FACULTAD DE INGENIERÍA
-Maestría en Ciencia de Datos-

Modelos de Aprendizaje Automático para la Optimización de la Gestión Farmacéutica.

Presentado por:

Bertha G. Botero Ortiz
Susan L. Botero Ortiz

Directores:

Anibal Sosa Aguirre
Ph. D en Computational Science
Andres A. Aristizabal Pinzón
Ph. D en Ingeniería de Sistemas

Cali, junio 22 de 2023

Tabla de contenido

Índice de Figuras

Índice de tablas

RESUMEN

En este documento se aborda la implementación de un modelo de aprendizaje automático para la optimización de la gestión farmacéutica. El objetivo de este modelo es fortalecer la prevención y promoción de estilos de vida saludables, así como prevenir complicaciones asociadas a enfermedades existentes de manera oportuna. Se busca mejorar la calidad de la atención médica, garantizar la seguridad y eficacia en la prescripción de medicamentos, y optimizar el uso de los recursos de salud.

El proyecto se centra en 19.362 pacientes con diabetes de la cohorte cardiovascular de una entidad promotora de salud. El propósito es formular una alternativa para mejorar los resultados clínicos de estos pacientes a través de la optimización de la gestión farmacéutica. Se busca crear una solución analítica y funcional que permita mantener buenos resultados de salud en los pacientes, al tiempo que se logra eficiencia operacional en el servicio y control presupuestario de los tratamientos médicos.

El proyecto utilizó la metodología CRISP-DM (Cross-Industry Standard Process for Data Mining). Esta metodología se ha estandarizando sobre todo en el sector de la industria para proyectos de data mining (Wirth, 2000) y continúa siendo ampliamente utilizado en proyectos analíticos debido a su versatilidad. Por otra parte, fue usado el algoritmo XGBoost (Extreme Gradient Boosting) para iterar diferentes arquitecturas y transformaciones de variables, además se realizó ingeniería de variables y se incluyo entre ellas, el número de pacientes.

Entre los principales resultados, se observó que variables como la edad, el sexo, el sobrepeso, la obesidad, el índice de masa corporal y la hemoglobina glicosilada (HbA1c) están estrechamente relacionadas con la diabetes y pudieran ser incluidas en el modelo de pronóstico. Se realizaron los pronósticos para la cantidad de dosis de 5 medicamentos que hacen parte de la prescripción del tratamiento de la cohorte.

Se concluye que el aprendizaje automático es una herramienta eficaz para la optimización de la gestión farmacéutica, mediante la predicción de la cantidad de dosis de medicamentos en pacientes con diabetes de la cohorte cardiovascular. Los resultados indican que el algoritmo XGBoost tuvo el mejor desempeño en comparación con otros algoritmos evaluados, como LSTM y LightgBM.

Palabras claves: Prescripción de medicamentos, Diabetes, Gestión del riesgo en salud, pronósticos, XGBoost, LSTM y LightgBM.

1. INTRODUCCIÓN.

1.1. Contexto, Antecedentes y Justificación

Una entidad aseguradora y prestadora de salud implementa un modelo moderador de riesgo como estrategia integral de prevención y promoción, brindando a los usuarios mecanismos para fortalecer estilos de vida saludables y prevenir complicaciones asociadas a enfermedades existentes de forma oportuna, con la mejor calidad humana y tecnología disponible (Ministerio de Salud y Protección Social [MinSalud], 2023). En el marco de este modelo, es necesario fortalecer la gestión farmacológica con el objetivo de lograr tres resultados: mejorar la calidad de la atención médica, garantizar la seguridad y eficacia en la prescripción de medicamentos, y optimizar el uso de los recursos de salud.

El costo de los medicamentos representa el 34% de los costos totales asociados al cuidado de la salud de la población afiliada, siendo el tercer rubro más costoso en orden descendente. Utilizando el indicador Pareto, también conocido como la ley 80/20 (Abdallah, 2020), se observa que los tres primeros rubros del costo médico representan más del 80% de los costos totales. Dentro del rubro de medicamentos, las enfermedades de alto costo constituyen el 83% del gasto total en medicamentos, y la cohorte cardiovascular representa el 54% de las enfermedades de alto costo. El objetivo de utilizar esta herramienta es priorizar y lograr resultados tempranos al controlar eficazmente las enfermedades. Para las entidades aseguradoras y proveedores de servicios de salud, es crucial garantizar el control de los costos asociados con el manejo de enfermedades. El objetivo no solo implica controlar la enfermedad, sino hacerlo de manera eficiente desde una perspectiva económica (Entidad Promotora de Salud [EPS], 2022).

De acuerdo con la priorización anterior, este trabajo se centra en analizar a los pacientes con diabetes de la cohorte cardiovascular de una Entidad Promotora de Salud (EPS). Se plantea como propósito del trabajo formular una alternativa para mejorar los resultados clínicos de los pacientes a través de la optimización de la gestión farmacéutica, creando una solución analítica y funcional que permita mantener buenos resultados de salud en los pacientes con diabetes de la cohorte cardiovascular, al mismo tiempo que se logra eficiencia operacional en el servicio y control presupuestario de los tratamientos médicos.

1.2. Planteamiento del Problema

El cumplimiento terapéutico se considera fundamental para obtener resultados clínicos, económicos y humanísticos en salud. Sin embargo, se ha observado que los profesionales de la salud enfrentan limitaciones en la atención, como la falta de tiempo y recursos para brindar tratamientos pertinentes. Además, están expuestos

a la presión de las compañías farmacéuticas para prescribir sus tecnologías, lo que puede afectar la efectividad de los tratamientos y aumentar sus costos (Arredondo A. y De Icaza E., 2011).

Los altos costos en los tratamientos médicos pueden atribuirse a diversas causas, como ineficiencias administrativas, médicas u operativas, baja adherencia de los pacientes a los tratamientos debido a deficiencias en el monitoreo y bajo conocimiento de los pacientes sobre sus tratamientos, y falta de controles en la prescripción médica. (Arredondo A. y De Icaza E., 2011)

Adicionalmente, la mezcla de estas causas generan un mayor riesgo de desenlaces evitables en los pacientes con diabetes, empeorando así los resultados de salud, aumentando los costos médicos, generando insatisfacción en los usuarios y provocando pérdida de productividad laboral e insostenibilidad financiera de las entidades de salud (Organización Panamericana de la Salud [PAHO], 2022).

Estas causas se relacionan directamente con la cantidad de recursos que los países destinan, para mantener una vida digna para las personas que padecen diabetes, lo que afecta los costos indirectos y la asignación eficiente de los recursos de salud. Se estima que los costos médicos directos e indirectos de la diabetes causarán pérdidas en el producto interno bruto (PIB) en todo el mundo de $1,7 trillones de dólares entre el 2011 y 2030 (World Health Organization [WHO], 2021).

En el caso específico de la entidad objeto de estudio, se puede constatar que los costos de los diagnósticos y tratamientos de la diabetes se han incrementado en aproximadamente un 51% entre el año 2021 al 2022. Se estimó un incremento del 2% en los desenlaces evitables en el mismo periodo, además, se identificó un 3% del sobrecosto de medicamentos en la prescripción médica (EPS, 2022)

Ante esta situación, surge la siguiente pregunta de investigación: ¿Cómo anticipar los costos de medicamentos por paciente de la cohorte cardiovascular con diagnóstico de diabetes, que permita a la EPS realizar controles presupuestales de sus tratamientos médicos y mejorar la negociación con la red de dispensación?

1.3. Objetivo General

Pronosticar los costos en medicamentos por paciente de la cohorte cardiovascular con diagnóstico de diabetes, que permita a la EPS realizar controles presupuestales de sus tratamientos médicos y mejorar la negociación con la red de dispensación.

1.4. Objetivos Específicos

1. Caracterizar el comportamiento de los costos de medicamentos de los pacientes de la cohorte cardiovascular con diagnóstico de diabetes en el año 2022.

2. Seleccionar las variables de mayor impacto en los costos de medicamentos de los pacientes de la cohorte cardiovascular con diagnóstico de diabetes en el año 2022.

3. Formular el modelo de pronóstico de los costos de medicamentos considerando las variables seleccionadas.

4. Evaluar el modelo de pronóstico de los costos de medicamentos considerando las variables seleccionadas.

2. ANTECEDENTES

2.1. Marco Teórico

2.1.1. Dominio del Problema.

La Organización Panamericana de la Salud (PAHO, 2022) en su informe “Panorama de Diabetes en las Américas” del 2022, define la Diabetes como una enfermedad metabólica crónica caracterizada por niveles elevados de glucosa en sangre (o azúcar en sangre), que con el tiempo conduce a daños graves en el corazón, los vasos sanguíneos, los ojos, los riñones y los nervios. La más común es la diabetes tipo 2, que generalmente ocurre en adultos cuando el cuerpo se vuelve resistente a la insulina o no produce suficiente insulina. Durante las últimas tres décadas, la prevalencia de la Diabetes Mellitus tipo 2 (DM tipo 2) ha aumentado significativamente en países de todos los niveles de ingresos (Wesson & Naude 2022). Se estima que 62 millones de personas en las Américas y 422 millones de personas en todo el mundo tienen diabetes, la mayoría vive en países de ingresos bajos y medianos, y 244 084 muertes (1.5 millones en todo el mundo) se atribuyen directamente a la diabetes cada año. La mortalidad por diabetes ha aumentado en un 70,0 % desde 2000, ubicándose entre las 10 principales causas de muerte en todo el mundo (World Health Organization [WHO], 2020).

La diabetes con el tiempo puede traer serios desenlaces e impactos en la salud. En particular, es fuente de enfermedades como la retinopatía diabética, causa importante de ceguera, la neuropatía (daño a los nervios) en los pies, que combinado con un flujo sanguíneo reducido aumenta la posibilidad de úlceras en el

pie, infección y eventual necesidad de amputación de una extremidad, daños en el corazón, los vasos sanguíneos y los riñones. Los adultos con diabetes tienen un riesgo dos o tres veces mayor de sufrir ataques cardíacos y accidentes cerebrovasculares (PAHO, 2022)

En consecuencia, la carga de esta enfermedad y sus complicaciones pueden afectar seriamente la calidad de vida de las personas con diabetes, sus familias y la sociedad en su conjunto (Wesson & Naude 2022). Pero se convierte también en una sobrecarga en el sistema de salud. Los altos costos asociados con su tratamiento son una carga económica para pacientes, entidades y el sistema de salud, y amenazan con impedir el desarrollo social y económico de los países. Se estima que los costos médicos directos e indirectos de la diabetes causarán pérdidas en el producto interno bruto (PIB) en todo el mundo de $1,7 millones de dólares entre el 2011 y 2030 (Organización Panamericana de la Salud [PAHO], 2022). Asimismo, asumiendo el supuesto de que la prevalencia de la diabetes se encuentra determinada por los cambios demográficos y la urbanización, se espera que los costos directos e indirectos lleguen a los 2,25 billones de dólares en 2030 (World Health Organization [WHO], 2021).

Cuando nos remitimos al caso colombiano, el sistema de salud se encuentra regulado por la Ley Estatutaria 1751 de 2015 del Congreso. En la misma el sistema de salud se define como “El conjunto articulado y armónico de principios y normas; políticas públicas; instituciones; competencias y procedimientos; facultades, obligaciones, derechos y deberes; financiamiento; controles; información y evaluación, que el Estado disponga para la garantía y materialización del derecho fundamental de la salud” (MinSalud, 2023).

En el marco de la misma Ley, se define que el Sistema General de Seguridad Social en Salud - SGSSS está integrado, primero, por el Estado, a través del Ministerio de Salud y Protección Social, quien actúa como organismo de coordinación, dirección y control; segundo, por las Entidades Promotoras de Salud (EPS), responsables de la afiliación y el recaudo de las cotizaciones y de garantizar la prestación del Plan Obligatorio de Salud a los afiliados; tercero, por las instituciones prestadoras de salud (IPS), que son los hospitales, clínicas y laboratorios, entre otros, encargadas de prestar la atención a los usuarios. Igualmente en la ley se expresa que también hacen parte del SGSSS las Entidades Territoriales y la Superintendencia Nacional de Salud, como entes de control y vigilancia. Así mismo, el SGSSS es el encargado de regular el servicio público esencial de salud y crear condiciones de acceso para toda la población (MinSalud, 2023).

La ley también declara que es el Estado a través de la acción coordinada con las EPS las responsables de intervenir el mercado de medicamentos, dispositivos médicos e insumos en salud con el fin de optimizar su utilización, evitar las inequidades en el acceso y asegurar la calidad de los mismos. Adicionalmente define que las EPS son responsables de cumplir con las funciones del

aseguramiento en salud, las cuales comprenden entre otras, la administración del riesgo financiero y la gestión del riesgo en salud (MinSalud, 2023).

Por una parte el Ministerio de Salud y Protección Social [MinSalud], 2022. Señaló que para el año 2022 el presupuesto para el aseguramiento y otros programas se proyectan en cerca de $73 billones. Un 53% desde el Presupuesto General de la Nación, y un 36 % desde las cotizaciones. Por otra parte, la OCDE estima que "el gasto promedio per cápita en el país en salud al año es de cerca de $1,3 millones, mientras que lo que financian tanto empresas como el Estado es de 4,8 millones. Eso es un total de $6,1 millones en el promedio per cápita del gasto en salud de cada persona en el país" (El Diario, LR La República [LR La República], 2022). Junto a esto, el DANE reportó en el área de política social en salud en el año 2021, una participación de 41,4%, con un crecimiento anual de 15,0%. También encontró que la salud es financiada por el sector público en 86,2%, seguido de la financiación privada voluntaria con el 10,0% y finalmente la financiación privada obligatoria con el 3,8% (Departamento Administrativo Nacional de Estadística [DANE], 2022).

Respecto a la financiación pública, los servicios médicos, hospitalarios y farmacéuticos del régimen contributivo y del régimen subsidiado registran las mayores participaciones con el 50,6% y el 34,3%, respectivamente; frente al año 2020 se registró un crecimiento de 16,7% y 12,7%, respectivamente. Estas participaciones dependen de diferentes factores, muchos de ellos se encuentran muy relacionados a los tratamientos médicos de las enfermedades (DANE, 2022).

2.1.2. Dominio de la Solución

En el contexto del acápite anterior y en la consideración de la Ley Estatutaria 1751 de 2015 del Congreso, que establece como marco de funcionamiento del SGSSS al Estado, las EPS y las IPS como responsables de la afiliación, el recaudo de las cotizaciones y de garantizar la prestación del Plan Obligatorio de Salud a los afiliados (MinSalud, 2023). Se plantea en el presente trabajo proponer un modelo de pronóstico de los costos en medicamentos por paciente de la cohorte cardiovascular con diagnóstico de diabetes de acuerdo al estadio de su enfermedad, con los cuales se espera que la EPS pueda realizar controles presupuestales de sus tratamientos médicos y mejorar la negociación con la red de dispensación.

Como se ha dicho, en el modelo de aseguramiento de salud en Colombia (SGSSS) las Entidades Promotores de Salud – EPS, entre otras funciones, están encargadas de "…en cada régimen son responsables de cumplir con las funciones indelegables del aseguramiento. Se entiende por aseguramiento en salud, la administración del riesgo financiero, la gestión del riesgo en salud, la articulación de los servicios que garantice el acceso efectivo, la garantía de la calidad en la prestación de los servicios de salud y la representación del afiliado ante el prestador y los demás actores sin perjuicio de la autonomía del usuario…" (MinSalud, 2023).

Es así que, la gestión del riesgo en salud, implica el uso de los sistemas de salud y otros sectores para identificar, evaluar, medir, intervenir, rastrear y monitorear los riesgos para la salud de las personas, las familias y las comunidades a fin de lograr resultados en salud y bienestar de la población. Además de lograr mejores niveles de salud en la población, una mejor experiencia de los usuarios durante el proceso de atención y unos costos acordes a los resultados alcanzados (Ministerio de Salud y Protección Social [MinSalud], 2018).

Hay que mencionar, además, los riesgos relacionados con el cumplimiento de las funciones del aseguramiento. Las EPS deben gestionar todos los riesgos que se presenten en su operación, los cuales dependen de la discrecionalidad y organización de cada EPS. Sin embargo, para efectos de Inspección, Vigilancia y Control y de la supervisión basada en riesgos, la Superintendencia Nacional de Salud - SNS definió categorías para "la identificación de potenciales riesgos de salud, financieros y operativos que enfrentan las entidades vigiladas, así como su capacidad para medir, gestionar y monitorear dichos riesgos" (Tomando de la referencia realizada en el documento de trabajo del MinSalud, 2018, sobre la Resolución 4559 de 2018 de la SNS, 2018). Entre los cuales es importante resaltar, el **riesgo actuarial**, el cual se define como:

"La posibilidad de incurrir en pérdidas económicas debido a la ocurrencia de diferentes sucesos futuros e inciertos, por ejemplo

Riesgo de incrementos inesperados en los índices de morbilidad y en los costos de atención: corresponde a la probabilidad de pérdida de un periodo contable que se genera como consecuencia de diferencias apreciables entre las condiciones de morbilidad asumidas y las actuales, así como pérdidas derivadas de incrementos inesperados en los costos de atención" (MinSalud, 2023).
De manera que, los costos de medicamentos al hacer parte de los costos de atención, se convierten en una prioridad para estas entidades y sus procesos de evaluación, medición, intervención y monitoreo. Así que, para los administradores de salud es importante contar con herramientas de toma de decisiones que se anticipen a los eventos futuros, a los incrementos inesperados en los costos de atención, a las enfermedades y los traumatismos para que éstos no se presenten o se detecten y se traten tempranamente para disminuir, acortar su evolución y consecuencias, en especial, sobre la prescripción y costos de medicamentos para el tratamiento de los pacientes de la cohorte cardiovascular con diagnóstico de diabetes.

Se ha demostrado la necesidad de construir una solución que permita anticiparse a los hechos, que posibilite un mejor tratamiento y un incremento en la eficiencia del gasto (disminución del costo de tratamiento) y que permita contribuir a la generación de diferentes acciones para mantener la estabilidad financiera y la salud de los pacientes en el sistema de salud.

Por lo que se refiere a anticiparse, es la habilidad de actuar conociendo lo que va a pasar en el futuro, actuar es tomar decisiones que permitan lograr los objetivos propuestos (EAE Business School Online - Blended.[EAE Bussiness School], 2023). En la antigua Grecia cuando alguien tenía que tomar una gran decisión, políticos, militares, reyes acudían a los oráculos, hoy los tomadores de decisiones que lideran las corporaciones encuentran en la explotación de los datos a sus oráculos.

Por otro lado, el desarrollo acelerado de las tecnologías de la información ha hecho emerger en los últimos años, tópicos que parecían solamente posibles en las películas de ciencia ficción. La inteligencia artificial, las plataformas digitales globales, la ciencia de datos, reconfiguran una realidad de convergencia del mundo físico y el mundo digital. En efecto, la ciencia de datos aparece con un rol fundamental en las organizaciones contemporáneas. La ciencia de datos permite a las organizaciones tomar decisiones basadas en evidencias y en hechos concretos, al analizar y comprender grandes cantidades de datos. Es una manera también eficiente para identificar patrones y tendencias ocultas en los conjuntos de datos. Permite la mejora de la eficiencia operativa de las organizaciones, posibilitando la optimización de los procesos internos, mediante el análisis de datos sobre el rendimiento, los tiempos de producción, los costos y otros factores. La ciencia de datos se ha convertido en una fuente muy importante para la detección de fraudes y riesgos al interior de las organizaciones, en tanto que, ayuda a identificar patrones de comportamientos sospechosos y detectar fraudes o riesgos potenciales. (Dhar, V., 2013)

Dicho lo anterior, se define la ciencia de datos como la encargada del estudio y explotación de los datos con el fin de extraer información significativa para las empresas. “Es un enfoque multidisciplinario que combina principios y prácticas del campo de las matemáticas, la estadística, la inteligencia artificial y la ingeniería de computación para analizar grandes cantidades de datos” (Amazon Web Services [AWS], 2023). La combinación de dichos campos, brindan herramientas y métodos de análisis estadístico y aprendizaje automático, que permiten resolver preguntas como ¿qué ocurrió?, ¿por qué ocurrió?, pero también, realizar pronósticos de diferentes fenómenos de estudio.

Por su parte, cuando nos referimos a los pronósticos, se hace referencia al conocimiento anticipado de lo que sucederá en el futuro, a través de elementos científicos, es un proceso de estimación en situaciones de incertidumbre, que se realiza mediante algoritmos analizando las diferentes variables que influyen en el fenómeno estudiado. (Montemayor Gallegos J. E., 2013). Los métodos de pronóstico se clasificaban en dos grandes grupos en métodos cualitativos y cuantitativos, con el crecimiento de los datos y avances de la tecnología entra un tercer grupo, los métodos de la inteligencia artificial (Jaramillo Ramirez L., 2012). A continuación se describen los principales conceptos de estos tres grandes grupos.

1. Cualitativos:

Los métodos cualitativos son principalmente subjetivos y se apoyan en el juicio humano. Son apropiados sobre todo cuando la información histórica no está disponible o existen muy pocos datos; o bien, cuando los expertos cuentan con resultados de investigación del mercado (market intelligence) que pueden afectar el pronóstico.Tales métodos pueden también ser necesarios para pronosticar la demanda a varios años en el futuro de una nueva industria.

2. Cuantitativos:

Cuando se dispone de información histórica, los métodos cuantitativos son los más comúnmente empleados para hacer pronósticos. Dentro de esta categoría se encuentran tanto los métodos univariados como los multivariados. Los primeros se basan en la suposición de que la variable que se está estudiando está influenciada por sus valores previos, mientras que los segundos parten del supuesto de que es posible prever el comportamiento de dicha variable a partir de los niveles de otras variables que se encuentran bajo control (Montemayor Gallegos J. E., 2013).

3. Inteligencia artificial

En la actualidad, existen diversos métodos para realizar pronósticos en distintas industrias. Dichos métodos se basan en la aplicación de algoritmos matemáticos, estocásticos de series de tiempo, regresivos y computacionales que utilizan técnicas de inteligencia artificial (IA) para hacer pronósticos (Perdigón y González, 2021). Estos métodos han demostrado ser más eficaces que los métodos cuantitativos y cualitativos, ya que tienen la capacidad de modelar sistemas complejos no lineales (Jaramillo Ramirez L., 2012).

Las herramientas de solución de la situación problemática para este caso aplicado están enmarcadas desde las técnicas de inteligencia artificial.

Entre las técnicas de inteligencia artificial más utilizadas se identificaron:

1. Técnicas basadas en Redes Neuronales Artificiales (ANN, por sus siglas en inglés) (Sugiono, Soenoko & Riawati, 2017; Gorgulu, 2018; Machado et al., 2019; Torres-Inga et al., 2019; Liseunea et al., 2020; Zhang et al., 2016; 2019; 2020; Zhang W. et al., 2020)

2. Técnicas basadas en Máquinas de Soporte Vectorial (SVM, por sus siglas en inglés) (Yan et al., 2015; Nguyen et al. 2020; Saha & Bhattacharyya, 2020)

3. Técnicas basadas en Árboles de Decisión (DT, por sus siglas en inglés) (Eyduran et al., 2013; Nguyen et al. 2020; Piwczyński et al., 2020; Kliś et al., 2021)

Estas técnicas muestran un comportamiento apropiado para la realización de pronósticos en diferentes ámbitos porque carecen de limitaciones para el manejo de grandes cantidades de datos y variables de entrada y poseen la capacidad de identificar, aprender y aproximar las características de los datos simulando las relaciones intrínsecas y no lineales existentes en estos (Dongre et al.,2012; Oscullo & Haro, 2016; Kaygisiz & Sezgin, 2017). Además, posibilitan obtener resultados precisos con ahorro de tiempo y recursos de cómputo (Nguyen et al., 2020).

2.1.3. Estrategia de validación de modelos

Existen diversas técnicas de validación de modelos de pronóstico que permiten analizar la precisión y capacidad de generalización de un modelo. A continuación, se describen algunas de las estrategias más utilizadas en esta tarea. (Brownlee, J., 2017)

Validación cruzada: La validación cruzada implica dividir los datos en varios conjuntos de entrenamiento y prueba, ajustando el modelo en cada conjunto de entrenamiento y evaluando su rendimiento en el conjunto de prueba. Evalidasta estrategia ayuda a reducir el riesgo de sobreajuste y a evaluar la capacidad de generalización del modelo.

División temporal de datos: En la división temporal de datos, se utiliza una parte de los datos para ajustar el modelo y otra parte para evaluar su rendimiento en un período de tiempo posterior. Esta estrategia es adecuada para conjuntos de datos con tendencias y estacionalidad, ya que el modelo puede ser evaluado en un período en el que aún no se había producido.

Validación en línea: La validación en línea implica la actualización del modelo con nuevos datos a medida que están disponibles, y la evaluación continua de su rendimiento en tiempo real. Esta estrategia es adecuada para modelos que se utilizan para hacer predicciones a corto plazo, y permite la identificación temprana de problemas en el modelo y la mejora continua.

Validación por ventana deslizante: En la validación por ventana deslizante, se ajusta el modelo utilizando un subconjunto de datos y se evalúa su rendimiento en un período posterior de la serie de tiempo. Luego, la ventana se desplaza hacia adelante en el tiempo y se repite el proceso. Esta estrategia ayuda a evaluar la capacidad del modelo para hacer predicciones a corto y largo plazo.

2.1.4. Métricas de evaluación de modelos de pronósticos

La precisión de un pronóstico depende de qué tan cercano está a lo que sucede en la realidad, y generalmente se evalúa utilizando una medida promedio de los errores del pronóstico. El error se considera una medida de seguimiento y se expresa comúnmente como una proporción. Existen diversas formas de medir el error, y estas se utilizan para evaluar el rendimiento de los modelos de inteligencia artificial y seleccionar el mejor modelo para cada tipo de elemento que se pronostica (Liu, Y. et al., 2015).

En la siguiente tabla se relacionan algunas métricas de evaluación de los modelos de pronóstico, una breve descripción y su interpretación:

TABLA 1. MÉTRICAS DE EVALUACIÓN DE LOS MODELOS DE PRONÓSTICO

Métrica	Breve descripción	Interpretación
MAE (Mean Absolute Error) o Error Absoluto Medio	Es la media de las diferencias absolutas entre las predicciones y los valores observados en el conjunto de datos.	Cuanto menor sea el valor de MAE, mejor será la precisión del modelo. Se expresa en las mismas unidades que la variable de pronóstico.
RMSE (Root Mean Squared Error) o Raíz del Error Cuadrático Medio	Se calcula como la raíz cuadrada de la media de los errores al cuadrado entre las predicciones y los valores observados en el conjunto de datos. En otras palabras, mide la raíz cuadrada de la varianza de los errores de pronóstico.	Cuanto menor sea el valor de RMSE, mejor será la precisión del modelo. Tiende a penalizar más los errores grandes que el MAE, ya que eleva los errores al cuadrado antes de calcular la media. Se expresa en las mismas unidades que la variable de pronóstico.
MAPE (Mean Absolute Percent Age Error) o Error Porcentual Absoluto Medio	Se calcula como el promedio de las diferencias porcentuales absolutas entre las predicciones y los valores observados en el conjunto de datos. En otras palabras, mide el promedio de los errores porcentuales absolutos de pronóstico, expresados como un porcentaje del valor observado.	Cuanto menor sea el valor de MAPE, mejor será la precisión del modelo. Es independiente de la escala de la variable de pronóstico. Puede tener problemas cuando los valores observados son cercanos a cero o cuando hay valores atípicos en el conjunto de datos.

Fuente: Elaboración propia

2.2. Estado del Arte

Para analizar los estudios que sean referentes del desarrollo que se ha conseguido en el tema que nos ocupa en el proyecto de grado, se utilizaron las siguientes palabras para su búsqueda: pronósticos, costos, diabetes.

2.2.1. Trabajos seleccionados

Design of a CPFR (collaborative planning forecasting and replenishment), location, inventory and routing approach to diabetes and high blood pressure medicines supply network planning (Chuchoque-Urbina F. A., Caro-Gutierrez M. P., and Montoya C. E., 2021). Este estudio tuvo como objetivo diseñar un modelo CPFR para la entrega de medicamentos para diabetes e hipertensión arterial desde una compañía de seguros de salud (EPS) a un proveedor de salud (IPS) y comparar el desempeño de esta cadena colaborativa con la tradicional a través de sus correspondientes costos de la cadena de suministro. Entre sus resultados logró demostrar el impacto del modelo propuesto, al reducir en un 11,2% en promedio, el costo mensual total de la operación de la cadena. Adicionalmente destacó los ahorros alcanzados en la compra y distribución de medicamentos desde el laboratorio a los centros de distribución.

Using information visualization to support the Self-Management of Type 2 Diabetes Mellitus - DM, (Wesson & Naude 2022). Esta investigación tuvo como propósito mejorar la presentación de las variables asociadas a la DM para apoyar a los pacientes con DM y ayudar en el proceso de identificación de patrones de salud. Esto podría, a su vez, conducir a una conciencia de causa y efecto para los pacientes con DM y permitirles autogestionar su DM de manera costo - efectiva para los servicios de salud a los cuales se encuentran afiliados. Para lo anterior utilizó un método para representar los datos en una forma gráfica interactiva no tradicional, denominado la visualización de información (IV). Los métodos de análisis visual e interactivos podrían generar cambios profundos en los programas de salud personal, la prestación de atención médica y la autogestión de los pacientes con DM a través de una presentación eficaz de la información de salud.

Adherencia terapéutica y variables relacionadas en adultos con diabetes mellitus tipo 2 en un hospital público, (Raraz Vidal J. & Raraz Vidal O., 2022). Este estudio tuvo como objetivo determinar la relación entre la adherencia terapéutica y variables relacionadas en adultos con Diabetes Mellitus tipo 2 en el Hospital Sergio E. Bernales, fue un estudio correlativo, prospectivo entre sus hallazgos encontró en los pacientes con Diabetes Mellitus una edad promedio de 58,78 años, siendo la obesidad (44 %) y la hipertensión arterial (47,7 %) sus

enfermedades más frecuentes e identificó entre las variables relacionadas con la adherencia al tratamiento, la mala relación médico paciente, un mal entorno ambiental y al menos una comorbilidad. En síntesis mostraron, para el caso de pacientes con diagnóstico de diabetes, entre los factores más relevantes a considerar están: la edad del paciente, la duración de la diabetes, el nivel de control glucémico, las comorbilidades, la presencia de complicaciones y la adherencia al tratamiento, entre otros.

Costos directos de atención médica en pacientes con diabetes mellitus tipo 2 en México: análisis de microcosteo, (Rodríguez Bolaños RA et al, 2010). El objetivo de este estudio fue calcular los costos directos del tratamiento de pacientes con diabetes mellitus tipo 2 (DM2) en el Instituto Mexicano del Seguro Social (IMSS). Esto se hizo utilizando la metodología de costeo de la enfermedad (CDE) basada en la prevalencia, que evalúa los costos actuales asociados con la enfermedad. Uno de los hallazgos fue el costo anual total del tratamiento de pacientes con DM2 del Instituto Mexicano del Seguro Social (IMSS). correspondiente al 31% de los costos de operación. Los pacientes con complicaciones incurren en costos promedio más altos que los pacientes sin complicaciones. Los servicios con mayor costo fueron los relacionados con las estancias hospitalarias y las estancias en unidades de cuidados intensivos.

Costos en Diabetes tipo II, (Bucca j., Damiani L., Esterkin G., García Dieguez M., Marcos E., Rodriguez M. 1994) El propósito de este estudio fue describir y estimar los gastos asociados con la diabetes tipo II. Los llamados costos directos tienen en cuenta el control, diagnóstico y tratamiento de la enfermedad. Las pruebas de laboratorio como la glucemia en ayunas, la hemoglobina glicosilada, el perfil de lípidos que incluye colesterol total, triglicéridos, colesterol/HDL, creatininemia, orina completa, proteinuria, electrocardiograma y consulta con cardiología se encuentran entre los indicadores cruciales que se establecen durante el diagnóstico y el seguimiento para evitar que un paciente desarrolle complicaciones que resulten en un aumento de los costos. Otro costo directo es el tratamiento, el cual tiene un costo variable dependiendo del medicamento seleccionado, sin embargo produce un gasto muy inferior al de la insulina. Los precios de los medicamentos varían de un país a otro, pero suelen ser hasta un 50 % más bajos en los países en desarrollo que en los países desarrollados.

Predictive Modeling of Total Healthcare Costs Using Pharmacy Claims Data: A Comparison of Alternative Econometric Cost Modeling Techniques, (Powers C. A., Meyer C. M., Roebuck M. C. and Vaziri B., 2005), Este trabajo realizó la evaluación a varios enfoques de modelos estadísticos para predecir los costos de salud anuales totales prospectivos (médicos más farmacia) de los participantes de un plan de salud que utilizaba Pharmacy Health Dimensions (PHD). Los modelos examinados fueron regresión de mínimos cuadrados ordinarios (OLS), regresión de OLS transformada logarítmicamente con estimador de smearing, y 3 modelos de dos partes usando regresión OLS, regresión log-OLS con estimador de smearing, y modelado lineal generalizado (GLM), respectivamente. para la validación del

modelo, tomaron una muestra aleatoria del 10%, evaluando mediante el r^2 ajustado, error de predicción absoluto medio, especificidad y valor predictivo positivo. Entre los hallazgos más relevantes se encontró que la mayoría de las categorías de medicamentos PHD fueron predictores independientes significativos de los costos totales. Entre los modelos probados, el modelo OLS tuvo el error de predicción absoluto medio más bajo y el r^2 ajustado más alto.

Forecasting the amount and cost of medicine to treat type 2 diabetes mellitus in Nepal using knowledge on medicine usage from a developed country, (Khanal S., Veerman L., Nissen L. and Hollingworth S., 2018). Este trabajo tuvo como objetivo pronosticar la cantidad y el costo de los medicamentos necesarios para el tratamiento de personas con diabetes mellitus tipo 2 (DM2) en Nepal durante 30 años. Para pronosticar el costo de medicamentos, consideraron el cambio en la población con DM2 en Nepal y el cambio en los costos de medicamentos para la DM2 en el tiempo. El modelo desarrollado contemplaba tres componentes: 1. número de pacientes con DM2 (todos los casos de DM2 en Nepal), 2. cantidad de medicamentos para la DM2 utilizados en el tratamiento (perfil de uso en Australia) y 3. el precio individual de los medicamentos para la DM2 (costos en Nepal). Realizaron análisis de sensibilidad de bidireccional (uno y dos vías). Modelando la carga financiera de los medicamentos para la DM2 estimando el costo de los medicamentos para tratar todos los casos en función de la prevalencia de la DM2 en Nepal durante tres décadas. En los resultados más destacados se encontró que la tendencia actual de la prevalencia de la DM2, le costaría entre 63 y 95 millones de dólares estadounidenses en 2013 comprar medicamentos para pacientes con DM2 en Nepal, usando la misma combinación de medicamentos que se utiliza en Austria para pacientes con DM2. Este costo representa una cuarta parte del presupuesto total de salud de Nepal (US$308 millones).

2.2.2. Matriz resumen de trabajos

TABLA 2. RESUMEN DE LOS CRITERIOS DE COMPARACIÓN ENTRE LOS ARTÍCULOS SELECCIONADOS Y EL PROYECTO DE GRADO.

Dimensión de Caracterización	Fecha Publicación	Aporte	Variables	Tipos de Modelos
F. A.Chuchoque-Urbina, M. P. Caro-Gutierrez, and C. E. Montoya, Design of a CPFR, location, inventory and routing approach to diabetes and high blood pressure medicines supply network planning	2021	Metodológico	- Niveles de interacción entre la aseguradora de salud, el prestador de servicios de salud, los laboratorios farmacéuticos proveedores y los pacientes. - Las previsiones de demanda - Ubicación de los centros de distribución - Estrategias de distribución de medicamentos orientadas a la minimización de costos a lo largo de la cadena.	CPFR (collaborative planning forecasting and replenishment)
Jarvis Raraz-Vidal1 , Omar Raraz-Vidal. Adherencia terapéutica y variables relacionadas en adultos con diabetes mellitus tipo 2 en un hospital público	2022	Teórico	- La mala relación médico paciente - Un mal entorno ambiental - Al menos una comorbilidad.	- Coeficiente de correlación de Spearman.
Janet Wesson1andMeggan Naude. Using information visualization to support the self-management of type 2 Diabetes Mellitus.	2022	Metodológico	- Consumo de agua, - El gasto de kilojulios - La medicación - El peso	- Visualización de información (IV). Método para representar datos en una forma gráfica interactiva no tradicional
Christopher A. Powers, Christina M. Meyer, M. Christopher Roebuck and Baze Vaziri. Predictive Modeling of Total Healthcare Costs Using Pharmacy Claims Data: A Comparison of Alternative Econometric Cost Modeling Techniques	2005	Teórico	- Médicos más farmacia de los participantes del plan de salud utilizando Pharmacy Health Dimensions (PHD) - Índice de riesgo basado en reclamos de farmacia.	- Regresión de mínimos cuadrados ordinarios (OLS) - Regresión de OLS transformada logarítmicamente con estimador de smearing - Modelado lineal generalizado (GLM)

Dimensión de Caracterización	Fecha Publicación	Aporte	Variables	Tipos de Modelos
Khanal S., Veerman L., Nissen L. and Hollingworth S., (2019). Forecasting the amount and cost of medicine to treat type 2 diabetes mellitus in Nepal using knowledge on medicine usage from a developed country	2019	Metodológico	- Número de pacientes con DM2 - Cantidad de medicamentos utilizados en el tratamiento - Precio individual de los medicamentos para la	- Análisis de sensibilidad bidireccional - Prevalencia

Fuente: Elaboración propia

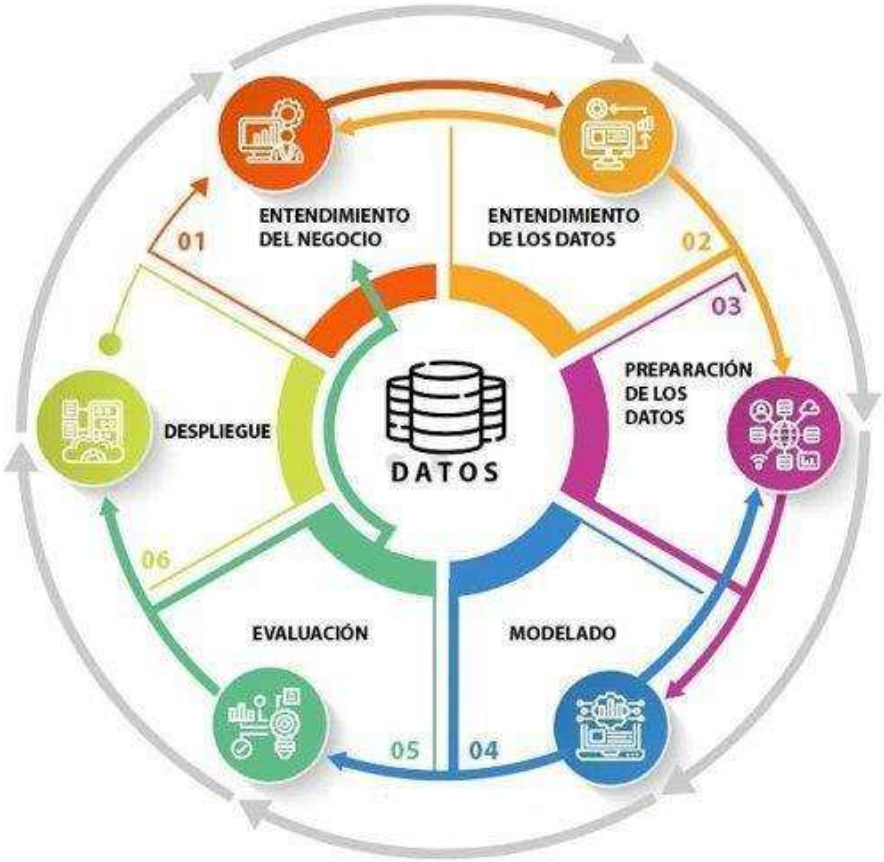
3. METODOLOGÍA

3.1. Conceptual

Este proy MODELOS DE APRENDIZAJE AUTOMÁTICO PARA LA OPTIMIZACIÓN DE LA GESTIÓN FARMACÉUTICAecto utilizará la metodología CRISP-DM (Cross-Industry Standard Process for Data Mining). La cual fue desarrollada por el consorcio constituido entre las empresas NCR Systems Engineering Copenhagen (USA and Denmark), DaimlerChrysler AG (Germany), SPSS Inc. (USA) y OHRA Verzekeringen en Bank Groep B.V (The Netherlands), y publicada en el año 2000 (Chapman et al., 2000). La metodología CRISP-DM se ha estandarizando sobre todo en el sector de la industria para proyectos de data mining (Wirth, 2000) y continúa siendo ampliamente utilizado en proyectos analíticos debido a su versatilidad.

CRISP-DM se basa en un modelo de procesos que maneja una estructura jerárquica. El primer nivel define las fases estándar del modelo. En el segundo nivel se encuentran las tareas genéricas, llamadas así porque buscan cubrir todas las situaciones que presente cualquier proyecto de data mining. El tercer nivel lo conforman las tareas específicas, que pueden ser propias de cada problemática de proyecto y el cuarto nivel corresponde a las instancias de proceso, donde se encuentran las actividades, acciones, decisiones y resultados específicos de cada tema y punto tratado dentro de los niveles superiores (Chapman et al., 2000).

FIGURA 1. METODOLOGÍA CRISP-DM



Fuente: https://www.iic.uam.es/innovacion/metodologia-crisp-dm-ciencia-de-datos/

3.1.1. Fases de la Metodología CRISP

Las fases vienen dadas en un orden secuencial dentro del proceso genérico, sin embargo, su orden y ejecución pueden variar dependiendo de las características propias del proyecto y de sus actividades. (Wirth, 2000)

Entendimiento del Negocio: Fase inicial del ciclo de vida de un proyecto de analítica. En esta fase se busca conocer los requerimientos y objetivos desde el punto de vista del negocio, con el fin de realizar la formulación del problema a resolver y desarrollar un plan preliminar para alcanzar los objetivos propuestos.

Entendimiento de los Datos: Esta fase busca la familiarización con los datos que van a ser objeto de análisis, por medio de una adquisición inicial de datos, la revisión de calidad y posibles hallazgos que permitan iniciar con una declaración de hipótesis que puedan ser validadas posteriormente.

Preparación de los datos: Aquí se busca trabajar con los datos crudos hasta convertirlos en datos que puedan ser leídos por los modelos. Incluye todas las actividades de selección, limpieza y transformación de datos.

Modelado: Durante esta etapa se seleccionan y ejecutan diferentes modelos sobre los datos preparados. Adicionalmente, los parámetros de cada modelo son calibrados con el fin de obtener los mejores resultados. Es probable que, dependiendo de los modelos, los datos deban tener una preparación específica, es por eso que hay una relación hacia la etapa previa de preparación de datos.

Evaluación: En esta fase, se busca evaluar los modelos con el fin de validar si cumplen con los objetivos del negocio. Se revisa tanto el modelo como los pasos que llevaron a su construcción, con el fin de detectar cualquier tema faltante necesario para el negocio. Es la etapa previa al despliegue del modelo y en ella se decide cuál es el paso siguiente de acuerdo con los resultados obtenidos.

Despliegue: Como etapa final, se busca poder compartir y difundir los resultados del proyecto, para que puedan ser usados por los usuarios del negocio. El despliegue de un proyecto de analítica/data mining depende de los resultados y del objetivo inicial del mismo y abarca desde la presentación de un informe detallado, hasta la implementación de un proceso de analítica inmerso en el proceso de toma de decisiones de una organización. En nuestro caso, el alcance propuesto comprende la producción de un reporte final, donde se encuentren los resultados obtenidos y las experiencias alcanzadas.

4. DESARROLLO

La presente metodología se desarrolló en el contexto de la necesidad de mejorar la eficiencia y la efectividad de la gestión farmacéutica de una EPS. La empresa enfrenta una serie de desafíos en esta gestión, incluyendo diferentes ineficiencias administrativas y médicas en la dispensación de medicamentos, generando sobrecostos. El propósito de esta metodología es proporcionar una herramienta práctica que permita identificar y priorizar acciones posibles de mejora, dentro del proceso de mejora continua establecido por la entidad. El alcance de la metodología se basa en la combinación de técnicas de la ciencia de datos para la toma de decisiones. A continuación se describe el proceso de aplicación de la metodología CRISP, se detallan las herramientas y técnicas utilizadas.

4.1. Comprensión del negocio

4.1.1. Tipo de estudio

El presente estudio es de naturaleza no experimental y cuenta con un diseño cohorte retrospectivo que permite analizar los cambios en la variable de estudio y su evolución en el tiempo. La información utilizada en el estudio es retrospectiva, ya que tanto la variable de estudio como las variables de caracterización han ocurrido antes del inicio del estudio.

Se selecciona los individuos de la cohorte cardiovascular del mes febrero del año 2023, se observa el comportamiento de la prescripción médica desde el mes de diciembre del 2015 para 46 medicamentos prescritos. Adicionalmente se realiza la caracterización de los individuos de la cohorte de estudio en un momento histórico específico. El período de estudio está comprendido entre diciembre 2015 a marzo del año 2023, la caracterización incluye la última actualización de los resultados de exámenes de laboratorio, que indican adherencia al tratamiento.

4.1.2. Población de estudio

La población de estudio corresponde a los afiliados de una EPS Colombiana que pertenecen al programa cardiovascular con diagnóstico de Diabetes Mellitus identificados E10 - E14 del CIE-10 de 2015 al mes de marzo del año 2023.

4.1.3. Criterios de inclusión y exclusión

Se incluyeron pacientes hombres o mujeres, mayores de edad (18 años) afiliados a la EPS y que reciben sus servicios desde una IPS en particular, durante el período 01 diciembre de 2015 – 31 de marzo de 2023, pertenecientes al programa cardiovascular con diagnóstico de diabetes mellitus en el mes de febrero de 2023. También se seleccionaron las prescripciones de medicamentos en el ámbito ambulatorio, cuyas clasificaciones dentro del sistema internacional ATC - Anatomical Therapeutic Chemical Classification System (Sistema de Clasificación

Anatómica, Terapéutica, Química) de la WHO Collaborating Centre for Drug Statistics Methodology ([WHOCC], 2023), correspondiera aquellos indicados para los individuos de la cohorte cardiovascular con diagnóstico de diabetes mellitus, los códigos ATC fueron los siguientes:

A10A: Insulinas y análogos de la insulina

A10B: Agentes hipoglucemiantes orales, excluyendo insulinas

A10BA: Sulfonilureas

A10BB: Biguanidas

A10BD: Tiazolidindionas

A10BF: Inhibidores de alfa-glucosidasa

A10BG: Inhibidores de la dipeptidil peptidasa IV (DPP-4)

A10BH: Agonistas del receptor de GLP-1

A10BJ: Inhibidores de SGLT2 (transportador de sodio-glucosa tipo 2)

A10BX: Otros agentes hipoglucemiantes, excluyendo insulinas

Se excluyeron los medicamentos prescritos por otros diagnósticos, además de los medicamentos que no tuvieran asociado un código de medicamento.

4.1.4. Variables de estudio

Como variable dependiente o variable de interés se encuentran las unidades prescritas por medicamento de forma mensual. El costo de la prescripción de medicamentos se definió como la cantidad del medicamento por el precio de venta del mismo, el precio de venta que se utilizó es el precio de venta modal o más frecuente dispensado.

Los principales factores determinantes del costo de medicamentos fueron seleccionados como resultado de la revisión de estudios anteriores, se tuvo en

cuenta las características demográficas y condiciones médicas. Las cuales se describen en el Anexo 1. Operacionalización de las Variables.

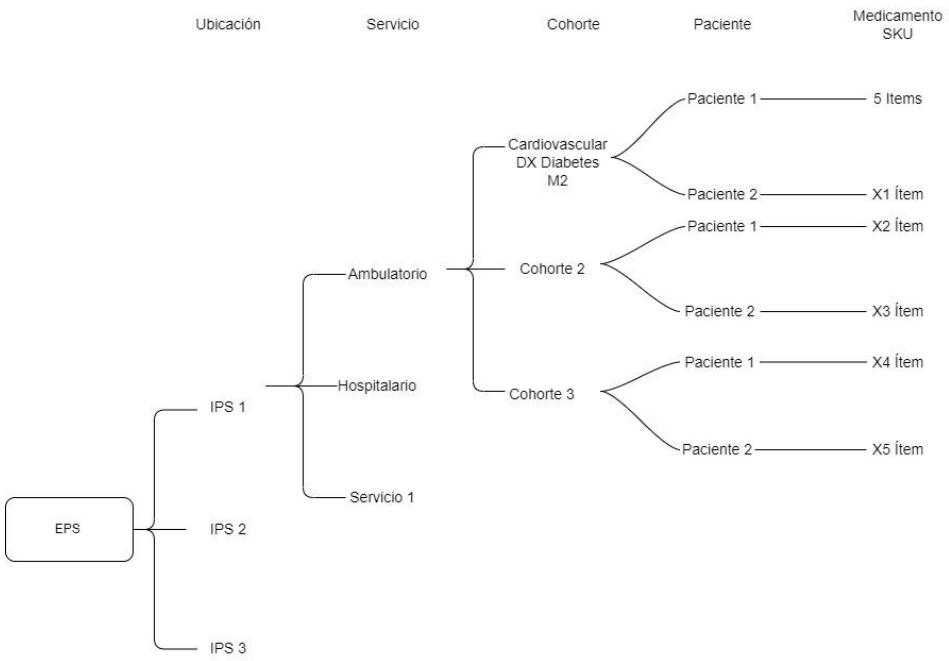
4.2. Comprensión de los datos

Este apartado tiene como propósito identificar las fuentes de información para la recolección de los datos, el manejo y control de la calidad de datos, la preparación de los datos para su entendimiento y exploración.

4.2.1. Estructura de los datos

Se tiene un histórico de datos mensuales durante 7 años, desde diciembre del año 2015 al mes de marzo del año 2023. La estructura de los datos nos enfrenta a un problema jerárquico, en este contexto significa que los datos se agrupan en diferentes niveles: por medicamento, por paciente, por cohorte, por servicio y ubicación. En la figura 2 se presenta el esquema del problema jerárquico de los datos.

FIGURA 2. ESTRUCTURA JERÁRQUICA DE LOS DATOS



Fuente: Elaboración propia

4.2.2. Recolección de datos

La información fue obtenida a través de las siguientes fuentes secundarias:

- Prescripción de medicamentos.
- Cohorte Cardiovascular.
- Precios de medicamentos.
- Datos maestros de medicamentos

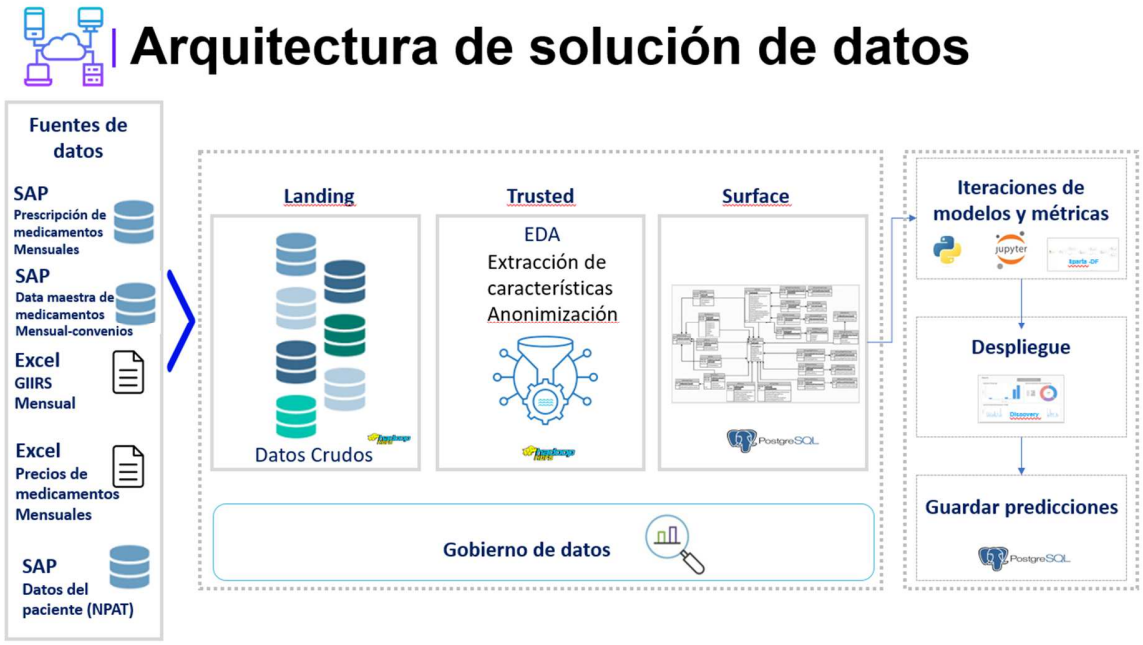
4.2.3. Manejo y control de la calidad de datos

La empresa donde se obtuvieron los datos cuenta con procesos existentes de gestión de datos, lo cual nos garantizan inicialmente que existen procesos de recolección, almacenamiento, limpieza y gestión de los mismos. Adicionalmente mediante el análisis exploratorio de datos - EDA se realizó la exploración para el control de calidad de los datos

4.2.4. Creación modelo de datos

A continuación se esquematiza la arquitectura de la solución analítica:

FIGURA 2B. ARQUITECTURA DE SOLUCIÓN DE DATOS



Fuente: Elaboración propia

Esta arquitectura de la solución analítica se desarrolló en la plataforma Augmented Data Fabric de la compañía Stratio alojada en la nube de Azure.

El diagrama anterior describe las principales fases de manera consecutiva: La primera fase hace referencia al proceso de ingesta, extraer a través del conector nativo JDBC, los datos de las bases de datos Oracle del aplicativo SAP y alojando los datos crudos en la zona landing, los archivos planos se ingestan para continuar a una segunda fase, donde pasan a la zona trusted en la cual sólo se extraen algunas características de los datos crudos, se realiza el protocolo EDA, proceso de anonimización, por último, se construye la base de datos con la estructura necesaria para empezar analizar la variable de estudio y los posibles factores determinantes de la misma. A partir de esta base y poder de procesamiento de la nube de Azure, se realiza el proceso de formato de características, modelación, entrenamiento, evaluación y selección de modelo y validación, se despliega a través de discovery los resultados y se guardan los pronósticos en la tabla.

4.2.5. Métodos de transformación.

Existen varios métodos de transformación de variables cuantitativas utilizadas en el preprocesamiento de datos, entre los cuales se encuentra el método MinMaxScaler. Este método escala los datos en un rango específico, comúnmente entre 0 y 1, para normalizar los datos y lograr que estén en la misma escala. La transformación de los datos se realiza restando el valor mínimo de los datos de cada valor en el conjunto de datos y luego dividiendo cada valor por la diferencia entre el valor máximo y el valor mínimo. Este método es útil para la preparación de datos, especialmente en modelos de aprendizaje automático, debido a que puede mejorar el rendimiento del modelo al asegurar que todos los valores estén en la misma escala, como por ejemplo las redes neuronales y los algoritmos de agrupamiento. (scikit-learn.org [scikit-learn.org], 2023).

4.3. Preparación de los datos

Este apartado consta de dos partes: la primera tiene como propósito conocer la cohorte de pacientes, las variables asociadas a su estado de salud y realizar una aproximación a posibles asociaciones con los costos de medicamentos; y la segunda parte se centra en la identificación de los insumos relevantes para la etapa de modelación.

La EPS suministró la data original, la cual consta de 19362 registros y 61 campos de los cuales se seleccionaron 20 variables.

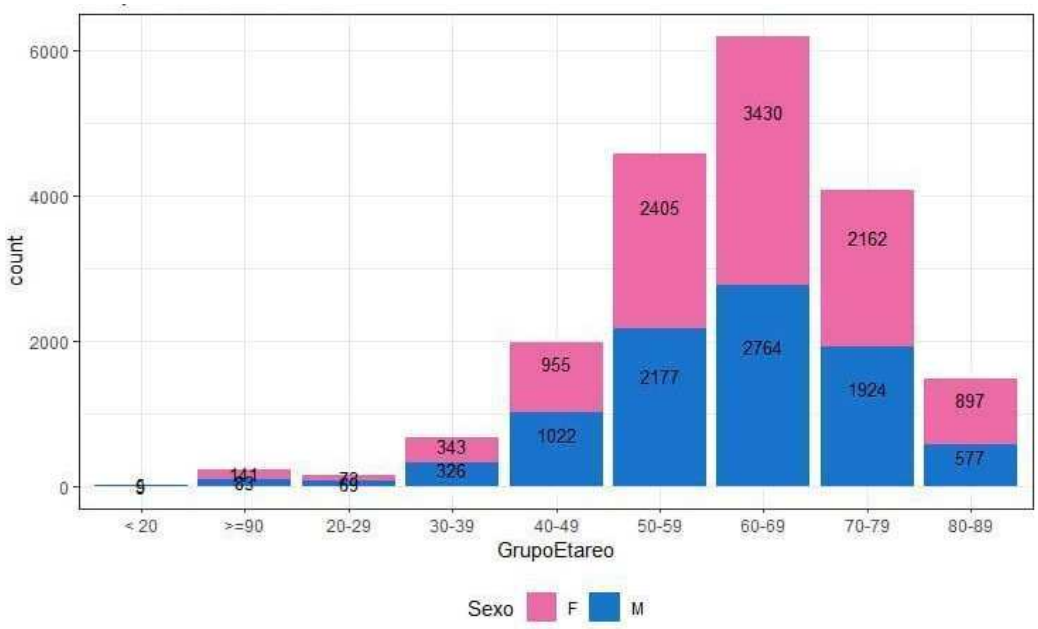
4.3.1. Análisis exploratorio de los datos – EDA

La Organización Panamericana de la Salud - OPS, en su informe ***Panorama de Diabetes en las Américas*** menciona que: "La diabetes es una enfermedad metabólica de etiología multifactorial en la que interactúan elementos genéticos, sociodemográficos y ambientales, junto con factores de riesgo como la obesidad, la inactividad física y las dietas poco saludables…" (PAHO, 2022).
Por lo anterior, se realizó el análisis exploratorio de las variables disponibles, describiéndolas por medio de gráficas de frecuencia o de dispersión, en dos frentes de los mencionados por la OPS, elementos sociodemográficos y factores de riesgo. Adicionalmente, se utilizan gráficos de tendencia para observar el comportamiento de las variables costos, medicamentos y dosis de manera general.

Análisis sociodemográfico

En cuanto a las variables sociodemográficas del paciente se encontró que en la cohorte el 54,25% son mujeres y 45,75 hombres. El 86,36% de los pacientes superan los 50 años. ver Figura 3.

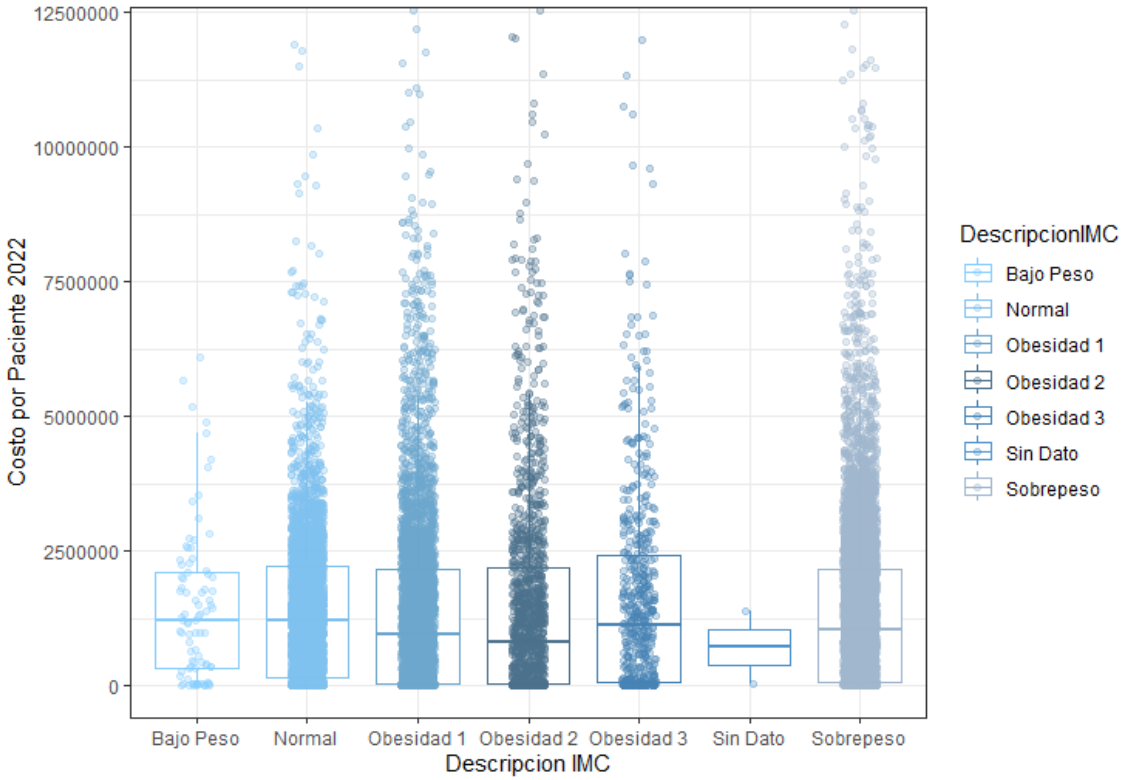
FIGURA 3. DISTRIBUCIÓN DE GRUPOS ETAREO POR SEXO



Análisis de las características de salud.

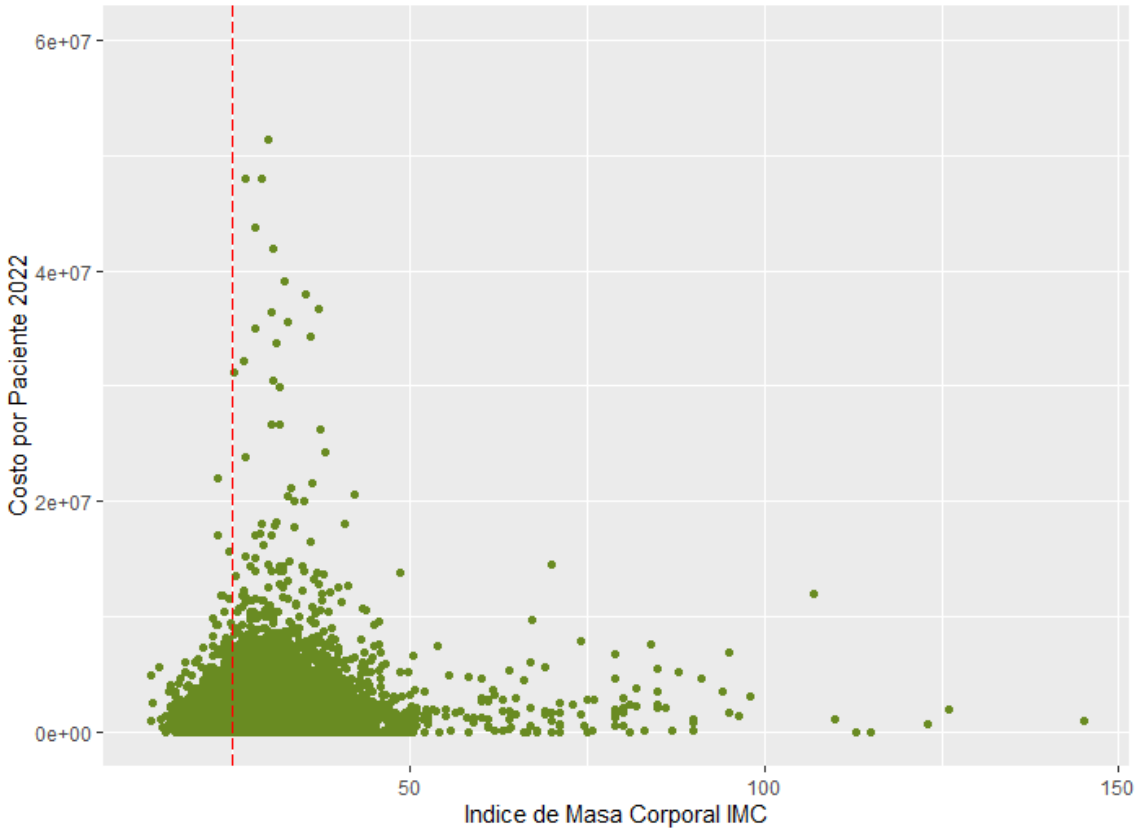
Con respecto a las características de salud de los pacientes de la cohorte, se observa una medida general del estado de salud Figura 4, el índice de masa corporal - IMC, el cual indica cuando un individuo tiene un exceso de grasa corporal, lo cual afecta su estado de salud y desencadena una serie de enfermedades asociadas al sobrepeso e impacta en el incremento de los costos médicos, como los medicamentos (PAHO, 2022). El 77,74% de la cohorte de pacientes tienen un IMC que se ubica en los rangos de sobrepeso o en alguno de los tres grados de obesidad. La distribución de los costos en los diagramas de cajas por categoría deja ver que para el sobrepeso y las categorías de obesidad 1,2, y 3 tienen mayor dispersión en el 25% de los costos más altos.

FIGURA 4. DISTRIBUCIÓN DE LOS COSTOS DE MEDICAMENTOS EN EL 2022 POR CATEGORÍA DE PESO



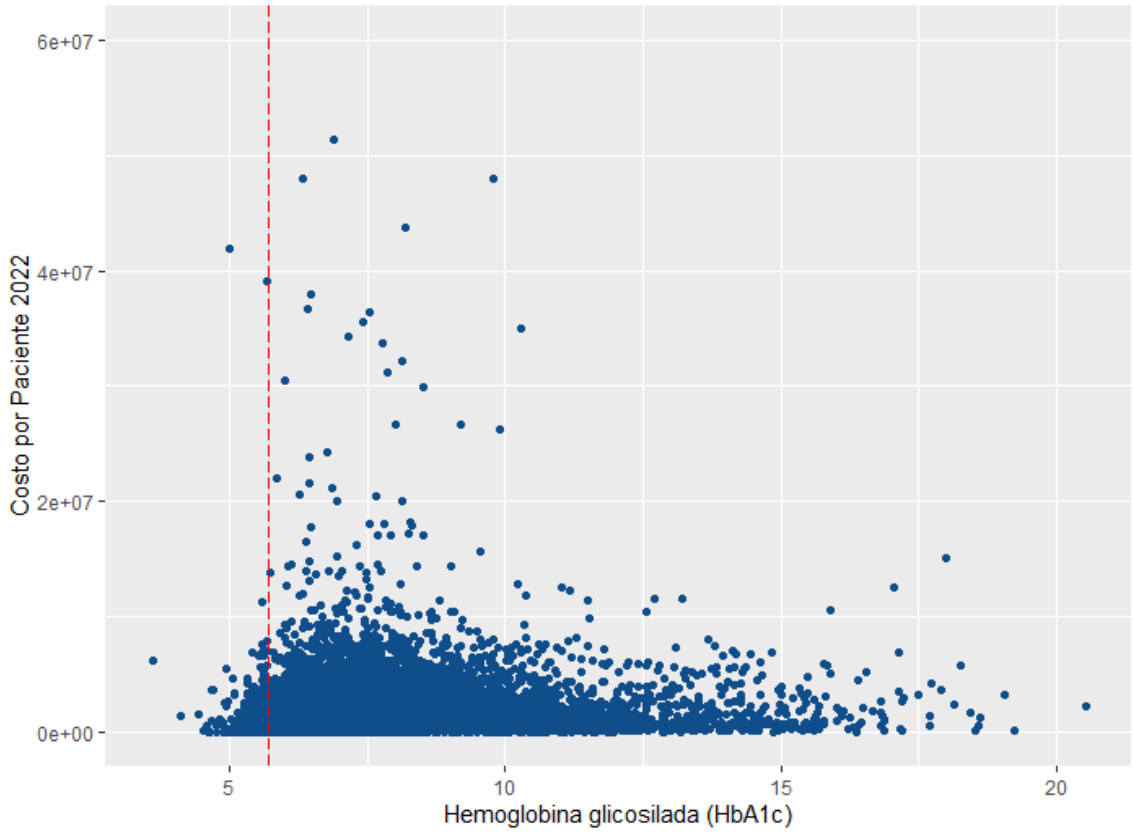
Según la OPS, las Américas tienen las tasas más altas de sobrepeso, obesidad e inactividad en el mundo, todos los cuales son factores de riesgo para la diabetes, es previsible el aumento de la tendencia en la prevalencia de la diabetes en los próximos años. Por lo tanto, es importante investigar la relación entre el IMC y los costos relacionados en 2022. Dado que por definición el valor del IMC a partir del cual se considera que una persona se encuentra en sobrepeso es de 24,9, se eligió como el punto crítico para analizar. A partir de este punto, la Figura 5 muestra el aumento de los costos de algunos pacientes en 2022.

FIGURA 5. DISTRIBUCIÓN DE LOS COSTOS DE MEDICAMENTOS EN EL 2022 Y EL ÍNDICE DE MASA CORPORAL IMC



Otro indicador relevante para describir el estado de salud de pacientes de la cohorte cardiovascular con diagnóstico de diabetes, es la medición de la hemoglobina glicosilada (HbA1c), la cual permite realizar el monitoreo y manejo del control de la glucemia en pacientes con diabetes mellitus, es un indicador que si se conserva en un nivel óptimo (HbA1c < 5,7%), tiene como propósito evitar que los personas con diagnóstico de diabetes lleguen a complicaciones a largo plazo, entre ellas neuropatía, retinopatía y nefropatía, las cuales son enfermedades que aumentan la carga asociada a la enfermedad y agudizan drásticamente la calidad de vida de las personas con diabetes (PAHO, 2022). De acuerdo a la dispersión de los costos versus el indicador de la Hemoglobina glicosilada (HbA1c), se observa que a partir del 5.7% del HbA1c que los costos de medicamentos para varios pacientes de la cohorte aumentan (ver Figura 6).

FIGURA 6. DISTRIBUCIÓN DE LOS COSTOS DE MEDICAMENTOS EN EL 2022 Y LA HEMOGLOBINA GLICOSILADA (HBA1C)



Análisis de la tendencia

Una de las premisas que se plantearon fue el desconocimiento de la cantidad de dinero que gasta el sector de la salud para manejar los problemas relacionados con la diabetes y las ineficiencias administrativas. A continuación en las Figuras 7 y 8 se muestra la evolución de los costos y las dosis de los pacientes de la cohorte, respectivamente. Se evidencia un aumento importante en ambos a partir del año 2021 y poca variación en los costos entre el 2016 al 2020. En el caso de las dosis se observa una disminución durante el año 2019.

FIGURA 7. EVOLUCIÓN DE LOS COSTOS

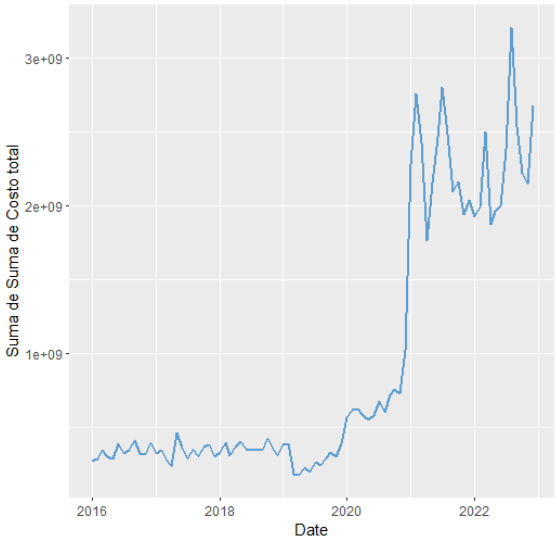
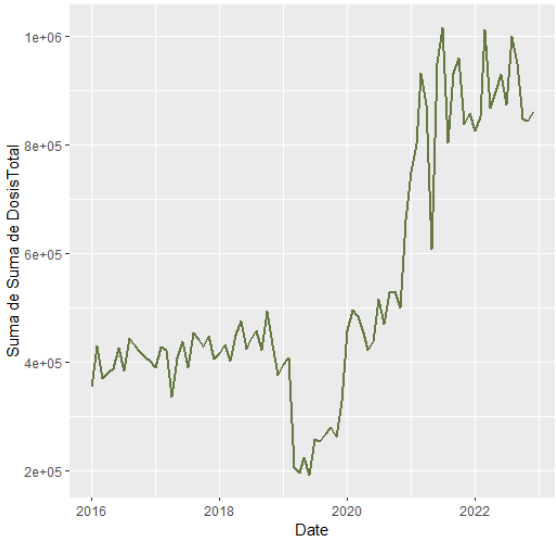
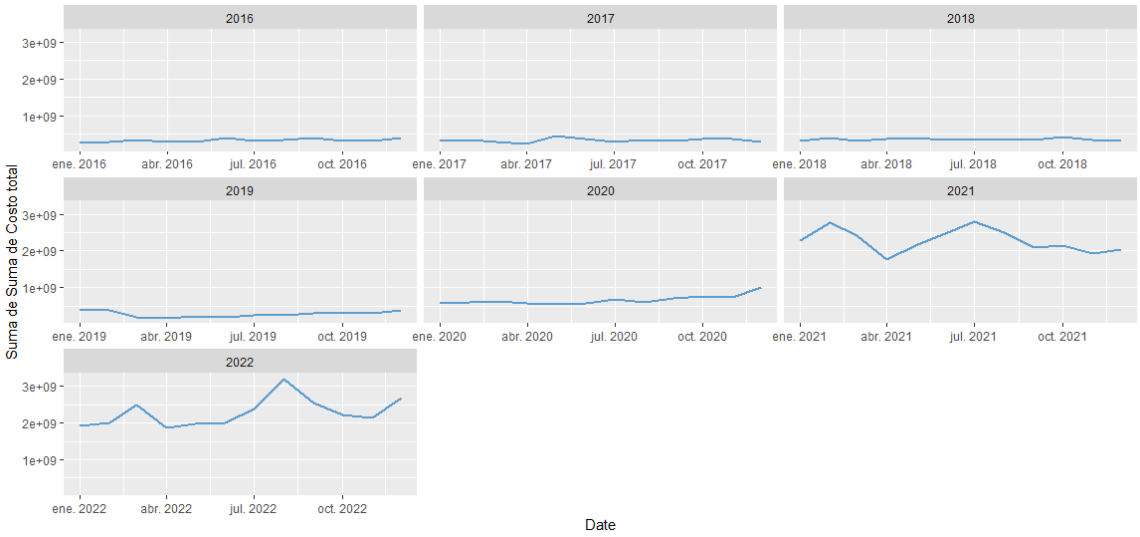


FIGURA 8. EVOLUCIÓN DE DOSIS



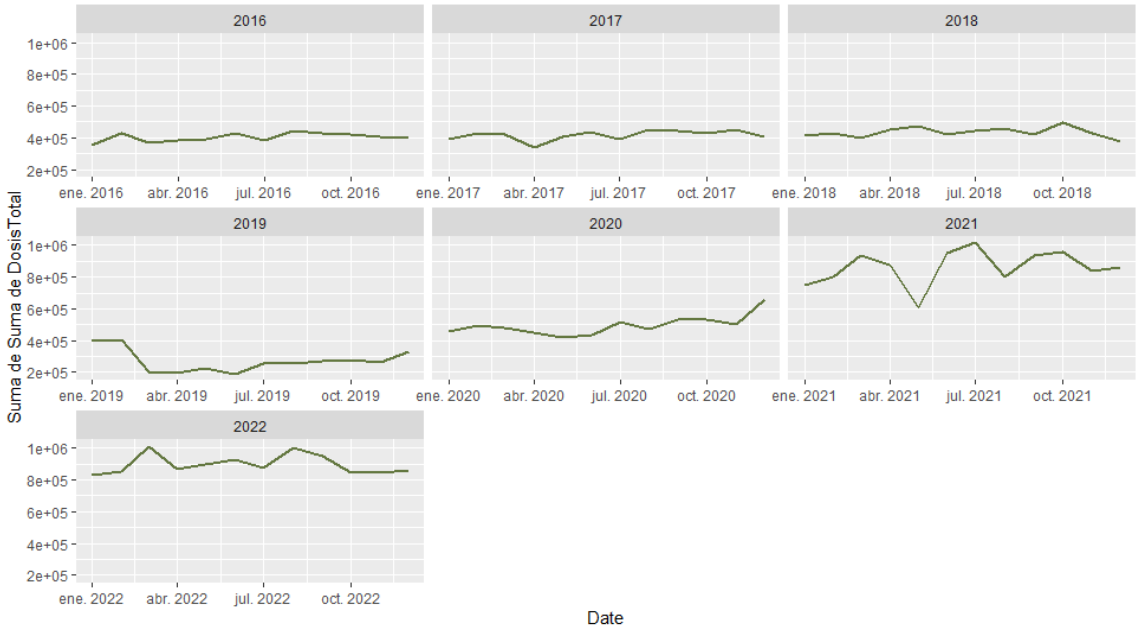
Para observar mejor la realidad de esa variación y aumento se revisó la evolución a detalle por año. La Figura 9 muestra claramente que si hay un aumento en los costos de medicamentos de los pacientes de la cohorte para el año 2021 y se mantiene en el 2022.

FIGURA 9. EVOLUCIÓN DE LOS COSTOS DESAGREGADOS POR AÑO



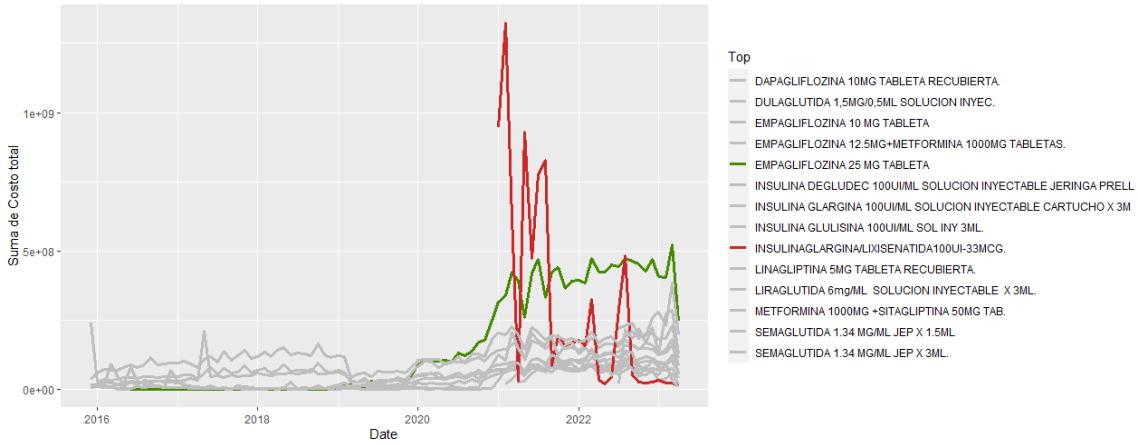
Del mismo modo, se revisaron las dosis, confirmándose, como se muestra en la Figura 10, el aumento de las dosis de los pacientes de la cohorte en 2021 y 2022. También es posible ver la disminución en 2019, luego de pasar de una dosis muy similar en la prescripción y sin variación significativa en los años 2016 a 2018.

FIGURA 10. EVOLUCIÓN DE LAS DOSIS DESAGREGADAS POR AÑO



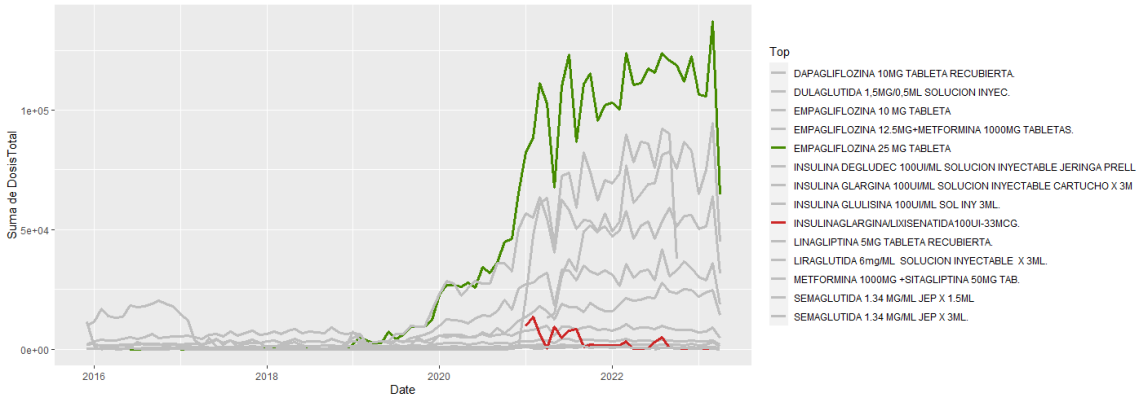
Los medicamentos prescritos con mayor frecuencia se incluyeron en un análisis descriptivo adicional que se consideró importante revisar. Los medicamentos más caros utilizados para tratar a los pacientes de la cohorte de la EPS se muestran en la Figura 11 aquí. Podemos ver que *Insulinaglargina/Lexisenatida 100UI-33MCG*, que es el fármaco más caro, se recetará a estos pacientes a partir de 2021. El comprimido de *Empagliflozina de 25 mg* es el siguiente medicamento más caro, pero se ha utilizado con frecuencia en el tratamiento desde hace varios años.

FIGURA 11. EVOLUCIÓN DE LOS COSTOS DE LOS MEDICAMENTO POR AÑO



El medicamento más costoso no es necesariamente el que tiene el mayor número de unidades prescritas. Como se mencionó anteriormente, la *Insulinaglargina/Lexisenatida 100UI-33MCG* se incluye en la lista de medicamentos para esta población de pacientes desde 2021. El segundo medicamento más costoso utilizado para estos pacientes, la tableta de *25 mg de Empagliflozina*, se prescribió con frecuencia entre 2016 y 2022, y el número de dosis prescritas comenzó a aumentar a partir de 2020.

FIGURA 12. EVOLUCIÓN DE LAS DOSIS DE LOS MEDICAMENTO POR AÑO



Transformaciones de la serie de dosis

El preprocesamiento de datos utiliza una variedad de técnicas de transformación de variables cuantitativas y con diferentes fines. Como se mencionó anteriormente. Se utilizó el método de transformación MinMaxScaler para normalizar la serie a la luz de los resultados del análisis de la tendencia de las dosis de medicamentos y a la expectativa de aplicar las técnicas de Machine Learning, este procedimiento permitió escalar los valores para estandarizar los rangos, mejorando la estabilidad numérica y su interpretación (Kuhn M. y Johnson K., 2013). El modelado se lleva a cabo utilizando varios algoritmos de Machine Learning sobre la serie transformada.

Análisis tradicional de la serie temporal

Como es conocido una serie temporal es una colección de puntos de datos en el tiempo. Las series temporales, aparte de depender del tiempo, muestran unas características específicas que deben cumplir, tendencia, estacionalidad y ruido blanco. Adicionalmente el enfoque tradicional de los modelos auto-regresivos, requieren de realizar el análisis de correlaciones temporales mediante los correlogramas, saber si cumlen ciertas propiedades estadísticas, analizando los test estadísticos y probando uno a uno todos los errores, para llegar a la construcción del mejor modelo matemático que describa la serie temporal, para posteriormente construir un modelo de predicción y probar su capacidad de predicción. Todo este

conjunto de tareas aquí descritas dificultan la automatización del modelo y sus resultados. Por lo anterior se aborda el problema de este proyecto de pronósticos de dosis de medicamentos para una EPS, mediante modelos de Machine Learning.

4.4. Modelado

Este apartado se enfoca en completar los objetivos del presente trabajo mediante el desarrollo y la evaluación de un modelo predictivo. A continuación se describe lo relacionado con la elección de las métricas de desempeño, la selección de algoritmos de Machine Learning, la definición de la estrategia de modelación, la validación para determinar el mejor modelo y la evaluación de los resultados.

4.4.1. Métricas seleccionadas

Para la evaluación del desempeño de los modelos candidatos, se utilizaron un grupo de métricas que se destacan en términos de interpretabilidad, complementariedad y robustez ante valores extremos, estas son: el error cuadrático medio (MSE), la raíz del error cuadrático Medio (RMSE), el error absoluto medio (MAE) y Error Porcentual Absoluto Medio (MAPE).

Una diferencia importante entre el MAE y el RMSE es que el primero se comporta de forma lineal, pues tratar el tamaño de los errores de forma indistinta, lo que lo hace relativamente intuitivo; mientras que el MSE y RMSE son más exigentes, porque castigan más los errores grandes que los pequeños, esto quiere decir, que los errores más grandes contribuirán más al MSE o al RMSE final. (Brownlee, J. 2021). Las cuatro métricas ayudaron a tener una idea de la calidad de las predicciones de los modelos.

4.4.2. Algoritmos seleccionados

De acuerdo al estado del arte, los antecedentes revisados en este trabajo y al conocimiento sobre la evolución de las capacidades predictivas que pueden ser comparadas mediante la utilización de algoritmos clásicos, algunos de ellos muy populares ya en la ciencia de datos. A continuación presentamos los algoritmos elegidos y la argumentación del porqué fueron elegidos.

- *LightGBM (Light Gradient Boosting Model).* Este algoritmo, relativamente nuevo, se ha destacado por su velocidad y eficacia de entrenamiento. Esto

resultaba conveniente para el enfoque iterativo previsto (Lipton, Z.C. at all, 2015).

- *LSTM (Long Short-Term Memory).* Al ser una variante de las redes neuronales conserva la misma capacidad de generalización, de capturar y recordar dependencias a largo plazo en los datos de series de tiempo, lo que lo convierte en un algoritmo apropiado para modelar patrones complejos (Lipton, Z.C. at all, 2015).

- *XGBoost (Extreme Gradient Boosting)*. Se tuvo en cuenta este algoritmo por su capacidad de capturar relaciones no lineales en los datos y cambios en las tendencias y patrones estacionales.

4.4.3. Proceso de modelación y validación

Para determinar objetivamente el mejor modelo, se definió una estrategia de modelación y validación que comprendía la división del conjunto de datos en tres particiones: entrenamiento, prueba y validación. Al tratarse de datos de serie de tiempo la técnica apropiada de validación es TimeSeriesSplit, al tener en cuenta la naturaleza secuencial y temporal de los datos. (Hyndman, R. J., & Athanasopoulos, G., 2018).

El proceso tiene los siguientes pasos:

a) Dividir el conjunto de datos en dos conjuntos uno de entrenamiento y otro de prueba en la proporcionalidad de 90% de registros para entrenamiento y 10% para prueba.
b) Utilizar el TimeSeriesSplit con una ventana de seis para entrenar el modelo con ventanas deslizantes en el tiempo de grupos de seis.
c) Utilizar el optimizador de hiperparámetros GridSearchCV para hallar el conjunto de hiperparámetros para entrenar el modelo con los datos de entrenamiento y ajustar el mejor modelo, teniendo en cuenta el paso anterior en el parámetro CV.
d) Realizar la optimización de hiperparámetros con el set de entrenamiento.
e) Ajustar el modelo con los mejores parámetros (best_params)

4.4.4. Selección de los modelos y validación

Se decidió que los modelos elegidos se obtendrían de la aplicación de cada uno de los tres algoritmos seleccionados sobre el conjunto de datos y para los siguientes 5 principios activos de los medicamentos (top 5) prescritos por la EPS:

- INSULINA GLARGINA 100UI/ML SOLUCION INYECTABLE CARTUCHO X 3M.
- INSULINA GLULISINA 100UI/ML SOL INY 3ML.
- INSULINA DEGLUDEC 100UI/ML SOLUCION INYECTABLE JERINGA PRELL.e
- INSULINA GLARGINA 100UI/ML VIAL 10ML.
- METFORMINA TAB 850 MG.

En consecuencia, se evaluaron en total 50 modelos candidatos. Para el procesamiento de estos modelos se utilizó Jupyter - Anaconda para luego ser incluida en la plataforma Datafabric.

En un futuro se deberá ser enriquecido el proceso de escritura del código con algunos recursos disponibles en dicha plataforma, incluyendo bibliotecas preinstaladas, lo cual simplificó el proceso de configuración del entorno de trabajo.

4.4.5. Comparación de los resultados

Al comparar los resultados obtenidos para cada uno de los modelos, se evidenció que el candidato de mejor desempeño es el modelo que corresponde al algoritmo *XGBoost* aplicado a los conjuntos de datos y diferentes medicamentos, según se evidencia y se identifica con los números 1, 4, 6, 8 y 10 en la Tabla 3. Se trata del mejor modelo después de haber realizado la optimización de los hiperparámetros, porque es la opción, donde las métricas evaluadas, que consideran que la precisión del modelo será la mejor, cuanto menor sea el valor de cada métrica, en el caso del MAPE la literatura comenta que esta evolución puede ser buena o aceptable, aceptable con un porcentaje entre 20 y 30 por ciento y buena por debajo del 20 por ciento.

TABLA 3. COMPARACIÓN DEL DESEMPEÑO DE LOS MODELOS CANDIDATOS.

División de la muestra	Medicamento	Modelo	Modelo	RMSE	MAE	MAPE	Modelo	Selección Modelo
80-20	**METFORMINA TAB 850 MG.**	1	**XGBoost**	**34.772**	**26.438**	**20,0**	1	X
80-20	METFORMINA TAB 850 MG.	2	LightGBM	78.861	73.431	44,2	2	
80-20	METFORMINA TAB 850 MG.	3	LSTM	73.294	65.948	43,2	3	
80-20	**INSULINA GLARGINA 100UI/ML VIAL 10ML.**	**4**	**XGBoost**	**11**	**9**	**37,9**	**4**	X
80-20	INSULINA GLARGINA 100UI/ML VIAL 10ML.	5	LSTM	21	18	63,3	5	
80-20	**INSULINA GLARGINA 100UI/ML SOLUCION INYECTAB**	**6**	**XGBoost**	**702**	**547**	**7,1**	**6**	X
80-20	INSULINA GLARGINA 100UI/ML SOLUCION INYECTA	7	LSTM	1.154	919	10,1	7	
80-20	**INSULINA GLULISINA 100UI/ML SOL INY 3ML.**	**8**	**XGBoost**	**240**	**176**	**5,7**	**8**	X
80-20	INSULINA GLULISINA 100UI/ML SOL INY 3ML.	9	LSTM	750	602	16,1	9	
80-20	**INSULINA DEGLUDEC 100UI/ML SOLUCION INYECTAB**	**10**	**XGBoost**	**321**	**233**	**11,7**	**10**	X
80-20	INSULINA DEGLUDEC 100UI/ML SOLUCION INYECTA	11	LSTM	454	342	16,0	11	
90-10	METFORMINA TAB 850 MG.	12	XGBoost	31.849	24.241	22,1	12	
90-10	METFORMINA TAB 850 MG.	13	LightGBM	73.294	65.948	43,2	13	
90-10	METFORMINA TAB 850 MG.	14	LSTM	59.861	54.173	35,9	14	
90-10	INSULINA GLARGINA 100UI/ML VIAL 10ML.	15	XGBoost	10	9	45,8	15	
90-10	INSULINA GLARGINA 100UI/ML VIAL 10ML.	16	LSTM	22	19	75,3	16	
90-10	INSULINA GLARGINA 100UI/ML SOLUCION INYECTA	17	XGBoost	692	576	8,4	17	
90-10	INSULINA GLARGINA 100UI/ML SOLUCION INYECTA	18	LSTM	897	750	9,5	18	
90-10	INSULINA GLULISINA 100UI/ML SOL INY 3ML.	19	XGBoost	244	162	6,4	19	
90-10	INSULINA GLULISINA 100UI/ML SOL INY 3ML.	20	LSTM	438	397	11,7	20	
90-10	INSULINA DEGLUDEC 100UI/ML SOLUCION INYECTA	21	XGBoost	345	268	13,6	21	
90-10	INSULINA DEGLUDEC 100UI/ML SOLUCION INYECTA	22	LSTM	337	267	12,7	22	

Fuente. Elaboración propia

4.5. Despliegue

A continuación se coloca el modelo en producción para agregar valor al negocio. En las siguientes imágenes se muestran los principales resultados de los modelos seleccionados para cada medicamento utilizando el algoritmo *XGBoost.* Estos se muestran en tres componentes: el primero describe la importancia o peso de las características que ingresan en el modelo para la predicción, el segundo la visualización de la serie real versus los pronósticos y finalmente los valores estimados e intervalos de confianza versus valores reales.

4.5.1. Uso del modelo con el mejor desempeño

En el Anexo 2 se observan las gráficas con los resultados, el comportamiento de la serie, el pronóstico de las dosis, el peso de las características ingresadas al modelo, que para el caso de todos los modelos, se observa que la de mayor peso, es la característica de los pacientes. Para el resto de modelos se muestran solamente los dos primeros resultados.

4.5.2. Visualización de los resultados

Para utilidad del cliente final y cumplir con los objetivos del negocio de mejorar los resultados en salud de los pacientes y la calidad de la atención médica, de garantizar el uso racional de los medicamentos en el tratamiento de las enfermedades, la seguridad y eficacia en la prescripción de los medicamentos utilizados y optimizar el uso de los recursos de la salud. Se elaboró un prototipo de Dashboard en la plataforma Power BI, herramienta utilizada por la compañía para la visualización de información y presentar los resultados en términos de las variables dosis, costos, principio activo del medicamento, valor del pronóstico de las dosis y su intervalo de confianza. Ver Anexo 3.

5. CONCLUSIONES

De acuerdo a lo mencionado por la literatura al priorizar los modelos de Machine Learning sobre los modelos de series de tiempo autorregresivos tradicionales, sacrificando la ventaja que tienen estos últimos, en la precisión de sus predicciones, sobre las ventajas de los modelos de aprendizaje automático para el pronóstico de dosis de medicamentos en la EPS, al obtener velocidad para mostrar resultados, la eficiencia al optimizar los recursos de procesamiento, y la utilidad de interacción con el usuario final.

A través del análisis descriptivo de la cohorte estudiada, se observó en las variables del estado de salud del paciente, que existen indicadores del estado de salud del paciente importantes y que son mencionados en la literatura como factores de riesgo estrechamente relacionados con la diabetes, como son la edad, sexo, sobrepeso, obesidad, medidas a través del índice de masa corporal y la hemoglobina glicosilada (HbA1c) que evidencia la presencia de azúcar en la sangra por largos periodos. Todas estas variables son potencialmente explicativas para representar el patrón de comportamiento de los cotos y las dosis y deben ser tenidas en cuenta para ser incluidas en el modelo de pronóstico.

Entre las fortalezas que se leen en la literatura sobre el algoritmo *XGBoost (Extreme Gradient Boosting)*, está la de permitir analizar el modelo de predicción incorporando variables explicativas, mediante árboles de decisión los cuales van seleccionando las mejores correlaciones e identificando cuales de esas variables explicativas aportan a la predicciones. Es posible continuar desarrollando el presente modelo y llevarlo a un mayor detalle de explicación del comportamiento de las dosis requeridas para los pacientes. Hasta el punto de pensar en las variables exógenas del fenómeno como lo es el desabastecimiento de medicamentos en el mercado.

En este trabajo se demostró que el aprendizaje automático es una herramienta eficaz para entregar información hacia el futuro que permita de forma automática anticiparse, identificar riesgos y mitigar la incertidumbre, aportando a la optimización de la gestión farmacéutica, mediante la predicción de la cantidad de dosis de medicamentos en pacientes con diabetes de la cohorte cardiovascular. Los resultados de la evaluación de los modelos indicaron que *XGBoost* tuvo un mejor desempeño, seguido del *LSTM* y por último *LightgBM*.

5.1. Trabajos futuros

Debido a los resultados obtenidos por los tres modelos, se puede pensar en realizar un desarrollo para establecer un stack de modelos de pronóstico a demanda de la necesidad del cliente y el negocio, evaluando el costo beneficio de hacer este nuevo desarrollo, teniendo en cuenta la infraestructura tecnológica necesaria y el cumplimiento de los objetivos de negocio.

Se entrega a la compañía este producto mínimo viable y se debe iterar e incrementar para alcanzar el objetivo de la generación de 2.4 millones de pronósticos mensuales con la siguiente granularidad: por medicamento (SKU), por ubicación de Instituciones Prestadoras de Salud IPS y dispensarios, para los servicios hospitalario, ambulatorio y laboratorio.

Esta herramienta seguimiento, control y predicción deberá facilitar la comprensión del negocio, generando insigth para la toma de decisiones, principalmente a los profesionales de la salud, a los profesionales que gestionan los medicamentos y al personal administrativo que les permita Mejorar la calidad de la atención médica, garantizar la seguridad y eficacia en la prescripción de los medicamentos utilizados y optimizar el uso de los recursos de la salud.

6. ANEXOS

Anexo 1. Operacionalización de las Variables.

Variable	Definición operativa	Tipo	Valores
Variable dependiente			
Costos	Cantidad del medicamento por el precio de venta del mismo entre el año 2016 al 2022	Cuantitativa Continua Razón	pesos ($)
Dosis	Cantidad de unidades prescritas por medicamento de forma mensual entre el año 2016 al 2022	Cuantitativa Continua Razón	Cuantitativa
Variables identificación del paciente			
NoPaciente	Identificación del paciente	numérico	7 dígitos
Reno.Edad	Edad al momento de obtener los datos para el estudio. Fecha hoy menos la fecha de nacimiento.	Cuantitativa Continua Razón	Entre 19 a 102 años
Sexo	Sexo del paciente	Cualitativa Nominal Dicotómica	M. Masculino F. Femenino
Características de salud del paciente			
FechaIngresoProgramaAtencionRenal	Fecha de ingreso al programa de atención renal	Temporal	DD/MM/AAAA
DiagnosticoConfirmadoHipertensionHrterialHTA	El usuario tiene diagnóstico confirmado de Hipertensión Arterial - HTA (CIE-10 con códigos entre I10-I15)	Cualitativa Nominal Dicotómica	1. SI 2. NO
FechaDiagnosticoHipertensionArterial	Fecha del diagnóstico de hipertensiòn arterial	Temporal	DD/MM/AAAA
DiagnosticoConfirmadoDiabetesMellitusDM	El usuario tiene diagnóstico confirmado de Diabetes Mellitus- DM (CIE-10 con códigos entre E10-E14)	Cualitativa Nominal Dicotómica	1. Confirmado 2. No confirmado

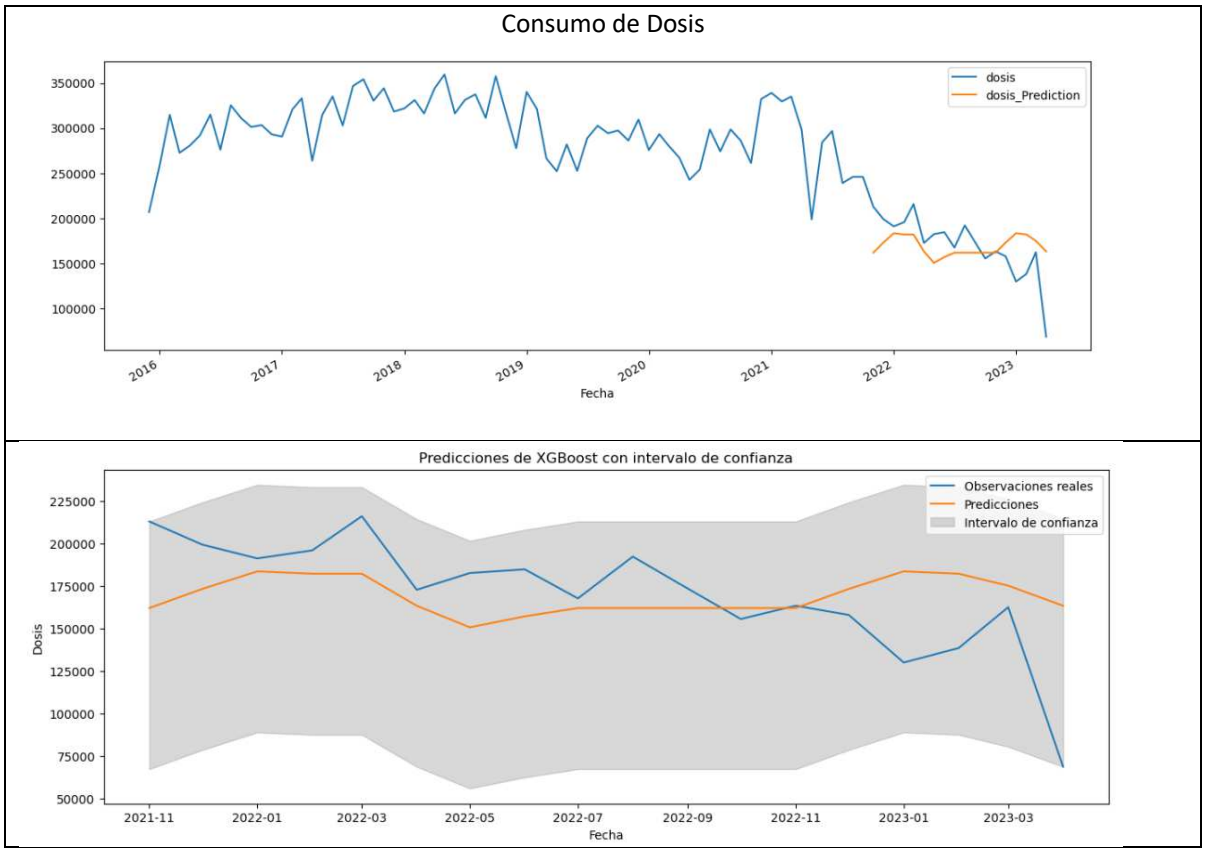
Variable	Definición operativa	Tipo	Valores
FechaDiagnosticoDiabetesMellitus	Fecha de diagnóstico de la Diabetes Mellitus	Temporal	DD/MM/AAAA
GrupoDiagnostico	Grupo Diagnóstico	Cualitativa Nominal Dicotómica	DM DM-HTA
Características físicas del paciente			
PesoKg	Peso del paciente en Kg	Cuantitativa Continua Razón	Entre 30 a 203 kgs
TallaM	Estatura del paciente	Cuantitativa Continua Razón	Entre 100 a 200 cms
IMC	Índice de masa corporal del paciente	Cuantitativa Continua Razón	Entre 12 a 145
DescripcionIMC	Descripción del Índice de masa corporal del paciente	Cualitativa Nominal	Bajo Peso Normal Obesidad 1 Obesidad 2 Obesidad 3 Sobrepeso Sin Dato
Características de laboratorio del paciente			
HemoglobinaGlicosilada	hemoglobina glicosilada (HbA1c) medida de la glucemia a largo plazo	Cuantitativa Continua Razón	Entre 3 y 21%
FechaUltimaHemoglobinaGlicosilada	Fecha de la última prueba de laboratorio de hemoglobina glicosilada		DD/MM/AAAA
VigenciaHBa1c2	Vigencia de la prueba de hemoglobina glicosilada (HbA1c)	Cualitativa Nominal	Vigente No vigente Sin toma de HBa1c
Albuminuria	Medición de Albuminuria, proteína indicadora de daño renal	Cuantitativa Continua Razón	Entre 0 y 13540 mg
FechaUltimaAlbuminuria	Fecha de la última prueba de laboratorio de hemoglobina glicosilada	Temporal	DD/MM/AAAA
ControlAlbum	Vigencia de la medición de Albuminuria	Cualitativa Nominal	Vigente No vigente Sin dato

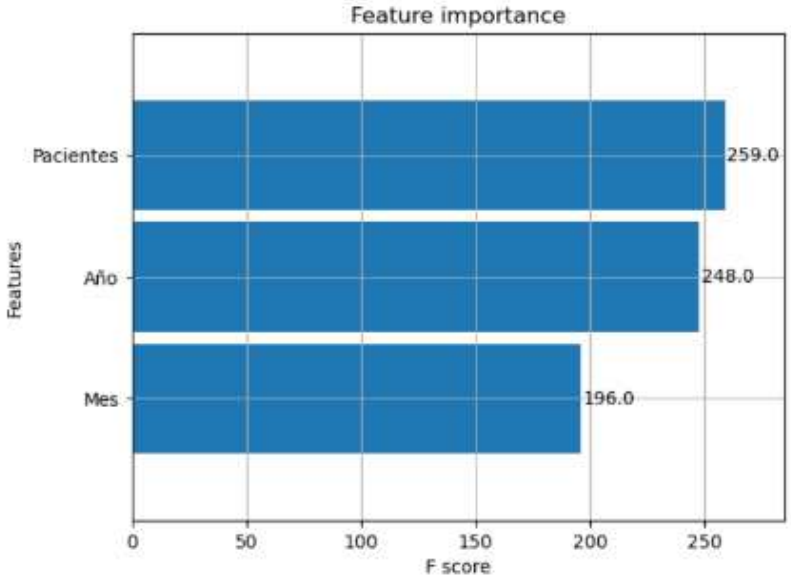
Variable	Definición operativa	Tipo	Valores
Creatinuria	Medición de los niveles de Creatinina en la sangre.	Cuantitativa Continua Razón	Entre 5 y 560 mg/dL
FechaCreatinuria	Fecha de la última prueba de laboratorio de la Creatinuria	Temporal	DD/MM/AAAA
ControlCreatinuria	Vigencia de la medición de Creatinuria	Cualitativa Nominal	Vigente No vigente Sin dato
AlbuminuriaCreatinuria	Cociente de microalbúmina y creatinina que compara la cantidad de albúmina con la cantidad de creatinina en la orina.	Cuantitativa Continua Razón	Entre 0 y 113 mg/dL
FechaUltimaAlbuminuriaCreatinuria	Fecha de la última prueba de laboratorio de la Albuminuria-Creatinuria	Temporal	DD/MM/AAAA
ControlAlbuminuriaCreatinuria	Vigencia de la medición de Albuminuria-Creatinuria	Cualitativa Nominal	Vigente No vigente Sin dato
ControlInasistenteHistorico	Control de inasistencia histórica	Cualitativa Nominal	Asistente Inasistente Inasistente Histórico
GrupoEtareo	Grupo de edad	Cualitativa Nominal	>=90 20-29 30-39 40-49 50-59 60-69 70-79 80-89

Fuente: Elaboración propia.

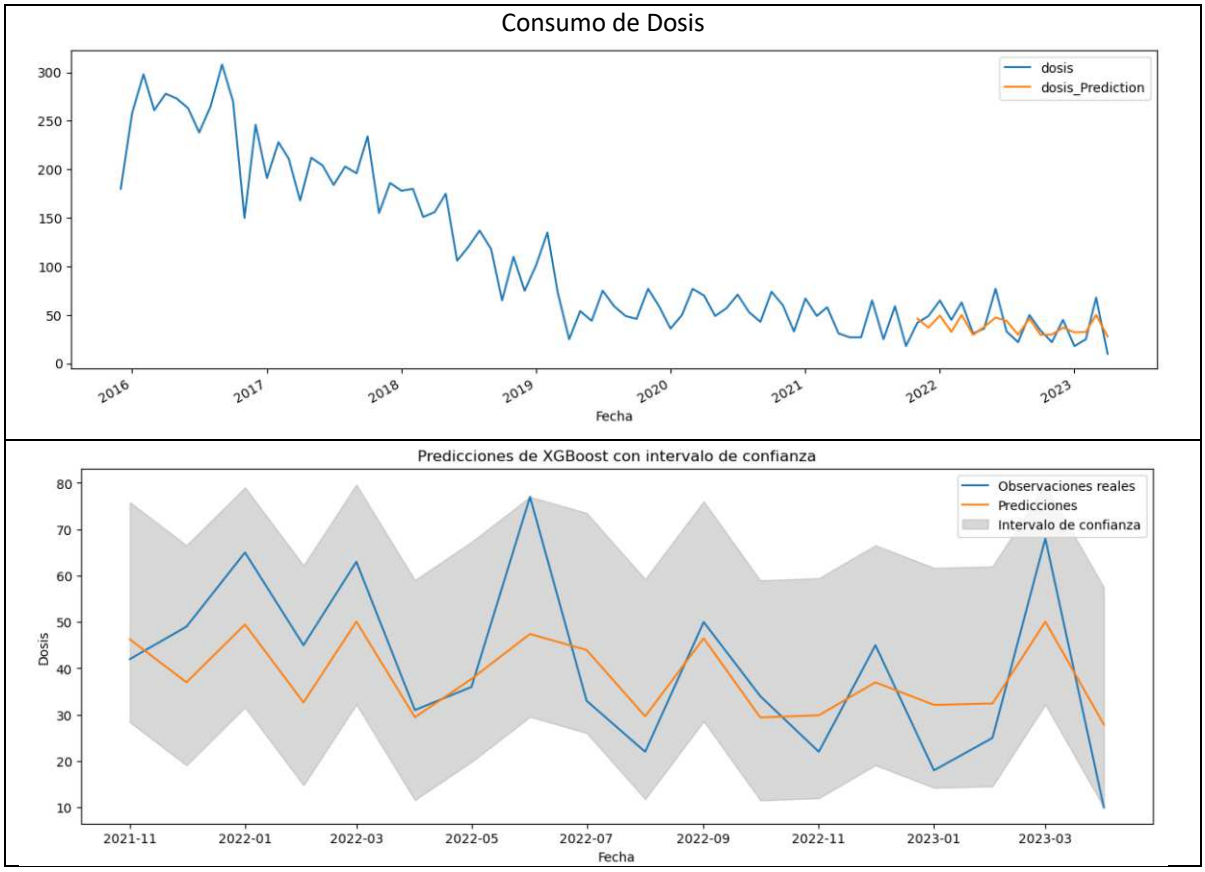
Anexo 2. Visualización de información de impacto para la toma de decisiones.

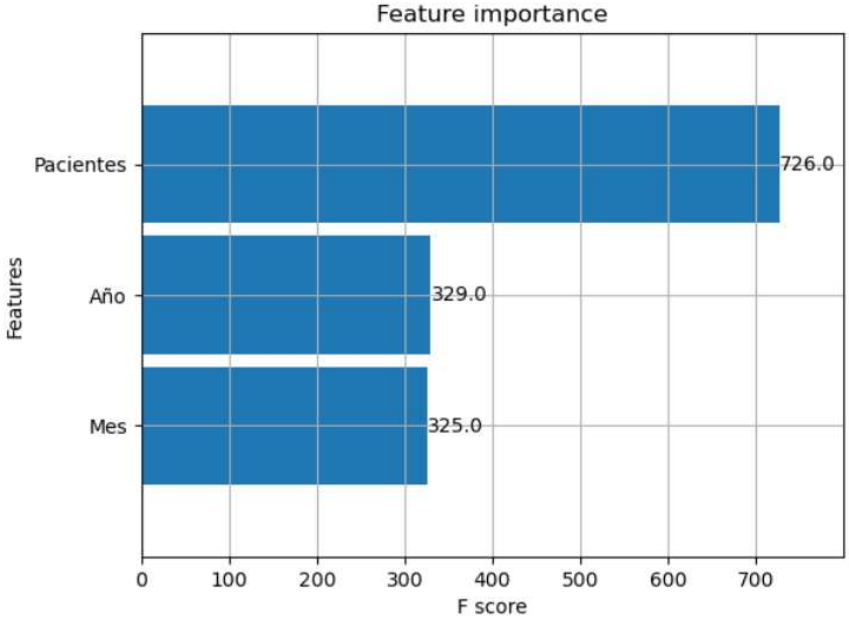
Modelo 1. METFORMINA TAB 850 MG.



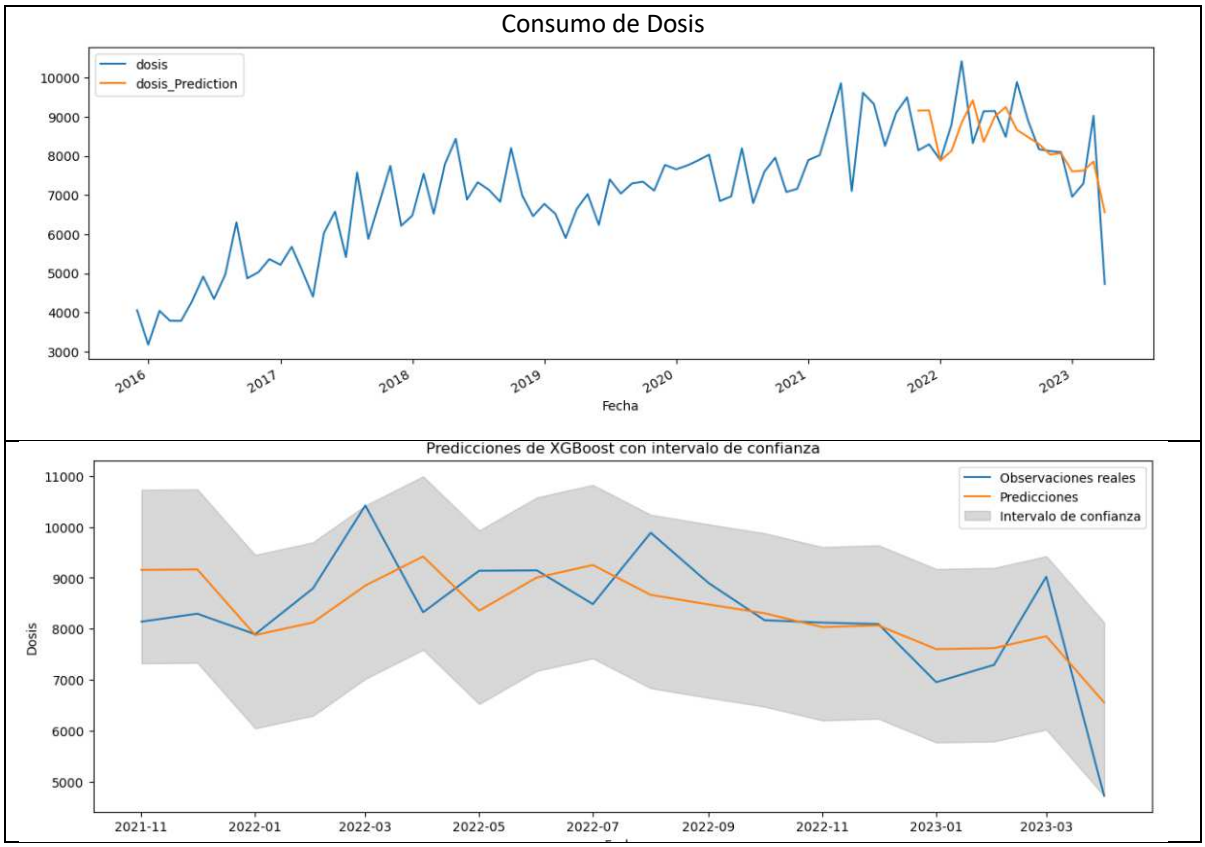


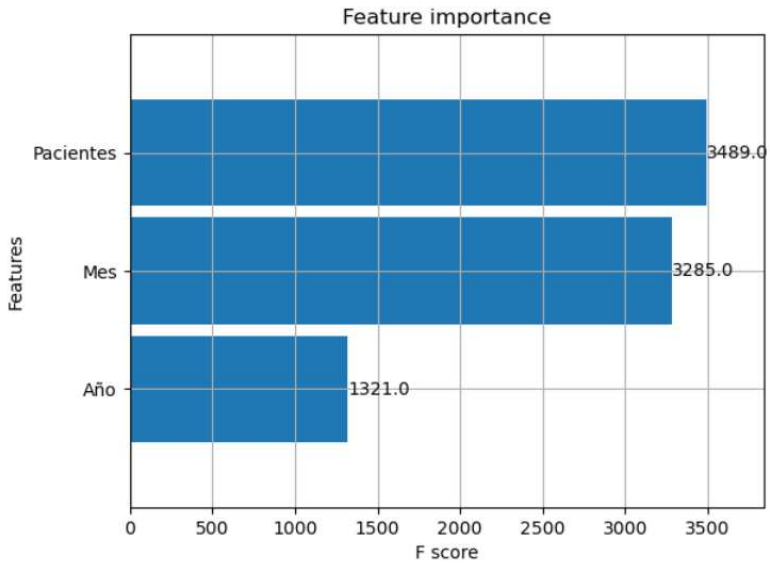
Modelo 4. INSULINA GLARGINA 100UI/ML VIAL 10ML.



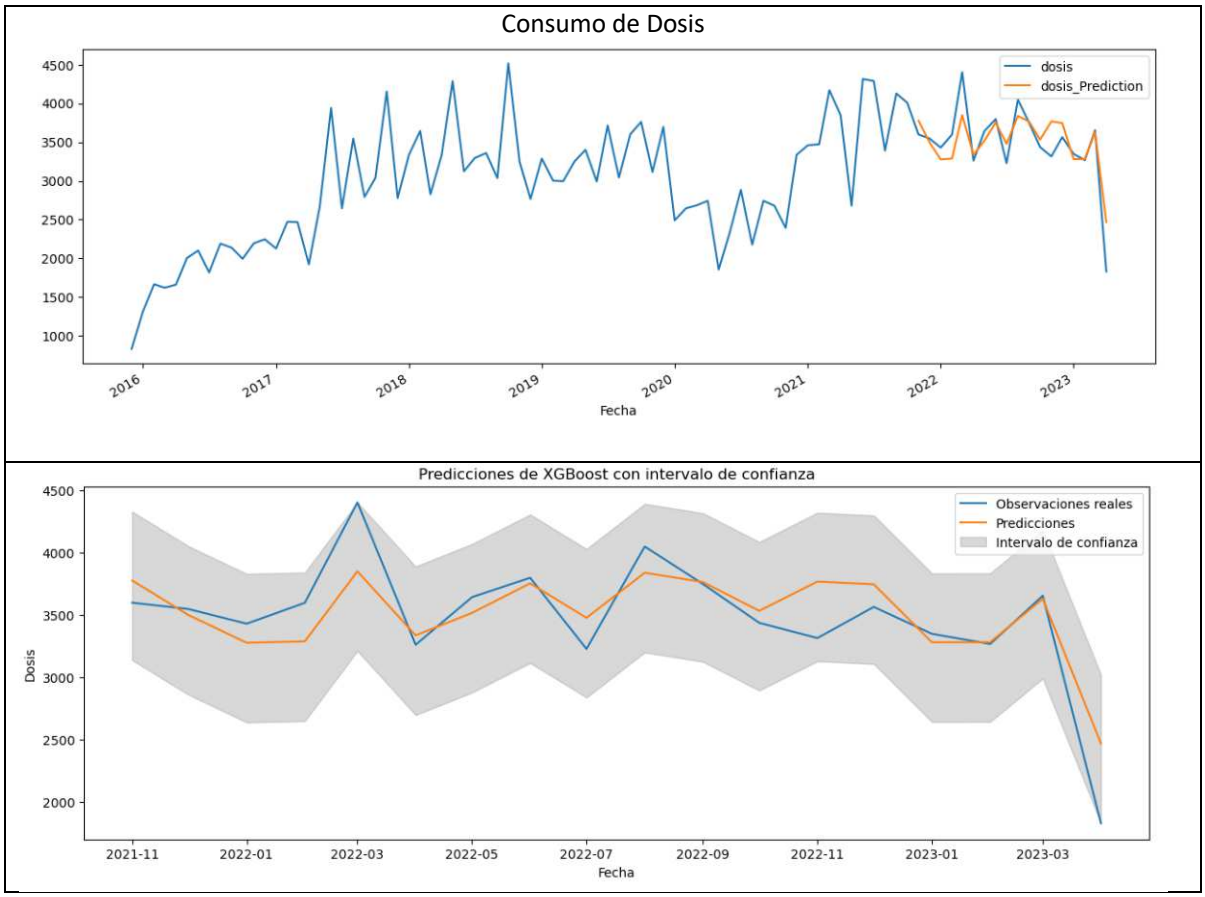


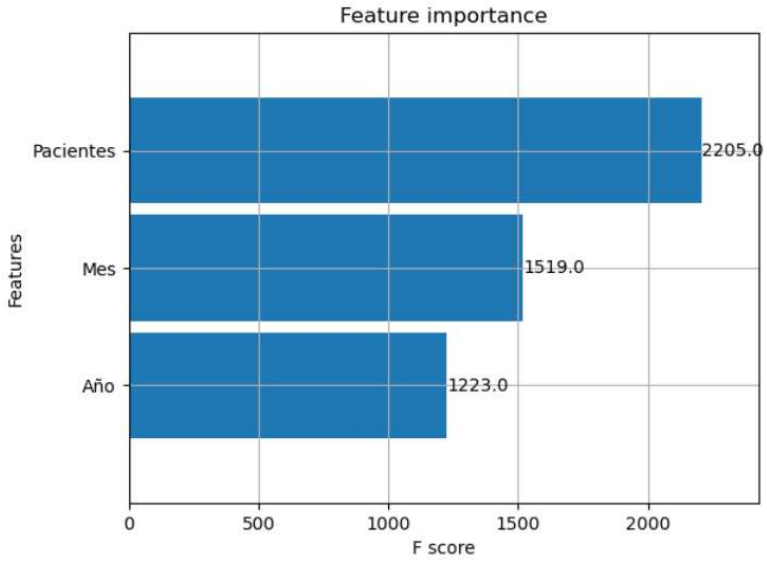
Modelo 6. INSULINA GLARGINA 100UI/ML SOLUCION INYECTABLE CARTUCHO X 3M



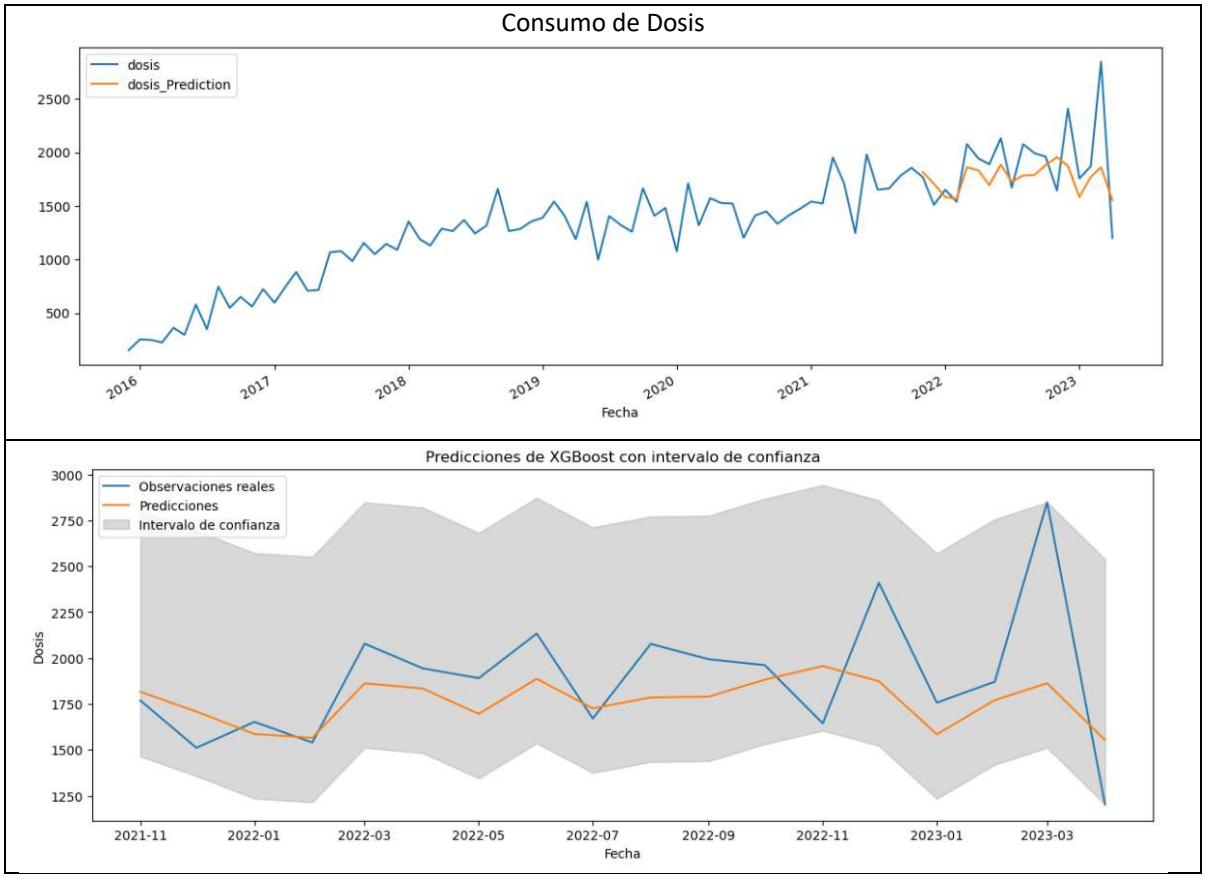


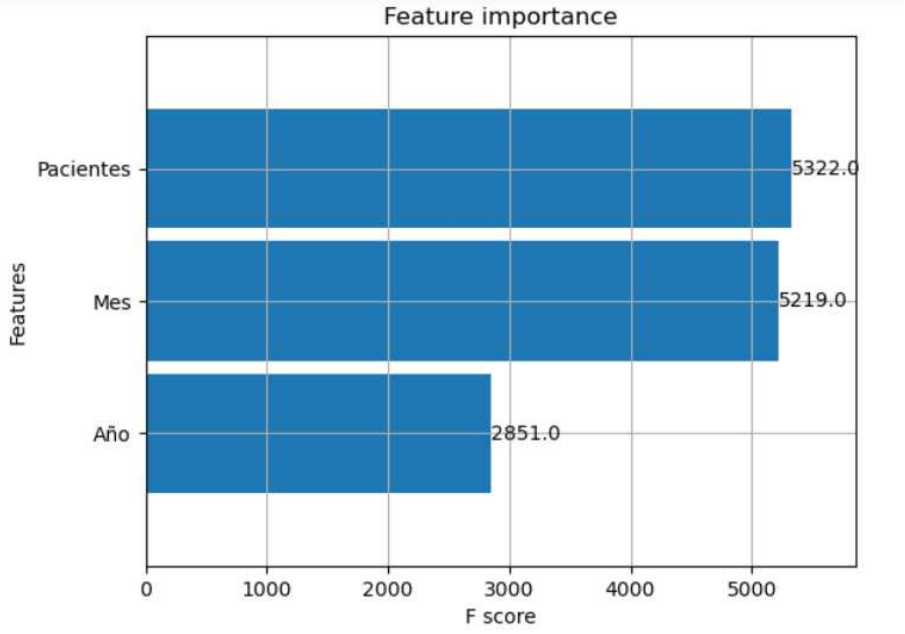
Modelo 8. INSULINA GLULISINA 100UI/ML SOL INY 3ML.



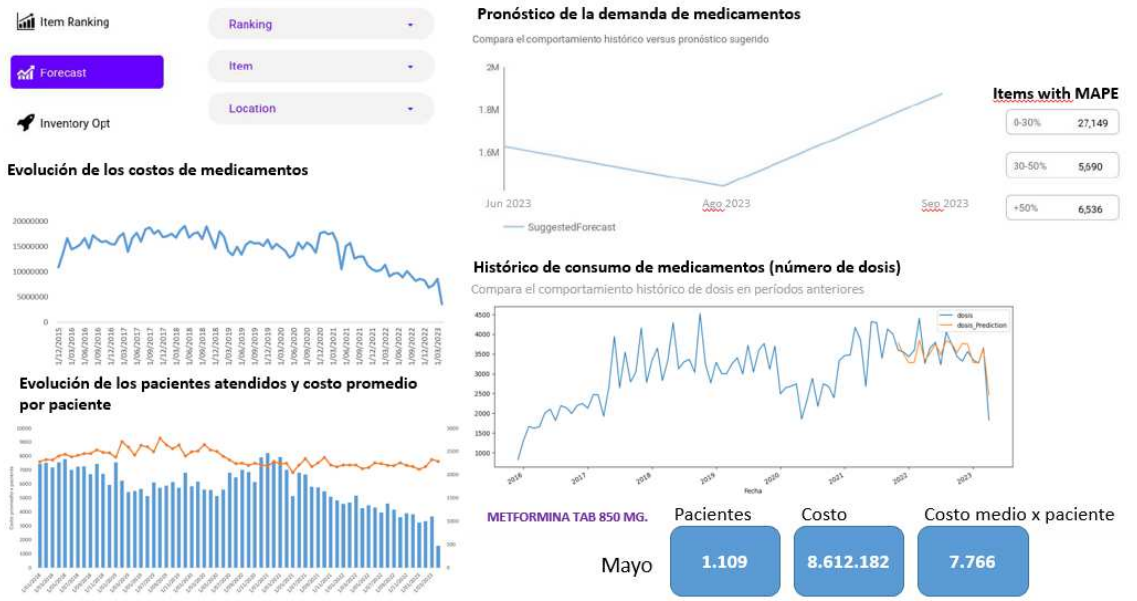


Modelo 10. INSULINA DEGLUDEC 100UI/ML SOLUCION INYECTABLE JERINGA PRELL





Anexo 3. Visualización mediante Dashboard.



7. BIBLIOGRAFÍA

Abdallah, A. A. (2020). Healthcare Engineering: A Lean Management Approach. Journal of Healthcare Engineering, 2020. https://doi.org/10.1155/2020/8875902

Amazon Web Services. (23 de marzo de 2023). ¿Qué es la ciencia de datos?. https://aws.amazon.com/es/what-is/data-science/

Arredondo A. y De Icaza E. (2011). Costos de la Diabetes en América Latina: Evidencias del Caso Mexicano.

Brownlee, J. (2017). Machine Learning Performance Improvement Cheat Sheet. 32 Tips, Tricks and Hacks to Make Better Predictions https://machinelearningmastery.com/choose-validation-method-for-machine-learning

Brownlee, J. (2021) Machine Learning Mastery With Python: Understand Your Data, Create Accurate Models and Work Projects End-To-End..

Bucca j., Damiani L., Esterkin G., García Dieguez M., Marcos E., Rodriguez M. (1994). Costos en Diabetes tipo II. Revista de la AMBB 1994; 2: 53-56

Clasificación Internacional de Enfermedades (2020). Secretaría General de Salud Digital, Información e Innovación del Sistema Nacional de Salud, Subdirección General de Información Sanitaria. www.mscbs.gob.es

Chopra, S., & Peter, M. (2008). Administración de la cadena de suministro. Pearson educación.

Chapman P., Clinton J., Kerber R., Khabaza T., Reinartz T., Shearer C., and Wirth R. The CRISP-DM consortium, (2000). https://maestria-datamining-2010.googlecode.com/svn-history/r282/trunk/dmct-teorica/tp1/CRISPWP-0800.pdf

Chuchoque-Urbina F. A., Caro-Gutierrez M. P., and Montoya C. E. (2021). Design of a CPFR (collaborative planning forecasting and replenishment), location, inventory and routing approach to diabetes and high blood pressure medicines supply network planning.

Dhar, V. (2013). Data Science and Prediction. Communications of the ACM, 56(12), 64-73. doi: 10.1145/2500499

Departamento Administrativo Nacional de Estadística - DANE. (21 de julio de 2022). Boletín Técnico Gasto Social Público y Privado (GSPP) 2021 Preliminar. https://www.dane.gov.co/files/investigaciones/boletines/pib/cuentas-nal-anuales/bol-socx-2021preliminar.pdf

Diario LR La República (14 de julio de 2022). *SALUD, El gasto promedio per cápita en el país en salud al año es de cerca de $1,3 millones*. https://www.larepublica.co/economia/el-gasto-promedio-de-los-colombianos-en-salud-al-ano-es-de-cerca-de-1-3-millones-3403198

EAE Business School Online - Blended.(25 de marzo de 2023). *Anticiparse a los problemas*. https://www.eaeprogramas.es/blog/negocio/empresa/anticiparse-los-problemas

Jaramillo Ramírez J. (2012). Pasantía de investigación. Pronósticos: Métodos Cualitativos y Cuantitativos vs. Métodos de Inteligencia Artificial. Colegio de Estudios Superiores de Administración CESA.

Khanal S., Veerman L., Nissen L. and Hollingworth S., (2019). Forecasting the amount and cost of medicine to treat type 2 diabetes mellitus in Nepal using knowledge on medicine usage from a developed country. Journal of Pharmaceutical Healt Services Reasearch - JPHS, 10; 91-99.

Kuhn M. y Johnson K., (2013) Applied Predictive Modeling. New York, NY:Springer, 2013. 10.1007/978-1-4614-6849-3.

Hyndman, R. J., & Athanasopoulos, G. (2018). Forecasting: principles and practice. pág. 135 -140

Lipton, Z. C., Berkowitz, J., & Elkan, C. (2015). A critical review of recurrent neural networks for sequence learning. arXiv preprint arXiv:1506.00019

Liu, Y., Kang, J., & Zhang, G. P. (2015). Interval forecast evaluation measures. International Journal of Forecasting, 31(1), 27-45.

Ministerio de Salud y Protección Social. (Junio de 2018). Dirección de Regulación de la Operación del Aseguramiento en Salud, Riesgos Laborales y Pensiones. *Gestión Integral del Riesgo en Salud. Perspectiva desde el Aseguramiento en el contexto de la Política de Atención Integral en Salud*. Documento de Trabajo, pág. 16, 22 – 24.

Ministerio de Salud y Protección Social. (29 de junio de 2022). *Colombia llegó al aseguramiento universal en salud al alcanzar el 99,6 %*. https://www.minsalud.gov.co/Paginas/Colombia-llego-al-aseguramiento-universal-en-salud-al-alcanzar-el 99.6.aspx#:~:text=Para%202022%2C%20el%20presupuesto%20para,un%2036% 20%25%20desde%20las%20cotizaciones)

Ministerio de Salud y Protección Social. (25 de marzo de 2023). *Aseguramiento al sistema general de salud*. https://www.minsalud.gov.co/proteccionsocial/Regimensubsidiado/Paginas/aseguramiento-al-sistema-general-salud.aspx

Montemayor Gallegos J.E. (2013). Métodos de Pronósticos para Negocios.D.R. Instituto Tecnológico y de Estudios Superiores de Monterrey, México.

Pan American Health Organization. (2022). Panorama of Diabetes in the Americas. Washington, D.C.: PAHO; Available from:https://doi.org/10.37774/9789275126332.

Perdigón LL. R., Gonzalez B. N. (2021). Comparison and selection of artificial intelligence techniques for forecasting bovine milk productions. Revista Cubana de Ciencias Informáticas, 2021, 15(2): p. 24-43

Powers C. A., Meyer C. M., Roebuck M. C. and Vaziri B. (2005). Predictive Modeling of Total Healthcare Costs Using Pharmacy Claims Data: A Comparison of Alternative Econometric Cost Modeling Techniques.

Raraz Vidal J., Raraz Vidal O., (2022). Adherencia terapéutica y variables relacionadas en adultos con diabetes mellitus tipo 2 en un hospital público.

Rodríguez Bolaños RA, Reynales Shigematsu LM, Jiménez Ruíz JA, Juárez Márquez SA, Hernández Ávila M. Costos directos de atención médica en pacientes con diabetes mellitus tipo 2 en México: análisis de microcosteo. Rev Panam Salud Publica. 2010;28(6):412–20.

Scikit-learn. Preprocessing data. (26 de abril de 2023). https://scikit-learn.org/stable/modules/preprocessing.html#scaling-features-to-a-range

Wirth R., Hipp D. (2000). CRISP-DM: Towards a standard process model for data mining.

Wesson J., Naude M., (2022) Using information visualization to support the Self-Management of Type 2 Diabetes Mellitus - DM.

World Health Organization Collaborating Centre for Drug Statistics Methodology (17 de abril de 2023). Anatomical Therapeutic Chemical Classification System - ATC/DDD Index 2023. https://www.whocc.no/atc_ddd_index/

World Health Organization. (2020). The top 10 causes of death. Available from: https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death

World Health Organization. (2021). Report of expert and stakeholder consultations on the WHO Global Diabetes Compact. Available from: https://apps.who.int/iris/handle/10665/340322