO
ESPECIALIDAD	MEDICO TRATANTE POR ESPECIALIDAD	TEXTO

Fuente: Elaboración propia

Anexo 4: RESULTADOS PARA CLASE 0 (ASISTENCIA), 1 (INASISTENCIA) MODELO PREDICTIVO

Modelo	Configuración	F1-score clase 0	F1-score clase 1	Precisión clase 0	Precisión clase 1	Recall Clase 0	Recall Clase 1	AUC-ROC	AUC-PR	FP (%)	FN (%)
Regresión logística	Todas las variables	85%	55%	93%	43%	79%	74%	85%	64%	22%	26%
Random Forest	Todas las variables	84%	56%	95%	42%	76%	81%	87%	66%	24%	19%
SVM	Todas las variables	87%	60%	94%	49%	81%	79%	88%	66%	19%	21%
XGBoost	Todas las variables	89%	66%	97%	53%	83%	89%	93%	79%	17%	11%
XGBoost (15 variables)	Variables Top 15	88%	65%	97%	52%	81%	90%	93%	77%	19%	10%
XGBoost (SMOTETomek threshold 0.322)	Variables Top 15, SMO	91%	68%	96%	58%	87%	82%	92%	78%	14%	18%
XGBoost (No balance threshold 0.703)	Variables Top 15, No b	93%	69%	94%	66%	92%	72%	92%	78%	8%	28%

Fuente Elaboración propia