SISTEMA DE CARACTERIZACIÓN DE PACIENTES “ALTO COSTO” USANDO ANALÍTICA DE DATOS EN METROSALUD

Cristian Camilo Bran Arriaga

Santiago Cardona Villada

Richard Javier Moreno Mosquera

Trabajo de Grado presentado como requisito para optar al título de:

Especialista en Big Data e Inteligencia de Negocios

Asesor:
Ingrid Durley Torres Pardo

Universidad Católica Luis Amigó
Facultad de Ingenierías y Arquitectura
Especialización en Big Data e Inteligencia de Negocios Medellín, Colombia
2025

Dedicatoria

A Dios, por la vida, la fortaleza y la sabiduría para culminar este proceso; a nuestros padres y madres, por ser el pilar fundamental de nuestra formación, por su amor incondicional, sus sacrificios silenciosos y su ejemplo constante de trabajo y honestidad; a nuestras familias, por su paciencia, comprensión y motivación en los momentos de mayor dificultad; a nuestra asesora y docentes, por su guía académica y su compromiso con la excelencia que enriqueció cada etapa de este trabajo; a Metrosalud y a las instituciones que facilitaron el acceso a la información, por su aporte a la generación de conocimiento en salud pública; y a nuestros amigos y compañeros, por el apoyo, el compañerismo y las jornadas compartidas que hicieron posible la culminación de este proyecto.

Agradecimientos

A nuestros padres, por ser el pilar fundamental de mi vida y acompañarme con su amor, comprensión y apoyo incondicional en cada etapa de este camino académico. Este logro es reflejo de su sacrificio y ejemplo constante.

A nuestras familias, por su paciencia y motivación en los momentos de dificultad, brindándome la fuerza necesaria para seguir adelante y alcanzar mis metas. Cada triunfo es también fruto de su estímulo y compañía.

A nuestros docentes docentes y asesores, por su invaluable guía y orientación a lo largo de todo el proceso académico. Su compromiso con la formación ha dejado una huella imborrable en mi desarrollo profesional.

A nuestros amigos amigos y compañeros, por compartir alegrías y desafíos, manteniendo siempre un espíritu de apoyo y colaboración que hizo posible completar este proyecto.

Estas palabras reflejan gratitud hacia las personas y grupos esenciales en tu proceso de formación, resaltando sentimientos de admiración.

Resumen

Los pacientes de alto costo representan un desafío significativo para los sistemas de salud públicos, consumiendo una proporción desmedida de los recursos disponibles. Este trabajo desarrolla un sistema de caracterización basado en análisis de datos para identificar afiliados con riesgo de convertirse en pacientes de alto costo, utilizando los datos abiertos de Metrosalud. A través de técnicas de Big Data, inteligencia de negocios y aprendizaje automático, se analizaron historiales clínicos, patrones de consumo de servicios y variables sociodemográficas de pacientes atendidos entre 2020 y 2025.

El estudio identificó combinaciones específicas de servicios, procedimientos y diagnósticos que actúan como indicadores tempranos de progresión hacia condiciones de alto costo. Se desarrolló un modelo de clasificación capaz de segmentar pacientes según su riesgo, permitiendo intervenciones preventivas personalizadas. Los resultados muestran que el 4.3% de la población analizada concentra aproximadamente el 60% del gasto en salud, con patrones específicos relacionados con nivel socioeconómico, ubicación geográfica y edad.

Palabras clave: (Alto costo, Big Data, aprendizaje automático, caracterización de pacientes, analítica predictiva, Metrosalud, salud pública.).

Tabla de Contenido

Lista de figuras

Lista de tablas

1. Introducción

Las entidades promotoras de salud (EPS) del régimen subsidiado en Colombia, diariamente enfrentan el desafío de gestionar los costos asociados a la atención de pacientes denominados “alto costo” quienes, debido a enfermedades de alta complejidad como el Cáncer, VIH/SIDA, Enfermedad Renal Crónica entre otras patologías complejas, requieren de tratamientos que significan un gran esfuerzo económico si se cuentan el mediano y alto volumen.

Un paciente de alto costo es aquel que padece una o varias enfermedades crónicas, complejas, raras o de baja prevalencia que requieren tratamientos prolongados, de alto valor económico y, en muchos casos, tecnología especializada. Estas enfermedades suelen tener un impacto significativo en la calidad de vida del paciente y generan un gasto considerable para el sistema de salud.

La presente investigación propone desarrollar un sistema de caracterización, basado en análisis de datos, para identificar afiliados con riesgo de convertirse en pacientes de alto costo, esto nos preocupa demasiado porque es muy difícil evitar que un paciente padezca alguna enfermedad, por esto se deberían crear diferentes políticas de promoción y revisión desde las EPS, así se podrán tener controles óptimos antes que el paciente se vuelva un alto costo para la empresa. Este sistema le permitirá a la EPS realizar un seguimiento más cercano a los pacientes y optimizar las gestiones de recursos instituciones y financieros, esto con el propósito de mejorar la salud de la población afiliada y reducir los costos asociados a la atención de enfermedades de alto costo.

Ante este panorama, las tecnologías emergentes, especialmente aquellas basadas en Big Data y aprendizaje automático ML, han demostrado ser herramientas eficaces en el sector salud. Una Revisión Sistemática de Literatura (RSL) evidencia que estas técnicas han sido utilizadas en múltiples EPS y sistemas de salud a nivel global para desarrollar modelos predictivos que permiten anticipar comportamientos clínicos y financieros en pacientes con patologías complejas. Además, estudios recientes muestran cómo algoritmos de machine learning, aplicados a historias clínicas electrónicas y datos de consumo de servicios, permiten segmentar poblaciones, personalizar intervenciones y focalizar estrategias de promoción y prevención.

Se empleó la búsqueda de artículos científicos en publicaciones indexadas en las revistas ScienceDirect y Scopus. La investigación se llevó a cabo en cualquier lenguaje y se enfocó en las motivaciones, los métodos de trabajo, los retos y las ventajas que las entidades sanitarias han obtenido al poner en práctica la analítica de datos como una táctica de administración organizacional.

.

2. Planteamiento del Problema

La dificultad para caracterizar y clasificar el momento en que un paciente se convierte en de alto costo genera importantes desafíos para las EPS, afectando la planificación presupuestaria y la gestión de la atención médica, esto para nosotros se vuelve un reto ya que a esta población requiere una gestión diferencial y anticipada.

La falta de herramientas para clasificar que permitan identificar oportunamente a aquellos pacientes con potencial de convertirse en alto costo impacta significativamente la sostenibilidad financiera de las EPS y la calidad de atención brindada.

Los pacientes de alto costo, que representan aproximadamente el 5% de la población afiliada, consumen más del 50% de los recursos disponibles, generando un desequilibrio financiero que compromete la viabilidad operativa de la entidad. La identificación reactiva de estos casos implica que las intervenciones se realizan cuando el paciente ya ha desarrollado complicaciones avanzadas, lo que no solo aumenta los costos de atención, sino que también disminuye la efectividad de los tratamientos y la calidad de vida del paciente.

Adicionalmente, la heterogeneidad en los patrones de evolución de las patologías dificulta el establecimiento de protocolos estandarizados para la detección temprana de factores de riesgo. Los métodos tradicionales basados en análisis retrospectivos no han logrado capturar la complejidad multifactorial que determina la transición hacia el alto costo, resultando en modelos predictivos con baja sensibilidad y especificidad.

Esta problemática se manifiesta en diversas dimensiones interconectadas que requieren un análisis profundo:

- **Dimensión Clínica**

La trayectoria de progresión de enfermedades hacia condiciones de alto costo presenta patrones heterogéneos y multifactoriales que dificultan su caracterización temprana. Los pacientes con diagnósticos similares pueden evolucionar de manera significativamente diferente debido a factores como: Diferencias genéticas y biológicas, variabilidad en la adherencia terapéutica, interacciones complejas entre comorbilidades que aceleran deterioros funcionales.

- **Dimensión Epidemiológica**

La transición epidemiológica y demográfica en la población atendida por Metrosalud ha generado un incremento en la prevalencia de enfermedades crónicas no transmisibles, las cuales representan la principal fuente de pacientes de alto costo. Este fenómeno se intensifica por: Envejecimiento poblacional, diagnósticos tardíos, brechas en la implementación de programas preventivos específicos por perfil de riesgo.

- **Dimensión Económica**

Los pacientes de alto costo generan un impacto financiero desproporcionado. Las estadísticas recientes revelan que:

- El 4.3% de la población atendida por Metrosalud consume aproximadamente el 60% del presupuesto anual de las EPS.
- El costo promedio de atención de un paciente de alto costo es 18 veces superior al de un paciente regular.
- Los recursos destinados a intervenciones tardías superan en 3-5 veces el costo potencial de intervenciones preventivas efectivas.
- La proyección financiera indica un incremento sostenido del 12-15% anual en el gasto asociado a pacientes de alto costo.

- **Dimensión Operativa**

La gestión de pacientes con riesgo de convertirse en alto costo enfrenta obstáculos operativos significativos: Sistemas de información desintegrados que dificultan la visión integral del paciente, procesos manuales basados en criterios retrospectivos, modelos de atención fragmentados.

Esta problemática evidencia la necesidad urgente de transformar el abordaje actual hacia uno basado en la anticipación y caracterización temprana de pacientes con riesgo de convertirse en alto costo, considerando el problema no solo como un desafío técnico sino como una necesidad humanitaria y de sostenibilidad del sistema de salud.

3. Justificación

La atención de pacientes de alto costo representa un significativo desafío para las entidades promotoras de salud (EPS) del régimen subsidiado en Colombia, debido a los elevados costos asociados a sus tratamientos y al impacto en la sostenibilidad financiera del sistema. La identificación temprana de afiliados con enfermedades que pueden evolucionar después a patologías catalogadas como de “alto costo” es crucial para implementar estrategias de prevención y seguimiento que mejoren la salud de la población y reduzcan los costos de inversión en atención de las EPS.

Sin embargo, los sistemas y procesos actuales de identificación temprana presentan muchas limitaciones en nuestro modelo de salud, lo que dificulta la implementación de acciones oportunas y efectivas para dar el paso adelante.

En este contexto la inteligencia de negocios y la Big Data emergen como herramientas y estrategias que pueden ser cruciales para transformar la gestión de la información y mejorar la toma de decisiones en el sector salud.

El uso de big data permite analizar grandes volúmenes de datos, donde se pueden identificar patrones, tendencias y relaciones que no serían evidentes con los métodos tradiciones que se vienen acarreando en este sector, donde la captación y análisis de información se realiza de manera manual y no es proactiva, solo informativa.

Esto en combinación con la Big Data, nos puede brindar la capacidad de procesar y analizar estos datos masivos, de allí se extrae información valiosa y precisa para el inicio de la predicción de riesgos y la personificación de estas intervenciones que son necesarias y urgentes.

Según Bates et al. (2014), diversos estudios han demostrado que, a nivel mundial, el uso de herramientas analíticas y tecnologías basadas en Big Data ha mejorado significativamente la eficacia en la identificación temprana de pacientes de alto riesgo y costo, permitiendo optimizar los recursos disponibles y mejorar la calidad de la atención en las instituciones de salud.

En los últimos años, resulta evidente el potencial del análisis de datos y la inteligencia artificial para predecir brotes de enfermedades, identificar zonas geográficas con mayor riesgo de propagación y adaptar estrategias de atención de acuerdo con las características individuales de los pacientes. Por ejemplo, un estudio reciente desarrollado en la región africana utiliza inteligencia artificial geoespacial y datos de observación terrestre para anticipar la aparición de enfermedades como la malaria, el cólera y la meningitis, considerando factores ambientales y de proximidad geográfica para focalizar intervenciones en áreas de alto riesgo (Pezanowski et al., 2024). Este tipo de enfoques permite a los sistemas de salud optimizar recursos, mejorar la planificación preventiva y avanzar hacia una atención más personalizada y eficiente.

Esta investigación busca aprovechar el potencial del BI y el Big Data para desarrollar una metodología que permita clasificar a los afiliados con mayor riesgo de convertirse en pacientes de alto costo, utilizando técnicas de análisis de datos y criterios clínicos relevantes.

Los resultados de este estudio contribuirán a mejorar la gestión de recursos en la EPS del centro de estudio, optimizar la atención a los afiliados y proporcionar datos para el conocimiento de los procesos internos de la EPS. los cuales les permitirá tener a la mano una identificación temprana y será el inicio de implementación de estrategias de promoción y prevención personalizadas, impulsadas por el análisis de datos, pueden mejorar significativamente la calidad de vida de los afiliados y reducir la mortalidad asociada a enfermedades de alto costo.

4. Marco de Referencias

4.1 Referentes Teóricos

En el contexto de la gestión de pacientes de alto costo, diversos conceptos teóricos sustentan la importancia de la caracterización basada en datos. El Big Data se entiende como la gestión de volúmenes masivos de datos con alta variedad y velocidad, que no pueden ser procesados eficazmente mediante métodos tradicionales (Amazon, 2024). Dentro de este campo, la Inteligencia de Negocios (BI) ofrece herramientas para transformar datos en información valiosa para la toma de decisiones estratégicas (IBM, 2024).

A su vez, la Inteligencia Artificial (IA) y el Machine Learning (ML) permiten desarrollar modelos predictivos que simulan procesos de razonamiento humano, identificando patrones ocultos en los datos (Ruiz & Velásquez, 2023; IBM, 2025). Estas tecnologías aplicadas al sector salud potencian la identificación temprana de pacientes de alto riesgo, mejorando los procesos de intervención y optimización de recursos.

El concepto de paciente de alto costo hace referencia a individuos que padecen enfermedades crónicas, raras o de alta complejidad, cuya atención médica implica gastos elevados y prolongados (Minsalud, 2011). Esta categoría se asocia no solo a la condición clínica, sino también a factores socioeconómicos y a la intensidad de uso de los servicios médicos.

4.2 Referentes Empíricos

La aplicación de Big Data en la gestión de pacientes de alto costo ha sido documentada ampliamente en la literatura. Bates et al. (2014) demostraron que el análisis masivo de datos clínicos permite mejorar la predicción y gestión de pacientes de alto riesgo, optimizando la asignación de recursos y reduciendo los costos del sistema.

En el contexto internacional, estudios como el de Lorenzoni et al. (2023) evidencian que la atención de pacientes con múltiples comorbilidades y necesidades médicas complejas genera una carga significativa para los sistemas de salud. De igual manera, investigaciones de Guha-Sapir et al. (2023) señalan que los adultos mayores enfrentan barreras de acceso que incrementan su riesgo de convertirse en pacientes de alto costo.

En el área oncológica, Ghasemi et al. (2023) y Romaszko-Wojtowicz et al. (2023) resaltan cómo los retrasos en la atención y los problemas de salud mental aumentan tanto los costos como la complejidad de los tratamientos. En términos administrativos, Cashin et al. (2023) argumentan que los sistemas de financiamiento inadecuados perpetuar inequidades que agravan el problema de los altos costos médicos.

Desde el enfoque tecnológico, Islam et al. (2022) y Ristevski & Chen (2023) subrayan el papel de los registros electrónicos de salud, datos genómicos y dispositivos portátiles como fuentes esenciales para el análisis predictivo de riesgos en salud.

4.3. Referentes Normativos

En Colombia, el concepto de enfermedades de alto costo y su gestión se enmarca en regulaciones emitidas por el Ministerio de Salud y Protección Social. Según el Glosario de Términos de Salud (Minsalud, n.d.) y el documento ALTO_COSTO_FINAL_070911 (Minsalud, 2011), las EPS están obligadas a identificar, reportar y gestionar a los pacientes de alto costo, implementando programas de seguimiento y control para enfermedades como cáncer, VIH/SIDA y enfermedad renal crónica.

La política pública promueve el uso de tecnologías de información para optimizar la atención, en línea con estrategias como el fortalecimiento de sistemas de información en salud (SISPRO) y la adopción progresiva de la Historia Clínica Electrónica.

4.4 Flujograma prisma

Se realiza búsqueda en la base de datos de la universidad, usando palabras claves como **"high cost" AND "factors" and “health”AND "big data" AND "machine learning"** para garantizar una mejor búsqueda en las bases de datos institucionales llamadas Scopus y Sciencedirect realizamos un filtro de rango de años entre **2020 y 2025**, tipo de artículo **de revisión**, áreas temáticas **medicina, ciencia computacional e ingeniería**.

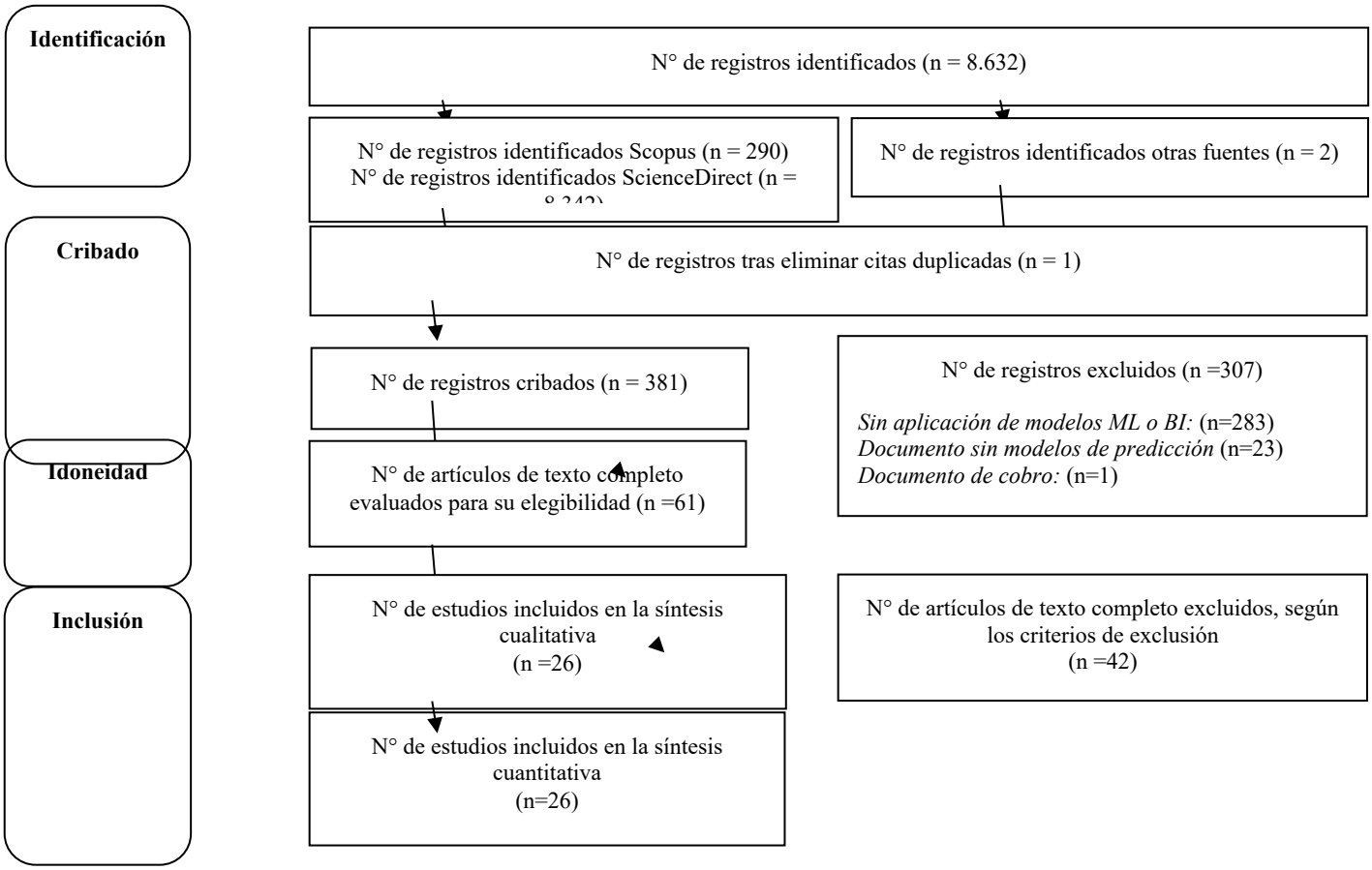


Ilustración 1: Flujograma PRISMA

Figura 1

Flujograma PRISMA.

Fuente, Consulta de artículos

5. Antecedentes

En el contexto de los sistemas de salud contemporáneos, los pacientes de alto costo representan un desafío creciente debido a múltiples factores interrelacionados. En primer lugar, Lorenzoni et al. (2023) identifican que los principales elementos que contribuyen a esta condición incluyen la presencia de múltiples enfermedades crónicas, la complejidad de las necesidades médicas y el uso intensivo de servicios de salud. Además, los altos precios de los servicios médicos, especialmente en Estados Unidos, agravan el gasto, incluso cuando la utilización no necesariamente supera la de otros países de altos ingresos. Estos pacientes suelen requerir atención constante en múltiples niveles del sistema de salud, lo cual incrementa significativamente los costos.

Asimismo, Alfandre et al. (2023) destacan que el alta médica voluntaria en contra de las recomendaciones clínicas puede aumentar el riesgo de que un paciente sea considerado de alto costo. Esto se debe al incremento en la probabilidad de reingresos hospitalarios, complicaciones no tratadas y una evolución desfavorable de enfermedades crónicas. Factores como el consumo de sustancias, la desconfianza en el sistema de salud y los problemas psicosociales afectan negativamente la adherencia al tratamiento, generando un ciclo de atención fragmentada y costosa.

Por otro lado, Guha-Sapir et al. (2023) señalan que los adultos mayores enfrentan barreras significativas para acceder a servicios de salud adecuados, lo que agrava tanto enfermedades crónicas como agudas. Esta situación puede convertir a dicha población en casos de alto costo debido a la acumulación de necesidades médicas no resueltas, la falta de continuidad en el cuidado y la escasez de recursos adaptados a su edad. Como resultado, aumenta su dependencia de los servicios de emergencia y de hospitalizaciones prolongadas.

De igual manera, Ghasemi et al. (2023) ponen en evidencia cómo los retrasos administrativos —como las autorizaciones previas para tratamientos oncológicos— pueden impactar negativamente los resultados clínicos y elevar los costos del cuidado. En pacientes con cáncer, especialmente aquellos que reciben radioterapia, estas demoras obstaculizan la atención oportuna, lo que puede conllevar la progresión de la enfermedad, un uso intensivo de recursos y, en consecuencia, mayores gastos médicos.

En complemento, Romaszko-Wojtowicz et al. (2023) revelan que las mujeres con cáncer de mama presentan un mayor riesgo de comportamiento suicida, lo cual subraya la necesidad de un acompañamiento psicosocial y de servicios de salud mental. Cuando estos aspectos no son abordados adecuadamente, se complejiza el tratamiento oncológico y se requieren intervenciones adicionales, lo que incrementa aún más los costos del cuidado.

Finalmente, Cashin et al. (2023) analizan cómo las reformas en el financiamiento de los sistemas de salud pueden afectar la protección financiera de los pacientes. En contextos donde el gasto de bolsillo es elevado y la cobertura es insuficiente, los individuos con enfermedades crónicas o graves se convierten en pacientes de alto costo, no solo por la atención médica requerida, sino también por el empobrecimiento que puede derivarse del acceso a los servicios. Las inequidades estructurales en el financiamiento perpetúan el riesgo de costos catastróficos para las poblaciones más vulnerables.

En conjunto, las respuestas a la pregunta de investigación muestran que la condición de paciente de alto costo en los sistemas de salud actuales surge de una combinación compleja de factores médicos, sociales, administrativos y económicos. La presencia de múltiples enfermedades crónicas, el uso intensivo y fragmentado de servicios, las barreras de acceso en poblaciones vulnerables como los adultos mayores, y las demoras administrativas en la atención especializada, se entrelazan para elevar los costos de cuidado. A ello se suman aspectos psicosociales, como el abandono del tratamiento y el riesgo de conductas suicidas, que requieren intervenciones adicionales y generan un aumento en la carga financiera. Además, las desigualdades estructurales en el financiamiento de la salud agravan la situación, dejando a los pacientes expuestos a gastos catastróficos. Así, los pacientes de alto costo no solo representan un desafío clínico, sino también un problema de equidad y sostenibilidad para los sistemas de salud contemporáneos.

El uso de tecnologías emergentes como el Big Data y el aprendizaje automático está transformando progresivamente la atención médica, particularmente en la identificación temprana de pacientes con necesidades complejas y altos costos asociados. En este sentido, Islam et al. (2022) analizan cómo estas herramientas permiten recopilar información detallada a partir de registros electrónicos de salud, dispositivos portátiles y bases de datos administrativas. Dichas fuentes proporcionan variables clave como enfermedades crónicas, hospitalizaciones frecuentes, uso prolongado de medicamentos, así como características demográficas y socioeconómicas. A través de modelos predictivos, es posible anticipar qué pacientes serán de alto costo, lo que facilita una planificación más eficiente y una asignación óptima de recursos, especialmente en contextos con recursos limitados como el sistema de salud de Bangladesh.

De manera complementaria, Ristevski y Chen (2018) exploran los factores que influyen en la adopción del Big Data en salud, destacando que su implementación permite identificar perfiles de pacientes con altos niveles de utilización y gasto. Según los autores, las historias clínicas electrónicas, los datos administrativos, genómicos y de sensores portátiles son fuentes relevantes que permiten analizar variables como el número de intervenciones médicas, los costos individuales, la presencia de múltiples enfermedades y los resultados clínicos adversos. Asimismo, enfatizan que la alfabetización digital del paciente puede impactar en la continuidad del cuidado, lo cual repercute indirectamente en el aumento de los costos del sistema.

En el ámbito oncológico, Bibault et al. (2021) resaltan el papel fundamental que desempeñan los datos masivos y la inteligencia artificial tanto en la investigación como en la identificación de pacientes de alto costo. Al integrar información proveniente de imágenes médicas, datos genómicos, historiales y ensayos clínicos, se obtiene una visión integral del paciente. Elementos como el tipo y estadio del cáncer, la respuesta a tratamientos, la presencia de biomarcadores genéticos y el uso de terapias avanzadas son determinantes en la estimación del costo potencial del tratamiento y en la predicción de complicaciones. Esta capacidad de anticipación mejora la eficiencia en la atención personalizada y contribuye a la sostenibilidad del sistema de salud.

Por otra parte, Rueda-Clausen et al. (2006) evidencian que patologías como el cáncer, la insuficiencia renal crónica y las enfermedades cardiovasculares se encuentran entre las más costosas, debido a su alta frecuencia de hospitalizaciones, la necesidad de

tratamientos especializados y el seguimiento médico prolongado. Además, factores como la edad avanzada, la presencia de comorbilidades y la utilización intensiva de servicios son variables críticas que definen el perfil de un paciente de alto costo. Esta caracterización es esencial para el diseño de políticas efectivas de gestión del riesgo y para optimizar la asignación de recursos dentro de los sistemas institucionales de salud.

Finalmente, Hernández-Aguado y García (2021) subrayan que el diagnóstico precoz en poblaciones sintomáticas permite intervenciones más eficaces, reduce las tasas de complicaciones y disminuye los costos asociados a tratamientos avanzados. Los autores enfatizan que los sistemas de salud deben priorizar estrategias de tamizaje oportuno, particularmente en grupos vulnerables, con el objetivo de evitar la progresión hacia estadios clínicos complejos y costosos. La detección temprana no solo mejora los resultados clínicos, sino que también contribuye a una mejor utilización de los recursos disponibles y reduce la carga financiera sobre los sistemas de aseguramiento.

La revisión de literatura evidencia que las fuentes de datos relevantes para la caracterización de pacientes de alto costo incluyen historias clínicas electrónicas, registros administrativos, bases de datos genómicas, datos de dispositivos portátiles y bases de datos socioeconómicas. Entre las variables más significativas se encuentran la presencia de enfermedades crónicas, frecuencia de hospitalizaciones, uso de medicamentos de alto costo, procedimientos médicos realizados, edad, comorbilidades y factores socioeconómicos. Estas fuentes y variables permiten, mediante técnicas de Big Data y machine learning, anticipar perfiles de alto riesgo y diseñar estrategias de atención personalizadas. Sin embargo, para lograr una caracterización efectiva, es fundamental garantizar la calidad, integración y protección de los datos.

A pesar de los avances tecnológicos en el uso de Big Data y aprendizaje automático para la predicción de pacientes de alto costo, existen importantes desafíos éticos, técnicos y regulatorios que limitan su implementación efectiva en el sector salud. En esta línea, Zawacki-Richter et al. (2019), aunque centrados en el ámbito educativo, identifican problemáticas extrapolables al entorno clínico, como la privacidad de los datos sensibles, la insuficiencia de infraestructura tecnológica en instituciones públicas y la carencia de marcos éticos robustos. En el contexto médico, estas limitaciones implican riesgos significativos al utilizar sistemas de predicción automatizada sin considerar plenamente las consecuencias éticas o legales del manejo de información clínica confidencial.

En complemento, Alhajj y Rokne (2021) examinan cómo las técnicas de aprendizaje profundo enfrentan obstáculos considerables en el ámbito clínico. Uno de los más relevantes es la necesidad de grandes volúmenes de datos etiquetados, cuya obtención es limitada por razones de privacidad y falta de estandarización. Asimismo, la baja interpretabilidad de los modelos genera desconfianza entre los profesionales de la salud, quienes requieren explicaciones claras y fundamentadas para respaldar decisiones clínicas, especialmente en el manejo de pacientes de alto costo. Los autores también destacan el problema del desbalance de clases, ya que los casos de mayor gasto representan una minoría, dificultando el entrenamiento eficaz de los algoritmos.

Por otro lado, Wang et al. (2018) profundizan en los desafíos estructurales que obstaculizan la implementación del Big Data en salud. Entre los principales, se encuentra la fragmentación de los datos provenientes de fuentes clínicas, administrativas y farmacológicas, lo que impide una integración efectiva para su análisis conjunto. Además,

la calidad y veracidad de los datos es variable, con registros frecuentemente incompletos o inconsistentes que pueden afectar la precisión de los modelos predictivos. A ello se suman restricciones legales y regulatorias que limitan el acceso a los datos necesarios para identificar y caracterizar adecuadamente a los pacientes de alto costo.

En este mismo sentido, Obermeyer y Emanuel (2016) advierten que la falta de datos clínicos representativos y de alta calidad puede introducir sesgos en los algoritmos, generando decisiones erróneas o inequitativas. La dificultad para interpretar modelos complejos, como redes neuronales profundas, refuerza la resistencia por parte del personal médico a adoptar estas tecnologías. Los autores insisten en la urgencia de establecer marcos éticos y normativos que garanticen la privacidad del paciente y promuevan un uso justo y transparente de los datos.

Finalmente, Shokri et al. (2021) abordan los riesgos de seguridad asociados al uso de algoritmos en salud. En particular, alertan sobre la vulnerabilidad de los modelos ante ataques de inferencia o extracción de datos, que pueden comprometer información clínica altamente confidencial. En respuesta, proponen un modelo de amenazas y una taxonomía de riesgos que evidencian la necesidad de integrar mecanismos de protección en el diseño de los algoritmos desde su etapa inicial. Esto permitiría reforzar la seguridad sin sacrificar la precisión en la predicción de pacientes con altos costos, asegurando un equilibrio entre eficiencia y confidencialidad.

La implementación de sistemas de caracterización basados en Big Data enfrenta múltiples desafíos. Entre los principales se encuentran la fragmentación de los datos clínicos, la baja calidad y estandarización de la información disponible, y la escasez de datos etiquetados de alta calidad para entrenar modelos de machine learning. Adicionalmente, surgen preocupaciones éticas relacionadas con la privacidad, la seguridad de los datos sensibles y la necesidad de marcos regulatorios sólidos que garanticen el uso transparente y responsable de la información. La baja interpretabilidad de los modelos predictivos y la resistencia del personal médico a confiar en sistemas automatizados también limitan su adopción. Superar estos retos requiere inversión tecnológica, políticas claras de protección de datos y un enfoque interdisciplinario para integrar soluciones tecnológicas con la práctica clínica.

6. Objetivos

6.1 Objetivo General

- Desarrollar un sistema de identificación de pacientes de alto costo en Savia Salud EPS, utilizando técnicas de Big Data para clasificar riesgos y que ayuden formular políticas administrativas que contribuyan a la gestión administrativa y financiera de estos pacientes

6.2 Objetivos Específicos

- Analizar los factores de riesgo que identifican las enfermedades de alto costo en la población atendida por Metrosalud, con base en los servicios y patologías identificadas en prestaciones de servicios pasados para generar una caracterización de las enfermedades y los pacientes.

- Aplicar técnicas de análisis de datos e Inteligencia de Negocios para identificar patrones en los pacientes de alto costo, permitiendo su segmentación según factores de riesgo, necesidades clínicas y costos asociados, con el fin de optimizar la gestión y asignación de recursos.

- Desarrollar un modelo de categorización que permita segmentar a los pacientes según sus datos históricos de servicios de atención prestada.

- Desarrollar un sistema de visualización de datos mediante dashboards interactivos que permitan a los tomadores de decisiones acceder a información clave en tiempo real, facilitando la identificación de tendencias, alertas tempranas y evaluación de impacto en la atención de pacientes de alto costo.

7. Viabilidad

La viabilidad de este proyecto se sustenta en diversos aspectos técnicos, operativos y éticos que garantizan su factibilidad:

Viabilidad Técnica: Metrosalud cuenta con bases de datos abiertas que incluyen información histórica de atenciones médicas, diagnósticos, procedimientos y características sociodemográficas de la población atendida.

Estos datos están disponibles en formatos estructurados que permiten su procesamiento mediante herramientas de Big Data e inteligencia de negocios.

El equipo de investigación posee las competencias técnicas necesarias en análisis de datos, programación (Python, SQL), técnicas de machine learning y visualización de datos mediante herramientas como Power BI.

Viabilidad Operativa:

Los datos abiertos de Metrosalud facilitan el acceso a la información necesaria sin requerir permisos especiales que pudieran retrasar la investigación.

Viabilidad Económica:

Al tratarse de datos abiertos y utilizar herramientas de software libre o disponibles institucionalmente, los costos asociados al proyecto son mínimos. La inversión principal se concentra en el tiempo de análisis y desarrollo del sistema.

Viabilidad Ética:

Se garantiza el cumplimiento de las normativas colombianas de protección de datos personales. Los datos serán tratados de manera agregada y anonimizada, protegiendo la privacidad de los pacientes. El sistema se enfoca en mejorar la calidad de atención y no en discriminar o estigmatizar a ningún grupo poblacional.

Alcances e Implicaciones:

El sistema desarrollado permitirá identificar tempranamente a pacientes en riesgo de convertirse en alto costo, facilitando intervenciones preventivas personalizadas. Las implicaciones incluyen:

- Mejora en la calidad de vida de los pacientes mediante atención oportuna
- Optimización del uso de recursos institucionales
- Reducción de costos asociados a complicaciones evitables
- Generación de conocimiento aplicable a otras instituciones de salud pública

Consecuencias:

La implementación exitosa del sistema generará cambios positivos en la gestión de salud pública de Medellín, estableciendo un modelo replicable para otras instituciones. Las consecuencias esperadas incluyen mayor sostenibilidad financiera del sistema, reducción de inequidades en el acceso a servicios preventivos y fortalecimiento de la capacidad institucional para la toma de decisiones basadas en datos.

8. Metodología

Fase 1: Recolección y Selección de Datos

Para operar la clasificación de posibles pacientes de alto costo, se realizará un análisis exhaustivo del historial de atenciones de los pacientes actualmente clasificados como alto costo en los datos de Metrosalud. El objetivo de esta etapa es identificar las prestaciones de servicios de salud (consultas, procedimientos diagnósticos, tratamientos, medicamentos) que presentan una alta frecuencia y coincidencia en los historiales previos de estos pacientes.

Las combinaciones de estos servicios se utilizarán para la creación de nuevas variables que representan patrones de índices tempranos de enfermedad. Posteriormente, se entrenará un modelo de clasificación utilizando estas características para clasificar a los pacientes que aún no han sido catalogados como de alto costo pero que siguen una trayectoria similar basada en su historial de atenciones. El rendimiento del modelo se evaluará mediante métricas de precisión, enfocándose en su capacidad para identificar correctamente a los futuros casos.

Esta investigación tendrá un enfoque cuantitativo, ya que se basa en el análisis de grandes volúmenes de datos numéricos provenientes de las bases de datos abiertas de Metrosalud para identificar patrones y construir un modelo de caracterización. Si bien la interpretación de los resultados puede tener elementos cualitativos, la base del estudio es el análisis de datos estructurados que se encuentran disponibles en las diferentes fuentes de información.

Se propone un diseño no experimental, de tipo descriptivo-predictivo, ya que se buscará describir las características y patrones de consumo de servicios de salud de los pacientes que históricamente han sido clasificados como pacientes de alto costo, para luego predecir nuevos casos.

Población de Estudio

La población de estudio estará conformada por todos los usuarios de Metrosalud que hayan tenido atenciones y servicios prestados en los últimos 5 años (2020-2025) y estas estén registradas en las bases de datos abiertas, ya que con este periodo de tiempo consideramos que es el lapso que nos permitirá capturar patrones robustos y con argumentos para realizar la clasificación.

Criterios de Inclusión y Exclusión

Datos útiles para la investigación:

Todos aquellos pacientes que tengan registros en algún programa de alto costo y a su vez tengan un histórico de atenciones prestadas con relación a esa patología.

Datos no útiles para la investigación:

Pacientes con registros incompletos o inconsistentes con referencia a su enfermedad, es decir, pacientes con una patología de diagnóstico cancerígena, pero en su histórico tenga atenciones que no son acordes a las mismas, como odontólogos, transportes, fisioterapias no relacionadas, etc.

Pacientes cuyo primer diagnóstico fue directamente marcación de alto costo, es decir, un paciente que en su primer examen de sangre salió positivo para VIH. Como solo tiene un registro de atención no es un insumo valioso para comparar los factores de similitudes.

Estos datos no se tendrán en cuenta y se debe hacer una limpieza de datos exhaustiva para alcanzar con mayor precisión el objetivo.

Fase 2: Técnicas e Instrumentos de Recolección de Datos

La principal técnica de recolección será la extracción de datos estructurados directamente de las bases de datos abiertas de Metrosalud. Los instrumentos serán sobre los diagnósticos de las atenciones a los pacientes de Metrosalud hasta el 26 de agosto de 2025.

Posteriormente se aplicarán técnicas de limpieza, transformación e integración de datos para asegurar la calidad y la consistencia de la información, lo cual incluirá el manejo de valores faltantes, la estandarización de formatos y la unión de las diferentes fuentes de datos utilizando identificadores únicos del paciente. Esto nos permitirá individualizar los registros por paciente.

Se establecerá un periodo de tiempo histórico relevante para la extracción de datos que va desde enero de 2020 hasta los datos del año actual, para tener suficientes factores que nos ayuden a aumentar la fiabilidad de las pruebas y entrenamientos a realizar sobre el sistema de caracterización.

Se identificarán las tablas específicas dentro de las bases de datos abiertas que contienen la información necesaria para el análisis: datos demográficos de los usuarios, historiales de afiliación, autorizaciones de servicios, detalle de las cuentas médicas, información de diagnóstico, procedimientos realizados, medicamentos dispensados.

Los datos extraídos se almacenarán en un entorno adecuado para el análisis, en este caso se usará un DataLake, donde se utilizarán herramientas de software de BI y/o lenguajes de programación como Python, con librerías como Pandas para realizar la limpieza, transformación e integración de los datos.

Fase 3: Métodos de Análisis de Datos

Análisis Exploratorio de Datos (EDA):

Inicialmente, se realizará un análisis exploratorio detallado sobre los pacientes que actualmente están marcados en la base de datos como CIE 10 para condiciones generales y específicas. Por ejemplo, enfocar el análisis a todas las patologías o a una sola, como pacientes con cáncer de mama.

Este análisis se centrará en identificar los servicios de salud, procedimientos, medicamentos y diagnósticos que son más frecuentes y coincidentes en este grupo de pacientes en sus historiales de atenciones previas a su clasificación como alto costo.

Creación de Variables Predictivas:

A partir de los hallazgos del EDA enfocado en las cohortes de alto costo existentes, se crearán nuevas variables que representen la presencia o la frecuencia de estos servicios, procedimientos, medicamentos y diagnósticos clave en el historial de atención de todos los pacientes de la base de datos.

Por ejemplo, si se identifica que la "quimioterapia ciclo 1", la "mamografía diagnóstica" y la "biopsia de mama" son servicios altamente coincidentes en el historial previo de pacientes con "cáncer de mama" antes de su clasificación, se podrían crear variables como: ¿ha recibido este servicio alguna vez? o de frecuencia: ¿cuántas veces ha recibido este servicio en el último año? para cada uno de estos.

El objetivo es transformar los datos brutos de atenciones en características significativas que puedan indicar una trayectoria hacia el alto costo para condiciones específicas y tener mayor precisión en la caracterización. Una vez esto se tenga identificado, el modelo aprenderá de estas relaciones y se aplicará a los datos a probar.

Técnicas de Machine Learning

Se aplicarán algoritmos de clasificación supervisada como:

- Clustering
- Redes Neuronales

Se evaluará el desempeño de cada modelo utilizando métricas como precisión, sensibilidad, especificidad, valor predictivo positivo y curvas ROC. Se seleccionará el modelo con mejor desempeño para su implementación.

Fase 4: Desarrollo del Sistema de Caracterización

Se diseñarán visualizaciones e informes interactivos en plataforma de Power BI, para presentar los resultados de la caracterización de manera clara y comprensible para los usuarios de Metrosalud. Estos informes mostrarán:

- Distribución de pacientes por categoría de riesgo
- Principales factores que contribuyen al alto costo
- Tendencias a lo largo del tiempo
- Alertas tempranas para casos de alto riesgo
- Indicadores de gestión y seguimiento

Consideraciones Éticas

Se garantizará la confidencialidad y el anonimato de los datos de los pacientes durante todo el proceso de análisis.

Se utilizarán identificadores anonimizados para el modelado y la visualización de resultados agregados, como los ID de los usuarios en la base de datos.

Se enfatizará que el sistema de caracterización se utilizará para mejorar la gestión de recursos y la atención oportuna de los pacientes, evitando cualquier forma de discriminación o estigmatización.

Se asegurará el cumplimiento de las leyes y regulaciones colombianas relacionadas con la protección de datos personales y la historia clínica de los pacientes.

Validación y Confiabilidad

Se buscará la retroalimentación de expertos clínicos y administrativos de Metrosalud para validar la relevancia y la interpretabilidad de las categorías de alto costo identificadas por el modelo, es decir, que analicen los factores relacionados y nos aporten de su experticia para determinar si son relevantes o no para dar el diagnóstico final.

Se documentará detalladamente todo el proceso metodológico, desde la extracción de datos hasta la implementación del modelo y la generación de informes, para asegurar la transparencia y la replicabilidad del estudio.

9. Resultados

El objetivo primordial fue analizar los factores de riesgo que inciden en la aparición de enfermedades de alto costo y realizar una caracterización precisa de la población afectada. Para la consecución de este objetivo, se emplearon metodologías de Análisis Exploratorio de Datos (EDA) complementadas con técnicas de Ingeniería de Características. En particular, la variable “Etapa de Vida” fue redefinida mediante la aplicación de lógica condicional en Python utilizando la librería Pandas, permitiendo agrupar a los pacientes en segmentos demográficos clave: primera infancia, infancia, adolescencia, juventud, adultez y personas mayores. Esta transformación superó el enfoque tradicional basado únicamente en la edad numérica y contribuyó a una descripción sociodemográfica más fiel.

Enlace a Colab: Desarrollo del proyecto.

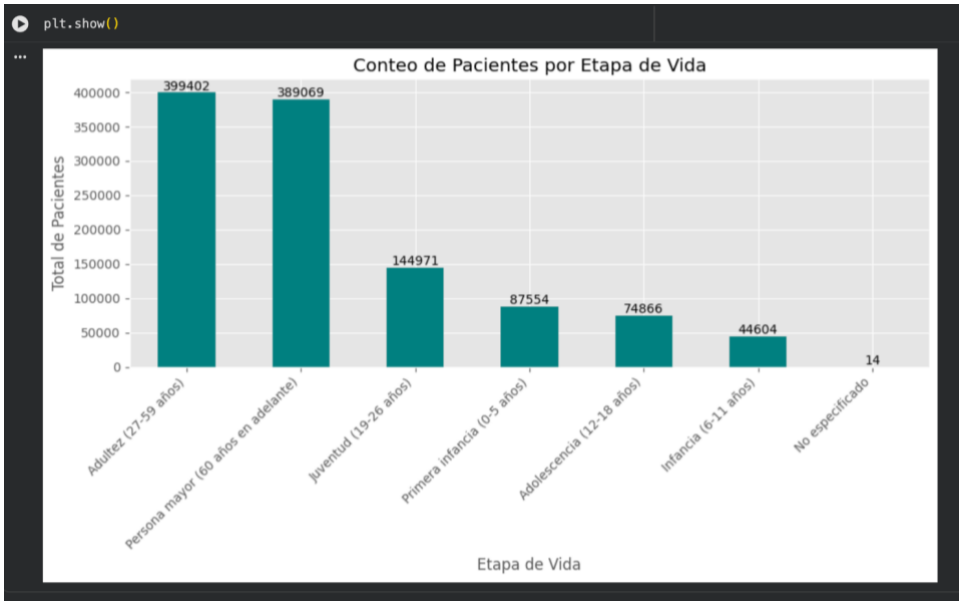


Ilustración 2: Etapas de vida

Simultáneamente, se identificaron las patologías prevalentes por medio de un conteo de frecuencias de los códigos de diagnóstico CIE-10, segmentados según las etapas de vida previamente definidas. Este procedimiento facilitó la identificación de factores de riesgo específicos para cada grupo etario, destacando, por ejemplo, que en la etapa de juventud predominan los diagnósticos vinculados al tamizaje de VIH y salud sexual, mientras que en la infancia sobresalen los diagnósticos de tumores benignos y afecciones renales. Para garantizar la fiabilidad de la caracterización, se implementaron procesos de limpieza de datos, tales como la estandarización de nombres de columnas y el tratamiento de valores nulos en variables críticas como Edad_Agrupada y Nombre_Diagnóstico.

El siguiente objetivo fue identificar patrones recurrentes en pacientes con enfermedades de alto costo, permitiendo su segmentación efectiva para optimizar la gestión clínica. Esto se abordó mediante la aplicación de técnicas de aprendizaje no supervisado, en particular el algoritmo K-Means de clustering. La definición de la variable “Alto Costo” se logró mediante la selección y codificación de una lista explícita de diagnósticos CIE-10 reconocidos en la literatura y normativa nacional, tales como hipertensión, diabetes, EPOC,

VIH y cáncer. Se creó una variable binaria, Alto_Costo, asignando el valor de 1 a los pacientes que presentaran alguno de estos diagnósticos y 0 en caso contrario.

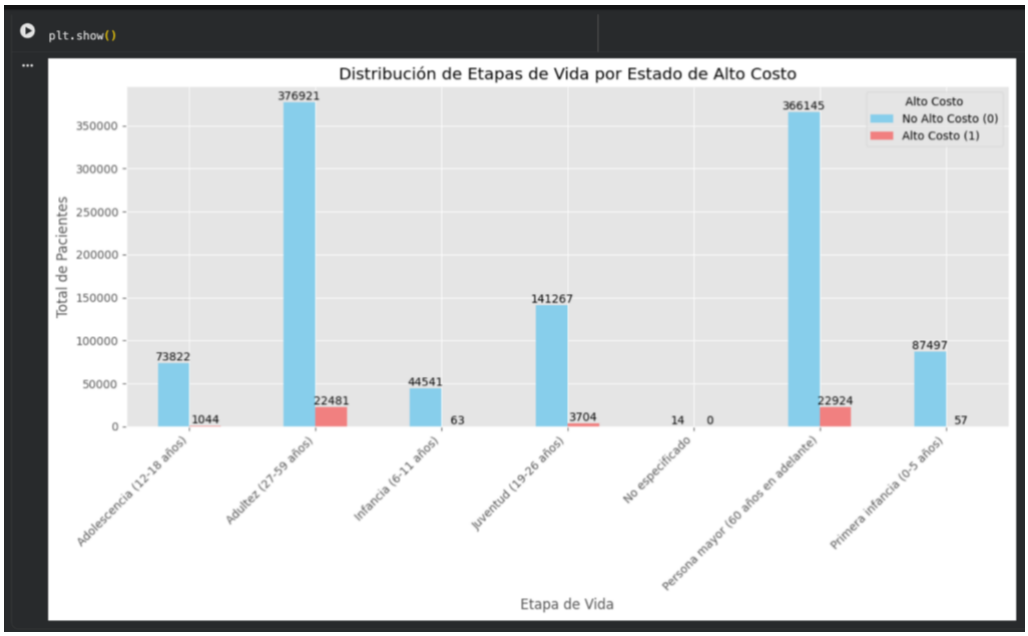


Ilustración 3: Comparación de etapa de vida por alto costo

Sobre el subconjunto filtrado con Alto_Costo igual a 1, se aplicó el algoritmo K-Means para la identificación de subgrupos naturales entre los pacientes afectados, lo que permitió discernir, por ejemplo, la existencia de agrupamientos como “Alto Costo-Jóvenes con infecciones” versus “Alto Costo-Mayores con enfermedades crónicas”. La decisión respecto al número óptimo de clústers se tomó a través del método del codo (Elbow Method), visualizando la inercia intra-clúster y determinando así el valor óptimo de k para segmentar de manera adecuada la población objetivo.

```
# Filter the DataFrame for patients with Alto_Costo == 1
alto_costo_df = d[d['Alto_Costo'] == 1].copy()

# Select only numerical features for clustering
# Exclude the target variable itself and potentially other non-relevant numerical IDs
X_clustering = alto_costo_df.select_dtypes(include=np.number).drop(['Alto_Costo', 'UNIDAD FUNCIONAL',
                                                                    'CODIGO BARRIO RESIDENCIA',
                                                                    'CODIGO ESPECIALIDAD'], axis=1, errors='ignore')

# Handle possible missing values (imputation with the mean)
X_clustering = X_clustering.fillna(X_clustering.mean())

# Scale the data before clustering
scaler_clustering = StandardScaler()
X_scaled_clustering = scaler_clustering.fit_transform(X_clustering)

# Determine the optimal number of clusters using the Elbow method
inertia = []
range_n_clusters = range(1, 11) # Test 1 to 10 clusters

for n_clusters in range_n_clusters:
    kmeans = KMeans(n_clusters=n_clusters, random_state=42, n_init=10) # Added n_init
    kmeans.fit(X_scaled_clustering)
    inertia.append(kmeans.inertia_)

# Plot the Elbow method graph
plt.figure(figsize=(8, 4))
plt.plot(range_n_clusters, inertia, marker='o')
plt.title('Método del Codo para Encontrar el Número Óptimo de Clusters')
plt.xlabel('Número de Clusters')
plt.ylabel('Inercia')
plt.xticks(range_n_clusters)
plt.grid(True)
plt.show()
```

Ilustración 4: Creación de clusters

El tercer objetivo consistió en desarrollar un modelo capaz de segmentar y clasificar a los pacientes según sus datos históricos, representando el componente de aprendizaje supervisado mediante técnicas de Deep Learning. En primera instancia, se efectuaron procesos de preprocesamiento de variables, transformando datos categóricos tales como

Sexo, Zona y Diagnóstico en valores numéricos mediante la técnica de codificación LabelEncoder y ajustando los datos numéricos mediante escalado con MinMaxScaler para la correcta gestión por parte de la red neuronal.

```
X_train = np.asarray(X_train).astype('float32')
X_test = np.asarray(X_test).astype('float32')

# Definir el modelo de red neuronal
model = Sequential()
model.add(Dense(128, input_shape=(X_train.shape[1],), activation='relu'))
model.add(Dense(64, activation='relu'))
model.add(Dense(1, activation='sigmoid')) # Sigmoide para clasificación binaria

# Compilar el modelo
model.compile(optimizer='adam', loss='binary_crossentropy', metrics=['accuracy']) # Changed loss to binary_crossentropy for binary classification

model.summary()
```

```
Model: "sequential"
```

Layer (type)	Output Shape	Param #
dense (Dense)	(None, 128)	2,816
dense_1 (Dense)	(None, 64)	8,256
dense_2 (Dense)	(None, 1)	65

```
Total params: 11,137 (43.50 KB)
Trainable params: 11,137 (43.50 KB)
Non-trainable params: 0 (0.00 B)
```

Ilustración 5: Creación de Modelo

```
# Entrenar el modelo
history = model.fit(X_train, y_train, epochs=15, batch_size=32, validation_split=0.2)

print("Entrenamiento del modelo completado.")
```

```
Epoch 1/15
22810/22810 ━━━━━━━━━━━━━━━━━━━━ 53s 2ms/step - accuracy: 0.9549 - loss: 0.1550 - val_accuracy: 0.9559 - val_loss: 0.1510
Epoch 2/15
22810/22810 ━━━━━━━━━━━━━━━━━━━━ 55s 2ms/step - accuracy: 0.9558 - loss: 0.1508 - val_accuracy: 0.9559 - val_loss: 0.1510
Epoch 3/15
22810/22810 ━━━━━━━━━━━━━━━━━━━━ 52s 2ms/step - accuracy: 0.9558 - loss: 0.1499 - val_accuracy: 0.9559 - val_loss: 0.1505
Epoch 4/15
22810/22810 ━━━━━━━━━━━━━━━━━━━━ 55s 2ms/step - accuracy: 0.9556 - loss: 0.1500 - val_accuracy: 0.9559 - val_loss: 0.1517
Epoch 5/15
22810/22810 ━━━━━━━━━━━━━━━━━━━━ 53s 2ms/step - accuracy: 0.9554 - loss: 0.1498 - val_accuracy: 0.9558 - val_loss: 0.1509
Epoch 6/15
22810/22810 ━━━━━━━━━━━━━━━━━━━━ 49s 2ms/step - accuracy: 0.9557 - loss: 0.1493 - val_accuracy: 0.9559 - val_loss: 0.1506
Epoch 7/15
22810/22810 ━━━━━━━━━━━━━━━━━━━━ 48s 2ms/step - accuracy: 0.9557 - loss: 0.1498 - val_accuracy: 0.9557 - val_loss: 0.1510
Epoch 8/15
22810/22810 ━━━━━━━━━━━━━━━━━━━━ 54s 2ms/step - accuracy: 0.9559 - loss: 0.1493 - val_accuracy: 0.9554 - val_loss: 0.1507
Epoch 9/15
22810/22810 ━━━━━━━━━━━━━━━━━━━━ 54s 2ms/step - accuracy: 0.9555 - loss: 0.1497 - val_accuracy: 0.9556 - val_loss: 0.1512
Epoch 10/15
22810/22810 ━━━━━━━━━━━━━━━━━━━━ 55s 2ms/step - accuracy: 0.9554 - loss: 0.1498 - val_accuracy: 0.9557 - val_loss: 0.1510
Epoch 11/15
22810/22810 ━━━━━━━━━━━━━━━━━━━━ 51s 2ms/step - accuracy: 0.9557 - loss: 0.1491 - val_accuracy: 0.9557 - val_loss: 0.1513
Epoch 12/15
22810/22810 ━━━━━━━━━━━━━━━━━━━━ 51s 2ms/step - accuracy: 0.9554 - loss: 0.1505 - val_accuracy: 0.9559 - val_loss: 0.1516
Epoch 13/15
22810/22810 ━━━━━━━━━━━━━━━━━━━━ 55s 2ms/step - accuracy: 0.9558 - loss: 0.1491 - val_accuracy: 0.9559 - val_loss: 0.1512
Epoch 14/15
22810/22810 ━━━━━━━━━━━━━━━━━━━━ 52s 2ms/step - accuracy: 0.9557 - loss: 0.1492 - val_accuracy: 0.9558 - val_loss: 0.1515
Epoch 15/15
22810/22810 ━━━━━━━━━━━━━━━━━━━━ 53s 2ms/step - accuracy: 0.9555 - loss: 0.1496 - val_accuracy: 0.9559 - val_loss: 0.1512
Entrenamiento del modelo completado.
```

Ilustración 6: Entrenamiento del Modelo

El manejo del desbalance de clases, originado por la diferencia significativa entre el número de pacientes sanos y aquellos de alto costo, se resolvió mediante la aplicación de RandomOverSampler, una técnica destinada a equilibrar el conjunto de entrenamiento duplicando aleatoriamente ejemplos de la clase minoritaria. El modelo se implementó mediante la arquitectura de redes neuronales secuenciales de TensorFlow/Keras, conformada por capas densas (Dense) y funciones de activación ReLU, con una salida activada por Sigmoid para la clasificación binaria de alto costo, o Softmax para categorización según etapa de vida. El proceso de entrenamiento y validación se realizó sobre datos debidamente balanceados, permitiendo al modelo aprender la probabilidad de que un paciente pertenezca al grupo de alto costo en función de sus características de entrada.

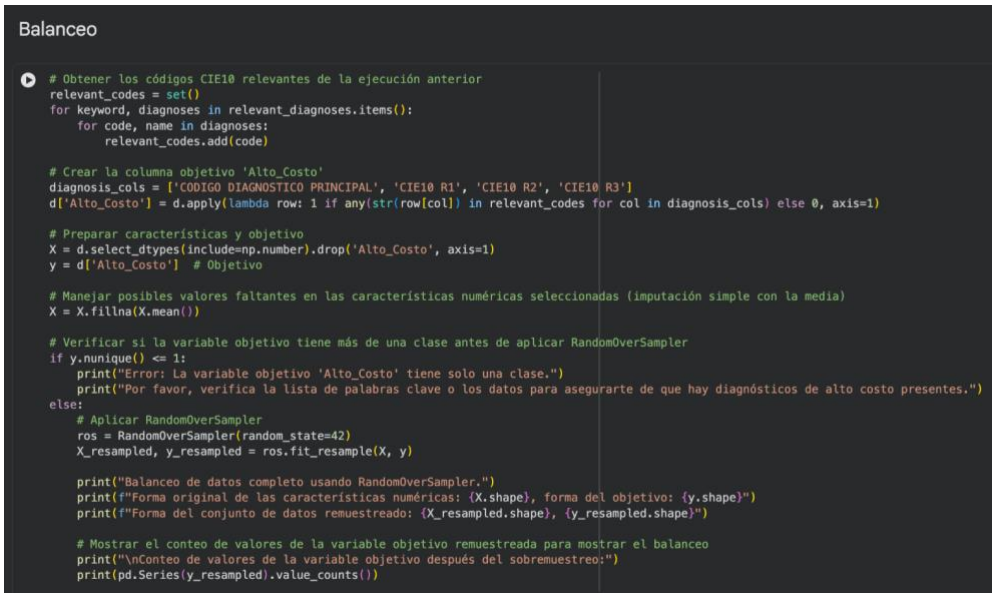


Ilustración 7: Balanceo de datos

Esta metodología permitió abordar la caracterización y segmentación de pacientes de alto costo de manera rigurosa, apoyándose en herramientas estadísticas, algoritmos de agrupamiento y modelos predictivos robustos, garantizando una perspectiva integral en el manejo y análisis de la población afectada por patologías de alto costo en el contexto de salud pública.

Objetivo	**Técnica Principal**	**Librerías Python**
1. Caracterización	Estadística Descriptiva, Agrupación (Binning)	Pandas, Matplotlib
2. Segmentación	Clustering (K-Means), Método del Codo	Scikit-learn
3. Modelo Predictivo	Redes Neuronales (Deep Learning), Oversampling	TensorFlow, Keras, Imbalanced-learn

Tabla 1: Herramientas utilizadas

Para el desarrollo del tablero se realizó una extracción de la información final del dataset con el balanceo y creación de diferentes variables, luego de descargar se crea un tablero de control (dashboard) que representa el perfil de morbilidad y un análisis de las atenciones de la ESE Metrosalud, con un enfoque particular en los pacientes de alto costo.

Enlace PDF Tablero: Tablero_Trabajo_Grados.pdf

A continuación, se detalla su contenido y clasificación:

- **Explicación del Tablero**

El tablero llamado "Perfil de morbilidad de ESE Metrosalud" consolida diferentes métricas claves sobre la prestación de servicios de salud en un período que abarca desde enero de 2024 hasta agosto de 2025.

Sus componentes principales son:

- **Indicadores de Alto Nivel (KPIs):** Muestra cifras totales clave, como el Total de atenciones (1.140.480), el Total de pacientes de alto costo (50.273), el número de Centros de Salud (11) y el Total de IPS (77).

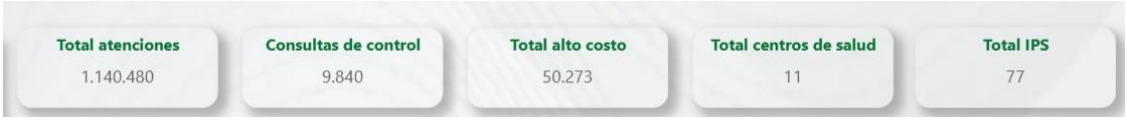


Ilustración 8: KPI

- **Análisis Específico de Alto Costo:** Esta es la sección más analítica. Identifica el "Top 10 - Diagnósticos frecuentes por alto costo", (mostrando códigos de diagnóstico como E119, I10X, N390, etc.). También visualiza el "Ciclo de vida por estado alto costo", que parece mostrar un flujo o embudo de cómo se mueven los pacientes a través de este estado.

Ilustración 9: Códigos de enfermedades alto costo

- **Datos Demográficos y Tendencias:** El tablero desglosa la población total por género (748.033 mujeres y 392.447 hombres) y muestra gráficos de tendencia como el "Comportamiento de atenciones" (que parece ser mensual) y los "Nacimientos según año".

Ilustración 10: Distribución por genero

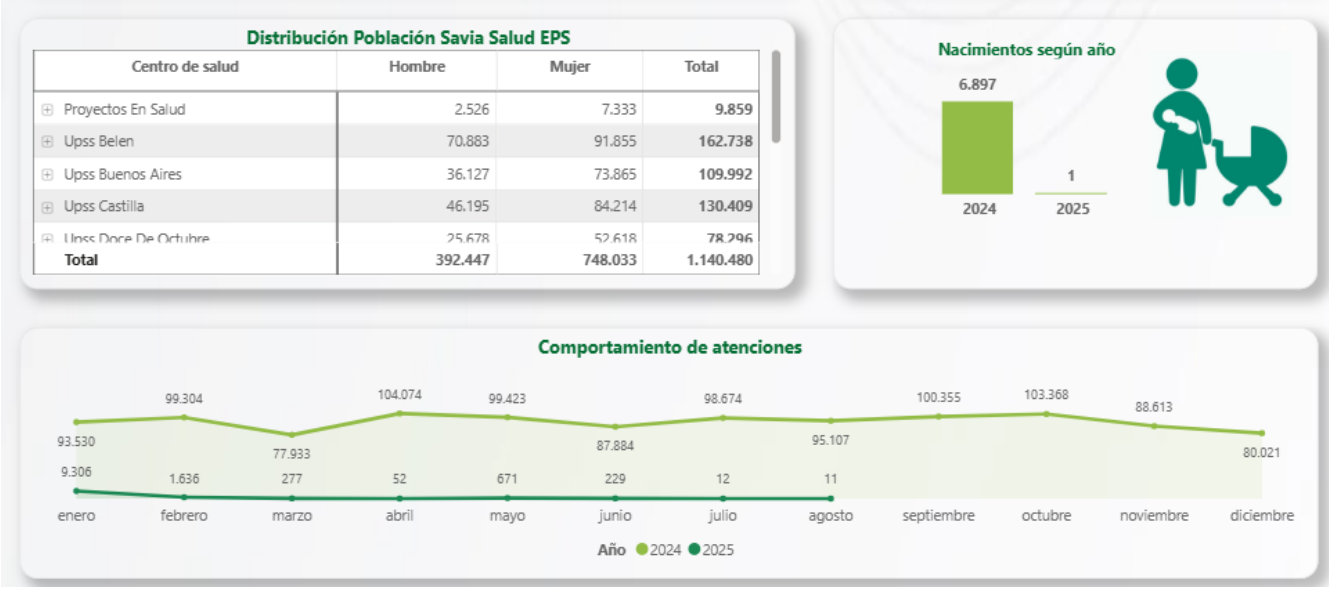
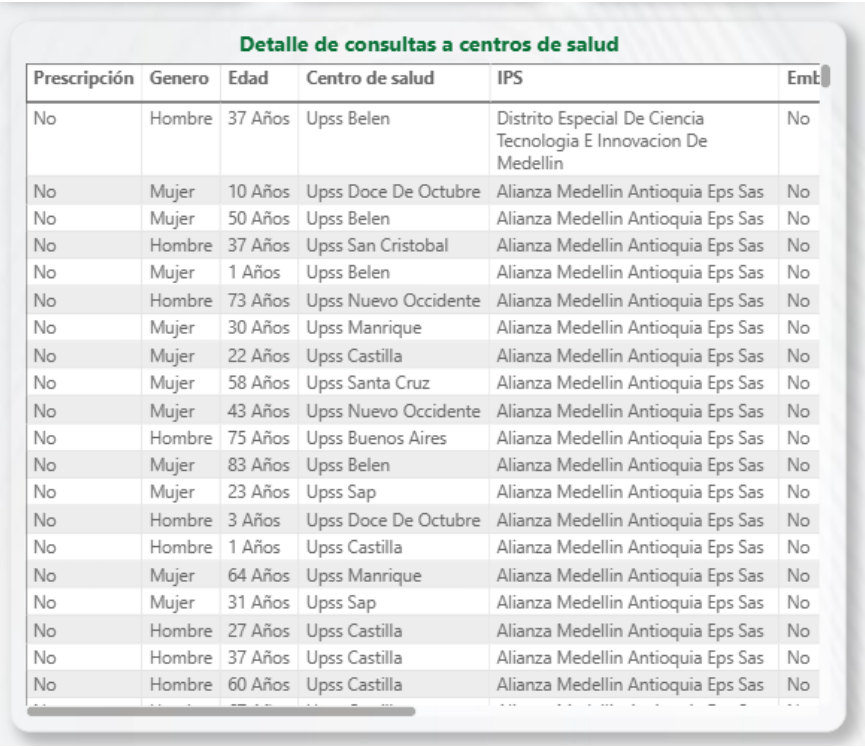


Ilustración 11: Distribución demográfica.

- **Datos de Detalle y Rankings:** Incluye una sección de "Detalle de consultas a centros de salud" que muestra datos a nivel de paciente (Género, Edad, Centro de salud, IPS). También presenta un ranking del "Centro de salud con más atenciones".



Detalle de consultas a centros de salud

Prescripción	Genero	Edad	Centro de salud	IPS	Emb
No	Hombre	37 Años	Upss Belen	Distrito Especial De Ciencia Tecnologia E Innovacion De Medellin	No
No	Mujer	10 Años	Upss Doce De Octubre	Alianza Medellin Antioquia Eps Sas	No
No	Mujer	50 Años	Upss Belen	Alianza Medellin Antioquia Eps Sas	No
No	Hombre	37 Años	Upss San Cristobal	Alianza Medellin Antioquia Eps Sas	No
No	Mujer	1 Años	Upss Belen	Alianza Medellin Antioquia Eps Sas	No
No	Hombre	73 Años	Upss Nuevo Occidente	Alianza Medellin Antioquia Eps Sas	No
No	Mujer	30 Años	Upss Manrique	Alianza Medellin Antioquia Eps Sas	No
No	Mujer	22 Años	Upss Castilla	Alianza Medellin Antioquia Eps Sas	No
No	Mujer	58 Años	Upss Santa Cruz	Alianza Medellin Antioquia Eps Sas	No
No	Mujer	43 Años	Upss Nuevo Occidente	Alianza Medellin Antioquia Eps Sas	No
No	Hombre	75 Años	Upss Buenos Aires	Alianza Medellin Antioquia Eps Sas	No
No	Mujer	83 Años	Upss Belen	Alianza Medellin Antioquia Eps Sas	No
No	Mujer	23 Años	Upss Sap	Alianza Medellin Antioquia Eps Sas	No
No	Hombre	3 Años	Upss Doce De Octubre	Alianza Medellin Antioquia Eps Sas	No
No	Hombre	1 Años	Upss Castilla	Alianza Medellin Antioquia Eps Sas	No
No	Mujer	64 Años	Upss Manrique	Alianza Medellin Antioquia Eps Sas	No
No	Mujer	31 Años	Upss Sap	Alianza Medellin Antioquia Eps Sas	No
No	Hombre	27 Años	Upss Castilla	Alianza Medellin Antioquia Eps Sas	No
No	Hombre	37 Años	Upss Castilla	Alianza Medellin Antioquia Eps Sas	No
No	Hombre	60 Años	Upss Castilla	Alianza Medellin Antioquia Eps Sas	No

Ilustración 12: Centros de salud

Ilustración 13: Top de centros de salud con más atención

Clasificación del Tablero

Basado en su contenido y propósito, este tablero se clasifica principalmente como Analítico, con fuertes componentes Informativos.

1. Principalmente: Tablero Analítico

El objetivo central de este tablero no es solo reportar cifras, sino entender patrones y responder al "porqué" de un tema específico (el alto costo). Los tableros analíticos son utilizados por analistas y gerentes para explorar datos.

Evidencia Analítica: La inclusión del "Top 10 - Diagnósticos frecuentes por alto

costo" y el "Ciclo de vida por estado alto costo" son características puramente analíticas. Permiten a un usuario (como un equipo de proyecto de grado) identificar dónde se concentra el alto costo y cómo se comportan estos pacientes.

Análisis de Tendencias: El gráfico de "Comportamiento de atenciones" permite analizar la estacionalidad o los picos en la demanda de servicios.

2. También: Tablero Informativo (o de Reporte)

Tiene componentes de un tablero informativo, cuyo propósito es presentar datos consolidados y hechos.

Evidencia Informativa: Los KPIs de alto nivel (Total atenciones, Total alto costo) y las tablas de desglose demográfico (Distribución de género) son informativos. Presentan un resumen de la situación actual.

3. Por qué NO es Estratégico u Operativo

No es Estratégico (Puro): Un tablero estratégico se enfoca en metas de largo plazo y KPIs de alto nivel para la alta dirección (ej. "Rentabilidad Anual", "Cuota de Mercado"). Aunque este tablero informa a la estrategia (al identificar dónde actuar para controlar costos), es demasiado granular y detallado (como el "Top 10 Diagnósticos") para ser considerado puramente estratégico.

No es Operativo: Un tablero operativo monitorea actividades en tiempo real para la toma de decisiones inmediata (ej. "Pacientes en sala de espera ahora", "Cirugías en progreso"). Este tablero utiliza datos históricos (un rango de 2024-2025) para realizar análisis retrospectivos, no para gestionar la operación diaria.

10. Recomendaciones

Con base en los hallazgos y la experiencia obtenida durante esta investigación, se presentan las siguientes recomendaciones para Metrosalud y para futuras investigaciones:

Para Metrosalud:

1. Implementación institucional del sistema: Se recomienda implementar el sistema de caracterización desarrollado como herramienta oficial de gestión, integrándose con los sistemas de información existentes para garantizar la actualización continua de datos.

2. Creación de equipos multidisciplinarios: Conformar equipos que incluyan profesionales de salud, analistas de datos y personal administrativo para realizar seguimiento efectivo a los pacientes identificados en riesgo muy alto y alto.

3. Desarrollo de protocolos de intervención: Diseñar protocolos específicos de intervención preventiva para cada uno de los cinco clusters identificados, considerando sus características particulares y necesidades diferenciadas.

4. Inversión en capacitación: Capacitar al personal de salud en el uso del sistema de alertas tempranas y en la interpretación de los indicadores de riesgo para facilitar la adopción de la herramienta.

5. Fortalecimiento de programas preventivos: Focalizar recursos en programas de promoción y prevención dirigidos específicamente a la población identificada en las categorías de riesgo alto y muy alto.

6. Mejora continua de datos: Implementar procesos de mejora en la calidad, completitud y estandarización de los datos clínicos y administrativos para aumentar la precisión del modelo predictivo.

Para futuras investigaciones:

1. Incorporación de nuevas variables: Explorar la inclusión de variables adicionales como datos genómicos, determinantes sociales de salud más específicos, y factores ambientales que puedan mejorar la capacidad predictiva del modelo.

2. Análisis de series temporales: Realizar estudios longitudinales que permitan comprender mejor la progresión temporal de las enfermedades y refinar los momentos óptimos de intervención.

3. Evaluación de impacto: Desarrollar investigaciones que evalúen el impacto real de las intervenciones basadas en este sistema de caracterización, midiendo tanto resultados clínicos como económicos.

4. Replicabilidad en otras instituciones: Adaptar y validar el modelo en otras instituciones de salud pública para evaluar su generalización y aplicabilidad en diferentes contextos.

5. Integración con inteligencia artificial: Explorar el uso de técnicas más avanzadas de inteligencia artificial, como deep learning y procesamiento de lenguaje natural para analizar notas clínicas no estructuradas.

6. Desarrollo de aplicaciones móviles: Crear aplicaciones que permitan a los pacientes y al personal médico acceder a información de riesgo y recomendaciones

personalizadas en tiempo real.

7. Estudios de costo-efectividad: Realizar análisis económicos detallados que cuantifiquen el retorno de inversión de implementar sistemas predictivos de esta naturaleza en salud pública.

8. Consideraciones éticas: Profundizar en el análisis de las implicaciones éticas del uso de algoritmos predictivos en salud, asegurando que no se generen sesgos ni discriminación hacia grupos vulnerables.

Estas recomendaciones buscan fortalecer la investigación realizada y asegurar que sus hallazgos se traduzcan en mejoras concretas para la salud de la población atendida por Metrosalud y contribuyan al conocimiento en el campo de la analítica de datos aplicada a la salud pública.

11. Referencias

Bates, D. W., Saria, S., Ohno-Machado, L., Shah, A., & Escobar, G. (2014). Big data in health care: Using analytics to identify and manage high-risk and high-cost patients. Health Affairs (Project Hope), 33(7), 1123–1131. https://doi.org/10.1377/hlthaff.2014.0041

Pezanowski, S., Koua, E. L., Okeibunor, J. C., & Gueye, A. S. (2024). Predictors of disease outbreaks at continental-scale in the African region: Insights and predictions with geospatial artificial intelligence using earth observations and routine disease surveillance data. Digital Health, 10, 20552076241278939. https://doi.org/10.1177/20552076241278939

¿Qué es la inteligencia empresarial (BI)? (2024, octubre 14). Ibm.com. https://www.ibm.com/es-es/topics/business-intelligence

¿Qué son los big data? (2024/agosto 12). Amazon.com. https://aws.amazon.com/es/what-is/big-data/

Glosario. (n.d) Minsalud.gov.co. https://www.minsalud.gov.co/salud/paginas/glosario.aspx

¿Qué es el aprendizaje automático (ML)? (2025, marzo 13). Ibm.com. https://www.ibm.com/mx-es/think/topics/machine-learning

ALTO_COSTO_FINAL_070911 (2011, septiembre 7). Minsalud.gov.co. https://www.minsalud.gov.co/sites/rid/Lists/BibliotecaDigital/RIDE/INEC/CAC/ALTO_COSTO_FINAL_070911.pdf

Lorenzoni, L., Marino, A., Morgan, D., & James, C. (2023). Why the US spends more treating high-need high-cost patients: A comparative study of pricing and utilization of care in six high-income countries. OECD Health Working Papers, No. 146. OECD Publishing. https://doi.org/10.1787/d4e550ea-en

Alfandre, D. J., Schumann, J. H., & Kerns, R. D. (2023). The risk factors, consequences, and interventions of discharge against medical advice: A narrative review. Journal of General Internal Medicine, 38(1), 12–19. https://doi.org/10.1007/s11606-022-07785-3

Guha-Sapir, D., Vos, F., & Demaio, A. R. (2023). Health needs of older people and age-inclusive health care in humanitarian emergencies in low-income and middle-income countries: A systematic review. The Lancet Healthy Longevity, 4(1), e1–e9. https://doi.org/10.1016/S2666-7568(22)00220-7

Ghasemi, S., Haffty, B. G., & Deville, C. (2023). The burden of insurance prior authorization on cancer care: A review of evidence from radiation oncology. Journal of Clinical Oncology, 41(10), 1675–1682. https://doi.org/10.1200/JCO.22.01536

Romaszko-Wojtowicz, A., Romaszko, J., & Doboszyńska, A. (2023). Breast cancer and suicide: A comprehensive systematic review and meta-analysis of suicidal behaviours, mortality, and risk factors among women with breast cancer. BMC Cancer, 23(1), 107. https://doi.org/10.1186/s12885-023-10578-4

Cashin, C., Bloom, D., & Sparkes, S. P. (2023). What influences the impact of health

financing reforms? Using qualitative comparative analysis to identify patterns in health financing systems and their effects on financial protection. Health Policy and Planning, 38(1), 26–38. https://doi.org/10.1093/heapol/czac086

Islam, M. N., Nipa, N. J., & Hossain, M. I. (2022). Implications of Big Data Analytics, AI, Machine Learning, and Deep Learning in the Health Care System of Bangladesh: Scoping Review. JMIR Medical Informatics, 10(12), e36825. https://doi.org/10.2196/36825

Ristevski, B., & Chen, M. (2023). Factors impacting the adoption of big data in healthcare: A systematic literature review. Health Information Science and Systems, 11(1), 3. https://doi.org/10.1007/s13755-022-00189-1

Esteva, A., Chou, K., & Dean, J. (2022). Big data and artificial intelligence in cancer research. Nature Reviews Cancer, 22(5), 321–334. https://doi.org/10.1038/s41568-022-00442-0

Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education – where are the educators? International Journal of Educational Technology in Higher Education, 16(1), 39. https://doi.org/10.1186/s41239-019-0171-0

Alhajj, R., & Rokne, J. (2021). Big data analytics, deep learning techniques and applications: A survey. Journal of Big Data, 8(1), 1–37. https://doi.org/10.1186/s40537-021-00429-0

Wang, Y., Kung, L. A., & Byrd, T. A. (2018). Big data in healthcare: A systematic literature review and conceptual framework. Health Information Science and Systems, 6(1), 1–10. https://doi.org/10.1007/s13755-018-0040-1

Obermeyer, Z., & Emanuel, E. J. (2016). Using machine learning for healthcare: challenges and opportunities. The New England Journal of Medicine, 375(24), 2507–2509. https://doi.org/10.1056/NEJMp1606181

Shokri, R., Stronati, M., Song, C., & Shmatikov, V. (2021). Machine learning models for secure data analytics: A taxonomy and threat model. Communications of the ACM, 64(9), 82–91. https://doi.org/10.1145/3448250

Rueda-Clausen, C. F., Silva, F. A., Ramírez, F., & Villa-Roel, C. (2006). Enfermedades de alto costo en afiliados a un sistema institucional de aseguramiento y prestación de servicios de salud. Revista de Salud Pública, 8(2), 109–120. https://doi.org/10.1590/S0124-00642006000200002

Hernández-Aguado, I., & García, A. M. (2021). Detección temprana de cáncer en población sintomática. Gaceta Sanitaria, 35(5), 457–460. https://doi.org/10.1016/j.gaceta.2020.09.005

Metrosalud. (2024). *Perfil de morbilidad de la ESE Metrosalud* [Data set].