



Machine Learning para predecir el costo anual de los servicios contratados por evento de los afiliados de la EPS Savia Salud

**Joan Mazo Solano
Laura Gallego Vélez
Sebastián Sandoval Valencia**

Especialización en Big Data e Inteligencia de Negocios
Facultad de Ingenierías y Arquitectura
Universidad Católica Luis Amigó
Medellín, Colombia
2022

**Machine Learning para predecir el costo anual de los
servicios contratados por evento de los afiliados de la EPS
Savia Salud**

**Joan Mazo Solano
Laura Gallego Vélez
Sebastián Sandoval Valencia**

Trabajo de grado, presentado como requisito para optar al título de:

Especialista en Big Data e Inteligencia de Negocios

Asesor:
Ph.D. Juan Camilo Giraldo Mejía

Especialización en Big Data e Inteligencia de Negocios
Facultad de Ingenierías y Arquitectura
Universidad Católica Luis Amigó
Medellín, Colombia
2022

Tabla de Contenido

1. Introducción.....	4
2. Motivación.....	6
3. Planteamiento del problema.....	7
4. Justificación.....	12
5. Objetivos.....	13
5.1. Objetivo general:.....	13
5.2. Objetivos específicos.....	13
6. Marco metodológico.....	14
7. Marco referencial.....	21
7.1. Marco teórico.....	21
7.2. Marco conceptual.....	22
7.3. Marco normativo.....	23
7.4. Estado del arte.....	25
8. Desarrollo del proyecto.....	29
8.1. Objetivo 1.....	29
8.2. Objetivo 2.....	36
8.3. Objetivo 3.....	45
8.4. Objetivo 4.....	58
8.5. Objetivo 5.....	59
9. Discusión.....	61
10. Conclusiones.....	62
11. Referencias.....	63
12. Anexos.....	68

Lista de figuras

Figura 1. Diagrama de espina de pescado	10
Figura 2. Proceso de liquidación de la prestación de servicios de salud	32
Figura 3. Arquitectura general de la solución	39
Figura 4. Descripción de variables cualitativas	41
Figura 5. Composición de variables cuantitativas	42
Figura 6. Resultados del Método del codo.....	49
Figura 7. Resultados Índice de Silhouette.....	49
Figura 8. Resultados de Componentes principales.....	51
Figura 9. Distribución de los grupos del costo médico	54
Figura 10. Balanceo de los grupos del costo médico	54
Figura 11. Resultados de la matriz de confusión	56
Figura 12. Resultados de la Curva ROC.....	56

Lista de tablas

Tabla 1. Resultados búsqueda bibliográfica.....	14
Tabla 2. Descripción de actividades de la Metodología CRISP-DM	15
Tabla 3. Rangos de interpretación del desempeño del modelo	20
Tabla 4. Conceptos del ingreso financiero	33
Tabla 5. Conceptos del egreso financiero operacional	34
Tabla 6. Sistemas de información Savia Salud EPS	36
Tabla 7. Resumen Diccionario de datos.....	38
Tabla 8. Descripción de la variable: Programas de alto costo	41
Tabla 9. Resultados de las medidas de tendencia central.....	43
Tabla 10. Categorización de la variable: Servicios	44
Tabla 11. Categorización de la variable: Diagnósticos	44
Tabla 12. Resultados de las variables dummy	46
Tabla 13. Descripción de los clústeres.....	50
Tabla 14. Resultados del modelo de Regresión lineal.....	52
Tabla 15. Creación de grupos y rangos según costo médico	53
Tabla 16. Resultados de desempeño de los modelos entrenados.....	55
Tabla 17. Resultado inicial de multicolinealidad	57
Tabla 18. Resultado final de multicolinealidad	57
Tabla 19. Actividades para el despliegue del modelo.....	60
Tabla 20. Diccionario de datos de la investigación.....	68

1. Introducción

Al inicio de la década de los 90, se realizó una reforma estructural al sistema de salud colombiano por medio de la Ley 100 de 1993. Con esta nueva normativa, el Estado asumió el rol de regulador del sistema y delegó la responsabilidad de la prestación de servicios a instituciones de carácter privado; de esta manera, se pretendía hacer frente a un sistema de atención en salud inequitativo con serios problemas de acceso y calidad vinculando a la población a través del aseguramiento obligatorio (Gaviria, 2013)(Otálvaro Castro et al., 2019)

En este contexto, se crean las EPS - Entidades Promotoras de Salud como instituciones responsables de la afiliación y del recaudo de los recursos. Además, de estar encargadas de organizar y garantizar la prestación de servicios de salud acorde a un plan de beneficios definido por el Estado (Ley 100 de 1993, 1993).

Desde la puesta en marcha de la Ley 100, diferentes autores han estudiado los logros, desafíos y problemas que enfrenta el país de cara a la actual estructura del sistema de salud. Si bien se han reconocido algunos logros, se han identificado problemas de accesibilidad, oportunidad, cobertura, equidad, corrupción, entre otros (Ayala-García, 2014) (Suárez Rozo et al., 2017)

La crisis se maximiza en un escenario donde la transición demográfica está aumentando la incidencia de enfermedades crónicas en todo el mundo; en los Estados Unidos, por ejemplo, los costos directos de las enfermedades crónicas para los sistemas de salud son de casi US \$214 mil millones al año. (Delpino et al., 2022).

De acuerdo con la junta directiva de la Asociación Colombiana de Empresas Sociales del Estado y Hospitales Públicos (ACESI), para diciembre de 2021, 17 EPS se encontraban en riesgo de ser liquidadas por el incumplimiento en sus indicadores financieros (la República, 2021).

En diferentes países se han realizado esfuerzos por utilizar herramientas de inteligencia artificial, con el fin de contribuir al diagnóstico temprano, el tratamiento de enfermedades, la predicción de las estancias hospitalarias y los costos asociados a la atención para aportar a la óptima gestión de sus recursos. (Delpino et al., 2022) (Navarro et al., 2018).

Con este proyecto, se pretende aportar en la elaboración del presupuesto anual de la EPS en lo concerniente a la proyección del costo de las atenciones contratadas por Evento, mediante la aplicación de técnicas de Machine Learning que permitan

explorar la relación de variables específicas con los costos operacionales de la entidad.

En la primera sección de este documento se encuentra la motivación, el planteamiento del problema y la justificación de esta investigación con base en los hallazgos obtenidos durante el estudio del negocio. Con los resultados de la primera sección, se obtienen los elementos necesarios para definir el objetivo general y los objetivos específicos establecidos entre los investigadores y la EPS.

En una segunda sección se plasman los diferentes marcos que sustentan y estructuran la investigación. En el Marco metodológico se describe las actividades propuestas y los entregables esperados, que se tienen como resultado de implementar la Metodología CRISP-DM. Luego, se presenta el Marco referencial del estudio compuesto por referentes teóricos, conceptuales, normativos y el estado del arte que contextualizan al lector.

Para la tercera sección, se describe de forma detallada el desarrollo del proyecto considerando cada una de las fases descritas en el Marco metodológico. En esta sección se presenta las necesidades del cliente, el estudio, análisis y comprensión de las variables y los datos explorados, la construcción y la evaluación de los modelos diseñados, y el despliegue de los resultados obtenidos en la entidad.

Finalmente, se exponen la discusión y las conclusiones de la investigación que plantean un panorama futuro para ser explorado por los expertos y los principales involucrados en la construcción del presupuesto anual de la EPS.

2. Motivación

La EPS Savia Salud nace como respuesta a la necesidad de asegurar a la población más vulnerable del departamento de Antioquia que cumple con los requisitos normativos para pertenecer al régimen subsidiario. En el año 2013, Savia Salud asume la responsabilidad de los afiliados que surgieron de otras EPS liquidadas, por no cumplir con los indicadores financieros y sanitarios establecidos por los Entes de control del sistema de salud. Debido a esto, Savia Salud no solo adquirió el compromiso de asegurar a la población antioqueña, sino que también contrajo las obligaciones financieras y deudas pasadas con las instituciones prestadoras (el Espectador, 2013).

En el año 2015, la Gobernación de Antioquia y el Consejo de la Ciudad de Medellín, generaron alertas sobre el déficit financiero que tenía la institución, ocasionado, entre otras cosas, porque sus gastos superan los ingresos. Más adelante en 2017, la Superintendencia Nacional de Salud - Supersalud, mediante la Resolución 2573, ordena vigilancia especial a la EPS con el fin de hacer seguimiento a los indicadores relacionados con la recuperación financiera de la institución (el Colombiano, 2015) (*Resolución Número 2573 de 2017*, 2017)

Tras varios años de prórrogas, la EPS continúa bajo medida de vigilancia especial debido al incumplimiento continuo de los indicadores financieros monitoreados por la Supersalud. La recuperación de la cartera con los hospitales ha sido adecuada hasta la fecha, pero aún no logra el punto de equilibrio esperado. Ante el riesgo de ser intervenida y, en el peor de los casos, liquidada; Savia Salud ha implementado estrategias para sufragar la cartera contraída con los hospitales desde el origen de la EPS, y a su vez, mantener el equilibrio financiero entre los nuevos gastos operacionales y no operacionales propios de la entidad (el Colombiano, 2022)).

Con base a la premisa anterior, el presente proyecto pretende aportar a las estrategias implementadas por la EPS para subsanar sus deficiencias financieras, aportando información que sea útil para la toma de decisiones en el modelo de contratación con las instituciones que prestan los servicios de salud.

El proyecto cuenta con concepto de viabilidad, acceso a la información y acompañamiento permanente de expertos del negocio, gracias a que uno de los miembros del equipo labora en la empresa. Además, para los investigadores resulta indispensable poner en práctica el conocimiento obtenido durante la especialización, y afianzar los conocimientos adquiridos en la construcción de modelos de aprendizaje automático para la solución de problemas de la sociedad.

3. Planteamiento del problema

El sistema de salud de Colombia tuvo una reforma importante en el año 1993 con la Ley 100 dictada por el Congreso y aprobada por el Gobierno vigente. Uno de los cambios estructurales más relevantes de dicha ley, fue la creación de una figura intermediaria entre el fondo de aportes y contribuciones destinadas a sostener financieramente la salud de la población, y las instituciones públicas y privadas asistenciales cuya misión es prestar servicios de salud de acuerdo con las necesidades de las personas (*Ley 100, 1993*).

En el esquema del Sistema General de Seguridad Social en Salud - SGSSS, se distinguen tres regímenes que atienden a la población de acuerdo con sus condiciones económicas y sociales. Para afiliar a la población con condiciones de vulnerabilidad se estableció el *régimen subsidiado* determinado por el Sisbén y los listados censales de grupos poblacionales específicos. La población que cuenta con un vínculo laboral o puede aportar a la financiación de su salud hace parte del *régimen contributivo*. Finalmente, los sectores que se rigen por las normas legales antes de la Ley 100 pertenecen al *régimen especial* (fuerzas militares, policía nacional, Ecopetrol, Magisterio, Universidades públicas, entre otros.) (*Ley 100, 1993*).

Dentro del sistema, hay tres entidades clave para comprender su funcionamiento, A) La Administradora de los Recursos del Sistema General de Seguridad Social en Salud - ADRES, cuyo fin es administrar las fuentes de financiación., B) Las Entidades Prestadoras de Salud - EPS, son las responsables de garantizar el aseguramiento en salud de la población y permitir la prestación de servicios de acuerdo a un Plan de Beneficios en Salud - PBS, anteriormente denominado Plan Obligatorio de Salud – POS y C) Las Instituciones Prestadoras de Servicios de Salud- IPS, como su nombre lo indica, tienen como finalidad la prestación de servicios de salud a la población mediante la atención asistencial (V. V. González, 2021) (Ministerio de Salud y Protección Social, n.d.) (*Ley 10 de 1990, n.d.*).

Las entidades anteriores interactúan de la siguiente manera: La ADRES le reconoce a las EPS la UPC por cada afiliado asegurado. Este ingreso le permite a las EPS responder a sus obligaciones administrativas, y contratar con las IPS los servicios y las tecnologías en salud con el cual será atendida la población. Como figura intermediaria, las EPS deben administrar de manera eficiente los recursos obtenidos a través de la ADRES, equilibrando las fuerzas del ingreso y el gasto financiero. En aras de mantener solvencia económica, las EPS conciertan mecanismos de pago con las IPS mediante los modelos de contratación adoptados

por el sistema de salud (Prada-Ríos et al., 2017).

De acuerdo con el Decreto 4747 de 2007, en el artículo 4, los mecanismos de pago aplicables a la compra de servicios de salud son (Decreto Número 4747 de 2007, n.d.):

a. Pago por capitación: Pago anticipado de una suma fija que se hace por persona que tendrá derecho a ser atendida durante un periodo de tiempo, a partir de un grupo de servicios preestablecido. La unidad de pago está constituida por una tarifa pactada previamente, en función del número de personas que tendrían derecho a ser atendidas.

b. Pago por evento: Mecanismo en el cual el pago se realiza por las actividades, procedimientos, intervenciones, insumos y medicamentos prestados o suministrados a un paciente durante un período determinado y ligado a un evento de atención en salud. La unidad de pago la constituye cada actividad, procedimiento, intervención, insumo o medicamento prestado o suministrado, con unas tarifas pactadas previamente.

c. Pago por caso, conjunto integral de atenciones, paquete o grupo relacionado por diagnóstico: Mecanismo mediante el cual se pagan conjuntos de actividades, procedimientos, intervenciones, insumos y medicamentos, prestados o suministrados.

Las diferentes modalidades de contratación, que hacen parte de los sistemas de pago mediante el cual las aseguradoras y los prestadores negocian, permiten fijar los precios de la atención en salud que requiera la población. Los modelos de contratación se determinan de acuerdo con el momento en el que se efectúa el pago, por lo cual son llamados pagos prospectivos y pagos retrospectivos.

En los sistemas de pagos prospectivos, la tarifa que la aseguradora establece con el prestador se determina antes de la prestación del servicio. Las modalidades más comunes dentro del sistema de pago prospectivo son la Capitación y los Presupuestos Globales. Por otro lado, en los sistemas de pagos retrospectivo, el pago que la aseguradora asume se calcula después de la prestación del servicio. La modalidad más frecuente dentro del sistema de pago retrospectivo es el pago por actividad o evento. El presente estudio, se enfoca en los sistemas de pago por evento (García-Lacalle et al., 2009).

Como lo define el decreto 4747, en la modalidad por evento, la unidad de pago se determina por cualquier actividad, intervención, procedimiento, insumo o medicamento, que hace parte de cada prestación de un servicio de salud, cuya tarifa fue concertada previamente entre la aseguradora y el prestador. En el modelo de contratación por evento, el riesgo es asumido por el financiador debido a la variabilidad de los costos (*Decreto Número 4747 de 2007*, n.d.).

El pago por evento aumenta la producción de atenciones e incentiva la productividad de los profesionales en salud. Es un modelo altamente aceptado por los prestadores debido a la rentabilidad que genera. Sin embargo, fácilmente se puede caer en la priorización de la cantidad de atenciones realizadas, por encima de la calidad del servicio o la búsqueda real de resultados en salud. Además, el pago por evento puede inducir al incremento de los costos médicos y administrativos, por el aumento de la demanda y la selección adversa de procedimientos e intervenciones con las mayores tarifas contratadas. Por otro lado, este mecanismo de pago requiere procesos extensos de auditoría y conciliación entre la aseguradora y el prestador, afectando directamente los tiempos de pago y la liquidez financiera de ambas instancias (Buglioli et al., 2002) (Álvarez et al., 2000) (Castro et al., 2014).

Como se señaló anteriormente, el modelo de contratación por evento es ampliamente empleado por las EPS que hacen parte del sistema de salud de Colombia (Castro et al., 2014). En Savia Salud, hace parte del macroproceso de Gestión de Recursos, y está constituido por veintidós (22) procedimientos relacionados a los componentes de Tesorería, Cartera, Contabilidad y presupuesto, y Cuentas médicas.

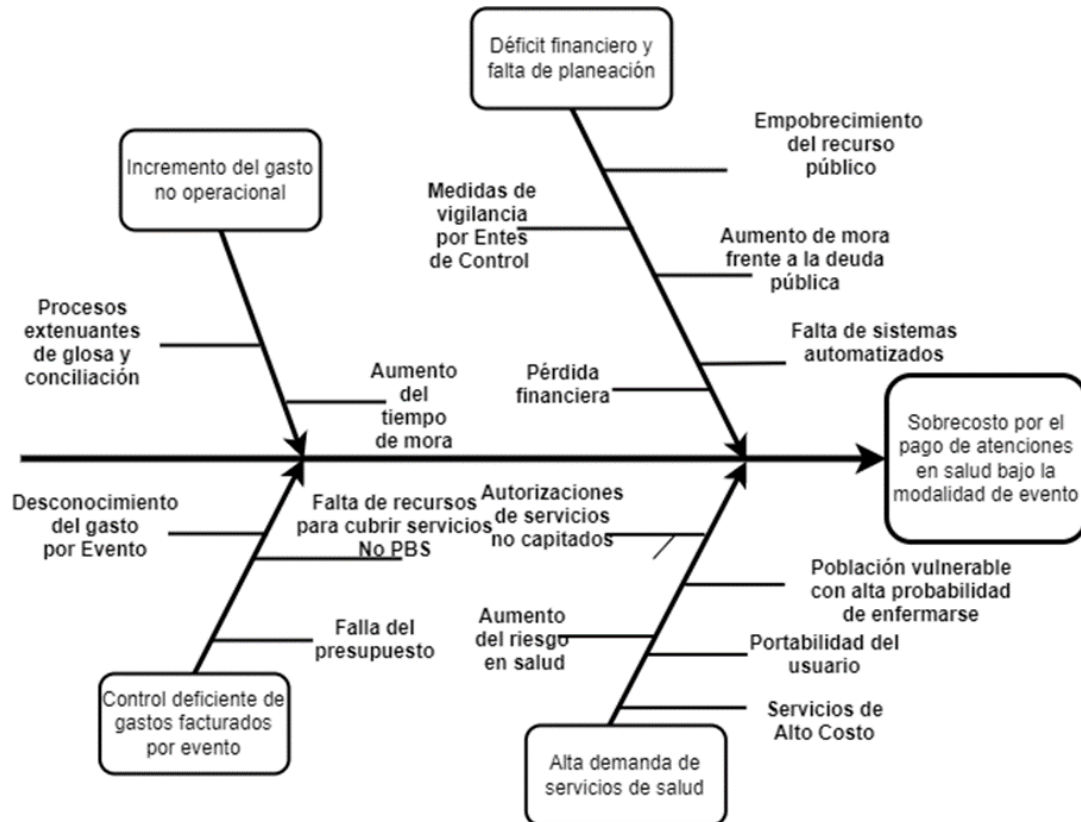
Dentro de los objetivos principales de la Gestión Financiera, se encuentra el garantizar el equilibrio financiero de la EPS por medio de la administración adecuada de los ingresos girados por la ADRES según la UPC de los afiliados, el gasto operacional y no operacional de la entidad. Para ello, se elabora el presupuesto de la EPS teniendo en cuenta el ingreso potencial y el costo de las atenciones, de acuerdo con las diferentes formas de contratación.

Aproximadamente, el 70% del costo de las atenciones en salud que son asumidas por Savia Salud está contemplado dentro del sistema de pago por evento. Este tipo de modelo de contratación implica para la EPS procesos extensos de conciliación aumentado el gasto no operacional de la entidad. Además, la variabilidad del costo de las atenciones y la dificultad para ejercer control sobre esto incrementa el gasto

operacional proyectado en los presupuestos anuales(Gonzalez Fuentes & others, 2017) (Savia Salud, n.d.) .

El problema anteriormente descrito, se puede asociar a ciertas causas identificadas en el siguiente esquema:

Figura 1. Diagrama de espina de pescado



Fuente: Elaboración propia

El déficit financiero existente y la falta de planeación e implementación de acciones de control afecta la proyección de costos relacionados con el sistema de pago por evento. Como se mencionó anteriormente, desde sus inicios, la EPS adquirió deudas y obligaciones financieras que han causado estragos en el patrimonio de la entidad, por lo cual la reserva técnica y de inversión no han sido suficientes para respaldar aquellas atenciones en salud que están fuera de los modelos de pagos prospectivos.

Sumado a lo anterior, es importante considerar el empobrecimiento del recurso público y la crisis económica latente en el país e incrementada por la recién pasada emergencia sanitaria, que perjudica directamente a las entidades que dependen de los tributos estatales, como es el caso de Savia Salud, cuya población es 92% subsidiada sin capacidad de pago (Banco Mundial, n.d.).

Otra situación que causa el sobre costo financiero a la EPS es la alta demanda de servicios de salud debido al tipo de población que atiende. Por ser una aseguradora subsidiada, dedicada a garantizar la prestación de servicios de salud a personas de bajos recursos económicos, clasificadas en grupos y subgrupos de acuerdo a la Metodología IV del Sisben, Savia Salud debe responder por el aseguramiento de usuarios en pobreza extrema, pobreza moderada y población vulnerable (Departamento Nacional de Planeación, n.d.)

La calidad de vida de los afiliados de la EPS está condicionada por los determinantes sociales de la salud, en donde la pobreza influye directamente en la salud de la población. El bienestar individual y colectivo de los territorios está ligado a factores como la oportunidad de poseer una vivienda digna, el disfrutar de servicios básicos, la garantía de una educación con calidad, el acceso seguro al transporte, la vida dentro de un medio ambiente sano, el desarrollo de la persona en medio de condiciones familiares y sociales adecuados, entre otros. De dichos factores adolece la población atendida en Savia Salud, por lo cual, la incidencia de las enfermedades y la ocurrencia de situaciones adversas es mayor (OMS, n.d.).

Por otro lado, el sobre costo que genera el modelo de contratación por evento, también resulta del incremento de los gastos no operacionales que genera el extenuante proceso de conciliación entre la EPS y los prestadores a la hora de revisar los eventos facturados relacionados a un usuario. Además, se carece de controles suficientes que auditen y garanticen alguna rentabilidad de los pagos retrospectivos, en los cuales, el riesgo es asumido por la aseguradora, y no es fácil predecir el conjunto de siniestros que puede acontecer a la población (Gonzalez Fuentes & others, 2017) (Savia Salud, n.d.).

Considerando las apreciaciones anteriores, y después de conocer la necesidad específica de la entidad frente al sobre costo generado de las atenciones pagadas bajo el modelo de contratación por evento, resulta importante para la EPS contar con más mecanismos y herramientas que le permitan obtener información relacionada al problema evidenciado, con el fin de tomar decisiones financieras y contractuales eficientes que aseguren la estabilidad económica de la entidad.

4. Justificación

El Sistema de salud depende de la interacción entre las EPS, la ADRES y las IPS. ADRES garantiza el adecuado flujo de los recursos financieros obtenidos a través de las contribuciones del gobierno y los aportes de los empleados y los empleadores, y los redistribuye a las entidades aseguradoras mediante el reconocimiento y pago de la UPC. Con este ingreso obtenido, las EPS se encargan de contratar los servicios de salud ofrecidos por las IPS, y se establecen sistemas de pago mediante unos modelos de contratación para la prestación y calidad de los servicios que den respuesta a las necesidades de la población (V. V. González, 2021).

Es importante, tanto para las EPS y las IPS, concertar un sistema de pago adecuado que garantice la prestación de los servicios de salud a la población objetivo, y propenda a su sostenibilidad financiera. Uno de los modelos de contratación más empleado dentro del sistema de salud es el pago relacionado directamente con el esfuerzo realizado, es decir, cada actividad asistencial o procedimiento clínico es convertido en un precio fijo establecido en un manual tarifario. El pago por actividad o evento conlleva problemas de monitoreo y control, y el riesgo del sobre costo es asumido por la aseguradora. Además, puede inducir en la selección adversa de servicios más costosos de parte del prestador, incluso, minimizando acciones preventivas de salud por su bajo costo en el mercado (V. V. González, 2021).

Para Savia Salud, más del 70% de los pagos efectuados a las IPS se realiza mediante el modelo de contratación por evento, y se ha percibido que los ingresos por los afiliados no compensan el sobre costo generado por las atenciones resultantes, por lo cual la rentabilidad y sostenibilidad de la entidad se ven comprometidas. De esta premisa, se desprende la necesidad de implementar controles más eficientes al pago efectuado por evento, y realizar análisis sobre la conveniencia de implementar otros modelos de contratación, o adoptar determinadas tarifas que no desgasten tanto el sistema financiero de la EPS (Molina-Marín et al., 2010) (Savia Salud, n.d.-a).

El presente proyecto pretende proponer un modelo para calcular el costo por evento de los afiliados, de acuerdo con las variables de interés que hacen parte de la tarificación anual de la UPC. Este estudio busca que Savia Salud pueda contar con una herramienta de apoyo para la construcción anual de los presupuestos, y pueda aportar a los controles financieros actualmente aplicados por la entidad.

5. Objetivos

5.1. Objetivo general:

Construir un modelo predictivo para calcular el costo anual por evento de los afiliados de la EPS Savia Salud de la ciudad Medellín.

5.2. Objetivos específicos

1. Identificar las principales características del modelo de contratación por evento para el pago de los servicios de salud de los afiliados, y su relación con los sobrecostos asumidos por la EPS.
2. Aplicar técnicas de Machine Learning para Identificar las variables relacionadas a la predicción del costo por evento de los afiliados de la EPS.
3. Construir un modelo de aprendizaje automático para predecir el costo por evento de los afiliados de la EPS
4. Evaluar el desempeño y la calidad del modelo construido mediante la aplicación de métricas de aprendizaje automático.
5. Elaborar un listado de recomendaciones para la implementación del modelo construido.

6. Marco metodológico

Al tratarse de un proyecto de minería de datos, se revisaron investigaciones publicadas en revistas científicas, con el fin de identificar las metodologías más utilizadas para proyectos de esta naturaleza, la búsqueda permitió concluir que se destacan: CRISP-DM (2006) y SEMMA -Sample, Explore, Modify, Model, Assess- (SAS, 2009b) (Moine et al., 2011).

Posteriormente, se realizó una búsqueda en las bases de datos Scopus y Sciencedirect de las investigaciones que utilizaron CRISP-DM y SEMMA obteniendo lo siguiente:

Tabla 1. Resultados búsqueda bibliográfica

Keyword	Resultados	Base de datos bibliográfica
CRISP-DM	627	Scopus
CRISP-DM	4.479	Sciencedirect
SEMMA	82	Scopus
SEMMA	390	Sciencedirect

Fuente: Elaboración propia

En ambas bases de datos bibliográficas, fue evidente que se usa con mayor frecuencia la metodología CRISP-DM. Este resultado, es consistente con la afirmación de (Sharma et al., 2017), quien señala que CRISP-DM es el modelo de referencia estándar más utilizado para el desarrollo de cualquier tipo de proyecto de analítica de datos. En esta metodología, se divide el ciclo de vida de un ejercicio de minería de datos en seis fases diferentes; en cada fase, se definen las tareas y el resultado esperado; así, el siguiente paso depende del resultado del paso anterior, lo permite volver a él y es menos rígido que un modelo de cascada tradicional utilizado en el desarrollo de software (Sharma et al., 2017).

En la siguiente tabla se describe la articulación de cada una de las fases con las actividades propuestas para el cumplimiento de cada objetivo específico.

Tabla 2. Descripción de actividades de la Metodología CRISP-DM

Objetivo específico	Fase CRISP-DM / Actividad	Entregable
<p>Objetivo 1: Identificar las principales características del modelo de contratación por evento para el pago de los servicios de salud de los afiliados, y su relación con los sobre costos asumidos por la EPS</p>	<p>Fase 1. Conocimiento de necesidades del cliente:</p> <p>Actividad 1: -Revisar la documentación publicada en el Sistema de Gestión de la Calidad de la EPS y su página web.</p> <p>Actividad 2: -Entrevistar los líderes de los siguientes procesos:</p> <ul style="list-style-type: none"> • Aseguramiento • Cuentas médicas • Gestión del riesgo • Acceso a servicios • Analítica y TI <p>Actividad 3: -Identificar las principales características del proceso actual de proyección de ingreso y gastos operacionales.</p> <p>Actividad 4: -Definir la situación actual de la EPS y los aportes de la investigación.</p>	<p>-Documentación de la situación actual de los sistemas de pago empleados por la EPS, frente al ingreso percibido por la UPC de cada afiliado</p> <p>-Diagrama de los principales procesos de la EPS relacionados con la atención de pacientes por evento</p>
<p>Objetivo 2: Aplicar técnicas de Machine Learning para Identificar las variables relacionadas a la predicción del costo por evento de los afiliados de la EPS.</p>	<p>Fase 2. Estudio y comprensión de los datos:</p> <p>Actividad 1: -Revisar los sistemas de información de la EPS, y la información administrada en cada sistema.</p> <p>Actividad 2: -Solicitar y revisar los datos para construir el modelo.</p>	<p>-Diccionario de datos de las variables seleccionados para el estudio</p> <p>-Dataframe de afiliados cuyos servicios médicos fueron facturados por evento</p>

Objetivo específico	Fase CRISP-DM / Actividad	Entregable
	<p>Actividad 3: -Describir los criterios de selección de la muestra</p> <p>Actividad 4: -Construir el diccionario de datos.</p>	
<p>Objetivo 2: Aplicar técnicas de Machine Learning para Identificar las variables relacionadas a la predicción del costo por evento de los afiliados de la EPS.</p>	<p>Fase 3. Análisis de los datos y selección de características</p> <p>Actividad 1: -Integrar los datos de los afiliados en registros con la misma granularidad.</p> <p>Actividad 2: -Eliminar las variables irrelevantes para el modelo, y limpieza de datos (búsqueda de datos nulos, datos atípicos y datos con formatos no consistentes).</p> <p>Actividad 3: -Realizar un análisis descriptivo de los datos</p>	<p>-Dataframe con los registros definitivos que serán parte del modelo predictivo</p>
<p>Objetivo 3: Construir un modelo de aprendizaje automático para predecir el costo por evento de los afiliados de la EPS</p>	<p>Fase 4. Construcción de los modelos:</p> <p>Actividad 1: - Desarrollar un modelo de Clusterización de los datos (Modelo de aprendizaje no supervisado)</p> <p>Actividad 2: -Desarrollar diferentes modelos de clasificación (Aprendizaje supervisado)</p>	<p>Modelos de machine learning con sus respectivas métricas Predicción del costo de la atención por evento de los afiliados de la EPS</p>
<p>Objetivo 4: Evaluar el desempeño y la</p>	<p>Fase 5. Evaluación de resultados:</p>	<p>Cuadro comparativo de métricas de los modelos</p>

Objetivo específico	Fase CRISP-DM / Actividad	Entregable
calidad del modelo construido mediante la aplicación de métricas de aprendizaje automático.	Actividad 1. -Evaluar las métricas de cada uno de los modelos de clasificación	
Objetivo 5: Elaborar un listado de recomendaciones para la implementación del modelo construido.	Fase 6. Socialización y despliegue: Actividad 1. Elaborar un listado de recomendaciones para la implementación del modelo en la EPS	Acta de cierre firmada y recomendaciones

Fuente: Elaboración propia

Con la implementación de la metodología descrita anteriormente, se espera obtener la documentación y el algoritmo que sustenta el modelo predictivo relacionado con el costo por evento de los afiliados seleccionados para el estudio. Dicho modelo, permite a la entidad comprender la relación de variables asociadas al gasto operacional bajo el modelo de contratación por evento, y sentará las bases para continuar explorando el tema y las acciones que mitiguen el sobrecosto financiero relacionado.

Para el desarrollo de cada una de las fases y las actividades se llevarán a cabo las siguientes actividades:

1. *Conocimiento de necesidades del cliente*

Con el fin de conocer las necesidades del negocio de la EPS Savia Salud, se realizará una serie de entrevistas semiestructuradas con los líderes con las áreas de Aseguramiento, Planeación y TI, con el fin de conocer el funcionamiento de los procesos, con especial énfasis en las actividades, procedimientos y tareas que se derivan de la atención de los pacientes que son atendidos bajo el modelo de contratación por evento.

Cada una de las entrevistas, permitirá elaborar un diagrama de en el que se describen los detalles de cada uno de los procesos, y se observan puntos clave donde se trabaja para alcanzar objetivos de sostenibilidad financiera y el saneamiento de cartera que permitirá superar la vigilancia especial a la que es sometida la EPS.

2. *Estudio y comprensión de los datos*

La fase de conocimiento de las necesidades cliente, finaliza con la caracterización de los principales datos que se gestionan en el área de aseguramiento y que servirán para la construcción del modelo predictivo indicado en el objetivo del proyecto.

Para la caracterización de los datos, se firmará un acuerdo institucional que nos permitirá acceder a las bases de datos con fines estrictamente académicos, respetando la custodia, privacidad y confidencialidad de la información. Una vez se tenga acceso a las bases de datos institucionales donde hay datos de los afiliados, se harán consultas utilizando lenguaje SQL y se consolidará la información final en archivos con formato csv.

3. *Análisis de los datos y selección de características*

Para el desarrollo de esta fase, se aplicarán los 6 pasos recomendados para la limpieza y preparación de los datos:

A. Integración de los datos

Se integrarán datos de los diferentes archivos, con fin de tener un archivo único con todas las variables de interés, la integración se realizará búsquedas por clave en Excel.

B. Eliminar variables irrelevantes y redundantes

Con los datos consolidados, se eliminarán variables irrelevantes para el modelo como la identificación del paciente, el teléfono, la dirección, el estado del afiliado y régimen de afiliación.

C. Descripción estadística de los datos

Para la descripción estadística, se utilizarán librerías de código abierto de Python, la descarga y procesamiento del lenguaje se realizará en la herramienta de uso libre de Google – Colab; allí se podrá cargar el set de datos y se aplicarán medidas de resumen, de tendencia central y se elaborarán gráficos descriptivos para las variables categóricas.

D. Limpieza de datos

Para garantizar la integralidad de los datos que se utilizarán en la construcción del modelo, se eliminarán datos nulos, se imputarán los valores atípicos y se verificará para cada una de las variables que los valores estuvieran acorde con las demás respuestas y el tipo de variable.

E. Transformaciones

Si se tienen variables que representan muchas categorías, se realizará una transformación, con el fin de reducir la dimensionalidad de la misma, en caso de ser necesario, se discretizarán las variables cuantitativas y se generarán columnas calculadas si el modelo lo requiere.

F. Análisis de correlaciones - (Análisis de colinealidad)

Con el fin de verificar si las variables independientes se encuentran correlacionadas entre sí (colinealidad), se aplicará un algoritmo con el fin de verificar el factor de inflación de varianza (VIF). De acuerdo con el resultado de la prueba, se eliminarán todas las variables del modelo que tengan un VIF mayor a 5 y se entrenará nuevamente los modelos para verificar el resultado.

4. *Construcción de los modelos*

Para el modelado de machine learning, se utilizarán las librerías de código abierto de Python, dividiendo los datos en un conjunto de entrenamiento y prueba de datos a partir de la información histórica suministrada por la EPS.

Para empezar, se realizará un modelo de regresión lineal, que nos permita identificar la capacidad explicativa de las variables independientes sobre la variable dependiente, de acuerdo con el resultado, se evaluará la posibilidad de construir un modelo de clasificación utilizando los siguientes algoritmos: A) Árboles de decisiones, B) Máquinas de soporte vectorial, C) Naive Bayes y D) Regresión Logística. Cada uno de los modelos será probado utilizando el método Split (80-20)

5. *Evaluación de resultados*

Con el fin de seleccionar el modelo con mejor comportamiento en la predicción, se realizará una evaluación de los principales estadísticos. Para el modelo de

regresión, se evaluará el R cuadrado¹ y el F-Score.

Los modelos de clasificación por su parte, se evaluarán con la revisión de los estadísticos de exactitud, precisión y sensibilidad. Además, en caso de ser necesario se revisará el F1 Score; esta medida nos facilitará la interpretación del modelo y nos permitirá concluir de acuerdo con el siguiente rango la clasificación que puede otorgarse tanto en el conjunto de entrenamiento como de pruebas.

Tabla 3. Rangos de interpretación del desempeño del modelo

Puntuación F1	Interpretación
> 0,9	Muy bueno
0,8 - 0,9	Bueno
0,5 - 0,8	OK
<0,5	No es bueno

Fuente: Elaboración propia

6. Despliegue

Para el proceso de despliegue del modelo y la eventual incorporación del mismo al proceso de planificación de la EPS, se realizarán un conjunto de sugerencias, con el fin de continuar su desarrollo e integración a los procesos de estimación de costos de atenciones por evento de los afiliados. En este apartado, se listan las responsabilidades y actividades en un cronograma de tal manera que se siga explorando la incorporación de técnicas de machine learning dentro de los procesos financieros de la institución.

El informe final de los resultados, se entregará a los principales interesados de la entidad, con el fin de que continúen avanzando en el desarrollo del mismo y lo utilicen para apoyar la toma de decisiones.

Finalmente, se realizará un acta de cierre, en la cual se hará entrega de los productos que quedaron consignados en el acuerdo con la EPS.

¹ El R-cuadrado es una medida estadística de qué tan cerca están los datos de la línea de regresión ajustada. También se conoce como coeficiente de determinación, o coeficiente de determinación múltiple si se trata de regresión múltiple, entre más cerca esté del 1, indica que el modelo explica toda la variabilidad de los datos de respuesta en torno a su media.

7. Marco referencial

7.1. Marco teórico

Las Instituciones Prestadoras de Salud poseen estructuras administrativas similares a las organizaciones cuyo fin es transformar y/o comercializar bienes y servicios, así mismo su funcionamiento se impacta con factores como: flujo y optimización de procesos, manejo de costos y de recursos, etc. Todos los procesos inherentes de la operación generan volúmenes de información útil para tomar decisiones que permitan maximizar la rentabilidad, o minimizar los costos de operación (Russom, 2011).

La prestación de servicios de salud a la población se puede considerar como un modelo de producción donde deben ser muy claros los procesos y procedimientos que garantizan la cobertura y accesibilidad en salud de los territorios. Al considerar los servicios de salud como cadenas de producción, se pueden implementar técnicas como el modelamiento de escenarios que puedan simular infinidad de posibles decisiones y resultados, esto sin el riesgo de romper el flujo de los servicios en situaciones reales. Estos ejercicios proveen de información valiosa que beneficia a las áreas administrativas, y permiten identificar fallas en los procesos, así también generan información que se pueda comparar con diferentes entidades del sector o consigo misma (Boyle et al., 2021).

Existen diferentes posibilidades en el sector salud para la aplicación de métodos y ejecución de soluciones a medida. El sector de la salud y su exponencial crecimiento se ha convertido en un gran generador de información, así también las técnicas para transformar y generar conocimiento nuevo a través de esa información le conocemos como Big data. El Big data y la Analítica abren un abanico de posibilidades que van desde la reducción de costos en el sistema médico hasta la optimización del capital humano (Bates et al., 2017).

Con el uso de métodos de captura y transformación de datos, aumentan las oportunidades de integración de la información entre diversos dataframes, esto en el futuro podría permitir identificar historial de pacientes, enfermedades y costos asociados sin importar en qué lugar o entidad haya recibido atención médica. Mediante el mapeo de términos de diferentes conjuntos de datos en un esquema coherente se puede lograr mediante el uso de machine learning para la clasificación de textos (Ferreira Alfaya & Zarzuelo Romero, 2022).

En la actualidad se empiezan a aplicar técnicas de machine learning a datasets capturados por medio de dispositivos HEOR (health economics and outcomes research), el HEOR es un método que a su vez se dedica a la investigación del costo de las intervenciones médicas y el resultado de esas intervenciones en los pacientes. Esta información se usa para controlar y monitorear los resultados del sector salud actual pero podría ser usado para pronosticar resultados futuros (Lee et al., 2022).

7.2. Marco conceptual

Para comprender el contexto, el desarrollo de los objetivos y la metodología planteada en este proyecto es necesario precisar las definiciones de algunos términos, marcos de trabajos y tecnologías utilizadas.

Inteligencia artificial: Es el conjunto de ciencias, teorías y técnicas con el propósito de reproducir en una máquina las habilidades cognitivas de un ser humano” (Iranzo Jiménez, 2021).

Machine Learning: Es el área de la inteligencia artificial que engloba un conjunto de técnicas que hacen posible el aprendizaje automático a través del entrenamiento con grandes volúmenes de datos (Russo et al., 2016).

Aprendizaje supervisado: En el aprendizaje supervisado, se entrena a la máquina con datos etiquetados a modo de ejemplo, aportando así una guía de aprendizaje. Por otra parte, en el aprendizaje no supervisado los datos no están etiquetados, por lo que la máquina busca similitudes y forma grupos en función de estas (Iranzo Jiménez, 2021).

Big data: Big Data es un término que describe un gran volumen de datos complejos y variables que requieren técnicas y tecnologías avanzadas para permitir la captura, almacenamiento, distribución, gestión, y análisis de la información (Aldana et al., 2018).

Analítica: La analítica de datos podría definirse como un subproceso en la extracción de la información de un conjunto de datos, en el que se adquiere inteligencia de esta para la toma de decisiones (Gandomi & Haider, 2015).

Python: Python es un lenguaje de programación ampliamente utilizado en las aplicaciones web, el desarrollo de software, la ciencia de datos y el machine

learning (ML) (AWS, n.d.).

Colab: Colaboratory o "Colab" para abreviar, es un producto de Google Research. Permite a cualquier usuario escribir y ejecutar código arbitrario de Python en el navegador. Es especialmente adecuado para tareas de aprendizaje automático, análisis de datos y educación (Google, n.d.).

CRISP-DM: Es un marco de trabajo, creado por el grupo de empresas SPSS, NCR y Daimlerchrysler en el año 2000, con el pasar del tiempo, se ha convertido en la guía de referencia más utilizada en el desarrollo de proyectos de minería de datos, contemplado el proceso en las siguientes seis fases: A) Entendimiento del negocio, B) Entendimiento de los datos, C) Preparación de los datos, D) Modelado, E) Evaluación, F) Evaluación y G) Despliegue (Moine et al., 2011) (Gómez, n.d.)

7.3. Marco normativo

Durante el desarrollo de esta investigación, se tuvieron en cuenta las siguientes normas:

Ley 100 de 1993: Por la cual se crea el Sistema de Seguridad Social Integral y se dictan otras disposiciones (*Ley 100, 1993*)

Decreto 050 de 2003. Por el cual se adoptan unas medidas para optimizar el flujo financiero de los recursos del régimen subsidiado del Sistema General de Seguridad Social en Salud y se dictan otras disposiciones(Savia Salud, 2018).

Decreto 3260 de 2004. Por el cual se adoptan medidas para optimizar el flujo de recursos en el Sistema General de Seguridad Social en Salud (Savia Salud, 2018).

Decreto 4747 de 2007: por medio del cual se regulan algunos aspectos de las relaciones entre los prestadores de servicios de salud y las entidades responsables del pago de los servicios de salud de la población a su cargo, y se dictan otras disposiciones (Savia Salud, 2018).

Ley 1122 del 2007: Por la cual se hacen algunas modificaciones en el Sistema General de Seguridad Social en Salud teniendo como prioridad el mejoramiento en la prestación de los servicios a los usuarios. Con este fin se hacen reformas en los aspectos de dirección, universalización, financiación, equilibrio entre los

actores del sistema, racionalización, y mejoramiento en la prestación de servicios de salud, fortalecimiento en los programas de salud pública y de las funciones de inspección, vigilancia y control y la organización y funcionamiento de redes para la prestación de servicios de salud, y se dictan otras disposiciones (Savia Salud, 2018).

Ley 1438 del 2011: Tiene como objeto el fortalecimiento del Sistema General de Seguridad Social en Salud, a través de un modelo de prestación del servicio público en salud que en el marco de la estrategia Atención Primaria en Salud permita la acción coordinada del Estado, las instituciones y la sociedad para el mejoramiento de la salud y la creación de un ambiente sano y saludable, que brinde servicios de mayor calidad, incluyente y equitativo, donde el centro y objetivo de todos los esfuerzos sean los residentes en el país. Incluye disposiciones para establecer la unificación del Plan de Beneficios para todos los residentes, la universalidad del aseguramiento y la garantía de portabilidad o prestación de los beneficios en cualquier lugar del país, en un marco de sostenibilidad financiera (Savia Salud, 2018).

Ley 1608 de 2013: Por medio de la cual se adoptan medidas para mejorar la liquidez y el uso de algunos recursos del sector salud (cuentas maestras) (Savia Salud, 2018).

Decreto 1683 de 2013: por medio del cual se reglamenta el artículo 22 de la Ley 1430 de 2011 sobre portabilidad nacional en el Sistema General de Seguridad Social en Salud (Savia Salud, 2018).

Decreto 3047 de 2013: por medio del cual se establecen las reglas sobre movilidad entre regímenes para afiliados focalizados en los niveles I y II del SISBÉN (Savia Salud, 2018).

Resolución 5592 de 2015: por la cual se actualiza integralmente el Plan de Beneficios en Salud con cargo a la Unidad de Pago por Capitación-UPC del Sistema General de Seguridad Social en Salud —SGSSS y se dictan otras disposiciones (Savia Salud, 2018).

Resolución 1479 de 2015: por la cual se establece el procedimiento para el cobro y el pago de servicios y tecnologías sin cobertura en el plan obligatorio de Salud suministradas a los afiliados al Régimen Subsidiado (Savia Salud, 2018).

Ley 1712 de 2014: Por medio de la cual se crea la Ley de Transparencia y del Derecho de Acceso a la Información Pública Nacional y se dictan otras disposiciones (Savia Salud, 2018).

Ley 1581 de 2012: Por la cual se dictan disposiciones generales para la protección de datos personales (Savia Salud, 2018).

7.4. Estado del arte

Para el desarrollo de este proyecto, se realizó una búsqueda de publicaciones científicas en revistas indexadas en las bases de datos bibliográficas ScienceDirect y Scopus con el fin de conocer el panorama actual de las instituciones del sector salud frente a la implementación de estrategias de analítica, establecer puntos de referencia, metodologías y buenas prácticas adoptadas en proyectos de gestión de los datos generación de conocimiento.

De acuerdo con (Murdoch & Detsky, 2013), la implementación de la analítica en el sector salud ha resultado ser un desafío que enriquece la generación de conocimiento y la toma de decisiones en las organizaciones que administran y prestan servicios de salud a la población. Mediante el análisis automatizado de cifras, resultados e indicadores, las instituciones han descubierto nuevo conocimiento que permite una comprensión integral de los procesos involucrados en la promoción, prevención y asistencia en salud (Murdoch & Detsky, 2013).

Los modelos predictivos y prescriptivos que resultan de la analítica se han convertido en el insumo más importante para los ejercicios de planeación estratégica dentro de las organizaciones, permitiendo realizar gestión del riesgo, priorización de poblaciones y el uso óptimo de los recursos (Groves et al., 2013).

En Pakistán, por ejemplo, un grupo de investigadores utilizó los datos de diferentes hospitales para realizar un análisis descriptivo, predictivo y prescriptivo durante la pandemia por COVID-19. Para el procesamiento de los datos, construyeron una red neuronal que permitió detectar si un paciente era portador del virus y sus probabilidades de recuperación de acuerdo con los síntomas. Con este modelo, lograron identificar que, en la mayoría de los pacientes diagnosticados positivamente, los síntomas fueron gripe, fiebre, dolor de cabeza, tos y dolor de garganta entre otros; con ello, optimizaron el proceso de atención, ya que hicieron recomendaciones tempranas para el inicio de tratamiento sin esperar los resultados

de laboratorio (Ahmed et al., 2021).

En el continente europeo, también se encontraron investigaciones que han utilizado técnicas diferentes de minería de datos y aprendizaje automático. En el artículo “Intelligible Support Vector Machines for Diagnosis of Diabetes Mellitus”, los autores elaboraron un modelo híbrido para el muestreo y la construcción de un algoritmo de máquinas de soporte vectorial, para el diagnóstico y predicción de Diabetes Mellitus. El modelo fue validado con datos reales de 4682 sujetos de más de 20 años, posibilitando clasificar el riesgo de futuros pacientes a partir de los datos clínicos, demográficos y de sus medidas antropométricas (Barakat et al., 2010).

En otro estudio realizado en Reino Unido cuyo objetivo fue predecir la aparición de neumonía tras una infección de las vías respiratorias, también se utilizaron estrategias híbridas, en este caso, para seleccionar de un conjunto de variables candidatas aquellas que harían parte del modelo predictivo. Los investigadores emplearon técnicas como la regresión logística y bosques aleatorios, con las que seleccionaron 31 variables que se incluyeron en el modelo predictivo; para este último, utilizaron algoritmo CART que hace parte de los árboles de decisiones logrando predecir el riesgo de reingreso de los pacientes durante los primeros 30 días a los servicios de salud por neumonía, posterior a una infección del tracto respiratorio (Lee et al., 2022).

Otras investigaciones, optan por el desarrollo de diferentes modelos de aprendizaje automático, con el fin de seleccionar el que obtenga mejores resultados. En Filadelfia Estados Unidos, por ejemplo, se realizó una investigación para predecir la obesidad infantil desde la edad > 2 a ≤ 7 años utilizando los registros de las historias clínicas de los pacientes pediátricos. En este estudio, desarrollaron 7 modelos diferentes: Árboles de decisiones, Gaussian Naive Bayes, Bernoulli Naive Bayes, regresión logística, redes neuronales, máquinas de soporte vectorial y XGBoost. Después de realizar las respectivas comparaciones, los autores concluyeron que las predicciones del modelo XGB fueron significativamente diferentes de todos los demás, logrando identificar otras variables que influyen en la obesidad infantil y que son potencialmente nuevas como el perímetro cefálico, la temperatura corporal y la frecuencia respiratoria (Pang et al., 2021).

Las técnicas de aprendizaje automático descritas anteriormente, permiten identificar patrones difíciles de encontrar en grandes volúmenes de datos, por ello cada vez son más utilizadas en el sector salud. Para la construcción de los modelos, no solo suelen utilizarse datos de las historias clínicas de hospitales; también, se utilizan bases de datos de portales científicos, un equipo de

investigadores por ejemplo, utilizó los registros de cáncer de colon disponibles en la plataforma GDC (Genomic Data Commons Data Portal) con el fin de identificar patrones genéticos en pacientes con esta enfermedad; durante este proceso, utilizaron máquinas de soporte vectorial, bosques aleatorios y árboles de decisiones. Con los resultados del modelo, los autores pretenden aportar al diagnóstico del cáncer de colon, cuya detección se realiza principalmente por medio de colonoscopias y otros exámenes más tradicionales, con ello, esperan impulsar más modalidades de tratamiento y proporcionar bases para terapias dirigidas (Su et al., 2022).

La utilización de fuentes de datos alternativas, también se ha realizado en otras esferas de la salud, en China, realizaron un estudio utilizando datos de la red social WeChat con el fin de detectar el riesgo de suicidio mediante un algoritmo de aprendizaje automático. En este proceso, construyeron un modelo de bosques aleatorios con el fin de aportar información al personal médico y de entornos escolares en la detección temprana de conductas asociadas al suicidio (McMullen et al., 2021).

Además de los objetivos clínicos previamente descritos, las técnicas de analítica de datos también han sido empleadas para fortalecer la sostenibilidad financiera de los sistemas de salud. La transición demográfica está aumentando la incidencia de enfermedades crónicas en todo el mundo; en los Estados Unidos, por ejemplo, los costos directos de las enfermedades crónicas para los sistemas de salud son de casi US \$214 mil millones al año; a esto, se le suman los costos por pérdida de productividad laboral que ascienden a US\$ 138 mil millones. Por ello, en diferentes países se han realizado esfuerzos por desarrollar modelos de aprendizaje automático que apoyen el diagnóstico temprano y el tratamiento de enfermedades (Delpino et al., 2022).

Navarro et al (2018), por ejemplo, utilizó un modelo de machine learning aplicando Naive Bayes con el fin de pronosticar la duración de la estancia y el costo en pacientes sometidos a artroplastia total de rodilla. Para la construcción del algoritmo, empleó datos como la edad, el origen étnico, el sexo, tipo de ingreso, el riesgo de mortalidad y la gravedad de la enfermedad. Con los resultados, se propuso al Programa de Cobertura de Seguridad Social de Estados Unidos (Medicare) un mecanismo de pago basado en el riesgo (Navarro et al., 2018).

Delpino et al (2022) encontró al menos 40 estudios realizados en países como China, India, Estados Unidos, Brasil, Canadá, Finlandia, Taiwán, Omán, Irak, Pakistán, Egipto, Suiza, Japón, Italia, Alemania, Australia y Jordania en los que

utilizaron modelos de aprendizaje automático para el diagnóstico o tratamiento de enfermedades crónicas. De acuerdo con el autor, las enfermedades más evaluadas fueron la diabetes y la enfermedad renal crónica y, los modelos más utilizados fueron las máquinas de soporte vectorial, los bosques aleatorios y la regresión logística (Delpino et al., 2022).

En otro estudio realizado en Tailandia, el objetivo fue comparar el resultado de diferentes modelos de aprendizaje automático, con el fin de predecir los costos de visitas al hospital por enfermedades crónicas. Para ello, probaron 8 tipos de modelos diferentes utilizando máquinas de soporte vectorial, redes neuronales, bosques aleatorios y XGBoost. Los hallazgos permitieron concluir que los bosques aleatorios presentaron un mejor resultado respecto de los demás modelos y con los resultados, podrían sugerir estrategias de salud pública para la población objetivo que tiene mayor riesgo (Thongpeth et al., 2021).

En Colombia, también se encontraron investigaciones que utilizaron técnicas de aprendizaje automático para la detección y tratamiento de enfermedades. Castro-Villarreal et al (2021) utilizó el algoritmo XGBoost para la estimación de la prevalencia y los costos incrementales del lupus eritematoso sistémico. Gonzalez Rodriguez (2019), usó la regresión logística y bosques aleatorios para identificar las principales variables relacionadas con la predicción del gasto farmacéutico asociado a la diabetes (Castro-Villarreal et al., 2021) (Gonzalez Rodriguez et al., 2019).

8. Desarrollo del proyecto

8.1. Objetivo 1: Identificar las principales características del modelo de contratación por evento para el pago de los servicios de salud de los afiliados, y su relación con los sobrecostos asumidos por la EPS

Fase 1. Conocimiento de necesidades del cliente

Actividad 1.

Como parte del acercamiento inicial que se hizo a la entidad, se revisó la información oficial de la EPS que reposa en el Sistema de Gestión de la Calidad y los informes de Rendición de cuentas publicados en la página web. Mediante estas dos fuentes, se examinó la documentación de todos los procesos, procedimientos, indicadores, manuales y demás registros que respaldan la misión y visión institucional. La revisión documental permitió conocer la situación actual de la EPS, y la relación del sobrecosto financiero con los diferentes sistemas de pago contratados con las instituciones prestadoras de servicios de salud, como a continuación se detalla.

Savia salud tiene como objetivo principal la prestación de servicios de salud que garanticen el cuidado continuo y la atención integral de la población afiliada. Para ello requiere conformar una Red de Servicios con base al número de afiliados, la cobertura en los municipios, la capacidad financiera de la EPS, el perfil epidemiológico y gestión de riesgo de la población, y la oferta de las instituciones prestadoras de servicios de salud por nivel de complejidad y especialidad (MD-RS-01 Modelo de Red de Servicios, n.d.)

La conformación de una Red de Servicios suficiente es una de las metas más importantes de la entidad, y pone a prueba la capacidad administrativa, gerencial y financiera de la EPS. La entidad realiza todo un proceso de planeación, identificación de la oferta y la demanda, negociación del sistema de pago, y contratación de servicios con los prestadores disponibles. A partir del año 2020, y mediante el Plan de modernización, la Red de Servicios se enfocó en cuatro (4) modelos importantes (Savia Salud, n.d.-b)

- a) La *gestión del riesgo* mediante acciones de prevención y promoción de la salud,
- b) la *prestación de servicios* ambulatorios o en el domicilio, centrados en el autocuidado, c) la administración del *riesgo financiero* cuyo fin es administrar

correctamente los ingresos percibidos por la UPC, d) la incorporación del *Gobierno corporativo* mediante la articulación de la EPS, los prestadores y las autoridades sanitarias (Savia Salud, n.d.-b)

Dichos modelos están sustentados en dos componentes necesarios para la conformación de la Red de Servicios:

- *Componente primario*: Es el responsable de las acciones definidas como primarias y básicas que satisfacen las necesidades más frecuentes de la población, mediante la garantía de servicios de baja y mediana complejidad. Dicho componente tiene un costo bajo y controlado para la entidad, y cuenta con una amplia oferta de prestadores disponibles contractualmente.

Las actividades propias del componente primario se contratan mediante un modelo de capitación y presupuesto global prospectivo, con un sistema de incentivos para el cumplimiento de metas orientadas a la prevención y promoción de la salud.

- *Componente complementario*: Es el responsable de las acciones de mayor complejidad y de la atención a situaciones de salud que requieren tecnología y competencias más especializadas, denominados servicios de mediana y alta complejidad. Dicho componente tiene un mayor costo para la entidad, y tiene una oferta escasa en el mercado.

Por otro lado, el componente complementario tiene más particularidades a considerar a la hora de conformar la Red de Servicios. En primer lugar, los prestadores de servicios de mediana y alta complejidad están ubicados fundamentalmente en centros urbanos de determinados municipios y en el Valle de Aburrá. Los servicios son contratados mediante invitaciones públicas, deben cumplir ciertos requisitos técnicos, financieros y administrativos, y son mediados mediante sistemas de pago de tipo evento, paquetes integrales de salud y grupos relacionados de diagnóstico (PD-RS-11 Procedimiento de Contratación Con Prestadores de Servicios de Salud, n.d.)

La Red de Servicios establecida por la EPS debe, además de garantizar cobertura y accesibilidad a la población afiliada, ser sostenible financieramente a través del tiempo, teniendo en cuenta la relación entre los ingresos operacionales y el costo

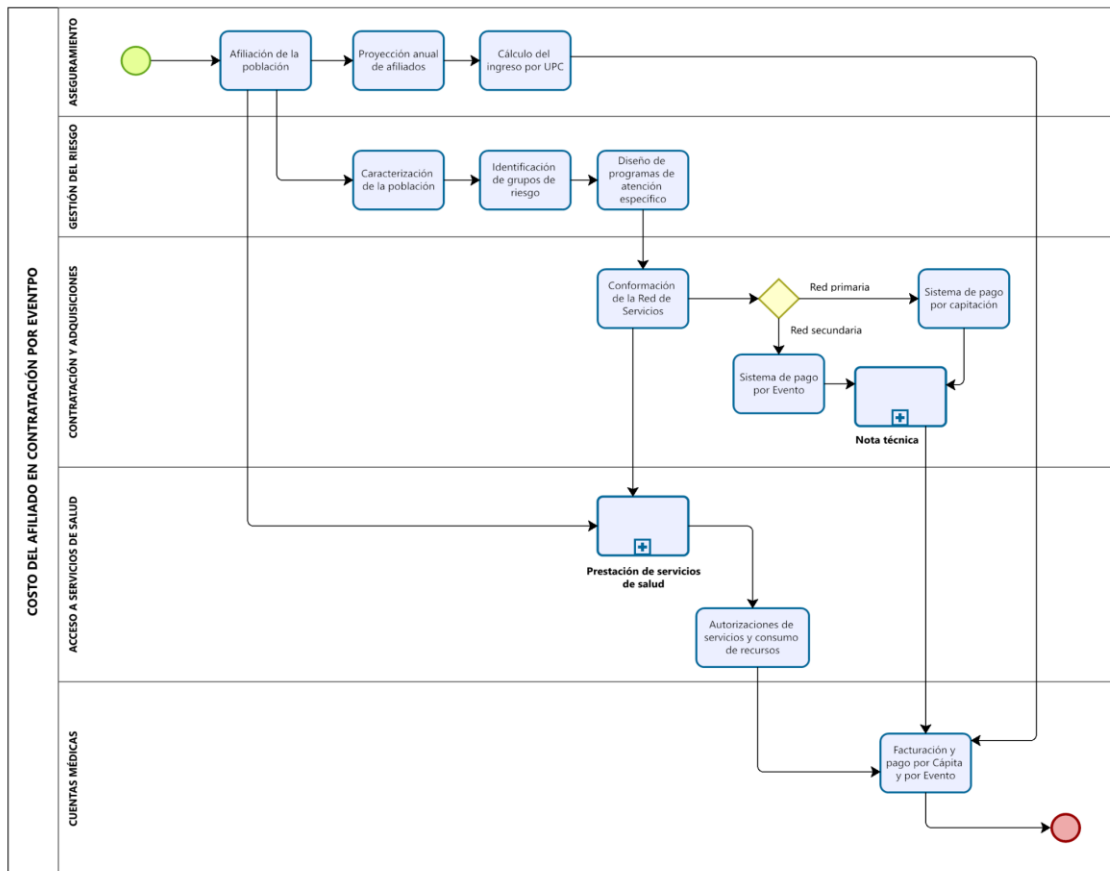
de los servicios contratados. Para ello, la entidad cuenta con una Nota técnica que reúne los lineamientos necesarios para determinar los costos de los servicios que se planean contratar sin importar el sistema de pago que se defina según el componente. La Nota técnica tiene su sustento normativo en el Decreto 441 del 2022 (Presidente de la República de Colombia, 2022).

La Nota técnica está compuesta por información de: población, frecuencias de uso, cohorte, prevalencia, incidencia, servicios y/o tecnologías, tarifas, costo usuario, monitoreo y evaluación, modalidad del contrato, temporalidad y caracterización poblacional. Las tarifas que se definen se basan en los manuales tarifarios vigentes, tarifas de referencia con el mercado, tarifas contractuales actuales, costo del talento humano, y otras variables propias de los servicios a contratar (Ministerio de Salud y Protección Social, n.d.)

La Red de Servicios en su componente complementario es negociado con los prestadores de mediana y alta complejidad, quienes ofertan sus servicios, y concilian con la EPS los precios que se establecerán de acuerdo con tarifas propias, o tarifas vigentes tales como: el SOAT, el ISS2000 y el ISS2001 (IN-RS-14 Instructivo de Asignación de Tarifas y Autorización de Servicios y Tecnologías En Salud No Contratadas, n.d.)

Cabe destacar que, a mayor nivel de complejidad el costo de los servicios aumenta, y en su mayoría de veces, no logra ser cubierto por los ingresos operacionales que la EPS recibe por cada afiliado. Debido al costo de los servicios de mediana y alta complejidad, Savia Salud contrata por evento dichas atenciones, ya que no logran ser captadas dentro de los modelos de contratación por cápita, y resulta más conveniente para las instituciones prestadoras recibir el pago por los eventos o actividades en salud, ya que el riesgo lo asume completamente la EPS.

El proceso descrito anteriormente se representa en el siguiente diagrama de proceso:

Figura 2. Proceso de liquidación de la prestación de servicios de salud

Fuente: Elaboración propia

Actividad 2.

Adicional a la información obtenida en la actividad anterior, se realizaron entrevistas con el Director del área de Aseguramiento, el Subgerente Financiero, el Coordinador de Analítica - TI, el Analista de información del área de Gestión del riesgo y los profesionales especializados de presupuesto. La información suministrada permitió conocer el proceso de ingreso por cada afiliado de acuerdo con los montos establecidos en la UPC, los sistemas de pago implementados en la EPS, el sobrecosto percibido en el modelo de liquidación por evento, y la relación de algunas variables con en el déficit financiero que presenta la entidad (enfermedades de alto costo, asignación de la portabilidad, incremento de atenciones en salud).

Con el conocimiento obtenido los acercamientos directos que se tuvieron con los principales involucrados de la EPS, se logró identificar el proceso actual realizado para la elaboración del presupuesto mensual. Con esta información, se dio paso al desarrollo de la siguiente actividad.

Actividad 3.

En la actualidad, la EPS calcula la proyección de ingresos y gastos operacionales y no operacionales que espera tener en la siguiente anualidad, teniendo en cuenta los históricos de los últimos cinco años, las metas establecidas por la Junta directiva, la siniestralidad percibida, y la situación económica y política del territorio departamental y nacional.

Para el cálculo de los ingresos, la Dirección de Aseguramiento proyecta el crecimiento de los afiliados para el siguiente año, teniendo en cuenta los datos de años anteriores, la proyección de habitantes en el departamento suministrado por el DANE, y la evaluación de la situación actual y esperada del aseguramiento en el territorio presentado por la Dirección Seccional de Salud de Antioquia - DSSA. Los conceptos más importantes que son considerados por la Dirección de Aseguramiento y la Subgerencia financiera son:

Tabla 4. Conceptos del ingreso financiero

Tipo ingreso	Concepto
INGRESOS PBS	<ul style="list-style-type: none"> → UPC Régimen subsidiado → UPC Régimen contributivo → Copagos y cuotas moderadoras → Incapacidades → Licencias de maternidad y paternidad → Restituciones → Otros
INGRESOS NO PBS	<ul style="list-style-type: none"> → Recobros por servicios, insumos y tecnologías NO PBS
OTROS INGRESOS	<ul style="list-style-type: none"> → No operacionales → Recuperación PGP → Recobros a la cápita

Fuente: Elaboración propia

Para el cálculo de los egresos, se contemplan los gastos de funcionamiento o no operacionales clasificados en: Gastos de personal directos y diversos, Gastos generales de bienes operativos y servicios, Gastos de provisiones y depreciaciones, y Gastos legales y otros financieros. Aproximadamente, los gastos de funcionamiento equivalen al 8% del presupuesto global.

En cuanto a los gastos correspondientes al aseguramiento en salud, es decir, los gastos operacionales se consideran los siguientes conceptos:

Tabla 5. Conceptos del egreso financiero operacional

Tipo egreso	Concepto
GASTO EN SALUD	Costo total de atención, discriminado por régimen subsidiado y régimen contributivo, de acuerdo a los siguientes ítems: → Capitación → PGP → Evento de primer, segundo y tercer nivel → Medicamentos ambulatorios → Alto Costo → Otros
INCAPACIDADES Y RECAUDO CONTRIBUTIVO	→ Incapacidades → Licencias de maternidad y paternidad → Servicio de Recaudo Contributivo

Fuente: Elaboración propia

Según la referencia histórica de la entidad, el 20% del gasto operacional se dirige a los contratos por capitación, y el 72% restante del gasto global se presupuesta para las atenciones pagadas por Evento.

Para mitigar el efecto de lo pagado por Evento, la EPS ha creado Rutas Integrales de Atención en Salud - RIAS dentro de la modalidad de contratación y pago PGP - Pago Global Prospectivo. Mediante las RIAS se asegura la prestación de servicios generales y especializados a los diversos grupos poblacionales según sus factores de riesgo, patología y especialidad médica.

Sin embargo, según los estados financieros de la entidad, el costo por evento de algunas atenciones supera el ingreso percibido por cada afiliado, que se determina por la UPC que la ADRES gira mensualmente a la EPS. Lo anterior conlleva, naturalmente, a un déficit financiero importante que pone en riesgo la sostenibilidad de la entidad, la reserva técnica y el plan de inversiones en salud (Savia Salud, n.d.-b).

Actividad 4.

Debido al déficit financiero de la entidad, producto de su rentabilidad negativa, desde el año 2017, la Superintendencia Nacional de Salud, como ente máximo de control del sistema de salud, estableció una medida preventiva de vigilancia especial en donde se hacía necesario la recuperación del patrimonio de la EPS y su solvencia económica. Tras varios años de prórrogas, auditorías y planes de mejoramiento, para el año 2022, Savia Salud se ve ante la posibilidad de una liquidación inminente.

Desde el 28 de septiembre del año actual, la empresa tiene cuatro (4) meses para desarrollar un Plan de Choque con seguimiento mensual, que busca intervenir sus principales deficiencias. Uno de los puntos críticos a resolver por la EPS, es el cumplimiento de las condiciones financieras y de solvencia que requiere para seguir operando en el departamento. Actualmente, la entidad no cumple con el indicador del capital mínimo y patrimonial adecuado, debido a que el costo en salud de la entidad figura en el 116%, influyendo directamente en su deterioro económico (J. I. González, 2022)

Teniendo en cuenta que más del 70% del gasto operacional que asume la EPS, está compuesto por el pago de los servicios contratados por evento, y este, a su vez, influye directamente en el sobrecosto en salud de la entidad, se justifica el continuo análisis frente al impacto del sistema de pago por evento para Savia Salud, y la búsqueda de herramientas que generen información de valor para la toma de decisiones, y la implementación de acciones correctivas frente a la situación expuesta. (J. I. González, 2022).

8.2 Objetivo 2: Aplicar técnicas de Machine Learning para Identificar las variables relacionadas a la predicción del costo por evento de los afiliados de la EPS.

Fase 2. Estudio y comprensión de los datos

Actividad 1.

Savia salud cuenta con un plan estratégico de tecnologías de información y comunicaciones (PETI) para el periodo 2021 - 2024. El PETI de Savia salud está alineado al logro de los objetivos estratégicos de la entidad, los cuales a su vez están consignados en el plan estratégico institucional. El PETI maneja la información de la institución regida por diversas normas y decretos del sector de la salud, y del sector de las TICs.

La EPS administra la información obtenida desde el proceso de afiliación y aseguramiento, los procesos de acceso a los servicios de salud (autorizaciones, referencia y contrarreferencia, urgencias, entrega de medicamentos), el proceso de gestión de riesgo, y los procesos administrativos operacionales y no operacionales de facturación, contratación y contabilidad; a través de cuatro (4) sistemas específicos:

Tabla 6. Sistemas de información Savia Salud EPS

Sistemas	Funcionalidad principal
Conexiones Savia	Sistema misional, propio e integrador de procesos de la EPS
Intrasavia	Sistema propio para la gestión de proyectos, asignación de tareas y radicación de solicitudes de desarrollo
SAP	Sistema comercial para la gestión empresariales de la EPS
Boxsalud	Sistema comercial que administra los procesos del régimen contributivo

Fuente: Elaboración propia

La EPS posee una estructura organizada y centralizada para la administración de los datos que se registran en los aplicativos anteriormente mencionados. Por medio

de un Acuerdo de Cooperación de investigación científica se logró acceder a la información necesaria y relevante para la formulación del modelo predictivo.

Los atributos para este proyecto fueron seleccionados después del desarrollo de la fase 1, y la asesoría externa de un experto del negocio. Para obtener la información, fue necesario realizar consultas en SQL sobre los datos gestionados por las siguientes áreas:

- Área de aseguramiento: caracterización de afiliados, ingreso por UPC y afiliados 2021.
- Área financiera: Costo facturación por cada evento de los afiliados 2021.
- Área gestión de riesgo: Afiliados de alto costo 2021.
- Área acceso: Relación de servicios y diagnóstico por afiliados 2021.

Actividad 2.

Por políticas de seguridad informática de la institución, no fue posible manipular directamente las bases de datos, ni conocer la estructura de estas; por ello, se realizó una solicitud interna al equipo de “Gestión de la Información y Analítica de Datos” para obtener un archivo con los atributos seleccionados para la investigación.

En respuesta a la solicitud, se recibió un archivo con extensión txt con un total de 576.378 registros, asociados a los usuarios que estuvieron activos durante el año 2021, y estuvieron en algún momento bajo el modelo de liquidación por Evento.

Actividad 3.

Tras evaluar los datos obtenidos con los analistas del área de Analítica y TI de Savia Salud, se decide seleccionar una muestra con los registros que corresponden a afiliados del municipio de Medellín, cuya permanencia en la EPS fue de todo el periodo 2021, por lo cual la UPC recibida corresponde al año completo.

Se decide ingresar al estudio solo a los afiliados de Medellín, porque este municipio equivale a más del 40% de la población de la EPS, y es el territorio que presenta un mayor gasto de los recursos del sistema de salud al ser la capital del departamento. Además, debido a que la concentración y peso de los datos se encontraban en los usuarios pertenecientes a Medellín, los datos correspondientes a los demás municipios se podrían ver afectados significativamente.

El archivo definitivo que hizo parte de la investigación tuvo un total de 21.115 registros que cumplían con las características preliminares para el modelo predictivo a desarrollar.

Actividad 4.

A manera de resumen los datos obtenidos para la investigación son:

Tabla 7. Resumen Diccionario de datos

Etiqueta	Naturaleza	Escala	Fuente
Serial_BDUA	Cuantitativa	De razón	Módulo Aseguramiento
Contrato_afil	Cuantitativa	De razón	Módulo Aseguramiento
ID_afiliado	Cuantitativa	De razón	Módulo Aseguramiento
tipo_documento	Cualitativa	Nominal	Módulo Aseguramiento
numero_documento	Cuantitativa	De razón	Módulo Aseguramiento
fecha_nacimiento	Cualitativa	Nominal	Módulo Aseguramiento
Sexo	Cualitativa	Nominal	Módulo Aseguramiento
Edad	Cuantitativa	De razón	Módulo Aseguramiento
fecha_afiliacion_ep	Cualitativa	Nominal	Módulo Aseguramiento
fecha_egreso_eps	Cualitativa	Nominal	Módulo Aseguramiento
municipio_divipola_afiliacion	Cualitativa	Nominal	Módulo Aseguramiento
zona	Cualitativa	Nominal	Módulo Aseguramiento
codigo_grupo_poblacional	Cualitativa	Nominal	Módulo Aseguramiento
grupo_Sisben	Cualitativa	Nominal	Módulo Aseguramiento
estado_afiliacion	Cualitativa	Nominal	Módulo Aseguramiento
modelo_liquidacion	Cualitativa	Nominal	Módulo Aseguramiento
UPC_netto	cuantitativa	De razón	Módulo Aseguramiento
dias_UPC_netto	cuantitativa	De razón	Módulo Aseguramiento
costo_afiliado	cuantitativa	De razón	Módulo Cuentas médicas
Pertenece_VIH	Cualitativa	Nominal	Módulo Gestión del riesgo
Pertenece_Cancer	Cualitativa	Nominal	Módulo Gestión del riesgo
Pertenece_Hemofilia	Cualitativa	Nominal	Módulo Gestión del riesgo
Pertenece_Reuma	Cualitativa	Nominal	Módulo Gestión del riesgo
Pertenece_Prog_Hepatitis	Cualitativa	Nominal	Módulo Gestión del riesgo
Pertenece_Prog_ERC	Cualitativa	Nominal	Módulo Gestión del riesgo
Pertenece_Prog_Huerfanos	Cualitativa	Nominal	Módulo Gestión del riesgo
otros_autoinmunes	Cualitativa	Nominal	Módulo Gestión del riesgo
otros_cardio	Cualitativa	Nominal	Módulo Gestión del riesgo
otros_pulmonar	Cualitativa	Nominal	Módulo Gestión del riesgo
otros_diabetes	Cualitativa	Nominal	Módulo Gestión del riesgo
otros_tuberculosis	Cualitativa	Nominal	Módulo Gestión del riesgo
portabilidad	Cualitativa	Nominal	Módulo Aseguramiento
servicios	Cualitativa	Nominal	Módulo Acceso
diagnósticos	Cualitativa	Nominal	Módulo Acceso

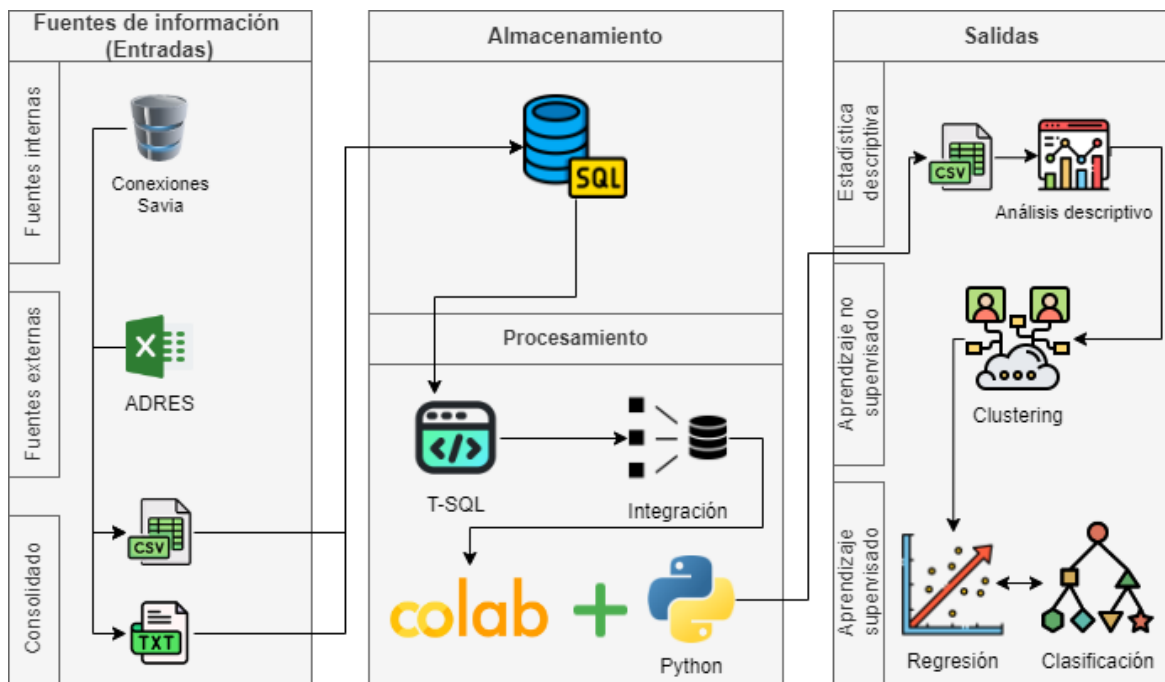
Fuente: Elaboración propia

La principal fuente de información de donde se hizo la extracción de datos fue Conexiones (aplicativo misional de la EPS). El diccionario de datos completo se visualiza en el Anexo 1 de este documento

Fase 3. Análisis de los datos y selección de características

Para realizar un adecuado análisis de los datos, se aplicaron un conjunto de pasos iterativos, es decir, a medida que evolucionó el análisis descriptivo, fue necesario regresar a las revisiones iniciales y, en algunos casos, repetir el ciclo. Las tecnologías utilizadas y el esquema general del proceso se resumen en el siguiente esquema :

Figura 3. Arquitectura general de la solución



Fuente: Elaboración propia

Actividad 1.

A) Integración de datos

Una vez se obtuvieron los datos, se encontró que los registros eran redundantes para cada uno de los afiliados, es decir, se contaba con 1 fila para cada una de las atenciones recibidas en algún centro de salud de la ciudad.

Por lo anterior, fue necesario transformar los datos, con el fin de no introducir un sesgo al modelo de Machine Learning, ya que no todos los afiliados tenían la misma cantidad de registros.

Para la manipulación y transformación de los datos, fue necesario utilizar el motor de bases de datos SQL Server 2019 Developer y el gestor SQL Server Management Studio. El objetivo, fue obtener un registro para cada uno de los afiliados, seleccionado la moda de cada atributo. Para ello, fue necesario agrupar los datos por afiliado, contar y ordenar los registros de manera descendente y seleccionar el registro con mayor número de apariciones. Con ello, se logró obtener un dataframe consolidado de 13.692 registros para la construcción del modelo.

Actividad 2.

B) Eliminación de variables irrelevantes o redundantes, y limpieza de datos

Después de realizar la integración, se obtuvo un archivo con extensión .xlsx, el cual se cargó en Google – Colab, para continuar realizando la preparación y análisis.

Se realizó un listado de las variables definitivas y se optó por eliminar las variables de identificación: Serial, Contrato_afil, ID_afil, tipo_documento ya que no aportan ningún elemento a la descripción estadística y al modelo de machine learning.

Posteriormente, se suprimieron las variables Zona, municipio_divipola_afiliacion y modelo_liquidacion, ya que al ser un estudio sólo de la ciudad de Medellín todos corresponden a zona Urbana y todos los afiliados solicitados recibieron atenciones por “Evento”.

Finalmente se eliminó la variable UPC_neto, ya que se identificó como redundante con dias_UPC_neto que suministraba la misma información.

Actividad 3.

C) Descripción estadística

Para la exploración inicial, se utilizaron las librerías matplotlib para realizar gráficos de barras de las variables cualitativas y seaborn para realizar gráficos de dispersión de las variables cuantitativas, posteriormente, se modificó el estilo de los gráficos en Excel.

Durante la realización del análisis descriptivo, se revisó el comportamiento de las 12 variables relacionadas con la pertenencia del afiliado a los programas de alto costo, encontrando lo siguiente:

Tabla 8. Descripción de la variable: Programas de alto costo

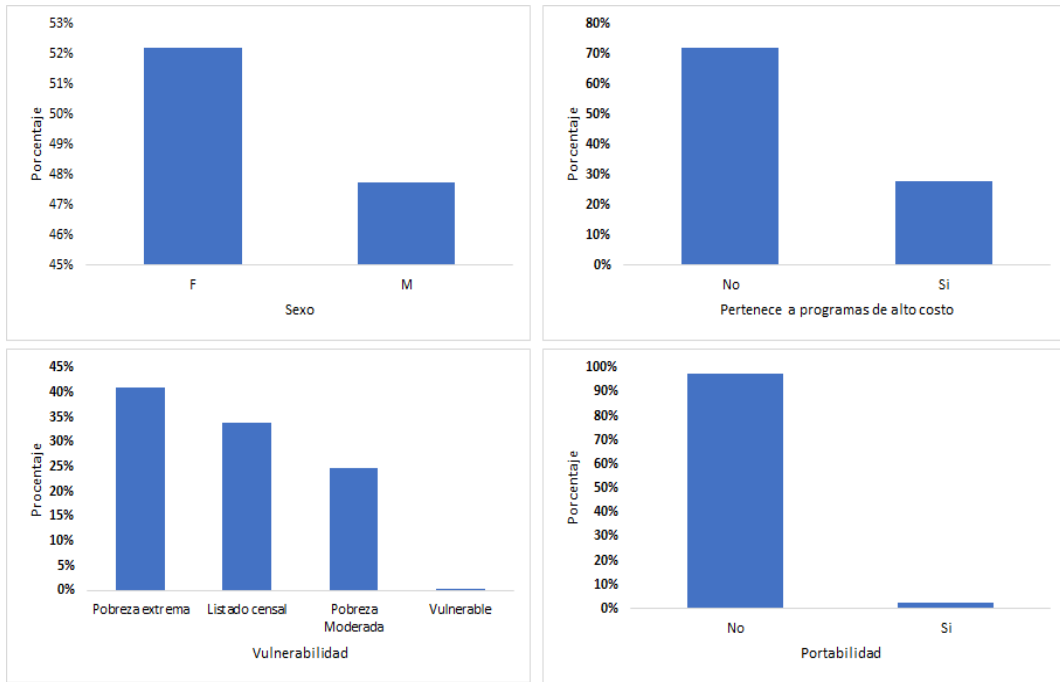
Programa	Si	No
VIH	1.8%	98.2%
Cáncer	2.4%	97.6%
Hemofilia	0.1%	99.9%
Reuma0	0.6%	99.4%
Hepatitis	0.0%	100.0%
ERC	0.4%	99.6%
Huérfanas	0.2%	99.8%
Autoinmunes	1.5%	98.5%
Cardiovascular	10.5%	89.5%
Pulmonar	2.0%	98.0%
Diabetes	0.1%	99.9%
Tuberculosis	0.5%	99.5%

Fuente: Elaboración propia

Debido a que los valores de pertenencia a un programa de alto costo fueron tan bajos y, considerando que algunos afiliados estaban matriculados en más de un programa, se integraron todas las variables en una que indica si pertenece algún programa de alto costo, de esta manera, se facilita el procesamiento de las variables para la construcción del modelo de machine learning y se facilita su entendimiento.

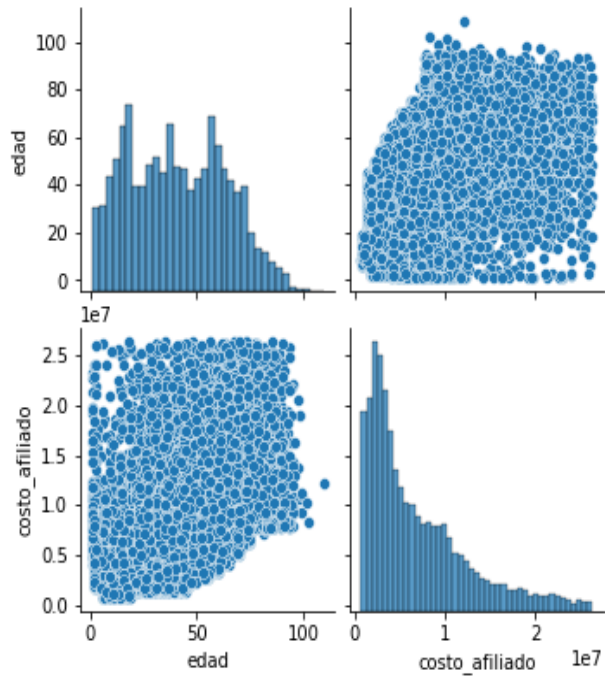
De este proceso se obtuvo lo siguiente:

Figura 4. Descripción de variables cualitativas



Fuente: Elaboración propia

Figura 5. Composición de variables cuantitativas



Fuente: Elaboración propia

Tabla 9. Resultados de las medidas de tendencia central

Medida/ Variable	edad	costo_afiliado
N	13692	13692
Missing	0	0
Mean	40.7	7.02e+6
Median	40.0	5.09e+6
Standard deviation	22.6	5.69e+6
Minimum	1	678652
Maximum	109	26409859

Fuente: Elaboración propia

Los afiliados de la ciudad de Medellín que fueron atendidos bajo el modelo de liquidación por evento y que permanecieron al menos 1 año completo en la EPS, fueron en su mayoría mujeres, con un promedio de edad de 41 años (\pm 23 años), cerca del 30% pertenece a programas de alto costo, aproximadamente el 70% es población altamente vulnerable al pertenecer al grupo de pobreza extrema o listado censal; solo 3% de los afiliados tiene el beneficio de portabilidad que les permite ser atendidos en otra institución que no es la del municipio.

En lo que corresponde al costo de las atenciones por evento, se encontró que el 50% de los afiliados tuvo un costo cercano a los 5 millones de pesos, mientras que el mayor costo fue de \$26,409,859.

En lo que corresponde a los servicios prestados a los pacientes y los diagnósticos asociados a la atención, en la base de datos original de más de 2 millones de registros, se encontraron 41.287 servicios y 34.167 diagnósticos diferentes.

Debido a la gran variedad de registros, se optó por reemplazar los servicios por la categoría a la que pertenecen según los Códigos Únicos de Procedimientos en Salud - CUPS (adoptados por el Ministerio de Protección Social) y, los diagnósticos por el grupo al que corresponden según la Clasificación Internacional de Enfermedades -CIE11. De esta forma, se obtuvo una variable nueva con 53 categorías de procedimientos y 23 grupos de diagnósticos, el top 10 de cada grupo se presenta a continuación:

Tabla 10. Categorización de la variable: Servicios

Servicios	Porcentaje
Medicamentos	31.21%
Laboratorio clínico	28.53%
Insumo y equipo médico	18.25%
Consulta especializada	8.08%
Consulta medicina general	5.03%
Programa	2.64%
Rayos x	2.17%
Consulta paramédica	1.75%
Odontología	1.48%
Ecocardiografía	0.87%
Total	100%

Fuente: Elaboración propia

Tabla 11. Categorización de la variable: Diagnósticos

Diagnósticos	Porcentaje
Factores que influyen en el estado de salud y contacto con los servicios de salud (Z00-Z99)	20.66%
Trastornos mentales y del comportamiento (F00-F99)	17.02%
Síntomas, signos y hallazgos anormales clínicos y de laboratorio, no clasificados en otra parte (R00-R99)	14.65%
Traumatismos, envenenamiento y algunas otras consecuencias de causas externas (S00-T98)	11.79%
Enfermedades del sistema circulatorio (I00-I99)	7.13%
Enfermedades del sistema digestivo (K00-K93)	6.36%
Códigos para propósitos especiales (U00-U99)	6.03%
Enfermedades del sistema genitourinario (N00-N99)	5.96%
Enfermedades del sistema respiratorio (J00-J99)	5.81%
Enfermedades del sistema osteomuscular y del tejido conjuntivo (M00-M99)	4.60%
Total general	100.00%

Fuente: Elaboración propia

Finalmente, las medidas de resumen para cada una de las variables permiten identificar una gran desviación de los datos respecto a su media, lo que sugiere la necesidad de normalizar los datos para la elaboración del modelo.

D) Limpieza de datos

Para ejecutar la limpieza de datos, se verificó la integridad de los datos verificando que ninguna variable tuviera datos nulos, alteraciones en las categorías de las variables o datos perdidos. De la misma manera, se identificaron los valores atípicos leves de las variables cuantitativas y se suprimieron.

Después de realizar el proceso descrito anteriormente el set de datos quedó con un total de 21.115 registros.

8.3 Objetivo 3: Construir un modelo de aprendizaje automático para predecir el costo por evento de los afiliados de la EPS

Fase 4. Construcción de los modelos

La construcción del modelo de Machine Learning, se dividió en dos actividades principales:

La primera, consiste en aplicar un algoritmo de clusterización que nos permita conocer con mayor profundidad como se agrupan los datos de los afiliados de la EPS. Con este proceso, se obtienen la cantidad de grupos o segmentos que se pueden identificar de la base datos, además, se obtiene un representante para cada grupo que nos permite conocer las características principales del mismo.

Basados en los resultados de la clusterización, se transformó manualmente la variable Costo; con ello, se pasó de un valor numérico a una categoría, con el fin de aplicar técnicas de aprendizaje supervisado.

Con la transformación descrita anteriormente, se realizó la segunda actividad, la cual consistió construcción de diferentes modelos de clasificación, que permitan predecir el costo de la atención por evento de un afiliado de la EPS, en un rango de valores previamente definido.

Los resultados se describen a continuación:

Actividad 1. Clusterización de los datos - Modelo de aprendizaje no supervisado

Para la primera actividad se realizaron los siguientes pasos:

I) Obtención de variables dummy: Después de ejecutar la limpieza de datos descrita en el ítem anterior, se convirtieron las variables cualitativas a variables “dummy”, con el fin de representar las categorías en valores sobre un rango continuo (variables cuantitativas).

Como resultado, cada una de las categorías de las variables cualitativas del siguiente dataframe, se convierten en una variable independiente aumentando la dimensionalidad de este.

Tabla 12. Resultados de las variables dummy

Variable	Tipo	Escala
edad	Cuantitativa	de razón
sexo_M	Cuantitativa	de razón
Vulnerabilidad_Pobreza Moderada	Cuantitativa	de razón
Vulnerabilidad_Pobreza extrema	Cuantitativa	de razón
Vulnerabilidad_Vulnerable	Cuantitativa	de razón
ProgramaAltoCosto_Si	Cuantitativa	de razón
Portabilidad_Si	Cuantitativa	de razón
CategoriaDX_Ciertas afecciones originadas en el periodo perinatal (P00-P96)	Cuantitativa	de razón
CategoriaDX_Ciertas enfermedades infecciosas y parasitarias (A00-B99)	Cuantitativa	de razón
CategoriaDX_Códigos para propósitos especiales (U00-U99)	Cuantitativa	de razón
CategoriaDX_Embarazo, parto y puerperio (O00-O99)	Cuantitativa	de razón
CategoriaDX_Enfermedades de la piel y del tejido subcutáneo (L00-L99)	Cuantitativa	de razón
CategoriaDX_Enfermedades de la piel y el tejido subcutáneo (L00-L99)	Cuantitativa	de razón
CategoriaDX_Enfermedades de la sangre y de los organos hematopoyeticos, y ciertos trastornos que afectan el mecanismo de la inmunidad (D50-D89)	Cuantitativa	de razón
CategoriaDX_Enfermedades del oído y de la apofisis mastoides (H60-H95)	Cuantitativa	de razón
CategoriaDX_Enfermedades del ojo y sus anexos (H00-H59)	Cuantitativa	de razón
CategoriaDX_Enfermedades del sistema circulatorio (I00-I99)	Cuantitativa	de razón
CategoriaDX_Enfermedades del sistema digestivo (K00-K93)	Cuantitativa	de razón
CategoriaDX_Enfermedades del sistema genitourinario (N00-N99)	Cuantitativa	de razón
CategoriaDX_Enfermedades del sistema nervioso (G00-G99)	Cuantitativa	de razón
CategoriaDX_Enfermedades del sistema osteomuscular y del tejido conjuntivo (M00-M99)	Cuantitativa	de razón
CategoriaDX_Enfermedades del sistema respiratorio (J00-J99)	Cuantitativa	de razón

Variable	Tipo	Escala
CategoriaDX_Enfermedades endocrinas, nutricionales y metabólicas (E00-E90)	Cuantitativa	de razón
CategoriaDX_Factores que influyen en el estado de salud y contacto con los servicios de salud (Z00-Z99)	Cuantitativa	de razón
CategoriaDX_Malformaciones congénitas, deformidades y anomalías cromosómicas (Q00-Q99)	Cuantitativa	de razón
CategoriaDX_Síntomas, signos y hallazgos anormales clínicos y de laboratorio, no clasificados en otra parte (R00-R99)	Cuantitativa	de razón
CategoriaDX_Trastornos mentales y del comportamiento (F00-F99)	Cuantitativa	de razón
CategoriaDX_Traumatismos, envenenamiento y algunas otras consecuencias de causas externas (S00-T98)	Cuantitativa	de razón
CategoriaDX_Tumores [neoplasias] (C00-D48)	Cuantitativa	de razón
CategoriaServicio_BIOPSIA	Cuantitativa	de razón
CategoriaServicio_CIRUGIA DE LA MANO	Cuantitativa	de razón
CategoriaServicio_CIRUGIA DERMATOLÓGICA	Cuantitativa	de razón
CategoriaServicio_CIRUGIA GENERAL	Cuantitativa	de razón
CategoriaServicio_CIRUGIA GINECOLÓGICA	Cuantitativa	de razón
CategoriaServicio_CIRUGIA ORTOPEDIA	Cuantitativa	de razón
CategoriaServicio_CIRUGIA OTORRINOLARINGOLOGÍA	Cuantitativa	de razón
CategoriaServicio_CIRUGIA PLÁSTICA	Cuantitativa	de razón
CategoriaServicio_CIRUGIA UROLOGÍA	Cuantitativa	de razón
CategoriaServicio_CIRUGIA VASCULAR	Cuantitativa	de razón
CategoriaServicio_CONSULTA ESPECIALIZADA	Cuantitativa	de razón
CategoriaServicio_CONSULTA MEDICINA GENERAL	Cuantitativa	de razón
CategoriaServicio_CONSULTA PARAMÉDICA	Cuantitativa	de razón
CategoriaServicio_DOPPLER / DUPLEX	Cuantitativa	de razón
CategoriaServicio_ECOCARDIOGRAFÍA	Cuantitativa	de razón
CategoriaServicio_ECOGRAFÍA	Cuantitativa	de razón
CategoriaServicio_ELECTROFISIOLOGÍA	Cuantitativa	de razón
CategoriaServicio_ENDOSCOPIA	Cuantitativa	de razón
CategoriaServicio_INSUMO Y EQUIPO MÉDICO	Cuantitativa	de razón
CategoriaServicio_INTERNACION ALTA COMPLEJIDAD	Cuantitativa	de razón
CategoriaServicio_JUNTA ESPECIALIZADA	Cuantitativa	de razón
CategoriaServicio_LABORATORIO CLÍNICO	Cuantitativa	de razón
CategoriaServicio_MEDICAMENTOS	Cuantitativa	de razón
CategoriaServicio_MEDICINA DOMICILIARIA	Cuantitativa	de razón
CategoriaServicio_MEDICINA NUCLEAR	Cuantitativa	de razón
CategoriaServicio_NEUROCIRUGÍA	Cuantitativa	de razón
CategoriaServicio_ODONTOLOGÍA	Cuantitativa	de razón
CategoriaServicio_OTRA TERAPIA DE REHABILITACION	Cuantitativa	de razón
CategoriaServicio_PROCEDIMIENTO ORTOPEDIA	Cuantitativa	de razón

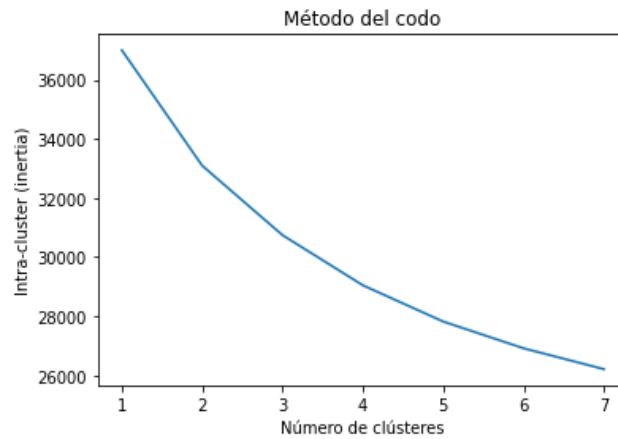
Variable	Tipo	Escala
CategoriaServicio_PROCEDIMIENTOS DX AUDIOLOGIA	Cuantitativa	de razón
CategoriaServicio_PROCEDIMIENTOS DX CARDIOLOGIA	Cuantitativa	de razón
CategoriaServicio_PROCEDIMIENTOS DX FISIATRIA	Cuantitativa	de razón
CategoriaServicio_PROCEDIMIENTOS DX NEUMOLOGIA	Cuantitativa	de razón
CategoriaServicio_PROCEDIMIENTOS DX NEUROLOGIA	Cuantitativa	de razón
CategoriaServicio_PROCEDIMIENTOS DX PSIQUE	Cuantitativa	de razón
CategoriaServicio_PROCEDIMIENTOS DX UROLOGIA	Cuantitativa	de razón
CategoriaServicio_PROCEDIMIENTOS ENFERMERIA	Cuantitativa	de razón
CategoriaServicio_PROCEDIMIENTOS TTO AUDIOLOGIA	Cuantitativa	de razón
CategoriaServicio_PROCEDIMIENTOS TTO CARDIOLOGIA	Cuantitativa	de razón
CategoriaServicio_PROCEDIMIENTOS TTO MEDICINA FISICA Y REHABILITACION	Cuantitativa	de razón
CategoriaServicio_PROCEDIMIENTOS TTO NEUROLOGIA	Cuantitativa	de razón
CategoriaServicio_PROCEDIMIENTOS TTO PSIQUE	Cuantitativa	de razón
CategoriaServicio_PROGRAMA	Cuantitativa	de razón
CategoriaServicio_QUIMIOTERAPIA	Cuantitativa	de razón
CategoriaServicio_RADIOLOGIA INTERVENCIONISTA	Cuantitativa	de razón
CategoriaServicio_RADIOTERAPIA	Cuantitativa	de razón
CategoriaServicio_RAYOS X	Cuantitativa	de razón
CategoriaServicio_RESONANCIA	Cuantitativa	de razón
CategoriaServicio_SERVICIOS DE OXIGENO	Cuantitativa	de razón
CategoriaServicio_TOMOGRAFIA	Cuantitativa	de razón
CategoriaServicio_TRANSPORTE ADMINISTRATIVO	Cuantitativa	de razón
Costo	Cuantitativa	de razón

Fuente: Elaboración propia

II) Normalización de variables cuantitativas: Las variables cuantitativas, se normalizaron con el propósito de convertirlas a una escala matemáticamente comparable; con este procedimiento, todos los atributos toman un valor entre cero y uno, de acuerdo con la siguiente fórmula:

$$z = (x - \text{mínimo } (x)) / (\text{máximo } (x) - \text{mínimo } (x))$$

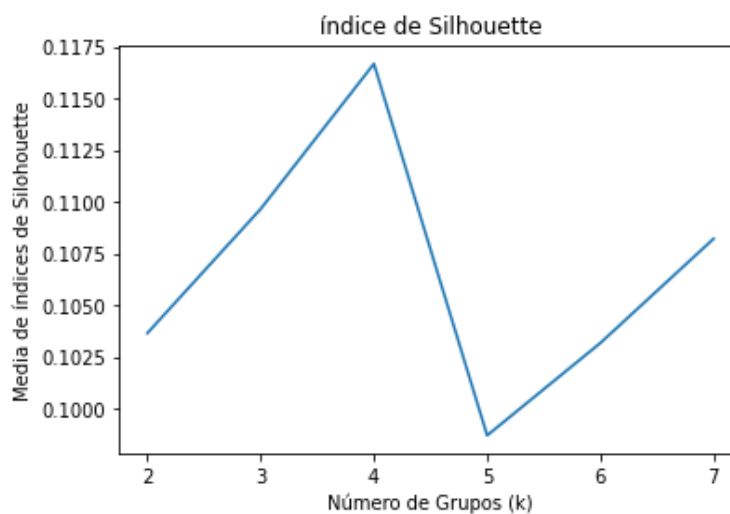
III) Identificar número de clústeres: Para identificar la cantidad de clústeres o grupos que se querían obtener de los datos, se aplicó el método Elbow, este método utiliza los valores de la inercia obtenidos tras aplicar K-means con diferentes números de Clústeres (desde 1 a N Clusters). Siendo la inercia la suma de las distancias al cuadrado de cada objeto del clúster a su centroide, el resultado fue el siguiente:

Figura 6. Resultados del Método del codo

Fuente: Elaboración propia

La gráfica anterior, permite apreciar un cambio brusco en la evolución de la inercia, donde la línea tiene una forma similar a la de un brazo, y su codo permite identificar la cantidad adecuada de clústeres. Debido a que el resultado no fue claro a la vista, se decidió aplicar el método silhouette, como una herramienta complementaria para obtener un valor más preciso.

Este último, nos permite visualizar un índice que toma valores entre -1 y 1; cuando esté más cercano al número negativo quiere decir que el número de clúster no es adecuado. De la aplicación de este método se obtuvo lo siguiente:

Figura 7. Resultados Índice de Silhouette

Fuente: Elaboración propia

De acuerdo con la gráfica anterior, procedemos a segmentar los datos en 4 clústeres, ya que es el pico que más se aleja de -1.

IV) Construcción de los clúster

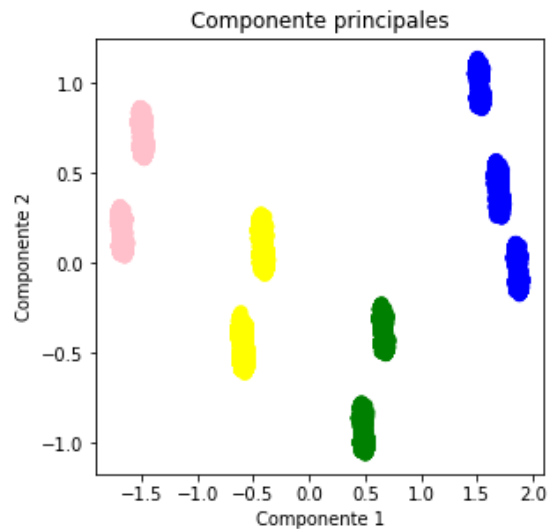
Aplicamos el algoritmo Kmeans con 4 clusters y obtenemos los siguientes centroides o representantes:

Tabla 13. Descripción de los clústeres

Clúster	0	1	2	3
Nombre clúster	Mujeres jóvenes en condición de pobreza extrema ligeramente costosas para la EPS	Hombres adultos jóvenes costosos para la EPS cuyo mayor servicio es la provisión de medicamentos	Adultos jóvenes con más de 2 veces el costo de la UPC por servicios relacionados con insumos y equipos médicos	Mujeres adultas jóvenes con más de 2 veces el costo de la UPC por servicios relacionados con el laboratorio clínico
Cantidad de registros	3005	4015	2348	4324
Edad	18	33	38	31
Sexo	F	M	M	F
Vulnerabilidad	Pobreza extrema	Listado censal	Listado censal	Listado censal
ProgramaAltoCosto	No	No	No	No
Portabilidad	No	No	No	No
CategoríaDX	Causas externas de morbilidad y de mortalidad (V01-Y98)	Causas externas de morbilidad y de mortalidad (V01-Y98)	Causas externas de morbilidad y de mortalidad (V01-Y98)	Causas externas de morbilidad y de mortalidad (V01-Y98)
CategoríaServicio	Laboratorio clínico	Medicamentos	Insumo y equipo médico	Laboratorio clínico
costo_afiliado	\$ 1,834,987.00	\$ 6,228,292.00	\$ 3,076,974.00	\$ 3,392,113.00

Fuente: Elaboración propia

Al representar los clústeres en dos componentes principales, se evidenció que la segmentación de estos es bastante amplia, es decir, existe suficiente separabilidad para identificar los grupos y cohesión entre ellos.

Figura 8. Resultados de Componentes principales

Fuente: Elaboración propia

Del ejercicio de clusterización, se puede destacar lo siguiente:

- El grupo con mayor costo es el *clúster 1*, este corresponde a hombres que poseen 33 años en promedio, a los cuales se le prestan servicios asociados a la entrega de medicamentos. En este grupo, se encuentra la población vulnerable del listado censal como desplazados por la violencia, habitantes en situación de calle, migrantes, privados de la libertad, entre otros.
- Las mujeres se ubican mayormente en los *clústeres 0 y 2*. El *clúster 0* está compuesto por mujeres jóvenes de 18 años en promedio, correspondientes al grupo de pobreza extrema, cuyo costo de atención supera ligeramente la unidad de pago por capitación y consumen con mayor frecuencia servicios de laboratorio.
- El *clúster 3* por su parte representa a las mujeres mayores de 30 años que también utilizan con mayor frecuencia los servicios de laboratorio, pero que le cuestan a la EPS más de 2 veces la unidad de pago por capacitación.
- El *clúster 2* representa a los hombres cercanos a los 40 años que cuestan más de dos veces la unidad de pago por capitación, que consumen servicios relacionados con insumos y equipos médicos.

Actividad 2.

A) Construcción de un modelo de aprendizaje supervisado

Inicialmente, se propuso construir un modelo de regresión que permitiera predecir el costo de la atención por evento de los afiliados en valores numéricos. Para este ejercicio se utilizó el dataframe preparado en el ítem anterior, excluyendo la variable donde se etiquetaron los clusters.

El objetivo del modelo de regresión, fue identificar la capacidad explicativa de las variables independientes, sobre la variable dependiente (Costo del afiliado), para ello se plantearon las siguientes hipótesis:

H0: Las variables independientes no explican la variable dependiente

Ha: Las variables independientes explican la variable dependiente.

Posteriormente, se dividió el conjunto de datos en dos grupos:

X: Conjunto de entrenamiento (80% de los registros)

Y: Conjunto de pruebas (20% de los registros)

Se construyó un modelo de regresión lineal y se obtuvo lo siguiente:

Tabla 14. Resultados del modelo de Regresión lineal

OLS Regression Results			
Dep. Variable	y	R-squared	0.281
Model	OLS	Adj. R-squared	0.276
Method	Least Squares	F-statistic	53.74
Date	Mon, 14 Nov 2022	Prob (F-statistic)	0.00
Time	20:18:03	Log-Likelihood	2791.2
No. Observations	10953	AIC	-5422
Df Residuals	10873	BIC	-4838
Df Model	79		
Covariance Type	nonrobust		

Fuente: Elaboración propia

Al interpretar los resultados, se identifica que el valor de *R-Cuadrado* es bajo. Este estadístico oscila entre 0 y 1, y entre más cercano se encuentre a 1 indica que el modelo explica toda la variabilidad de los datos de respuesta en torno a su media.

No obstante, al revisar con mayor detalle el valor de F^2 , encontramos que valor p es estadísticamente significativo ($p < 0.05$), por lo cual se rechaza la hipótesis nula y se concluye que las variables independientes explican la variable dependiente. El valor bajo del R^2 puede estar relacionado con variables que generan ruido al modelo o que el mismo no tiene un comportamiento lineal.

De acuerdo con lo anterior, se opta por discretizar la variable objetivo (costo del afiliado) y construir categorías tomando como base la información obtenida en el proceso de clusterización.

Debido a que en los clústeres las categorías de costo se solapan, se toman la misma cantidad de grupos (4), construyendo los rangos manualmente de la siguiente manera:

Tabla 15. Creación de grupos y rangos según costo médico

Grupo	Rango
Grupo 1	\$2.000.000 o menos
Grupo 2	Desde \$2.000.001 hasta \$4.000.000
Grupo 3	desde \$4.000.001 hasta \$6.000.000
Grupo 4	Más de \$6.000.000

Fuente: Elaboración propia

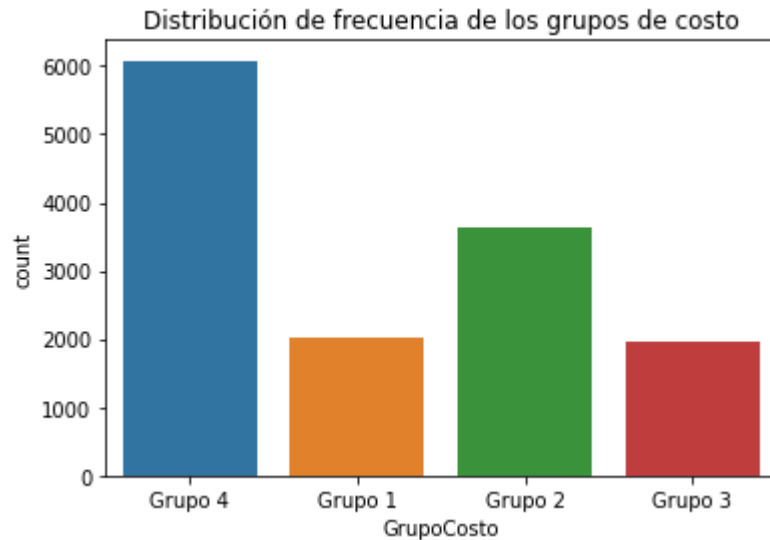
Con este proceso, se procede a construir un modelo de clasificación que permita predecir la categoría de costos de un afiliado, para ello se realizan los siguientes pasos:

I) Balanceo de los datos

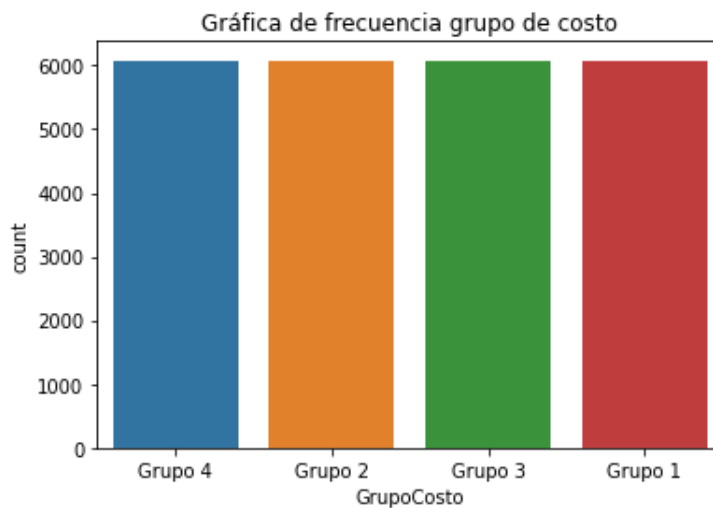
Debido a que cada categoría tiene una distribución de frecuencias diferente, se opta por utilizar la librería de balanceo resample, con el fin de generar registros sintéticos e igualar la cantidad de registros de cada grupo.

La distribución de los datos antes del balanceo era la siguiente:

² El estadístico F es una prueba que se utiliza para evaluar la capacidad explicativa que tiene un grupo de variables independientes sobre la variación de la variable dependiente.

Figura 9. Distribución de los grupos del costo médico

Para el balanceo, se igualaron todos los registros al grupo que contenía la mayor cantidad de datos, en este caso el grupo 4, el resultado se presenta a continuación:

Figura 10. Balanceo de los grupos del costo médico**II) Se dividieron los datos en dos grupos:**

X: Conjunto de entrenamiento (80% de los registros)

Y: Conjunto de pruebas (20% de los registros).

III) Entrenamiento de los modelos

Se utilizaron los siguientes algoritmos en la construcción de los modelos de clasificación con el fin de comparar los resultados y verificar cual tenía mejor comportamiento

- Máquinas de vectores de soporte (SVC)
- Regresión logística (LR)
- K-Nearest-Neighbor (KNN)
- Naive Bayes (NB)
- Árboles de decisiones (TreeClas)
- Random forest (RF)

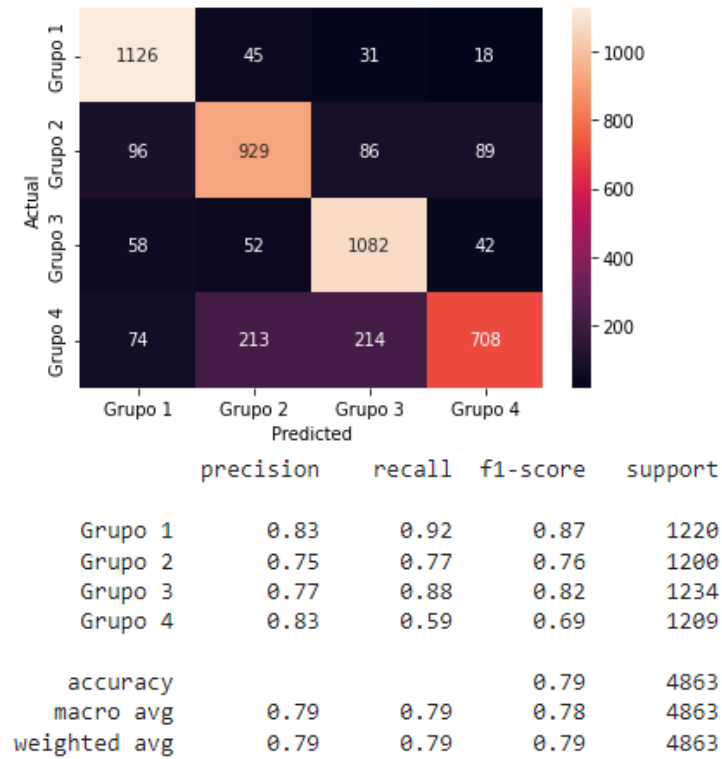
Los resultados se presentan a continuación:

Tabla 16. Resultados de desempeño de los modelos entrenados

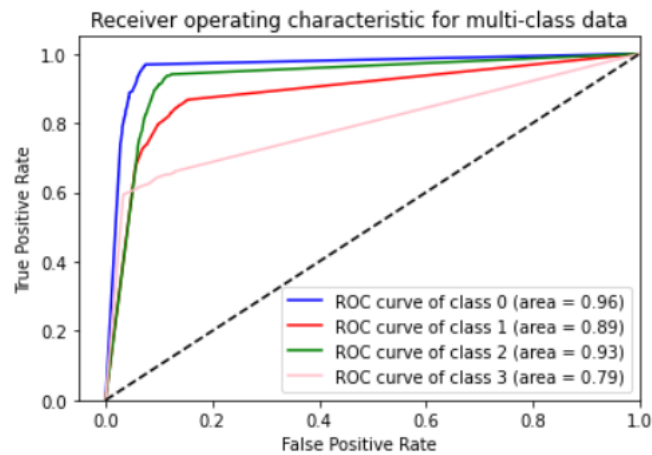
Algoritmo	Accuracy Score	Accuracy Error	Sensitivity Score	Precision Score	F1 Score
SVC	0,50	0,50	0,50	0,47	0,48
LR	0,49	0,51	0,49	0,47	0,47
KNN	0,50	0,50	0,50	0,49	0,49
NB	0,41	0,59	0,41	0,39	0,37
TreeClas	0,79	0,21	0,79	0,79	0,78
RF	0,56	0,44	0,56	0,56	0,54

Fuente: Elaboración propia

IV) Pruebas - matriz de confusión: Al realizar la matriz de confusión con el conjunto de datos de prueba para el modelo que mejor comportamiento tuvo en el entrenamiento (Árboles de decisiones), se obtuvo lo siguiente:

Figura 11. Resultados de la matriz de confusión

Fuente: Elaboración propia

VI) Curva ROC:**Figura 12.** Resultados de la Curva ROC

Fuente: Elaboración propia

VI) Revisión de multicolinealidad: Se aplicó un algoritmo para verificar si las variables independientes se encontraban relacionadas entre sí.

Tabla 17. Resultado inicial de multicolinealidad

Variable	VIF
CategoriaServicio_MEDICAMENTOS	24.15
CategoriaServicio_LABORATORIO CLINICO	22.23
CategoriaDX_Factores que influyen en el estado de salud y contacto con los servicios de salud (Z00-Z99)	15.16
CategoriaServicio_INSUMO Y EQUIPO MEDICO	14.49
CategoriaDX_Trastornos mentales y del comportamiento (F00-F99)	12.59

Fuente: Elaboración propia

El resultado indica que las variables con un VIF superior a 10 podrían estar correlacionadas. Por ello, se procede a eliminar la variable “CategoriaServicio_MEDICAMENTOS” y se verifica nuevamente la correlación entre las variables, el resultado fue el siguiente:

Tabla 18. Resultado final de multicolinealidad

Variable	VIF
CategoriaDX_Factores que influyen en el estado de salud y contacto con los servicios de salud (Z00-Z99)	2.38
Vulnerabilidad_Pobreza extrema	2.2
CategoriaDX_Trastornos mentales y del comportamiento (F00-F99)	2.19
Sexo_M	2.12

Fuente: Elaboración propia

Después de eliminar la variable no se identificaron problemas de colinealidad.

Aunque se entrenaron nuevamente los modelos, no se encontraron cambios en los resultados estadísticos para valorar el modelo.

8.4 Objetivo 4: Evaluar el desempeño y la calidad del modelo construido mediante la aplicación de métricas de aprendizaje automático.

Fase 5. Evaluación de resultados

Actividad 1.

De acuerdo con los resultados presentados en el ítem anterior, el algoritmo que realiza una mejor predicción de la categoría de costo del afiliado es los árboles de decisiones, en el conjunto de entrenamiento la exactitud, precisión y sensibilidad ascienden al 79%, es decir, 79 cada 100 veces, el modelo clasifica correctamente los verdaderos positivos y verdaderos negativos.

Al verificar el comportamiento del modelo en el conjunto de pruebas (20% de los registros), se obtuvo un 70% de instancias correctamente clasificadas, este dato es consistente con el F1-Score el cual mide el rendimiento del modelo mediante el cálculo de la media armónica de precisión, teniendo en cuenta tanto los falsos positivos como los falsos negativos.

Teniendo en cuenta lo anterior y los criterios de evaluación definidos en el apartado metodológico, los únicos modelos que se consideran buenos son aquellos donde se utilizaron los algoritmos Random Forest y Árboles de decisiones, siendo este último el que presenta mejores métricas.

Al evaluar el resultado del modelo seleccionado en el marco del problema previamente definido, se puede concluir la predicción del costo de las atenciones por evento en los rangos definidos, contribuye a los procesos de planificación orientados a disminuir los sobrecostos por las atenciones de pacientes que se encuentran en esta modalidad de contrato.

Si la EPS incorpora el modelo dentro de la estimación presupuestal anual, eventualmente, podría identificar la necesidad de cambiar el modelo de contratación para aquellos pacientes que representan costos por encima del promedio esperado. De esta manera, no solo fortalecen el proceso de planificación, también contribuyen a la disminución del déficit financiero, la previsión de servicios de alta demanda y el control a los procesos de facturación que generan las atenciones por evento.

8.5 Objetivo 5: Elaborar un listado de recomendaciones para la implementación del modelo construido.

Fase 6. Despliegue

Actividad 1.

Tras la socialización y la aceptación del modelo por parte de la Entidad, a continuación, se detalla el listado de recomendaciones que fueron construidos en articulación con expertos de la EPS para la correcta implementación del modelo, y las tareas a futuro por realizar con el propósito de continuar fortaleciendo el proceso de planificación presupuestal anual:

Con el fin de avanzar en la incorporación de este tipo de herramientas a los procesos de planificación de la institución, se listan a continuación un conjunto de recomendaciones:

-Consolidar un equipo experto para la proyección de presupuestos, compuesto por un integrante de cada área que actualmente aporta por separado los datos para el presupuesto anual.

-Continuar con la exploración de variables y la segmentación de datos que permitan fortalecer el modelo propuesto.

-Mediante los aportes del equipo experto conformado y la asesoría del grupo de Analítica y TI de la EPS, ajustar y actualizar el algoritmo predictivo, con el fin de encontrar más y mejores relaciones entre nuevas variables y el sobre costo de las atenciones pagadas por evento.

-Evaluar la posibilidad de inversión en herramientas de aprendizaje automático y la contratación de expertos relacionado a los procesos financieros, cuyo impacto sea el salvamento económico de la entidad.

-Definir lineamientos claros y documentados sobre la elaboración del presupuesto anual, y los ajustes presupuestales que se deben realizar mensualmente, mediante el uso de herramientas de aprendizaje automático.

-Negociar “paquetes de eventos” con las principales Instituciones Prestadores de Salud de acuerdo a la frecuencia de uso de los servicios más requeridos según la especialidad medida, con el propósito de contrarrestar el sobre costo generado por

el modelo de liquidación por Evento.

Como complemento a lo anterior, se proponen las siguientes actividades para fortalecer el modelo y garantizar su correcto despliegue:

Tabla 19. Actividades para el despliegue del modelo

Actividad	Responsable
Revisar la pertinencia de predecir el costo de atención por evento de cada afiliado	Área financiera
Evaluar la posibilidad de incorporar más variables relacionadas con el afiliado como el centro de atención, condiciones de vida, comuna, empleabilidad etc.	Área de aseguramiento
Ejecutar el entrenamiento del modelo con las nuevas variables	Área de aseguramiento
Seleccionar una muestra de afiliados y hacer un seguimiento durante 1 año para evaluar el comportamiento real de los costos de atención por evento	Área de aseguramiento
Comparar los resultados del modelo con los resultados reales	Área de aseguramiento
Diseñar un plan de mejora para el modelo	Área de aseguramiento

Fuente: Elaboración propia

9. Discusión

Tras la socialización de los hallazgos obtenidos con los involucrados de la EPS, se consideró valioso el aporte de la presente investigación, a la formulación de la proyección anual de gastos y la revisión mensual de la ejecución del presupuesto; considerando que en la actualidad ambas actividades se realizan mediante cálculos y formulas en Excel, desaprovechando el potencial del machine learning y la gran cantidad de datos con los que se cuenta para fortalecer los procesos de planificación.

Con este primer acercamiento, se concluye de manera conjunta que es necesario continuar explorando la relación de más variables con la predicción del costo de las atenciones cuya modalidad de pago se realiza por evento, debido a la complejidad de estimar la probabilidad de que un usuario o un servicio no incluido dentro del modelo de capitación, pueda ocurrir en cualquier momento.

En la EPS, es la primera vez que se realiza un ejercicio de predicción a este nivel de detalle. En los ejercicios actuales centrados en el modelo de capitación, se conoce los factores de riesgo, los servicios y las especialidades médicas, por lo que se logra estimar los gastos que se tendrá durante el año. Para la contratación por evento, sólo se contaba con cifras históricas para el cálculo presupuestal, por lo que investigaciones como estas, donde se centra la atención en el afiliado, representan un avance e innovación que sugieren la posibilidad de encontrar nuevos mecanismos para la planificación presupuestal.

Después de evaluar los resultados del modelo, se encontró un gran interés de la EPS en continuar explorando la utilización de técnicas de machine learning y en especial el modelo construido para la predicción del costo de atenciones por evento del afiliado, con el fin de desarrollar una estrategia que permita mejorar sus resultados e incorporarlo a los procesos de planificación interna.

Debido a que las morbilidades de los afiliados no inciden directamente con lo pagado por evento, resulta interesante explorar la frecuencia de uso de los servicios por prestador y especialidad para determinar una especie de “paquetes por evento” que puedan ser negociados con las instituciones prestadoras, y contrarresten el sobrecosto generado en dicha modalidad. Para esto, se pretende emplear la documentación y el algoritmo construido en esta investigación, ampliando la exploración de variables y la segmentación de datos, con el fin de fortalecer los planes de salvamento financiero actuales.

10. Conclusiones

Aunque la revisión bibliográfica sugiere un mayor interés de los investigadores del sector salud, por implementar de machine learning en el campo clínico; es necesario profundizar en el reconocimiento de otras experiencias en las que se haya utilizado un gran volumen de datos y algoritmos de aprendizaje máquinas, para fortalecer los procesos de planificación de las instituciones de este sector.

De acuerdo con lo expuesto en este proyecto de investigación, las técnicas de aprendizaje supervisado en algunas ocasiones, deben ser complementadas con la aplicación de algoritmos de aprendizaje no supervisado, así, facilitan la comprensión de los datos y permiten identificar grupos o segmentos que posteriormente pueden ser utilizados en algoritmos de clasificación.

Pese a que esta investigación es el primer ejercicio exploratorio en la EPS, en el que se utiliza machine learning para predecir el costo anual de atenciones por evento de un afiliado, el desempeño de los modelos fue aceptable; los resultados de las diferentes pruebas permiten concluir que las variables seleccionadas explican la variable objetivo. En este sentido, se podrían seguir explorando estrategias para fortalecer el modelo, incluyendo una mayor cantidad de datos, con más variables, con otras transformaciones y con un rango de temporalidad mayor.

En lo que corresponde al modelo de clasificación, es necesario realizar una predicción para el año 2023, seleccionar una muestra y hacer un seguimiento para comparar los resultados del modelo con el comportamiento de las atenciones por evento. Este proceso, servirá para tener una línea base que permita verificar la utilidad del modelo e identificar las estrategias para fortalecer el mismo en el futuro.

Este ejercicio de investigación, contribuye a la solución de los sobrecostos por pago de atenciones en salud bajo la modalidad de evento; en el futuro, después de verificar la utilidad del modelo dentro del proceso de planificación, la EPS podría cambiar la modalidad de contratación de un grupo de afiliados, si el modelo predice sobrecostos en la atención para el año siguiente a su ejecución.

11. Referencias

- Ahmed, I., Ahmad, M., Jeon, G., & Piccialli, F. (2021). A Framework for Pandemic Prediction Using Big Data Analytics. *Big Data Research*, 25, 100190. <https://doi.org/10.1016/J.BDR.2021.100190>
- Álvarez, B., Pellisé, L., & Lobo, F. (2000). Sistemas de pago a prestadores de servicios de salud en países de América Latina y de la OCDE. *Revista Panamericana de Salud Pública*, 8(1–2), 55–70.
- AWS. (n.d.). *¿Qué es Python?* Retrieved October 9, 2022, from <https://aws.amazon.com/es/what-is/python/>
- Ayala-García, J. (2014). La salud en Colombia: más cobertura pero menos acceso. *Documentos de Trabajo Sobre Economía Regional y Urbana; No. 204*.
- Banco Mundial. (n.d.). *Colombia: panorama general*. Retrieved October 5, 2022, from <https://www.bancomundial.org/es/country/colombia/overview>
- Barakat, N., Bradley, A. P., & Barakat, M. N. H. (2010). Intelligible support vector machines for diagnosis of diabetes mellitus. *IEEE Transactions on Information Technology in Biomedicine*, 14(4), 1114–1120. <https://doi.org/10.1109/TITB.2009.2039485>
- Bates, D. W., Saria, S., Ohno-Machado, L., Shah, A., & Escobar, G. (2017). Big Data In Health Care: Using Analytics To Identify And Manage High-Risk And High-Cost Patients. <https://doi.org/10.1377/Hlthaff.2014.0041>, 33(7), 1123–1131. <https://doi.org/10.1377/HLTHAFF.2014.0041>
- Boyle, L. M., Marshall, A. H., & Mackay, M. (2021). A framework for developing generalisable discrete event simulation models of hospital emergency departments. *European Journal of Operational Research*. <https://doi.org/10.1016/J.EJOR.2021.12.033>
- Buglioli, M., Gianneo, O., & Mieres, G. (2002). Modalidades de pago de la atención médica. *Revista Médica Del Uruguay*, 18(3), 198–210.
- Castro, H., Bardey, D., Castaño Yepes, R. A., López, Á., Estrada, C., Vélez, M., Rico, A., Arribas, D., Zuluaga, O., Palacios, L., Bolívar, P., & Barragán, S. (2014). *Modelos de contratación en servicios de salud*. 90. https://issuu.com/acemi/docs/modelos_de_contratacio__n__web_
- Castro-Villarreal, S., Beltran-Ostos, A., & Valencia, C. F. (2021). Estimation of Prevalence and Incremental Costs of Systemic Lupus Erythematosus in a Middle-Income Country Using Machine Learning on Administrative Health Data. *Value in Health Regional Issues*, 26, 98–104. <https://doi.org/10.1016/J.VHRI.2021.04.005>

- Ley 100, (1993) (testimony of Congreso de la República).
<https://www.minsalud.gov.co/sites/rid/Lists/BibliotecaDigital/RIDE/DE/DIJ/ley-100-de-1993.pdf>
- Ley 100 de 1993, (1993).
<https://www.minsalud.gov.co/sites/rid/Lists/BibliotecaDigital/RIDE/DE/DIJ/ley-100-de-1993.pdf>
- Decreto número 4747 de 2007.*
- Delpino, F. M., Costa, K., Farias, S. R., Chiavegatto Filho, A. D. P., Arcêncio, R. A., & Nunes, B. P. (2022). Machine learning for predicting chronic diseases: a systematic review. *Public Health, 205*, 14–25. <https://doi.org/10.1016/J.PUHE.2022.01.007>
- Departamento Nacional de Planeación. (n.d.). *Conoce el Sisben IV*. Retrieved October 5, 2022, from https://www.sisben.gov.co/Paginas/conoce_el_sisben.aspx
- el Colombiano. (2015). *Savia Salud perdería \$50 mil millones en 2015*.
<https://www.elcolombiano.com/antioquia/savia-salud-perderia-50-mil-millones-en-2015-XG1585231>
- el Colombiano. (2022, May 27). *Avanza plan para capitalizar la EPS Savia Salud para seguir atendiendo a 1,7 millones de afiliados*.
<https://www.elcolombiano.com/antioquia/como-va-el-plan-para-salvar-a-la-eps-savia-salud-FD17635220>
- el Espectador. (2013). *Nace primera EPS mixta de Colombia | EL ESPECTADOR*.
<https://www.elespectador.com/colombia/mas-regiones/nace-primera-eps-mixta-de-colombia-article-419147/>
- Ferreira Alfaya, F. J., & Zarzuelo Romero, M. J. (2022). Health literacy: A field for pharmaceutical intervention. *Research in Social and Administrative Pharmacy*.
<https://doi.org/10.1016/j.sapharm.2022.06.011>
- Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management, 35*(2), 137–144.
<https://doi.org/10.1016/J.IJINFOMGT.2014.10.007>
- García-Lacalle, J., Martín Vallespín, E., & Royo Montañés, S. (2009). La financiación de la sanidad pública. Efecto de los sistemas de pago prospectivos en el rendimiento de los hospitales. *Presupuesto y Gasto Público, 4*(57), 99–115.
- Gaviria, A. (2013). Exposición de motivos del Proyecto de Ley “Por el cual se redefine el Sistema General de Seguridad Social en Salud y se dictan otras disposiciones.” *Ministerio de Salud y Protección Social*.

- Gómez, O. S. (n.d.). *Propuesta de mejora sobre la primera etapa del modelo de proceso KDDM, CRISP-DM Related papers*. Retrieved October 5, 2022, from <http://www.sigkdd.org/>
- Gonzalez Fuentes, A. A., & others. (2017). *Comparación de sistemas de pago convenido en los costos de la prestación servicios ambulatorios y hospitalarios*. Universidad del Rosario.
- González, J. I. (2022, October 6). *Critican voluntad para capitalizar Savia Salud*. Teleantioquia Noticias. <https://www.youtube.com/watch?v=iKltX-oMoo>
- Gonzalez Rodriguez, J., Pinzon Espitia, O. L., Franco, C., & Augusto, V. (2019). PDB21 PREDICTING HEALTHCARE COSTS OF DIABETES USING MACHINE LEARNING MODELS. *Value in Health*, 22, S575. <https://doi.org/10.1016/J.JVAL.2019.09.903>
- González, V. V. (2021). Financiamiento del sistema de salud en Colombia: aspectos normativos y de gestión. *SUMMA. Revista Disciplinaria En Ciencias Económicas y Sociales*, 3(1), 1–34. <https://doi.org/10.47666/SUMMA.3.1.18>
- Google. (n.d.). *Google Colab*. Retrieved October 4, 2022, from <https://research.google.com/colaboratory/intl/es/faq.html>
- Groves, P., Kayyali, B., Knott, D., & Kuiken, S. van. (2013). *The big data revolution in healthcare*.
- Iranzo Jiménez, V. A. (2021). *Aplicación de Machine Learning en productos software mediante un enfoque de Model-Driven Development*.
- la República. (2021). *Las 17 EPS en riesgo de ser liquidadas en 2022 por incumplir indicadores financieros*. <https://www.larepublica.co/economia/las-17-eps-en-riesgo-de-ser-liquidadas-en-2022-por-incumplir-indicadores-financieros-3282257>
- Lee, G., Kim, J., & Lee, C. (2022). State-of-health estimation of Li-ion batteries in the early phases of qualification tests: An interpretable machine learning approach. *Expert Systems with Applications*, 197, 116817. <https://doi.org/10.1016/J.ESWA.2022.116817>
- Ley 10 de 1990. Retrieved September 12, 2022, from https://www.minsalud.gov.co/Normatividad_Nuevo/LEY%200010%20DE%201990.pdf
- McMullen, L., Parghi, N., Rogers, M. L., Yao, H., Bloch-Elkouby, S., & Galynker, I. (2021). The role of suicide ideation in assessing near-term suicide risk: A machine learning approach. *Psychiatry Research*, 304, 114118. <https://doi.org/10.1016/J.PSYCHRES.2021.114118>
- Ministerio de Salud y Protección Social. (n.d.). *ABECÉ Decreto 441 de 2022*.

- Moine, J. M., Haedo, A. S., & Gordillo, S. E. (2011). Estudio comparativo de metodologías para minería de datos. *XIII Workshop de Investigadores En Ciencias de La Computación*.
- Molina-Marín, G., Vargas-Jaramillo, J., Berrío-Castaño, A., & Muñoz-Marín, D. P. (2010). Características de la contratación entre aseguradores y prestadores de servicios de salud, Medellín, 2007-2008. *Gerencia y Políticas de Salud*, 9(18). <https://doi.org/10.11144/JAVERIANA.RGSP9-18.CCEA>
- Murdoch, T. B., & Detsky, A. S. (2013). The Inevitable Application of Big Data to Health Care. *JAMA*, 309(13), 1351–1352. <https://doi.org/10.1001/JAMA.2013.393>
- Navarro, S. M., Wang, E. Y., Haerberle, H. S., Mont, M. A., Krebs, V. E., Patterson, B. M., & Ramkumar, P. N. (2018). Machine Learning and Primary Total Knee Arthroplasty: Patient Forecasting for a Patient-Specific Payment Model. *Journal of Arthroplasty*, 33(12), 3617–3623. <https://doi.org/10.1016/J.ARTH.2018.08.028>
- OMS. (n.d.). *Determinantes sociales de la salud - OPS/OMS | Organización Panamericana de la Salud*. Retrieved October 5, 2022, from <https://www.paho.org/es/temas/determinantes-sociales-salud>
- Otálvaro Castro, G. J., Zuluaga Salazar, S. M., Jaramillo Delgado, G., Ternera Pulido, J. H., & Sandoval Valencia, S. (2019). Políticas de salud bucal en Colombia. Tendencias y puntos críticos para la garantía del derecho a la salud. *Universitas Odontologica*, 38(80). <https://doi.org/10.11144/Javeriana.uo38-80.psbcb>
- Padula, W. v., Kreif, N., Vanness, D. J., Adamson, B., Rueda, J. D., Felizzi, F., Jonsson, P., IJzerman, M. J., Butte, A., & Crown, W. (2022). Machine Learning Methods in Health Economics and Outcomes Research—The PALISADE Checklist: A Good Practices Report of an ISPOR Task Force. *Value in Health*, 25(7), 1063–1080. <https://doi.org/10.1016/J.JVAL.2022.03.022>
- Pang, X., Forrest, C. B., Lê-Scherban, F., & Masino, A. J. (2021). Prediction of early childhood obesity with machine learning and electronic health record data. *International Journal of Medical Informatics*, 150, 104454. <https://doi.org/10.1016/J.IJMEDINF.2021.104454>
- Prada-Ríos, S. I., Pérez-Castaño, A. M., & Rivera-Triviño, A. F. (2017). Clasificación de instituciones prestadores de servicios de salud según el sistema de cuentas de la salud de la Organización para la Cooperación y el Desarrollo Económico: el caso de Colombia. *Revista Gerencia y Políticas de Salud*, 16(32), 51–65.
- Presidente de la República de Colombia. (2022, March 28). *Decreto 441 de 2022* . <https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=184728>

- Resolución número 2573 de 2017*, (2017).
<https://www.normassalud.com/archivos/162fda1c3929b763e2d86af726d2bffe63c88d66c327b6eb9bd69d6042c8aefb>
- Russo, C., Ramón, H., Alonso, N., Cicerchia, B., Esnaola, L., & Tessore, J. P. (2016). *Tratamiento masivo de datos utilizando técnicas de Machine Learning*.
- Russom, P. (2011). Big data analytics. *TDWI Best Practices Report, Fourth Quarter*, 19(4), 1–34.
- Savia Salud. (n.d.-a). *Informe de gestión 2021*. Retrieved September 13, 2022, from <https://www.saviasaludeps.com/sitioweb/index.php/component/sppagebuilder/187-informe-de-gestion-2021>
- IN-RS-14 Instructivo de asignación de tarifas y autorización de servicios y tecnologías en salud no contratadas.
- MD-RS-01 Modelo de red de servicios.
- PD-RS-11 Procedimiento de contratación con prestadores de servicios de salud.
- Savia Salud. (n.d.-b). *Savia Salud EPS*. Retrieved November 8, 2022, from <https://saviasaludeps.com/sitioweb>
- Savia Salud. (2018, July 28). *Normograma*.
https://www.saviasaludeps.com/sitioweb/media/com_downloadmanager/protected/Normograma%20Actualizado%20Ago31.pdf
- Sharma, V., Stranieri, A., Ugon, J., Vamplew, P., & Martin, L. (2017). An agile group aware process beyond CRISP-DM: A hospital data mining case study. *ACM International Conference Proceeding Series, Part F130280*, 109–113.
<https://doi.org/10.1145/3093241.3093273>
- Su, Y., Tian, X., Gao, R., Guo, W., Chen, C., Chen, C., Jia, D., Li, H., & Lv, X. (2022). Colon cancer diagnosis and staging classification based on machine learning and bioinformatics analysis. *Computers in Biology and Medicine*, 145, 105409.
<https://doi.org/10.1016/J.COMPBIOMED.2022.105409>
- Suárez Rozo, L. F., Puerto García, S., Rodríguez Moreno, L. M., & Ramírez Moreno, J. (2017). La crisis del sistema de salud colombiano: una aproximación desde la legitimidad y la regulación. *Gerencia y Políticas de Salud*, 16(32), 34–50.
<https://doi.org/10.11144/JAVERIANA.RGPS16-32.CSSC>
- Thongpeth, W., Lim, A., Wongpairin, A., Thongpeth, T., & Chaimontree, S. (2021). Comparison of linear, penalized linear and machine learning models predicting hospital visit costs from chronic disease in Thailand. *Informatics in Medicine Unlocked*, 26, 100769. <https://doi.org/10.1016/J.IMU.2021.10076>

12. Anexos

Tabla 20. Diccionario de datos de la investigación

N°	Etiqueta	Descripción	Tipo	Naturaleza	Escala	Observación	Fuente
1	Serial_BDUA	Número único asignado a cada afiliado dentro de la ADRES	Numérica	Cuantitativa	De razón	Discreta	Conexiones Savia / Módulo Aseguramiento
2	Contrato_afil	Número único asignado a cada afiliado del sistema de la EPS	Numérica	Cuantitativa	De razón	Discreta	Conexiones Savia / Módulo Aseguramiento
3	ID_afiliado	Número único asignado a cada afiliado del sistema de la EPS en interoperabilidad con Boxalud	Numérica	Cuantitativa	De razón	Discreta	Conexiones Savia / Módulo Aseguramiento
4	tipo_documento	Tipo de documento que contiene la identificación de una persona	Categórica	Cualitativa	Nominal	Politómica	Conexiones Savia / Módulo Aseguramiento
5	numero_documento	Número único asignado a cada habitante del territorio colombiano	Numérica	Cuantitativa	De razón	Discreta	Conexiones Savia / Módulo Aseguramiento
6	fecha_nacimiento	Fecha registrada ante la Registraduría Nacional relacionada al nacimiento de una persona	Categórica	Cualitativa	Nominal	Politómica	Conexiones Savia / Módulo Aseguramiento
7	Sexo	Sexo/género que registra el afiliado (F,M)	Categórica	Cualitativa	Nominal	Dicotómica	Conexiones Savia / Módulo Aseguramiento

N°	Etiqueta	Descripción	Tipo	Naturaleza	Escala	Observación	Fuente
8	Edad	Tiempo de vida de un persona desde su fecha de nacimiento	Numérica	Cuantitativa	De razón	Discreta	Conexiones Savia / Módulo Aseguramiento
9	fecha_afiliacion_ep	Fecha en la que el usuario es afiliado a la EPS	Categórica	Cualitativa	Nominal	Politómica	Conexiones Savia / Módulo Aseguramiento
10	fecha_egreso_eps	Fecha en la que el usuario es retirado de la EPS	Categórica	Cualitativa	Nominal	Politómica	Conexiones Savia / Módulo Aseguramiento
11	municipio_divipola_afiliacion	Codificación DANE, según municipio de afiliación	Categórica	Cualitativa	Nominal	Politómica	Conexiones Savia / Módulo Aseguramiento
12	zona	Tipo de zona en la que vive una persona según la clasificación del DANE: Urbana (U), Rural (R)	Categórica	Cualitativa	Nominal	Dicotómica	Conexiones Savia / Módulo Aseguramiento
13	codigo_grupo_poblacional	Clasificación del grupo poblacional de una persona según la Resolución 4622 de 2016	Categórica	Cualitativa	Nominal	Politómica	Conexiones Savia / Módulo Aseguramiento
14	grupo_Sisben	Grupo de clasificación del Sisben	Categórica	Cualitativa	Nominal	Politómica	Conexiones Savia / Módulo Aseguramiento
15	estado_afiliacion	Relación del afiliado y la EPS	Categórica	Cualitativa	Nominal	Politómica	Conexiones Savia / Módulo Aseguramiento
16	modelo_liquidacion	Tipo de sistema de pago por el que está contratado el afiliado	Categórica	Cualitativa	Nominal	Dicotómica	Conexiones Savia / Módulo Aseguramiento
17	UPC_neto	Valor en pesos de cada afiliado según la UPC	Numérica	Cuantitativa	De razón	Discreta	Conexiones Savia / Módulo Aseguramiento

N°	Etiqueta	Descripción	Tipo	Naturaleza	Escala	Observación	Fuente
18	dias_UPC_net	Número de días por los que se percibió la UPC por cada afiliado	Numérica	Cuantitativa	De razón	Discreta	Conexiones Savia / Módulo Aseguramiento
19	costo_afiliado	Costo atribuido a la prestación de servicios de salud en la modalidad por Evento	Numérica	Cuantitativa	De razón	Discreta	Conexiones Savia / Módulo Cuentas médicas
20	Pertenece_VIH	Pertenece a programas por VIH	Categórica	Cualitativa	Nominal	Dicotómica	Conexiones Savia / Módulo Gestión del riesgo
21	Pertenece_Cancer	Pertenece a programas por Cáncer	Categórica	Cualitativa	Nominal	Dicotómica	Conexiones Savia / Módulo Gestión del riesgo
22	Pertenece_Hemofilia	Pertenece a programas por Hemofilia	Categórica	Cualitativa	Nominal	Dicotómica	Conexiones Savia / Módulo Gestión del riesgo
23	Pertenece_Reuma	Pertenece a programas por ERC (renales)	Categórica	Cualitativa	Nominal	Dicotómica	Conexiones Savia / Módulo Gestión del riesgo
24	Pertenece_Prog_Hepatitis	Pertenece a programas por hepatitis	Categórica	Cualitativa	Nominal	Dicotómica	Conexiones Savia / Módulo Gestión del riesgo
25	Pertenece_Prog_ERC	Pertenece a programas por enfermedad renal crónica	Categórica	Cualitativa	Nominal	Dicotómica	Conexiones Savia / Módulo Gestión del riesgo

N°	Etiqueta	Descripción	Tipo	Naturaleza	Escala	Observación	Fuente
26	Pertenece_Prog_Huerfanas	Pertenece a programas por alguna enfermedad huérfana	Categórica	Cualitativa	Nominal	Dicotómica	Conexiones Savia / Módulo Gestión del riesgo
27	otros_autoinmunes	Pertenece a programas por alguna enfermedad autoinmunes	Categórica	Cualitativa	Nominal	Dicotómica	Conexiones Savia / Módulo Gestión del riesgo
28	otros_cardio	Pertenece a programas por alguna enfermedad cardiovascular	Categórica	Cualitativa	Nominal	Dicotómica	Conexiones Savia / Módulo Gestión del riesgo
29	otros_pulmonar	Pertenece a programas por alguna enfermedad pulmonar	Categórica	Cualitativa	Nominal	Dicotómica	Conexiones Savia / Módulo Gestión del riesgo
30	otros_diabetes	Pertenece a programas por alguna enfermedad relacionada a diabetes	Categórica	Cualitativa	Nominal	Dicotómica	Conexiones Savia / Módulo Gestión del riesgo
31	otros_tuberculosis	Pertenece a programas por alguna enfermedad relacionada a tuberculosis	Categórica	Cualitativa	Nominal	Dicotómica	Conexiones Savia / Módulo Gestión del riesgo
32	portabilidad	Asignación de portabilidad del afiliado	Categórica	Cualitativa	Nominal	Dicotómica	Conexiones Savia / Módulo Aseguramiento

N°	Etiqueta	Descripción	Tipo	Naturaleza	Escala	Observación	Fuente
33	servicios	Relación de servicios de salud asociados a cada afiliado	Categórica	Cualitativa	Nominal	Politómica	Conexiones Savia / Módulo Acceso
34	diagnósticos	Relación de diagnósticos de salud asociados a cada afiliado	Categórica	Cualitativa	Nominal	Politómica	Conexiones Savia / Módulo Acceso

Fuente: Elaboración propia