ESCUELA SUPERIOR POLITÉCNICA DEL LITORAL

Facultad de Ingeniería en Electricidad y Computación

ESTIMACION DE GASTOS MEDICOS PARA CLIENTES CORPORATIVOS EN BENEFICIO A UNA AGENCIA DE SEGUROS MEDIANTE EL USO DE MODELOS DE APRENDIZAJE AUTOMATICO

PROGRAMA DE MAESTRÍA EN CIENCIA DE DATOS

TRABAJO DE TITULACIÓN PREVIO A LA OBTENCIÓN DEL TÍTULO DE MAGÍSTER EN CIENCIA DE DATOS

Presentado por:

Morales Mendoza José Francisco

Guevara Esteves Melanie Rommina

Guayaquil, Ecuador

2025

Dedicatoria

Francisco Morales:

Dedico este proyecto a mi madre, Celenita Mendoza, y a mis familiares y amigos, quienes siempre confiaron en mí y me motivaron a seguir creciendo personal y profesionalmente, acompañándome en este importante paso en mi formación académica.

Melanie Guevara:

A mi madre, por enseñarme con su ejemplo el valor de la educación y la importancia del esfuerzo constante. Su influencia ha sido fundamental en cada paso de este camino académico.

Agradecimientos

Francisco Morales:

Agradezco sinceramente a todos los profesores que contribuyeron a mi formación durante la maestría, por su dedicación y exigencia académica.

Extiendo también mi agradecimiento al tutor de este trabajo de titulación, por su orientación y apoyo durante el desarrollo de esta investigación.

Finalmente, a todas las personas que, de alguna manera, brindaron su respaldo a lo largo de este proceso, les expreso mi más sincera gratitud.

Melanie Guevara:

Quiero expresar mi sincero agradecimiento a los profesores, quienes con su dedicación, conocimiento y exigencia académica contribuyeron significativamente a mi formación.

Agradezco especialmente a mi compañero de tesis, por su compromiso, colaboración y constancia a lo largo de este proyecto.

A todos quienes, de una u otra forma, apoyaron este trabajo, les extiendo también mi gratitud.

Declaración expresa

Nosotros, José Francisco Morales Mendoza y Melanie Rommina Guevara Esteves acordamos y reconocemos que: La titularidad de los derechos patrimoniales de autor (derechos de autor) del proyecto de graduación corresponderá al autor o autores, sin perjuicio de lo cual la ESPOL recibe en este acto una licencia gratuita de plazo indefinido para el uso no comercial y comercial de la obra con facultad de sublicenciar, incluyendo la autorización para su divulgación, así como para la creación y uso de obras derivadas. En el caso de usos comerciales se respetará el porcentaje de participación en beneficios que corresponda a favor del autor o autores. El o los estudiantes deberán procurar en cualquier caso de cesión de sus derechos patrimoniales incluir una cláusula en la cesión que proteja la vigencia de la licencia aquí concedida a la ESPOL.

La titularidad total y exclusiva sobre los derechos patrimoniales de patente de invención, modelo de utilidad, diseño industrial, secreto industrial, secreto empresarial, derechos patrimoniales de autor sobre software o información no divulgada que corresponda o pueda corresponder respecto de cualquier investigación, desarrollo tecnológico o invención realizada por nosotros durante el desarrollo del proyecto de graduación, pertenecerán de forma total, exclusiva e indivisible a la ESPOL, sin perjuicio del porcentaje que nos corresponda de los beneficios económicos que la ESPOL reciba por la explotación de nuestra innovación, de ser el caso. En los casos donde la Oficina de Transferencia de Resultados de Investigación (OTRI) de la ESPOL comunique los autores que existe una innovación potencialmente patentable sobre los resultados del proyecto de graduación, no se realizará publicación o divulgación alguna, sin la autorización expresa y previa de la ESPOL.

Guayaquil, mayo del 2025.

Jose Francisco Morales Mendoza
Time Stamping Security Data

Morales Mendoza José Francisco

Melanie Rommina Guevara Esteves
Time Stamping Security Data

Guevara Esteves Melanie Rommina

Evaluadores

Firmado electrónicamente por:
SERGIO ALEX BAUZ OLVERA
Validar únicamente con FirmaEC

Sergio Alex Bauz Olvera

Tutor de proyecto

Maria Isabel Mera Collantes
Time Stamping Security Data

Ma. Isabel Mera Collantes

Revisor de proyecto

Resumen

Este proyecto desarrolló un modelo predictivo para estimar el gasto médico cubierto por aseguradoras, utilizando datos históricos de empleados pertenecientes a clientes corporativos. Mediante técnicas de preprocesamiento, reducción de dimensionalidad (UMAP), segmentación (KMeans) y modelos de regresión avanzados como Random Forest, Gradient Boosting y XGBoost, se construyó una herramienta precisa y escalable. El modelo permite anticipar con mayor precisión los gastos futuros y facilitar la planificación financiera, tanto para la aseguradora como para el asegurado, reduciendo la asimetría de la información en la gestión del riesgo. Además, se diseñó una herramienta de visualización interactiva que facilita la interpretación de los resultados por parte de los usuarios finales. Se concluye que la metodología es replicable y que existen claras oportunidades de mejora, como la incorporación de nuevas variables explicativas y el uso de algoritmos más sofisticados. Futuras versiones del modelo deberán integrar una mayor cantidad de datos para optimizar la predicción de ciertos tipos de afectaciones a la salud.

Abstract

This project developed a predictive model to estimate medical expenses covered by insurance companies, using historical data from employees of corporate clients. Through preprocessing techniques, dimensionality reduction (UMAP), clustering (KMeans), and advanced regression models such as Random Forest, Gradient Boosting, and XGBoost, a scalable and accurate tool was built. The model enables more precise forecasting of future medical expenses, improving financial planning for both insurers and policyholders by reducing information asymmetry in risk management. Additionally, an interactive visualization tool was developed to help end users interpret the results with ease. The methodology is proven to be replicable and shows clear opportunities for improvement, including the incorporation of new explanatory variables and the use of more sophisticated algorithms. Future versions of the model should integrate larger datasets to enhance the prediction of specific health conditions.

Abreviaturas

AWS: Amazon Web Services

CNN: Redes neuronales convolucionales

ETL: Extracción, Transformación y Carga

INEC: Instituto Ecuatoriano de Estadística y Censos

k-NN: k-Nearest Neighbors

MAE: Error Absoluto Medio

MSE: Error Cuadrático Medio

ML: Machine Learning

PCA: Análisis de componentes principales

RF: Random Forest

UMAP: Uniform Manifold Approximation and Projection

SNE: Stochastic Neighbor Embedding

Índice de contenido

Índice de tablas

Índice de figuras

1. Planteamiento de la problemática

1.1. Título

Estimación de gastos médicos para clientes corporativos en beneficio a una agencia de seguros mediante el uso de aprendizaje automático.

1.2. Descripción del Problema

Al momento de emitir o renovar una póliza de seguros, no es posible prever con exactitud los eventos que ocurrirán durante el período de vigencia. En esta incertidumbre inciden dos factores clave: por un lado, el comportamiento del asegurado en relación con su nivel de siniestralidad y las medidas que adopta para cuidar su salud; y por otro, el constante incremento en los costos de los servicios médicos, influenciado por la inflación, la demanda de tratamientos especializados y otros factores externos.

Las agencias de seguros, aunque tienen acceso a información relevante sobre reembolsos y gastos médicos de sus clientes corporativos, no siempre cuentan con una visibilidad completa de todos los costos incurridos. Esto se debe a que existen gastos que no pasan directamente por la agencia, lo que dificulta la trazabilidad total y la proyección precisa de los costos. Además, factores externos como el aumento en la morbilidad, la inflación en los servicios de salud y eventos imprevistos también influyen en los ajustes de las primas, dificultando que las agencias puedan anticipar con exactitud los incrementos en las pólizas.

Esta falta de previsibilidad puede traducirse en procesos poco transparentes para los asegurados, generando inconformidades y cambios constantes en los planes seleccionados. Claramente, esta metodología debe de estar relacionada con los gastos médicos cubiertos por la aseguradora.

(Manchester, 2014) destaca que Estados Unidos han incrementado los costos de la salud en general. Esto ha sido impulsado por factores como el aumento de enfermedades crónicas, el envejecimiento de la población y el encarecimiento de los tratamientos médicos. Esta situación también se refleja en Ecuador. Según (INEC, 2024), en abril del 2024, los precios relacionados al cuidado de la salud aumentaron un 2,78%. Este incremento está

enmarcado en un contexto de inflación general, influenciado por la reciente alza del IVA del 12% al 15%, lo cual ha impactado directamente en los costos del sector salud.

De igual manera, (Dutta, Bose, & Bandyopadhyay, 2023) señala que en los últimos años los gastos médicos han crecido exponencialmente. Por este motivo, es importante que exista más información sobre la siniestralidad de las personas, permitiéndoles escoger un plan que tenga la cobertura correcta y que puedan recibir la atención correspondiente. (Ismail, Stam, Portrait, & Van Witteloostuijn, 2024) concuerdan que la asimetría de la información puede conllevar a que ciertos actores queden excluidos de este mercado de seguros. A través de sus modelos pudieron determinar que técnicas de Gradient Boosting Machines (GBM) mejoran selectivamente la predicción para grupos con los costos médicos reales que se ubican en los extremos, mientras que el RF es más preciso para los otros casos.

Actualmente las aseguradoras se basen en los gastos históricos para calcular una nueva cuota anual no permite capturar todos los factores de impacto, tales como el incremento de precios, estilo de vida, entre otros. Es decir, el problema de asimetría de información se acentúa dado que no se conoce cómo el asegurado está manejando ciertos aspectos de su salud. Por ejemplo, (Mladenovic, y otros, 2020) señalan que el factor más importante para determinar el gasto médico de asegurados fue si eran fumadores o no.

(Sharma & Jeya, 2024) encontraron que variables como el tabaquismo, la relación entre estatura y peso, el número de hijos, la edad y el sexo explicaban hasta el 81,3% de los gastos médicos, según los resultados de su modelo de regresión lineal. No obstante, este estudio fue realizado durante la pandemia de COVID-19, un contexto que pudo haber alterado las dinámicas habituales del sector salud, lo cual limita la generalización de los resultados a períodos posteriores. Por su parte, (Sandra, y otros, 2024) propusieron una modificación en el modelo al sustituir el número de hijos por la región geográfica como variable explicativa. Utilizando un modelo de Random Forest, lograron un coeficiente de determinación R^2 de 86.43%. Es importante mencionar que el análisis se realizó en un contexto diferente, ya que se centró exclusivamente en población residente en India.

Por otra parte, (Matloob, y otros, 2021) buscan un enfoque diferente tratando de generar planes basados en necesidades individuales, en lugar de corporativos incluyendo en su estudio los planes que existen en mercado. De manera similar, (Chan, Lee, & Zainol, 2021) segmentaron por tipo de empleado, permitiendo considerar riesgos médicos que incurran en costos. Además, (Haiyan, Saeed, Hang, Renying, & Hongxia, 2024) establecen un modelo para un grupo específico de personas: las mujeres embarazadas que podrían desarrollar diabetes. En este escenario se miden constantemente métricas específicamente relacionadas a la enfermedad debido al estado de la mujer.

(Panda, Purkayastha, Das, Chakraborty, & Biswas, 2022) destaca que la importancia de un correcto seguro médico debe de considerar cuántos años en promedio viven las personas de una determinada población, además de cambios epidemiológicos y aumento de costos médicos. A diferencia de la mayoría de las modelos, no solo se consideran datos demográficos sino también el historial clínico y estado actual de la persona, permitiendo obtener una regresión polinómica.

Una propuesta que ha tomado fuerza entre aseguradoras y empleadores es el seguimiento continuo de la salud de los colaboradores, con el objetivo de detectar de manera temprana posibles enfermedades o condiciones médicas. Esto permitiría intervenir a tiempo y, en consecuencia, reducir significativamente los gastos médicos a largo plazo. Por ejemplo, (Amin, Liu, & Kensaku, 2020) desarrollaron un modelo basado en redes neuronales que analiza imágenes médicas para identificar señales tempranas de enfermedades.

Sin embargo, este tipo de iniciativas, aunque valiosas, no suelen ser implementadas directamente por las agencias de seguros. Estas agencias, que actúan como intermediarias y representan a los clientes corporativos ante las aseguradoras, tienen un enfoque comercial centrado en la satisfacción del cliente, más que en el desarrollo de soluciones tecnológicas complejas. Por tanto, asumir los costos y recursos necesarios para implementar este tipo de modelos no forma parte de su rol operativo ni de su estructura de negocio, lo que puede hacer inviable su aplicación desde el lado de la agencia.

En cuanto a la comparación de modelos, (Orji & Ukwandu, 2024) emplean el mecanismo de XGBoost y Gradient Boosting Machine, siendo el primero el de mayor precisión. Además, las enfermedades crónicas y antecedentes de cáncer son determinantes importantes en la prima. Al igual que los modelos anteriores, (Syarifah & Herdianto, 2023) encontraron que el modelo con un mejor desempeño fue el XGBoost al momento de predecir los reclamos por cuestiones médicas.

1.3. Justificación

La estimación de gastos deducibles a través de modelos de aprendizaje automático ofrece una perspectiva integral sobre el comportamiento promedio de los asegurados, fundamentándose en características demográficas, diagnósticos médicos, historias clínicas y tipo de servicio. Esta metodología permite a los clientes corporativos anticipar variaciones en los gastos, optimizar la gestión del pago de primas y establecer estrategias preventivas para evitar incrementos futuros en los deducibles o, en su defecto, mitigar la reducción de beneficios. Asimismo, contribuye a una mejor planificación de beneficios médicos al identificar

tendencias generales en las necesidades de salud de los empleados, promoviendo un uso más eficiente y equitativo de los servicios de atención.

Es importante destacar que esta solución no busca identificar o monitorear individualmente a empleados con determinadas condiciones médicas, sino analizar al conjunto de colaboradores como una sola entidad. El objetivo es extraer patrones generales a partir de datos agregados y generalizados, sin vulnerar la privacidad de ningún trabajador. Para asegurar el correcto uso de la información, la herramienta debe operar bajo estrictas restricciones de acceso y confidencialidad, evitando cualquier forma de discriminación o uso indebido de datos personales.

La estimación precisa de los gastos médicos permite a las empresas seleccionar pólizas que se ajusten al perfil de salud de su plantilla, garantizando cobertura para eventos relevantes. Este enfoque facilita el diseño de primas más alineadas con el riesgo real, optimizando la relación entre costo y beneficio de la cobertura. A su vez, permite una asignación más eficiente de recursos dentro del mercado asegurador, promoviendo decisiones más informadas tanto para las aseguradoras como para las compañías aseguradas. (Ridzuan, y otros, 2024) indica que el aumento de la contratación de seguros se debe a la sociedad es consciente de los gastos médicos y de cirugía, por ende, es necesario saber qué variables son importantes para pronosticar el precio de una prima de seguros.

Entre los beneficios destacados por (Mathauer & Oranje, 2024), la aplicación de modelos de la ciencia de datos en el ámbito de la salud permite una distribución más eficiente y transparente de las herramientas que se dispongan, el acceso a servicios médicos adecuados, protección financiera y una mayor satisfacción de las necesidades médicas. No obstante, es importante considerar que estos modelos predictivos pueden incorporar sesgos derivados de datos históricos incompletos o desbalanceados, lo que podría afectar la equidad en la toma de decisiones si no se implementan mecanismos de control y validación adecuados.

(Santos, Leal, & Balancieri, 2024) comentan que poder acceder a servicios de salud es un derecho, por lo que el sistema de salud privado puede llegar a compensar las ineficiencias que existen en algunos países en el sector público para así poder garantizar este derecho. Por este motivo, se destacan a los datos como un insumo de importancia, el cual permitirá identificar grupos altamente riesgosos.

A través de esta información es posible la elaboración de modelos predictivos. (Kwon, Park, Park, Park, & Baik, 2024) mencionan que para poder conocer el estado de un paciente tradicionalmente es necesario conocer su historia clínica pasada y sucesos por lo que se encuentra atravesando, lo que ocasiona demoras en los procesos. De manera similar sucede

en los seguros, por lo que un modelo de predicción permitiría ahorrar tiempo para una mejor gestión de los procesos que lleva a cabo un agente de seguros, lo que a su vez da paso a que el cliente tenga pronto acceso a un seguro. Adicionalmente, (Kuo, Yu, Chen, & Chan, 2018) señalan que a través de la predicción de gastos médicos, los hospitales pueden mejorar la eficiencia de la atención y en cuanto a términos financieros, mientras que (Hassan, y otros, 2021) también comentan que esto permite la creación de mejores políticas en el ámbito de los seguros de tal forma que se puedan crear planes más personalizados que beneficien a todas las partes interesadas.

Un aspecto crucial de la correcta estimación de los gastos médicos es la mitigación del impacto negativo del incremento de precios en el sector salud. Investigaciones previas han evidenciado que el aumento en los gastos puede disuadir a las personas de acceder a servicios médicos, resultando en un deterioro de su salud. Proporcionar una estimación precisa del gasto por siniestralidad, junto con un plan médico adecuado, fomenta una mayor frecuencia de controles médicos y un acceso más amplio a los servicios de salud en general, mejorando así la carga repartida entre la salud pública y privada.

En resumen, el uso de modelos de ciencia de datos para la estimación de gastos deducibles no solo optimiza la gestión financiera de las aseguradoras y sus clientes, sino que también promueve un acceso más equitativo y eficiente a los servicios de salud, beneficiando tanto a las empresas como a los asegurados. (Santos, Leal, & Balancieri, 2024)

1.4. Objetivos (General y Específico):

Objetivo general:

- Diseñar un modelo predictivo de gastos médicos para empleados de clientes corporativos en beneficio a una agencia de seguros usando modelos de aprendizaje automático para mejorar la satisfacción del cliente.

Objetivos específicos:

- Desarrollar un modelo de ciencia de datos para la estimación de los gastos médicos de empleados en beneficio a clientes corporativos mediante el ajuste de los hiper parámetros del modelo.
- Evaluar el desempeño de los modelos de aprendizaje automático para predecir los gastos médicos asociados mediante el uso de varias métricas de rendimiento.

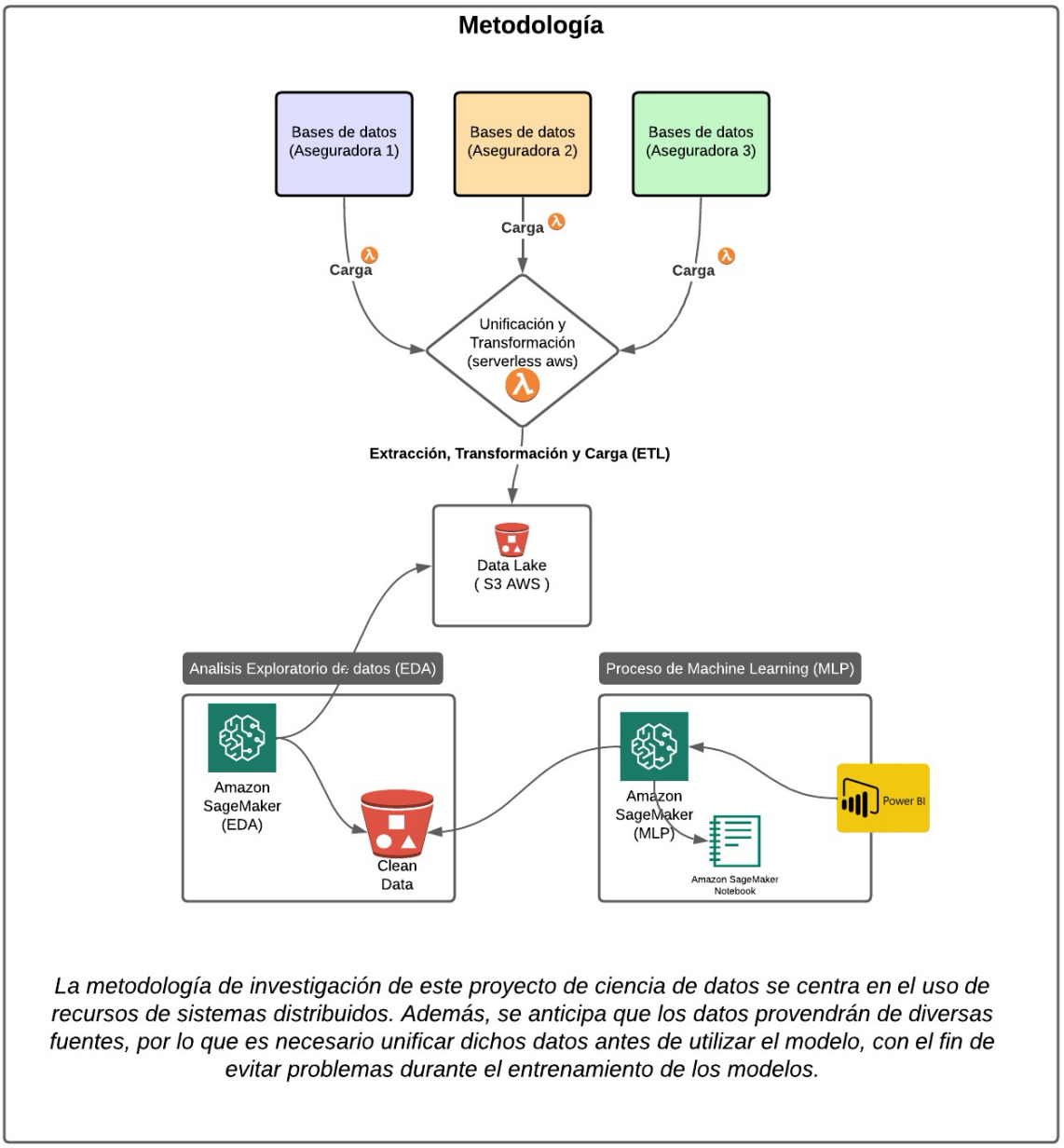
- Implementar una herramienta de visualización de datos dinámica e interactiva que use el mejor modelo para la estimación adecuada de los gastos médicos.

1.5. Metodología:

Durante el proyecto, se hará un análisis exhaustivo de varias fuentes de datos, incluidas las bases provenientes de las aseguradoras y los servicios atendidos a través de la red interna de la agencia de seguros. Dado que esta problemática involucra a múltiples emisores de seguros corporativos, es necesario seguir el siguiente esquema de forma secuencial:

- Crear un diccionario que contenga los nombres de las variables relevantes para el análisis. Esto asegurará que el modelo predictivo pueda interpretar correctamente las variables durante el proceso de entrenamiento. El objetivo es generalizar las variables para que, al fusionar la información que se encuentra distribuida, puedan ser ingresadas en un único almacén de datos con variables reconocidas por el modelo.
- En la figura 1 se presenta el proceso de Extracción, Transformación y Carga (ETL) en sistemas distribuidos, priorizando repositorios de datos masivos, para unificar registros (gastos médicos, demografía, sector laboral) de diversas fuentes en un único almacén de datos.
- Se realiza un análisis exploratorio de datos que identifica datos atípicos, realiza imputaciones, limpia datos, concatena, entre otros procesos, utilizando Jupyter Notebook en SageMaker.
- Con una base de datos consolidada, se procede a crear modelos de aprendizaje automático tipo ensamblador, así como también árboles de decisión, con el fin de determinar el modelo más preciso con respecto a los valores reales. (Kandula, Kalyanapu, Rayapalli, Modugumudi, & Kanikella, 2024) mencionan que un enfoque de Machine Learning son particularmente beneficiosos para poder estimar pacientes de altos costos y necesidades.
- Finalmente, el modelo que tenga mayor grado de precisión se utilizará para la estimación de gastos médicos basados en empleados de clientes corporativos reflejado en un tablero de control mediante el uso de Power BI.

Figura 1 Metodología propuesta



La metodología de investigación de este proyecto de ciencia de datos se centra en el uso de recursos de sistemas distribuidos. Además, se anticipa que los datos provendrán de diversas fuentes, por lo que es necesario unificar dichos datos antes de utilizar el modelo, con el fin de evitar problemas durante el entrenamiento de los modelos.

1.6. Resultados Esperados

La proyección de gastos médicos es crucial para permitir a las empresas anticipar y gestionar eficazmente futuros incrementos en los costos de facturación de clientes corporativos.

Como resultado de implementar un modelo predictivo de regresión, se anticipa lograr estimaciones precisas de gastos médicos que satisfagan las necesidades del cliente corporativo. Esto beneficiará a la agencia de seguros al mejorar la atención al cliente y

aumentar su confianza en los servicios proporcionados, los mismos que se detallan a continuación:

- Proporcionar a la agencia de seguros una herramienta tecnológica visual que integre un modelo de estimación de gastos médicos con una precisión mínima del 80%. Este modelo empleará variables demográficas, así como características del entorno laboral, como el sector de empleo, gastos médicos y diagnósticos presentados, para generar predicciones precisas para revelar insights estratégicos.
- Generar un informe detallado para el cliente con proyecciones de gastos médicos basadas en la demografía de sus colaboradores, gastos presentados y sector económico, para mejorar las decisiones corporativas seleccionando las condiciones de la póliza durante los procesos de contratación o renovación.

1.7. Dataset

En el proceso de preparación y validación del modelo de aprendizaje automático se consolida usando varios conjuntos de datos que respalden los comportamientos a analizar, como la información demográfica de los empleados, los gastos médicos asociados y el sector de la organización en el mundo laboral; hay más de 400 empresas de diferentes perfiles, la misma que tienen una póliza de seguros en las diferentes aseguradoras de coberturas nacionales e internacionales; por cada cliente corporativo se tiene más de 9000 registros médicos aproximadamente; considerando estos registros sobre el número de empleados, pero podría variar según los años de cobertura que el filial cuenta en el mercado de seguros.

El conjunto de variables a utilizar en el desarrollo del modelo de aprendizaje automático se puede apreciar en la Tabla 1.

Tabla 1 Datos a emplear

Variable	*Tipo de Dato*	*Tipo de Variable*	*Ejemplo*
Diagnósticos médicos	Carácter	Cualitativa nominal	Dermatitis Seborreica
Red hospitalaria	Carácter	Cualitativa nominal	Directa
Proveedor de salud	Carácter	Cualitativa nominal	Dermovasc S.A.
Tipo de servicio	Carácter	Cualitativa nominal	Ambulatorio
Valor presentado	Decimal	Cuantitativa continua	$ 80,00
Valor liquidado	Decimal	Cuantitativa continua	$ 64,00
Fecha registro	Fecha	Cuantitativa temporal	12/05/2023
Edad	Entero	Cuantitativa continua	41
Sexo	Carácter	Cualitativa nominal	F
Ciudad de residencia	Espacial	Cualitativa nominal	Galápagos
Sector económico	Carácter	Cualitativa nominal	Turismo

En el marco de la recopilación de la información; las aseguradoras al momento de cotizar y/o renovar nuevos acuerdos entre partes otorgan toda la información de siniestralidad del cliente corporativo, adicionando que se cuenta con los datos provenientes de la agencia de seguros que figura como intermediario entre el grupo corporativo y/o filial y la aseguradora. La actualización de esta se da periódicamente ya que esta información está disponible para fines empresariales.

El formato actual de la información es presentado en diferentes conjuntos de datos; mayoritariamente de hojas de cálculo (.xlsx) y de base de datos relacional (RDB), por la cual se plantea realizar el proceso ETL previo al análisis exploratorio de los datos y de entrenamiento a los modelos a emplear.

1.8. Consideraciones éticas

A partir del desarrollo de este trabajo de titulación, centrado en el acceso a seguros y tratamientos médicos de los empleados de clientes corporativos, es fundamental incorporar una sección de ética que asegure el acceso y la protección de los datos personales, en concordancia con lo dispuesto por la Ley Orgánica de Protección de Datos Personales propuesto por la (Asamblea Nacional del Ecuador, 2021).

A continuación, se presentan algunas consideraciones:

Promover el acceso equitativo e informado:

- **Relación con los Objetivos:** El modelo predictivo a desarrollar para estimar los gastos médicos podría influir en las decisiones de las empresas sobre la cobertura de seguros a sus empleados o la provisión de tratamientos médicos. Es esencial asegurar que este modelo no sea utilizado para negar injustamente el acceso a seguros o tratamientos a ciertos empleados.
- **Ética en la Implementación:** En línea con los principios de no discriminación de la Constitución del Ecuador (Constitución de la República del Ecuador, 2008; Asamblea Nacional del Ecuador, 2021) y la LOPDP, es importante evitar que el modelo sea utilizado para discriminar o perjudicar a determinados empleados con perfiles de alto riesgo o con condiciones preexistentes.

Garantizar la NO discriminación:

- **Relación con los Objetivos:** Al ajustar los hiperparámetros del modelo y evaluar su desempeño, es importante que el modelo trate de manera equitativa a todos los involucrados, sin considerar sus características demográficas o de condiciones médicas.
- **Ética en la Implementación:** Debe realizarse un análisis permanente para detectar posibles sesgos algorítmicos, conforme a principios de justicia algorítmica (Jobin, A., Ienca, M., & Vayena, E., 2019) que puedan afectar negativamente a determinados grupos.

Transparencia de la información:

- **Relación con los Objetivos:** La herramienta de visualización a implementar debe permitir a los usuarios finales a tomar decisiones más informada y responsable en base a los resultados del modelo predictivo.

- **Ética en la Implementación:** Es importante que los resultados del modelo sean comprensibles y auditables, siguiendo principios de explicabilidad en inteligencia artificial (Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D., 2018), para evitar decisiones arbitrarias o incomprensibles para los usuarios afectados.

Respetar la privacidad de los datos:

- **Relación con los Objetivos:** Al desarrollar e implementar el modelo, se debe considerar la privacidad de la información a visualizar, puesto que se les debe garantizar a los involucrados la privacidad de sus datos personales y gastos médicos acorde a lo establecido por la ley de protección de datos.
- **Ética en la Implementación:** La visualización de los resultados del modelo predictivo debe respetar la privacidad de los datos de los involucrados, evitando la divulgación de diagnósticos médicos específicos. En su lugar, se deben presentar datos de manera generalizada para permitir usar la información obtenida responsablemente sin comprometer la privacidad de los individuos (González, 2022).

Uso responsable de la herramienta:

- **Relación con los Objetivos:** Aunque el objetivo es mejorar la satisfacción del cliente, es crucial considerar las consecuencias que se podría presentar por el uso del modelo a largo plazo, especialmente en lo que respecta el promover el acceso equitativo e informado a seguros y tratamientos médicos.
- **Ética en la Implementación:** Se sugiere crear un delegado de los datos, experto en salud de la agencia de seguro, con la que junto al cliente corporativo se evalué cómo se usa el modelo y proponer mejoras si detecta efectos negativos. Así se asegura un uso justo y responsable de la herramienta (Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., ... & Vayena, E., 2018).

2. Estado del Arte

2.1. Marco teórico

2.1.1. Análisis descriptivo de datos

El análisis descriptivo de datos es un método utilizado para resumir e identificar las características de un conjunto de datos. Este análisis nos permite observar patrones en el comportamiento y posibles tendencias que pueden ser significativas, sin embargo, para esto es necesario complementar el análisis posteriormente. Esto se debe a que este tipo de métodos proporciona una comprensión de los datos observados sin hacer suposiciones adicionales o inferencias que puedan ser generalizaciones acerca de la población.

Entre los beneficios de este análisis se encuentra que permite una mejor interpretación rápida de los datos ya que permite comprender su estructura y ciertos patrones que refleja. Facilita la presentación de los datos ya que los agrupa de manera ordenada y comprensible.

Por otra parte, también simplifica la presentación de la información ya que la vuelve más manejable, permitiendo una mejor comprensión, así como destaca aspectos claves de la distribución y comportamiento de las variables observadas.

Entre las herramientas más comunes para poder tener una rápida comprensión de la información se encuentran los gráficos y las tablas, ya que es una forma efectiva de representar los datos. Estos se incluirán en el dashboard que se presentará como parte final de este proyecto.

Las medidas de tendencia central más empleadas se pueden expresar a través de las siguientes formulas (Mendenhall, 2002):

Mediana

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad (1)$$

Mediana

Para calcular la mediana es importante ordenar los datos de forma ascendente

$$Mediana = x_{\frac{n+1}{2}} \quad (2)$$

Moda

Es el valor más frecuente.

Varianza

$$s^2 = \frac{1}{n-1}\sum_{\{i=1\}}^{\{n\}}(x_i - \bar{x})^2 \quad (3)$$

Por otra parte, el análisis descriptivo es crucial en el proceso de limpieza de datos, ya que ofrece una visión clara y detallada de los datos, lo que facilita la identificación y corrección de anomalías. Ayuda a identificar valores atípicos o anomalías que podrían indicar errores en la recopilación de datos, como valores extremos o inconsistencias que deben de ser tomadas en cuenta y corregidas para que no afecte la calidad del estudio (Chan, Lee, & Zainol, 2021).

Por otra parte, también permite detectar datos faltantes y comprender su extensión y distribución. Dependiendo del caso, quien maneja los datos deberá de escoger el mecanismo para el tratamiento de los datos. Entre las formas en las que pueden manejar los mismos se encuentran:

Eliminación de las observaciones con datos faltantes: Se eliminan las filas que cuenten con datos faltantes. Entre las desventajas de este mecanismo se encuentra que se puede perder información importante sobre otras variables.

Eliminación de variables que cuenten con datos faltantes: Se reduce la posibilidad de que los resultados solo representen a un grupo de la muestra y no a toda la población, puesto que no todos cuentan con registros. Sin embargo, se podría estar eliminando una variable con alto poder explicativo para el fenómeno estudiado.

Imputación de datos: Este mecanismo consiste en aplicar medidas de tendencia central para poder completar los datos faltantes.

Emplear fuentes externas: En algunos casos se puede recopilar información de otras fuentes para poder completar la información incompleta.

Retomando el tema del análisis descriptivo, una vez que se ha realizado el debido tratamiento de los datos, este también facilita que los datos sean consistentes al comparar diferentes variables y asegurar que los datos sigan reglas lógicas o formatos esperados. Es decir, se encuentren en el tipo de datos correspondientes, así como tengan sentido una vez que sean ubicadas en un contexto.

Una vez que se hayan ajustado los datos, permite generar una base para poder extraer información estratégica y el camino a seguir para poder determinar el enfoque del estudio. Sin embargo, para esto también es importante observar la distribución de datos de tal forma que se pueda identificar en etapas tempranas cómo es la distribución de datos. De esta forma se podrá identificar sesgos o comportamientos atípicos que podrían afectar al estudio.

En resumen, el análisis descriptivo actúa como una herramienta fundamental en la limpieza de datos al ofrecer una comprensión inicial y detallada que ayuda a identificar y corregir errores, inconsistencias y problemas en el conjunto de datos. (Kandula, Kalyanapu, Rayapalli, Modugumudi, & Kanikella, 2024).

Debido a su naturaleza, permite obtener los hallazgos principales de datos. El análisis descriptivo de datos es esencial para iniciar la exploración de la información ya que facilita el entendimiento de los datos y crea una línea base para profundizar el estudio sobre ciertos factores o variables, así como realizar inferencias y estadísticas más elaboradas.

2.1.2. Métodos actuariales

Las ciencias actuariales desempeñan un papel relevante en el cálculo del valor de la prima de un seguro combinando métodos estadísticos y financieros. Estas herramientas son empleadas a la interna por parte de las aseguradoras para establecer planes para sus clientes.

Los actuarios utilizan datos históricos y modelos estadísticos para evaluar los riesgos asociados con la póliza de seguro en cuestión. Esto implica examinar factores como la probabilidad de un siniestro que representa una cobertura, la gravedad del evento y la frecuencia con la que se realizarán los reclamos.

Basándose en datos históricos y probabilidades de ocurrencia, los actuarios desarrollan modelos matemáticos para prever las tendencias de los reclamos y su comportamiento futuro. Estos modelos pueden incluir distribuciones de probabilidad para estimar la frecuencia y el gasto deducible que esto implica, así como ajustes por tendencias históricas y factores externos que puedan afectar las reclamaciones.

Una vez que se ha evaluado el riesgo, los actuarios utilizan técnicas estadísticas para determinar el monto de la prima que debe cobrarse a los asegurados. Esto implica considerar factores como el riesgo esperado, los costos administrativos y de gestión, los gastos generales y un margen que representen una reserva para asegurar la solvencia financiera de la aseguradora.

El precio del seguro es una compensación que se hace por parte del asegurado hacia la aseguradora por motivo de la transferencia de riesgo que esta empresa está asumiendo.

Los principios fundamentales actuariales, como el principio de suficiencia de la prima (la prima debe ser suficiente para cubrir las reclamaciones y gastos), el principio de equidad (los asegurados similares deben pagar primas similares), y el principio de prudencia financiera (asegurar la estabilidad financiera de la aseguradora a largo plazo), guían el proceso de cálculo de la prima.

Finalmente, se procede a realizar la revisión de las primas, puesto que estas pueden ajustarse periódicamente en función de cambios en los factores de riesgo, experiencia de reclamaciones y condiciones del mercado. Los actuarios revisan regularmente las primas para asegurar que sigan siendo adecuadas y competitivas.

En resumen, las ciencias actuariales permiten calcular el valor de la prima de un seguro al evaluar y cuantificar los riesgos asociados con la cobertura asegurada, utilizando técnicas estadísticas avanzadas y principios actuariales para determinar un precio justo y suficiente que equilibre los intereses de los asegurados y la solvencia de la aseguradora.

(Lurie, 2007) menciona mecanismos empleados por las aseguradoras, los cuales emplean tanto el conocimiento de los expertos como técnicas estadísticas. Entre ellos los que se muestran a continuación.

Método de la escalera de cadena (Chain ladder method)

Este método es uno de los más utilizados para estimar reservas de siniestros. Se basa en datos históricos de siniestros y ajustarlos para prever las reclamaciones futuras. Utiliza patrones observados en datos pasados para estimar cómo evolucionarán los siniestros aún pendientes de liquidación en el futuro. Estos patrones se aplican a los siniestros reportados, pero no cubiertos por el seguro, de tal forma que permita calcular las reservas necesarias. Es

particularmente efectivo cuando hay suficientes datos históricos y consistencia en los patrones de desarrollo de siniestros a lo largo del tiempo.

Métodos basados en pagos por reclamación incurrida (Incurred claim payments methods)

Estos métodos calculan las reservas basándose en los pagos reales que la aseguradora ha realizado por reclamaciones incurridas. Puede ser en tiempo ordinario, es decir, estima las reservas basándose en los pagos realizados hasta la fecha, ajustados por factores esperados. Otro caso es en tiempo operacional, lo que considera los pagos realizados y proyecta los pagos futuros esperados utilizando modelos de siniestralidad y otras técnicas actuariales.

Método del ratio de pérdidas definitivo

Es la relación entre las pérdidas incurridas y las primas ganadas después de que todos los siniestros han sido liquidados.

Bornhuetter-Ferguson

Combina aspectos del método del ratio de pérdidas definitivo con una estimación inicial de siniestros incurridos, pero no reportados (IBNR) utilizando una distribución de probabilidad.

Métodos basados en estimaciones por caso (Case estimate based methods)

Estos métodos utilizan juicio experto o modelos específicos para estimar individualmente cada caso de siniestro. Son útiles cuando los datos históricos son limitados o cuando se requiere una evaluación detallada de casos específicos que pueden ser atípicos o de alta complejidad.

La selección del método adecuado será acorde a la disponibilidad y calidad de los datos, así como de las características específicas del portafolio de seguros y las necesidades de la aseguradora.

2.1.3. Regresión lineal

La regresión lineal permite predecir, a través de relaciones, cuál será el comportamiento de una variable dependiente acorde a una o más variables independientes. Es decir, debido a cómo se encuentra construido el modelo, hay una relación entre las variables que puede ser descrita mediante una combinación lineal de las variables independientes (Andika, Putra, Citra Lesmana, & Purnaba, 2021).

La forma de esta regresión es la siguiente (Stock & Watson, 2002):

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon \quad (4)$$

Donde:

y = variable dependiente

β = coeficientes de regresión

x = variables independientes

ϵ = término de error

Adicionalmente, para calcular los coeficientes se debe de aplicar la siguiente fórmula:

$$\beta = (X^T X)^{-1} X^T y \quad (5)$$

Donde:

X = matriz de las variables independientes

y = vector de la variable dependiente

La regresión lineal se basa en una serie de supuestos que estrictamente deben de validarse para garantizar que los resultados del modelo sean válidos. Estos supuestos permiten obtener coeficientes estimados precisos y que los intervalos de confianza y pruebas de hipótesis sean válidos. Los principales supuestos de la regresión lineal son:

Entre los supuestos de la regresión lineal se encuentran:

Linealidad: Hay una relación lineal entre la variable dependiente y las independientes. Es decir, si una variable independiente cambia, esto implicaría un cambio proporcional en la variable dependiente.

Independencia de los Errores: Los errores o residuos del modelo no deben de estar relacionados entre sí en las observaciones.

Homocedasticidad: La varianza de los errores se debe de mantener constante para las observaciones en el caso de las variables independientes. Esto significa que los errores tienen una dispersión constante a lo largo del rango de las variables independientes.

Normalidad de los Errores: Los errores muestran un comportamiento de una distribución normal. Esto es importante para las pruebas de hipótesis y la construcción de intervalos de confianza.

No Multicolinealidad: No existe una fuerte correlación entre las variables independientes.

Especificación Correcta del Modelo: El modelo debe estar correctamente especificado, es decir, debe incluir todas las variables que puedan aportar al modelo. Un modelo mal especificado puede llevar a estimaciones sesgadas e ineficientes.

Una vez cumplidos estos supuestos, los coeficientes del modelo proporcionan una interpretación directa de cómo cada variable independiente puede tener un impacto en la variable dependiente.

Los algoritmos para ajustar modelos de regresión lineal son relativamente simples y rápidos en cuanto a términos computacionales.

Por otra parte, es importante observar la relación que guardan las variables ya que para que el modelo se ajuste correctamente, esta debe de ser lineal. El modelo de regresión lineal tiene problemas para poder capturar y predecir estructuras más complejas. Además, los valores extremos pueden influir significativamente en los parámetros del modelo, lo que puede afectar la robustez y precisión de las predicciones.

2.1.3.1. Métricas para Evaluar la Regresión Lineal

Entre las métricas empleadas para evaluar el desempleo de una regresión lineal simple se encuentran las siguientes:

Coeficiente de Determinación (R^2)

Evalúa la variabilidad total explicada por el modelo. Mientras más alto sea el indicador R^2, mejor es el ajuste del modelo.

Para usar un R^2 se debe de considerar que:

- El R^2 siempre aumenta cuando se agregan más variables independientes al modelo, incluso si estas variables no son realmente relevantes. Esto

puede dar una falsa impresión de que el modelo está mejorando simplemente por incluir más variables.

- El R^2 mide la relación entre dos variables, pero esto no implica la presencia de causalidad. Es decir, puede haber un R^2 alto, pero esto no implica que una variable está explicando a la otra o que una modificación en la variable independiente sea lo que está causando cambios en la variable dependiente.

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} \quad (6)$$

Error Cuadrático Medio (MSE)

Promedia los cuadrados de las diferencias entre los valores reales y los valores obtenidos a partir del modelo. Sirve para cuantificar la precisión del modelo en relación a la variabilidad de los errores. El hecho de elevar los errores al cuadrado permite obtener términos absolutos y que valores menores a cero no sean anulados por valores superiores.

Raíz del Error Cuadrático Medio (RMSE)

Explica la magnitud de los errores en las mismas unidades que la variable dependiente, calculando la raíz cuadrada del MSE. Esto permite generar mayor interpretabilidad y una mejor medición de la magnitud del error.

Error Absoluto Medio (MAE)

Promedia las variaciones absolutas entre los valores reales y los valores propuestos por el modelo. Ofrece una medida de la magnitud promedio de los errores.

Estadístico F

Utilizado para evaluar la hipótesis de los coeficientes del modelo son iguales a cero. Un valor alto del estadístico F sugiere que al menos una variable independiente tiene una relación significativa con la variable dependiente.

Pruebas de Hipótesis sobre Coeficientes

Se utilizan para analziar si los coeficientes individuales son significativamente distintos de cero, lo que indica que las variables independientes si mantienen una relación para determinar la variable dependiente.

Análisis de Residuos

Examina las diferencias entre los valores reales y los determinados por el modelo para identificar patrones que puedan sugerir problemas como no linealidad, heterocedasticidad (variación no constante de los errores), o la presencia de valores atípicos.

Estas métricas y características son fundamentales para evaluar la adecuación y la precisión de un modelo de regresión lineal, y para garantizar que el modelo sea robusto.

2.1.4. k-Nearest Neighbors

k-Nearest Neighbors (k-NN) es conocido por ser un modelo de aprendizaje supervisado el cual permite realizar clasificaciones como regresiones.

El algoritmo k-NN asigna una categoría a un dato basándose en el comportamiento de sus k vecinos más cercanos en el conjunto de entrenamiento. La etiqueta más frecuente entre estos vecinos es la asignada al punto. De esta forma, k-NN predice el valor de una observación calculando la media (o mediana) de los valores de las observaciones que se encuentran próximas. (Kafuria, 2022)

Este método no requiere estrictamente alguna distribución de los datos, lo que facilita su uso y accesibilidad para distintos conjuntos de datos. No asume ningún comportamiento específico para los datos. El modelo se adapta a la estructura de los datos.

Además, puede ser utilizado tanto para problemas de clasificación como de regresión, lo que amplía su uso. Además, de que es un modelo no paramétrico, lo que implica que no requiere una estructura rígida, por lo que el modelo puede captar relaciones complejas y no lineales.

El algoritmo k-NN utiliza directamente los puntos de datos del conjunto de entrenamiento. Las predicciones se basan en la similitud con estos puntos en función de la distancia.

Dado que k-NN no requiere un modelo ajustado previamente, puede adaptarse fácilmente a nuevos datos. Cuando se presentan nuevos puntos de datos, el algoritmo simplemente calcula la distancia a los puntos de entrenamiento y ajusta la predicción en función de los vecinos más cercanos. Esto facilita la actualización del modelo cuando existen nuevos datos en la muestra.

La eficacia del k-NN puede verse disminuida en alta dimensión ya que las distancias entre puntos tienden a asemejarse, lo que puede hacer que el algoritmo sea menos discriminativo y disminuya su rendimiento. En alta dimensión, las distancias entre puntos se vuelven menos diferenciadas e intuitivas.

El t-SNE (Stochastic Neighbor Embedding) permite reducir la dimensionalidad, conservando las relaciones de proximidad entre los puntos. Técnicas como esta o el Análisis de Componentes Principales (PCA) facilitan la visualización al ser un espacio más manejable.

Dado que la distancia entre puntos se vuelve similar, el algoritmo k-NN tiene dificultades para distinguir entre vecinos cercanos y lejanos. La capacidad del algoritmo para identificar y utilizar los vecinos más cercanos se ve reducida, lo que puede llevar a un desempeño deficiente.

Adicionalmente, los modelos de alta dimensionalidad afectan a la interpretabilidad del modelo. Las relaciones entre las variables no son fácilmente observables, lo que conduce a que la interpretación del modelo sea menos clara.

Una vez seleccionada la dimensión, también es importante considerar el número de vecinos. Pocos vecinos puede llevar a un modelo sobreajustado, mientras que un número demasiado grande puede generalizar el modelo y que pierda detalles importantes. Asimismo, la elección de la métrica de distancia (como Euclidiana, Manhattan, etc.) puede afectar el rendimiento del modelo. La selección debe estar alineada con la naturaleza de los datos.

Otro tema sensible para el modelo k-NN son los valores atípicos dado que estos pueden influir significativamente en la clasificación o predicción si están cerca del punto de consulta, sesgando los resultados del modelo. Por otra parte, el rendimiento del k-NN puede verse afectado si las características tienen escalas diferentes, lo que implicaría otra posible fuente de sesgo. Es crucial normalizar o estandarizar las características para asegurar que todas tengan una contribución equitativa en el cálculo de distancias.

Métricas para Evaluar el Rendimiento del k-NN

Precisión

Para problemas de clasificación, esta métrica brinda explica la capacidad para clasificar correctamente las observaciones, especialmente en situaciones donde los costos de los falsos positivos son altos.

Es la proporción de positivos correctamente estimados en relación al total de instancias clasificadas como positivas por el modelo. Es decir, el total de verdaderos positivos sobre todas aquellas observaciones marcadas como positivas.

Recall

Son las instancias positivas correctamente identificadas respecto al total de positivos que se han observado. En este caso, se calcula considerando también a los falsos negativos.

Su cálculo se basa en la división de positivos reales sobre número de positivos reales que existen, ya sea que el modelo los haya clasificado correctamente o no. En este caso se suele usar más cuando los falsos negativos generan alta preocupación.

F1 Score

Es una métrica útil cuando es necesario un balance entre precisión y recall, y cuando el desbalance entre clases puede hacer que una de las métricas por sí sola sea engañosa.Se emplea cuando se requiere evaluar el modelo combinando tanto el recall como la precisión dado que ambos son importantes.

Error Absoluto Medio (MAE) y Error Cuadrático Medio (MSE)

Al igual que en la regresión lineal, estas métricas miden la magnitud de los errores de predicción, proporcionando una idea clara de la precisión del modelo.

Matriz de Confusión

Permite una evaluación detallada del rendimiento del modelo, dado que permite visualizar qué proporción de positivos y negativos han sido clasificados adecuadamente, así como los errores que han existido. Se ubican tal como se muestra en la Tabla 2.

Tabla 2 Matriz de confusión

	Valores reales		
Valores predichos		Positivo	Negativo
	Positivo	Verdadero positivo	Falso positivo
	Negativo	Falso negativo	Verdadero negativo

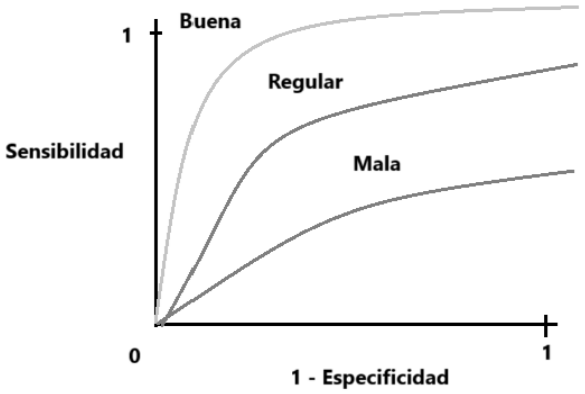
Fuente: (Syarifah & Herdianto, 2023)

Es una herramienta que analiza el desempeño de un modelo de clasificación al mostrar cómo se comparan las predicciones con las etiquetas reales.

Curvas ROC y AUC

La Curva ROC, la cual se puede apreciar en la figura 2, muestra, de forma visual, cómo el modelo puede identificar las clases positivas y negativas a partir de los umbrales de la clasificación. La curva muestra Tasa de Verdaderos Positivos (TPR) frente a la Tasa de Falsos Positivos (FPR) para diferentes umbrales.

Figura 2 Curva ROC



Fuente: (Burgos & Manterola, 2010)

Por otra parte, la curva AUC es una medida cuantitativa del rendimiento de la modelo basada en el área bajo la curva ROC.

Finalmente, se aplica la validación cruzada, lo que permite evaluar tanto su capacidad de generalización como su estabilidad.

En cuanto a los requerimientos computacionales, el rendimiento del algoritmo puede verse afectado cuando se trabaja con datos que sean extensos ya que se debe calcular distancias entre el punto de consulta y todos los puntos del conjunto de entrenamiento.

En resumen, k-NN es un método flexible y sencillo para clasificación y regresión, pero su eficacia puede verse limitada por la escalabilidad, la elección de parámetros y la dimensionalidad de los datos.

2.1.5. Árboles de decisión

Los árboles de decisión son una técnica empleada para poder realizar predicciones a partir de regresiones, así como clasificaciones. Se utilizan para modelar decisiones, visualizando el proceso en una estructura de árbol. (Ismail, Stam, Portrait, & Van Witteloostuijn, 2024)

El nodo inicial del árbol, donde comienza el proceso de decisión. Representa el conjunto completo de datos y se divide en nodos secundarios basados en una característica.

Los nodos internos representan decisiones basadas en una característica específica. Cada nodo interno realiza una división de los datos en función de una condición. Finalmente, los nos nodos finales del árbol que representan el resultado de la decisión. En clasificación, indican la clase a la que se asigna una instancia, y en regresión, el valor continúo predicho.

Las ramas conectan los nodos y representan las decisiones o reglas basadas en las características de los datos. Cada rama lleva a un nodo interno o a un nodo hoja.

La división en un nodo se realiza en función de una característica y un umbral que maximiza alguna medida de calidad, como la ganancia de información o el índice Gini. Cada división tiene como objetivo reducir la impureza o la variabilidad en las hojas.

El número de divisiones en un árbol de decisión tiene un impacto significativo en su rendimiento y capacidad de generalización. Un número inadecuado de divisiones puede llevar a problemas específicos en el modelo.

Cuando un árbol de decisión tiene demasiadas divisiones, puede capturar muy específicamente el comportamiento de los datos que ha observado, pero no el de los reales. Esto se traduce en un modelo que tiene un excelente rendimiento en el conjunto de entrenamiento, en el caso de los datos no vistos (de prueba) puede ser muy bajo debido a su incapacidad para generalizar. Adicionalmente, los árboles con muchas divisiones pueden volverse muy complejos y difíciles de interpretar. Esto puede hacer que el modelo sea menos accesible para la comprensión y el uso del modelo.

Por otra parte, también existe el problema de underfitting. Si un árbol tiene muy pocas divisiones, no capturará la complejidad del patrón de los datos. Esto resulta en un modelo que es demasiado simple para representar adecuadamente la relación entre las características y las etiquetas. Un árbol con pocas divisiones puede hacer predicciones demasiado generales, ignorando detalles importantes de los datos. Esto puede resultar en una baja precisión, especialmente si hay interacciones complejas entre las características. Lo mismo se traduce en una pérdida de información crucial que podría mejorar la precisión del modelo.

La sensibilidad de los árboles de decisiones es una característica importante que se debe de tener en cuenta antes de emplear este modelo. Se ha evidenciado que variaciones leves en los datos pueden resultar en árboles muy diferentes, ya que el modelo es sensible a los cambios en los datos de entrenamiento.

Otra posible limitante de los árboles de decisiones es que tienden a preferir características con más valores únicos, lo que puede no ser ideal si las características no son realmente relevantes. Los árboles de decisión pueden asimilar interacciones complejas y relaciones no lineales entre variables sin la necesidad de transformación explícita.

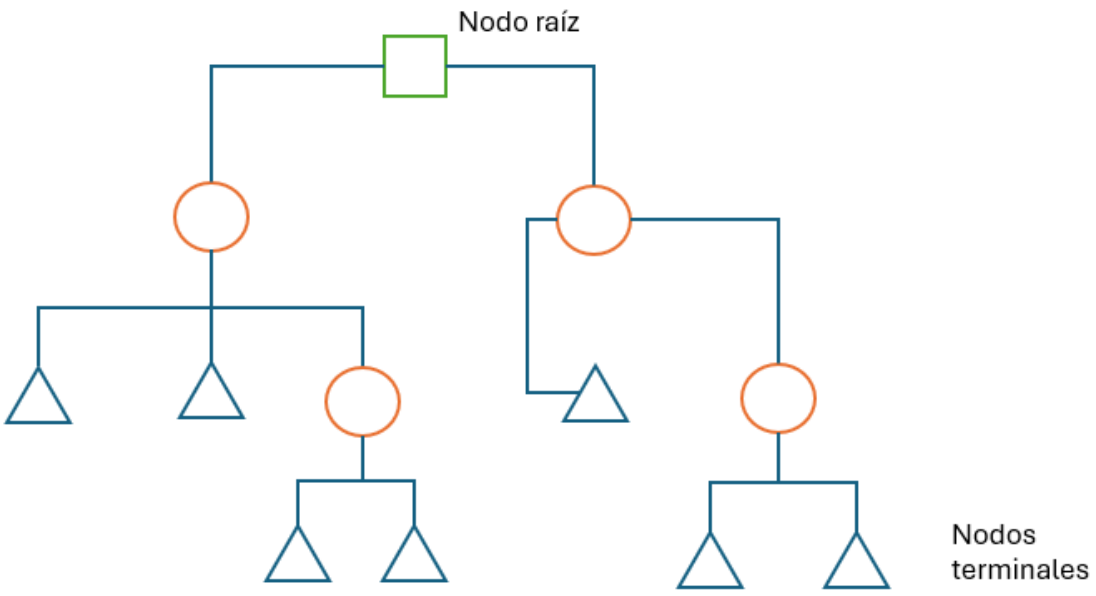
Para poder crear un árbol de decisión es fundamental tener claro los criterios de división. Cada nodo se divide en función de un criterio que mide la calidad de la división tal

como se muestra en la figura 3. Para clasificación, se utilizan medidas como la ganancia de información, índice Gini, o entropía. Para regresión, se utilizan medidas como la varianza. El fin es poder tener la mejor división de los datos en cuanto a esta métrica. De esta forma se podrán crear grupos más homogéneos que permitan clasificar a los datos.

El proceso de división se aplica recursivamente a cada nodo interno hasta que se cumplen ciertos criterios, como un máximo de niveles o una mejora leve en la métrica.

Finalmente, en un modelo de clasificación, cada observación se asigna a la clase que aparece con mayor frecuencia en la hoja a la que llega. Si el modelo es de regresión, se asigna el promedio de los valores de respuesta en las observaciones que llegan a esa hoja.

Figura 3 Árbol de decisión



Fuente: (Molano, Rodríguez, Valera, & Vanegas, 2017)

Cada nodo equivale a una consulta sobre las características de los datos, permitiendo que estos puedan ser clasificados y de esta forma determinar la probabilidad de ocurrencia.

Uno de los motivos por los que los árboles de decisión son altamente empleados es su facilidad de interpretación y visualización. Se pueden entender las decisiones del modelo siguiendo el camino de la raíz a las hojas, lo que facilita la explicación del modelo a los usuarios.

Por otra parte, en cuanto a su procesamiento, los árboles de decisión no necesitan que las variables sean escaladas o normalizadas, ya que las divisiones se basan en comparaciones de valor. Además, pueden manejar tanto variables categóricas como continuas. Es decir, es apto para diferentes tipos de datos.

Métricas para Evaluar Árboles de Decisión

Ganancia de Información

Mide la reducción en la entropía o la impureza de los datos después de una división.

Índice Gini

Evalúa la impureza de un nodo basado en la probabilidad de una clasificación incorrecta.

Error de Predicción

Calcula la tasa de error entre los datos de prueba, para estimar la capacidad de generalización del árbol.

Importancia de Características

Mide la contribución de cada característica en la toma de decisiones dentro del árbol.

Los árboles de decisión son un mecanismo poderoso y comprensible para clasificación y regresión, capaces de manejar diferentes tipos de datos y modelar relaciones complejas. Sin embargo, son propensos al sobreajuste y pueden ser inestables.

2.1.6. Random Forest

Random Forest es un algoritmo basado en el ensamblaje de varios árboles de decisión para realizar tareas de clasificación y regresión. Este modelo ayuda a incrementar la precisión y estabilidad de las predicciones. Permite manejar el desbalance entre los datos, evaluar la importancia de características y reducir el riesgo de sobreajuste. (Kemboi, Kasozi, & Nkurunziza, 2021)

Cada árbol se entrena con una muestra distinta de los datos observados. En cada nodo de un árbol, se selecciona aleatoriamente un grupo de características para decidir la mejor división. Esto permite reducir la correlación y aumentar la diversificación de los árboles.

Los árboles en el bosque se construyen utilizando su propio subconjunto de datos y subconjunto aleatorio de características.

Para problemas de clasificación, cada árbol selecciona a que categoría pertenece la observación. La clase con mayores características similares se muestra como la predicción final. En el caso de problemas de regresión, la predicción final es el resultado del promedio de los resultados por cada árbol.

Al promediar las predicciones de múltiples árboles, Random Forest suele ofrecer una mayor precisión y una mejor generalización en comparación con un solo árbol de decisión. Asimismo, esto permite que tenga una ventaja al momento de enfrentar problemas como el sobreajuste ya que se vuelve menos susceptible al mismo, además de que es más robusto frente a variaciones de los datos. La agregación de múltiples árboles suaviza las fluctuaciones y errores individuales.

Random Forest puede manejar mejor los datos desbalanceados, ya que la técnica de agregación ayuda a mitigar el impacto de las clases mayoritarias. Además, también puede tratar datos categóricos como continuos.

Al igual que en los árboles de decisión, se pueden emplear métricas de evaluación como la precisión, recall, las curvas ROC, AUC y la validación cruzada.

Para elaborar un Random Forest se necesita considerar varios parámetros, entre ellos el número de estimadores, features, profundidad máxima y mínimo para la división.

Se debe de especificar el número de árboles, dado que incrementarlos puede mejorar la precisión y estabilidad. En cuando al número de features, este se refiere al número máximo de rasgos consideradas para generar un nodo por árbol. Un valor más bajo introduce más aleatoriedad y puede mejorar la generalización. Sin embargo, un valor muy bajo puede llevar a que los árboles individuales no utilicen toda la información con la que se cuenta en los datos para tomar decisiones, reduciendo la capacidad del modelo para capturar patrones complejos que pueden explicar la realidad de los datos.

En cuanto a la profundidad máxima y el número mínimo de observaciones requeridas para crear una hoja del árbol, estas pueden controlar el sobreajuste si es escogida adecuadamente.

Estos factores permiten modelar la complejidad del modelo y por ende, su capacidad para explicar los datos de la vida real.

Entre las principales limitantes del Random Forest se encuentra la capacidad computacional que este requiere. Es decir, es complejo cuando se lo emplea en grandes cantidades de datos o con varias particiones, lo que podría complicar su actualización si se

introduce nueva información. Además, la predicción puede ser más lenta en comparación con modelos más simples, ya que requiere la evaluación de todos los árboles del bosque para cada instancia.

Otra limitante es que, a pesar de que los árboles individuales son fáciles de interpretar, el conjunto completo de Random Forest puede ser menos comprensible y más difícil de comprender debido a su complejidad.

2.1.7. XGBoost

XGBoost es una optimización del Gradient Boosting que utiliza árboles de decisión como modelos base. Cabe recalcar que el Gradient Boosting genera modelos secuencialmente de tal forma que se van corrigiendo los modelos anteriores mediante el ajuste de los residuos. La idea principal de boosting es combinar modelos generando un nuevo modelo que mejore la precisión y la capacidad de generalización.

XGBoost crea con un modelo base simple. A medida que aumentan las iteraciones, un nuevo árbol se construye para predecir los errores residuales del modelo anterior. El árbol es ajustado para corregir las predicciones erróneas y minimizar el error en los datos de entrenamiento.

Este modelo optimiza una función de pérdida específica mediante el gradiente descendente, ajustando los pesos de los árboles para minimizar el error de predicción.

La función de pérdida penaliza árboles complejos, de tal forma que evitar el sobreajuste. Esto ayuda a mejorar la capacidad de generalización.

Finalmente, los resultados de todos los árboles construidos se combinan para hacer la predicción final. Cada nuevo árbol corrige al anterior, mejorando la capacidad predictiva del modelo. Durante el entrenamiento, XGBoost ajusta los pesos de cada instancia para enfocar el aprendizaje en las instancias mal clasificadas por los árboles anteriores.

XGBoost a menudo supera a otros algoritmos en términos de precisión y rendimiento en una variedad de problemas. Su capacidad para combinar múltiples árboles y optimizar el ajuste es clave para este alto rendimiento. (Orji & Ukwandu, 2024)

Entre las ventajas para poder aplicar este algoritmo se encuentra su capacidad para ajustar numerosos parámetros para adaptarse a diferentes tipos de situaciones, contextos y datos.

La amplia gama de parámetros que se pueden ajustar puede hacer que la configuración de XGBoost sea diversa. Requiere tiempo y experiencia para encontrar la mejor combinación de parámetros que permitan encontrar al mejor modelo acorde a la situación que se busca representar. Aunque XGBoost incluye mecanismos de regularización, un ajuste inadecuado de los parámetros puede llevar a sobreajuste, especialmente con un número muy alto de árboles o una profundidad excesiva.

Aunque XGBoost proporciona medidas de importancia de características, el modelo en sí puede ser menos interpretable en comparación con modelos más simples como los árboles de decisión individuales.

Entre los parámetros que se toman en consideración se encuentra el número de árboles que se construirán. Cabe destacar genera un trade-off entre precisión y sobreajuste. Por otra parte, también está la tasa de aprendizaje. Mientras más baja, se necesitarán más árboles para converger, pero puede mejorar la precisión y evitar el sobreajuste. Por otra parte, profundidades mayores permiten capturar relaciones más complejas.

Al igual que en los árboles anteriores, se requiere un número mínimo de muestras para generar un nodo.

De manera similar, también se requieren de métricas de evaluación tales como precisión, recall, F1-score, y AUC-ROC en caso de clasificaciones. Para regresión, se utilizan métricas como el MSE o el coeficiente R^2. Se emplea la validación cruzada para medir el rendimiento y ajustar los parámetros. Esto ayuda a garantizar que el modelo generalice a datos no observados.

A diferencia de modelos anteriores, XGBoost es eficiente en términos de tiempo de entrenamiento y uso de recursos. Utiliza técnicas de optimización como la poda de árboles y la paralelización para acelerar el proceso de entrenamiento. Sin embargo, puede ser complejo de configurar y menos interpretable que modelos más simples.

2.1.8. Redes neuronales

Las redes neuronales están diseñadas para identificar patrones, realizar predicciones y aprender de datos complejos. Funcionan como sistemas de procesamiento de información compuestos por unidades básicas llamadas neuronas que se organizan en capas. Cada neurona está conectada a las neuronas en la capa siguiente s través de ponderaciones. Estas redes se utilizan para modelar y resolver problemas en áreas como la clasificación, regresión, reconocimiento de patrones y más. (Amin, Liu, & Kensaku, 2020)

Su componente principal, las neuronas, son unidades de procesamiento que reciben una serie de entradas, aplican una función de activación a las ponderaciones respectivas de estas y producen una salida.

Las capas de entrada son las que reciben los datos y los comparten a las siguientes capas. De esta forma llegan a las capas ocultas, que procesan esta información y pueden tener una o más capas en una red. Cada capa oculta aplica funciones de activación a las

combinaciones ponderadas de las entradas, para luego, en las capas de salida, brindar predicciones o clasificaciones.

Por su parte, las funciones de activación introducen no linealidades en el modelo, permitiendo a las redes neuronales generar relaciones que son más complejas de modelar. Ejemplos incluyen la función ReLU (Rectified Linear Unit), la función sigmoide y la tangente hiperbólica (*tanh*).

La función ReLU convierte todas las entradas negativas en cero y deja las entradas positivas sin cambios. En otras palabras, cualquier valor de entrada menor o igual a cero se convierte en cero, mientras que cualquier valor mayor que cero se mantiene igual.

La función sigmoide genera valores reales a un rango entre 0 y 1. A medida que el valor de entrada sea mayor, este se acercará mas a 1, caso contrario, será más próximo a 0.

La función *tanh* genera valores reales perteneciente al rango entre -1 y 1. Es similar a la función sigmoide, pero su rango puede ser beneficioso para capturar tanto valores positivos como negativos.

Los pesos se ajustan durante el entrenamiento para minimizar el error del modelo. La función de pérdida es la encargada de medir la diferencia entre las predicciones y las verdaderas etiquetas. El objetivo del entrenamiento es disminuir esta pérdida todo lo que sea posible.

El proceso de ajuste de los pesos de la red para minimizar la función de pérdida se realiza mediante algoritmos, como el descenso de gradiente y sus variantes (como Adam o RMSprop).

La ventaja de las redes neuronales frente a los modelos anteriores es que estas pueden adaptarse a una variedad de problemas y tipos de datos, incluyendo datos estructurados, imágenes, texto y secuencias. Las redes neuronales han demostrado un excelente rendimiento en una serie de tareas específicas, entre ellas el reconocimiento de voz y la visión por computadora.

Además, las redes neuronales profundas (con múltiples capas) pueden escalar bien con grandes volúmenes de datos y pueden mejorar su rendimiento con más datos y más recursos computacionales. Estas pueden ser difíciles de interpretar debido a su naturaleza de "caja negra". Entender cómo se toman las decisiones puede ser complicado.

El rendimiento de las redes neuronales puede depender en gran medida de la elección de hiperparámetros, y encontrar la combinación óptima puede ser un proceso complejo y

costoso, por ende, se busca encontrar un modelo lo suficientemente bueno para el fenómeno estudiado.

2.2. Fuentes de datos relacionadas al problema

En cuanto a los reembolsos médicos, las aseguradoras poseen los registros de los gastos de las atenciones hospitalarias, coordinaciones de medicinas, prestadores de servicios, entre otros, para llevar un control de los gastos incurridos durante la vigencia de la póliza de asistencia médica, posteriormente evaluados cuando se presenta la renovación de contratos, los mismos ayudan a determinar la siniestralidad del cliente.

Las aseguradoras ofrecen a las empresas aliadas en cuanto al análisis de datos, bases con información de las corporaciones a la que brinda servicios la agencia de seguros, que son útiles para la generación de visualizaciones previas a la toma de decisiones sobre los planes ofertarles y el precio de los mismos.

Las bases de datos otorgadas por las aseguradoras vienen dadas en hojas de cálculos con extensión xlsx, cada una de ellas mantiene un formato diferente; por lo que para una mejor compatibilidad de la información ofrecidas de todas las aseguradoras se espera tener un mismo almacén de datos.

S3 (Simple Storage Service)

Es un servicio de almacenamiento en la nube desarrollado por Amazon, utilizado para acceder a grandes cantidades de datos (Amazon Web Services, s.f.). Ofrece escalabilidad y durabilidad al 99.99999999%, adicionalmente ofrece múltiples capas de seguridad, convirtiéndola en un excelente referente para trabajos de datos masivos. Así mismo, el costo-beneficio que este otorga a sus clientes la posiciona como gran referente para la industria de ciencia de datos.

Athena (AWS)

Permite realizar consultas interactivas acerca de datos almacenados en Amazon S3 utilizando el lenguaje de consultas estructuradas (SQL). Al ser completamente sin servidor, no requiere la implementación de una arquitectura. Athena se integra con AWS Glue para simplificar la creación de catálogos de datos y la preparación de estos (Amazon Web Services, s.f.). Además, que Athena, es ideal para consultas ad-hoc y la creación de informes, su método de cobro la hace por consulta ejecutada, basándose en la cantidad de datos procesados convirtiéndola en una herramienta con gran impacto en la gestión de proyectos de datos masivos.

2.3. Descripción de los datos

En el contexto de la naturaleza de los datos a utilizar para el entrenamiento del mejor modelo de aprendizaje automático, se debe considerar las siguientes definiciones de variables:

Diagnósticos médicos

Los diagnósticos médicos son fundamentales en el proceso de reembolso de una póliza de seguros. Para que una reclamación sea aprobada, es necesario que el diagnóstico esté bien documentado y coincida con las coberturas de la póliza. Los asegurados deben presentar tanto los informes de diagnóstico como las facturas médicas para que la aseguradora pueda verificar la validez de los gastos y proceder con el reembolso.

Red hospitalaria

La red hospitalaria consiste en un grupo de hospitales y centros médicos asociados con una aseguradora o sistema de salud. A través de esta red, los asegurados pueden acceder a diversos servicios médicos y especialistas en establecimientos aprobados por la aseguradora. Usar los servicios dentro de esta red puede simplificar el proceso de reembolso y, en muchos casos, proporcionar una cobertura más amplia y costos más reducidos para los pacientes

Proveedor de salud

Un proveedor de salud es cualquier entidad o profesional que brinda servicios médicos, como médicos, clínicas, hospitales y laboratorios. Estos proveedores colaboran con aseguradoras y sistemas de salud para ofrecer atención a los asegurados. Al recibir tratamiento de un proveedor de salud, los pacientes pueden utilizar su póliza de seguros para cubrir los gastos, según las coberturas y acuerdos con la aseguradora.

Tipo de servicio

La clasificación de los servicios cubiertos por la aseguradora, el tipo de servicio puede ser ambulatorio, hospitalario, coordinación de medicinas, órdenes de atención, entre otras.

Valor presentado

El total de las facturas incluidas en un único reembolso a veces permite coordinar con los proveedores de salud. En estos casos, las aseguradoras tienen acuerdos con dichos proveedores, por lo que las facturas no se presentan directamente por el afiliado, sino que son gestionadas entre la aseguradora y el proveedor de salud.

Valor liquidado

Representa el monto total asumido por la que la compañía de seguro con respecto al gasto presentado por el afiliado, teniendo en cuenta los montos de coberturas y en cumplimiento de haberse superado el deducible aplicado acorde a los lineamientos de la póliza de seguros.

Fecha registro

La fecha de servicio indicada en la solicitud de reembolso médico debe ser considerada como la más antigua entre las facturas presentadas; en ocasiones, esta fecha también se conoce como fecha de incurrencia.

Edad

Representa la edad del asegurado, ya sea del titular de la póliza o de sus dependiente.

Sexo

Representa el sexo del asegurado, ya sea del titular de la póliza o de sus dependientes.

Ciudad de residencia

Representa la ciudad de residencia del asegurado, ya sea del titular de la póliza o de sus dependientes.

Sector económico

Un sector económico agrupa actividades similares en la economía, como la extracción de recursos (primario), la manufactura (secundario), y la prestación de servicios (terciario). También incluyen sectores como el cuaternario (conocimiento) y quinario (decisiones y políticas).

Morbilidad

El nivel de morbilidad de los diagnósticos médicos en una región ayuda a las aseguradoras a ajustar las primas según la prevalencia de determinadas enfermedades, a crear programas de prevención y bienestar que traten las causas más comunes de

hospitalización, y ofrecer coberturas flexibles acordes a grupos demográficos. El nivel de morbilidad de estas enfermedades favorece en la predicción del comportamiento de aquellos diagnósticos en determinados grupos y por ende conocer las posibles enfermedades a presentarse por los titulares de las pólizas corporativas.

2.4 Librerías y software a utilizar

Python

El lenguaje de programación por emplear es Python, versión 3.11.9. Entre los usos que se le da a Python se encuentran la aplicación de ciencia de datos, desarrollo web y automatización (Van Rossum & Drake, 2009). En este caso, esta herramienta permitirá realizar análisis de datos, para esto es importante iniciar por la limpieza de los datos y la consolidación de las bases que se van a utilizar. De igual manera, se debe de validar que toda la información se encuentre en el formato correcto. Luego de esto, se procederá a realizar análisis descriptivo que permitirá comprender mejor la data, así como explorar si existen posibles relaciones que deberán de ser estudiadas a continuación o si se presentan características inusuales que deberán de ser tratadas de forma diferente, es decir, realizar un preprocesamiento. Una vez que se cuente con la data lista, se procederá a implementar modelos de Machine Learning a través de las distintas librerías con las que cuenta Python.

Power BI

Para poder visualizar los datos y mostrar la información a los usuarios de una forma más resumida, didáctica e intuitiva se empleará Power BI. Se elaborará un dashboard que recopile la información más relevante de nuestro modelo, así como características y estadísticas descriptivas que sean de interés para el usuario. Por ejemplo, las relaciones entre ciertas variables demográficas, tendencias, entre otros.

Matplotlib

Esta librería de Python permitirá crear gráficos exploratorios que permitirá tener una comprensión profunda de las tendencias que inicialmente se puedan observar. Este análisis exploratorio es fundamental para poder comprender la data y establecer posibles relaciones que podrían surgir y deberán de ser estudiadas a profundidad a través de los modelos para poder determinar si existe efectivamente una relación o no.

Seaborn

La librería Seaborn permitirá la elaboración de gráficos preliminares para poder mostrar las observaciones de una forma más atractiva y así tener una mejor comprensión de estos. Seaborn permite entender la data de una forma más sencilla y visual.

Numpy

Permite manipular los datos y crear arreglos que serán útiles para la exploración de los datos, así como para su uso dentro del modelo predictivo.

Pandas

Esta librería permitirá tratar los datos de una forma más sencilla. A su vez, esto da paso a que la limpieza y preparación de datos sea posible y el modelo pueda contar con la data necesaria en el formato requerido.

Scikit-learn

Scikit-learn es una biblioteca de machine learning en Python que está compuesta por varios algoritmos de clasificación, regresión, clustering y más. Sería utilizada para entrenar y evaluar modelos predictivos que puedan predecir el gasto deducible o basado en características demográficas y empresariales.

AWS Glue

Es un servicio en la nube de Amazon Web Services (AWS) diseñado para integrar datos y facilitar su preparación para el análisis. Este servicio permite a los usuarios descubrir, catalogar y transformar datos de diversas fuentes. AWS Glue proporciona un entorno sin servidores para ejecutar tareas de ETL (Extracción, Transformación y Carga), automatizando así la preparación de datos para su análisis.

AWS SageMaker

Es un servicio de AWS que simplifica el proceso de generación, optimización y despliegue de modelos. Ofrece una plataforma integral que permite preparar datos, elegir y ajustar modelos, y entrenarlos a gran escala. SageMaker incluye algoritmos preintegrados y soporte para frameworks populares, así como herramientas para desarrollar modelos personalizados. También facilita el despliegue de modelos en producción y la realización de inferencias en tiempo real o por lotes.

AWS Lambda

Permite ejecutar código en la nube. Los usuarios cargan su código y definen eventos que lo activan, como cambios en Amazon S3 o mensajes en Amazon SNS. Este servicio maneja automáticamente la infraestructura y la escalabilidad, ejecutando el código en respuesta a eventos de manera eficiente.

3. Diseño e Implementación

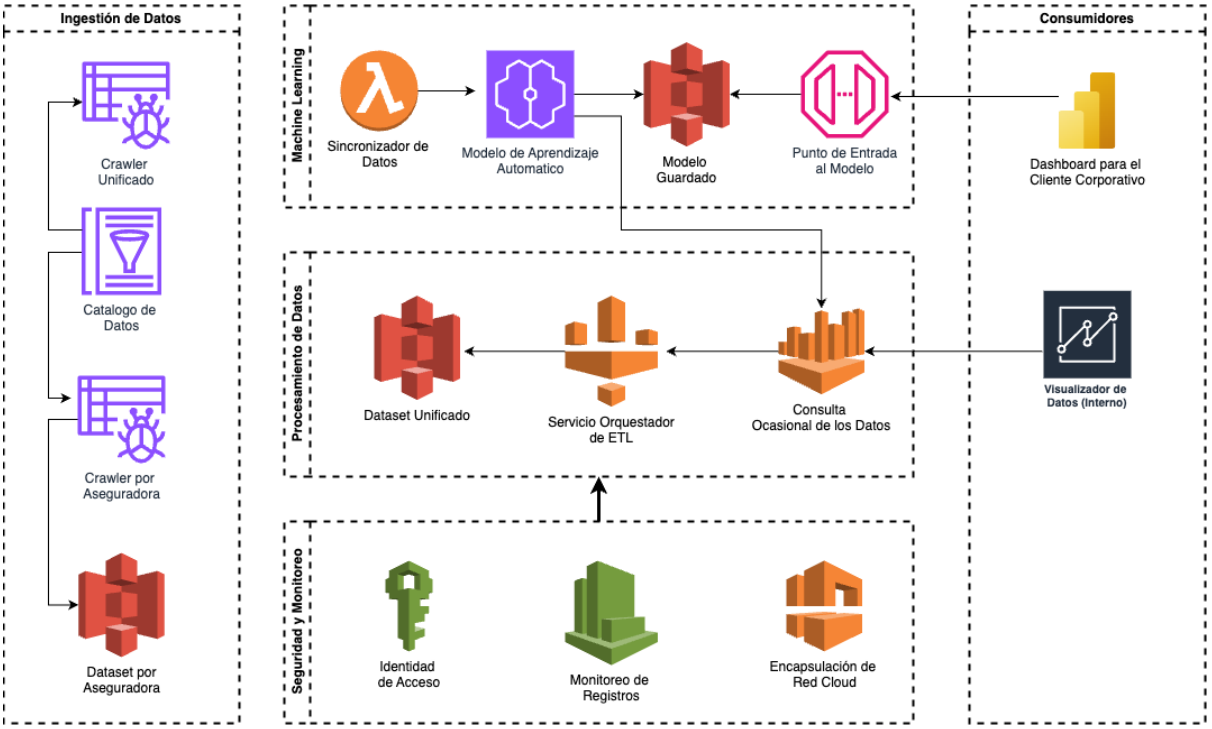
En el marco del diseño e implementación propuesta en el presente proyecto, se establece a continuación una arquitectura de referencia con enfoque a una solución Cloud-Native. Adicionalmente, se considera incluir todas las fases de entrenamiento, validación y prueba del conjunto de datos para la construcción del mejor modelo de aprendizaje automático, por otro lado se establece un flujo ETL para la correcta ingestión de datos en una laguna de datos limpio y estandarizado, con la finalidad de encontrar el mejor modelo de aprendizaje automático capaz de predecir los gastos médicos de los empleados en un cliente corporativo.

El diseño de la solución está basado en el uso de varios orígenes de fuentes de datos, la misma que conlleva a realizar el proceso de extracción, transformación y carga de los datos estructurados en un repositorio de datos previo al entrenamiento del modelo; se plantea la necesidad de monitorear la solución mediante servicios de AWS y el otorgamiento de permisos correctos.

3.1 Pipeline de la implementación

En la figura 4 se visualiza la arquitectura de referencia, la misma que representa un flujo de procesamiento de datos basado en infraestructura cloud, persistiendo el flujo en donde los datos se ingieren, procesan y consumen mediante herramientas de Inteligencia de Negocio.

Figura 4 Arquitectura de referencia.



La arquitectura de referencia integra diversas fuentes de datos mediante procesos de extracción, transformación y carga (ETL), centralizando la información en una única laguna de datos. Para ello, se emplean servicios de AWS como Glue, junto con componentes asociados como Crawler para la exploración automatizada de datos y la generación de tablas estructuradas.

Paralelamente, se realiza el entrenamiento y validación del modelo de aprendizaje automático más adecuado, el cual se almacena en un repositorio de objetos. Este modelo puede ser consumido a través de un servicio Gateway, que permite su integración con herramientas de visualización como Power BI, asegurando así el acceso controlado y eficiente a los resultados analíticos.

En el marco de la implementación de la solución se sigue un proceso totalmente distribuido, escalable y sin servidor; la misma que para la interpretación de la arquitectura de referencia se establece todo el ciclo de vida del dato, desde el origen hacia la interpretación.

Previo al análisis y exploración de los datos, se estandarizará los atributos de los diferentes conjuntos de datos logrando implementar un catálogo capaz de canalizar nuevos valores con el mismo formato ya categorizado.

Es importante mencionar que, la principal herramienta a utilizar son los servicios de Amazon Web Services, debido a la flexibilidad y a la familiaridad con el uso de estos servicios en la empresa beneficiaria de la solución logrando alcanzar el objetivo del proyecto.

Bajo lo dicho anteriormente, se establece que la implementación de esta solución mantiene 4 ejes principales:

1. Integración de datos
2. Búsqueda del modelo
3. Validación del modelo
4. Despliegue y Monitoreo de la solución

3.1.1 Integración de datos

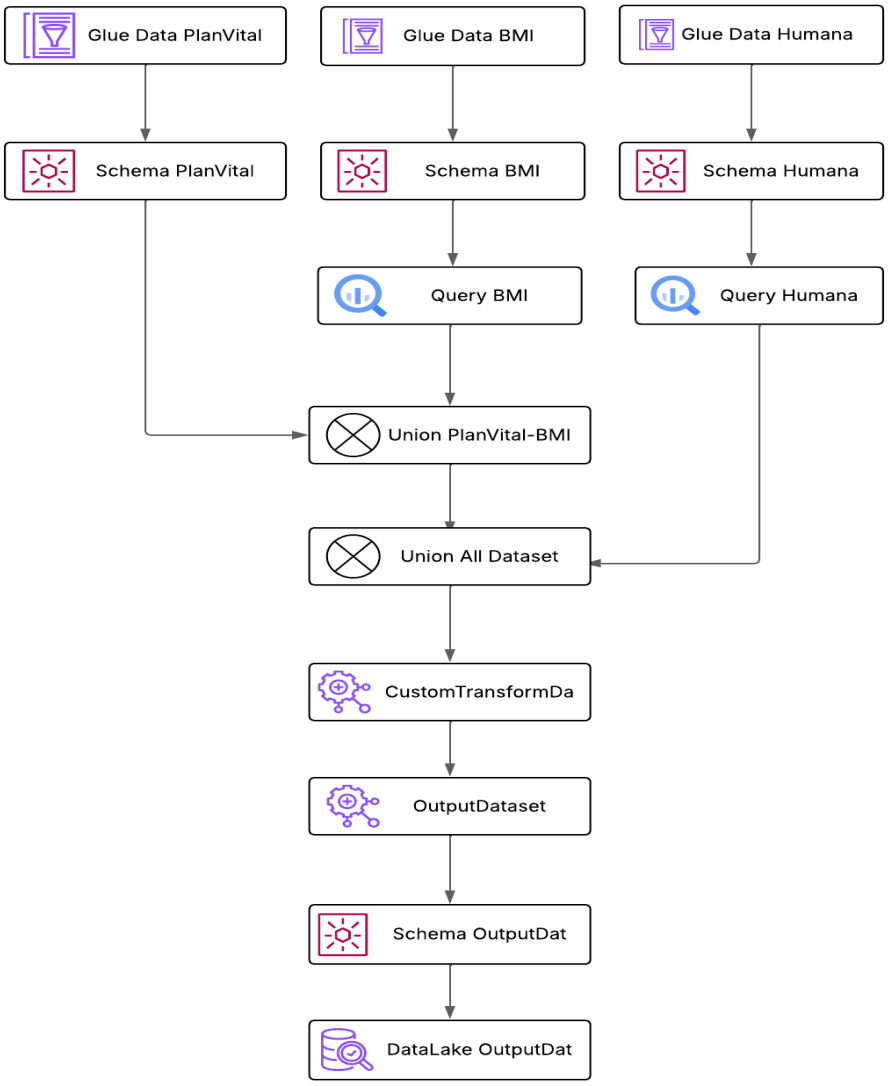
Como parte inicial de la solución, se realiza la carga de las fuentes de datos en una laguna de datos centraliza, para efecto de la arquitectura de referencia se utiliza el servicio AWS S3. Como cada aseguradora mantiene un formato de presentar sus datos de manera diferente, se crea una carpeta por cada una, para que al extraer la información se identifique con el nombre de la carpeta.

Posteriormente, se realiza la catalogación de las fuentes de datos, mediante el uso de AWS Glue; el mismo que recolectará los metadatos de los archivos y procesará la información recolectada para crear una tabla estructura. El proceso de recolección del dato es manual, más, sin embargo, existe un mecanismo de automatización en base a tiempos definidos.

Como cada conjunto de datos proviene de diferentes fuentes, se aplica un flujo de proceso de ETL, posteriormente cargados en un repositorio listo para su análisis. Las transformaciones que se aplican en esta fase del flujo son las siguientes: tipos de datos, nombres de las columnas, limpieza de datos, eliminación de datos nulos, impugnación, traducción de valores en cada fila para establecer un equilibrio entre todas las fuentes.

Posteriormente, se define un flujo ETL que se puede observar en la figura 5. Este es capaz de llevar a cabo cualquier base que mantenga la estructura analizada en el diagrama mostrado posteriormente.

Figura 5 Proceso ETL



Este diagrama representa un flujo ETL que integra datos provenientes de diversas aseguradoras con sede en Ecuador (BMI, Humana y Plan Vital), con las cuales la agencia de seguros mantiene alianzas estratégicas. El objetivo es generar un conjunto de datos unificado a partir de múltiples fuentes. El proceso inicia con la carga y transformación de los datos específicos de cada una de estas aseguradoras, siguiendo los pasos posteriores:

1. **Orígenes de datos:** Se establece el conjunto de datos de cada origen, previamente catalogado mediante el uso de AWS Glue.
2. **Transformación de esquema:** Se realiza la transformación de cambio de esquema para cada fuente, es decir se ajustan los datos con un formato uniforme predefinido para cada uno de los esquemas.
3. **Transformación de datos:** Se realizan consulta SQL de tal manera que se ajusten a las necesidades de cada fuente; se establece la cantidad de columnas, tipo de datos y agregación de nuevas columnas condicionales.
4. **Unificación de conjunto de datos:** Se combinan los datos de cada una de las fuentes, la misma deben tener el mismo esquema.
5. **Transformación personalizada:** Después de la unión de los conjuntos de datos, se aplica una transformación personalizada. En este paso se aplica una lógica específica, tales como cálculos avanzados, limpiezas, o modificaciones que no estaban cubiertas por las transformaciones estándar.
6. **Salida del conjunto de datos:** Se transforma nuevamente el esquema final para el conjunto de datos de salida.
7. **Esquema Final:** Se ajusta el esquema del conjunto de dato final para que coincida con el formato de salida esperado, adicionalmente los datos transformados se almacenan en nuevo repositorio de dato.

El proceso detallado anteriormente busca encontrar el mejor modelo de aprendizaje automático capaz de predecir el comportamiento de los gastos médicos de los afiliados de un cliente corporativo.

3.1.2 Búsqueda del modelo

Una vez desarrollado la etapa anterior y previamente obtenida un conjunto de dato unificado y consistente con el esquema del conjunto de dato final, se continua con el análisis de las variables y la distribución de los datos, bajo este enfoque se utiliza el servicio de SageMaker de AWS.

Para predecir los gastos médicos basados en las características mencionadas en capítulos anteriores, comenzamos seleccionando un modelo de aprendizaje automático adecuado que permita modelar relaciones complejas entre variables. A continuación, se detallan los pasos seguidos para la búsqueda del modelo óptimo.

1. Selección del modelo base

El primer paso fue entrenar un modelo base de Random Forest Regressor. Se elige este modelo debido a su capacidad de interpretación de variables frente a datos ruidosos y relaciones no lineales entre las variables de entrada y la variable objetivo (gastos médicos). Arboles aleatorios también ofrece una ventaja significativa: la capacidad de determinar la importancia de las variables en relación con el objetivo, lo que permite identificar qué características contribuyen más al modelo.

2. Preprocesamiento de datos

Previo al entrenamiento del modelo, se aplicó la eliminación de aquellas columnas que no aportaban valor predictivo, tales como filial, fechapago y sequence, con el objetivo de reducir el ruido en el conjunto de datos.

A continuación, se transformaron las variables categóricas —como aseguradora, proveedor y tiporeclamo— en variables numéricas mediante técnicas de codificación, lo que permitió su utilización en algoritmos de aprendizaje automático. Paralelamente, se normalizaron las variables numéricas empleando la clase StandardScaler de la biblioteca scikit-learn. Esta técnica ajusta las variables para que presenten una desviación estándar de 1, manteniendo sus medias originales (es decir, sin centrado), lo cual mejora la estabilidad y el rendimiento de ciertos algoritmos.

Una vez normalizados los datos, se utilizó UMAP (Uniform Manifold Approximation and Projection), una tecnica no lineal que valida tanto la estructura local como global del conjunto de datos. UMAP permitió proyectar los datos en un espacio bidimensional (2D), facilitando su interpretación visual sin perder características significativas.

El resultado de esta reducción fue un nuevo conjunto de datos en dos dimensiones, representado como un DataFrame de pandas. Esta estructura simplificada es particularmente útil para detectar patrones, relaciones latentes y posibles agrupamientos dentro del conjunto de datos, aspectos esenciales para el análisis exploratorio y la validación del modelo.

3. Importancia de características

Tras entrenar el modelo base, se analizó la importancia las variables significativas generada por el algoritmo de Random Forest. Este análisis permitió identificar qué variables tienen mayor peso en la predicción de los gastos médicos. Las características más relevantes incluyeron:

- **Edad del paciente:** Directamente relacionada con los gastos médicos.
- **Tipo de afiliado (titular o dependiente):** Variable con alta significancia acorde a los gastos médicos.
- **Diagnóstico y tipo de atención:** Variables clave en la predicción de los costos.

Algunas características mostraron menor relevancia y fueron descartadas en iteraciones posteriores para reducir la complejidad del modelo.

4. Ajustes de Hiperparametro

Para optimizar el rendimiento del modelo, se realiza un ajuste de hiperparámetros utilizando la técnica de búsqueda *GridSearchCV*. Los hiperparámetros evaluados son:

- **Número de árboles (n_estimators)**: Se evalúa entre 100 y 300 árboles.
- **Profundidad máxima del árbol (max_depth):** Se evalúa con 10, 20, y 30 para ajustar la complejidad del modelo.
- **Tamaño mínimo de hojas:** Se evalúa los valores de min_samples_split y min_samples_leaf para evitar el sobreajuste.

Este proceso resulta ayuda significativamente en el desempeño del modelo, aumentando la precisión (R^2) y reduciendo el error cuadrático medio (RMSE).

3.1.3 Proceso de validación del modelo

En esta etapa de la implementación en donde se evalúa el modelo predictivo, se observa que el modelo final mejora en cuanto a precisión y capacidad predictiva en comparación con el modelo base. Las métricas obtenidas fueron:

- **R^2 (coeficiente de determinación):** Indica que el modelo abarca un alto porcentaje de la variabilidad en los gastos médicos.

- **RMSE (error cuadrático medio):** Indica una disminución en el error promedio de las predicciones.

Este proceso iterativo de búsqueda y ajuste permite lograr estimar el mejor modelo para predecir los gastos médicos, proporcionando una herramienta predictiva que puede usarse para tomar decisiones más informada en seguros médicos.

3.1.4 Despliegue y monitoreo de la solución

El modelo de aprendizaje automático entrenado en AWS SageMaker se guarda utilizando *joblib* y se almacena en un repositorio de AWS S3. En base a la presentación de los resultados en un tablero de visualización se incorpora una función sin servidor (Lambdas), la misma que carga el modelo entrenado y almacenado en S3 con la finalidad que realice las predicciones enviadas a través de la consulta por la puerta de enlace (API Gateway).

AWS API Gateway se puede usar para exponer la función Lambda como un servicio RESTful. Por otro lado, la herramienta de visualización se configuró para visualizar los datos y las predicciones generadas por el modelo, este se integra con AWS Athena.

Con respecto al monitoreo de la solución, se utiliza AWS CloudWatch para monitorear el desempeño de AWS Glue, Athena, y Lambda. Por otro lado, se configura alarmas para detectar cualquier anomalía en el proceso ETL o en el rendimiento de la API.

En base a todo lo anterior, con respecto a los procesos de implementación y diseño de la arquitectura de referencia en el siguiente capitulo se abordarán nuevos descubrimientos relacionados a los resultados obtenidos de los datos.

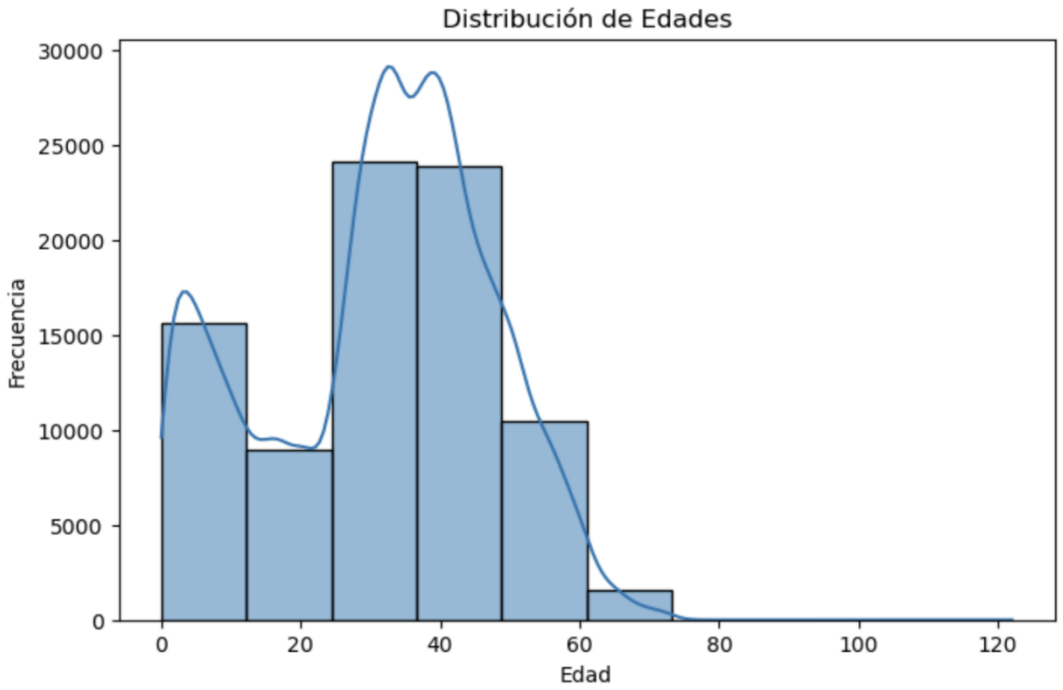
4. Análisis De Resultados

4.1 Análisis exploratorio de datos

El análisis exploratorio de los datos (EDA) es parte fundamental para comprender el comportamiento de las características y distribución del conjunto de datos. Bajo esta premisa, se han realizado varias visualizaciones que proporcionan una visión detallada de la distribución y relaciones entre variables clave.

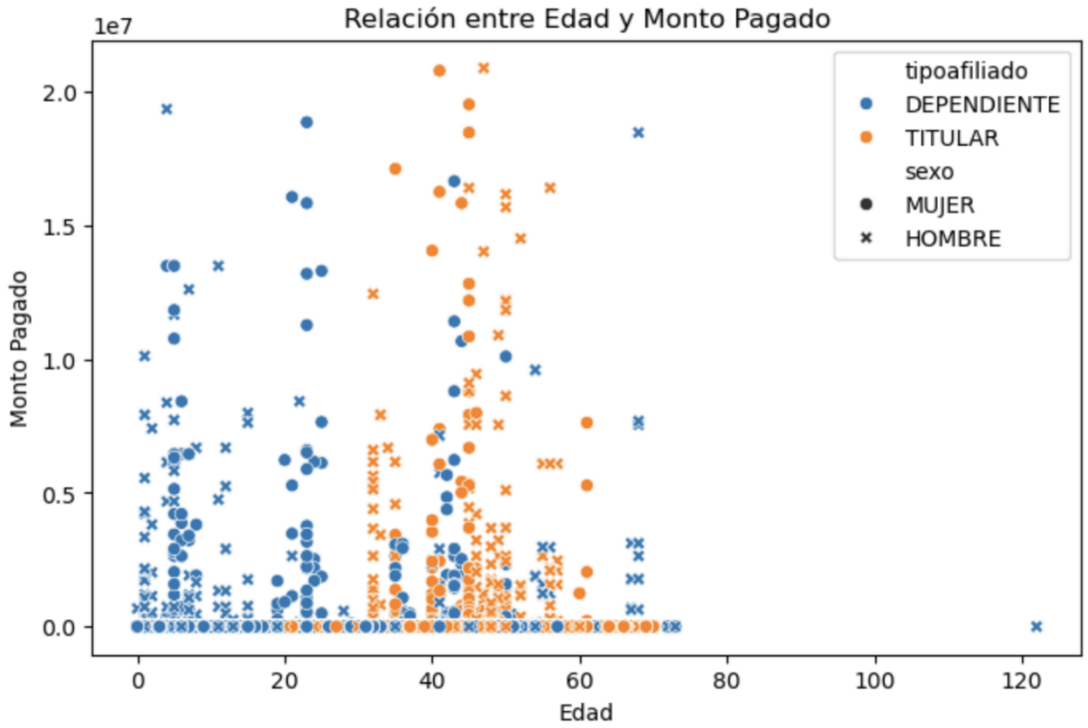
Por otro lado, es importante conocer el comportamiento de cada variable categórica y el grado de incidencia en la categorización del valor en cada registro del conjunto de dato. Como se puede observar, se encuentran 3 aseguradoras, 20 filiales, 2025 proveedores de salud, 2 tipos de reclamos, 2727 diagnósticos representados por su código cie 10, 2 tipos de atención, 2 tipos de afiliados y 2 sexo.

Figura 6 Distribución de la variable Edad



En la figura 6 se muestra se visualiza el comportamiento de la variable edad en el conjunto de datos. La distribución de edad puede revelar si hay una concentración de gastos médicos en ciertos grupos etarios. En este contexto, se visualiza que la mayoría de las observaciones pertenecen a grupos de dades entre los 30 a 50 años, lo que puede influir en los patrones de gastos médicos asociados a dicha edad.

Figura 7 Relación entre edad y monto pagado.



En la figura 7 se visualiza el comportamiento de la edad y el monto pagado, tomando en referencia el tipo de afiliado y el sexo de la persona. Gráficamente se observa que la mayoría de los dependientes tienden a ser menores de 20 años con un balance homogéneo entre hombres y mujeres por otro lado los titulares de la póliza presentan mayor gasto presentado y suelen ser personas adultas.

En esta etapa del análisis se evidencia el comportamiento de los datos sobre la variable de interés, que es el gasto médico de cada persona respecto a su edad, sexo, tipo de atención, entre otras variables; ya que se necesita conocer la importancia de la variable para entrenar el modelo, se establecen inferencias y correlaciones de cada columna mencionadas en la sección posterior.

4.1.1. Estadísticas descriptivas de los datos.

Las estadísticas descriptivas permiten obtener una visión general de las características más relevantes de los datos, utilizando medidas como el promedio, la mediana, la desviación estándar, los valores mínimos y máximos, así como los percentiles. A continuación, se detallan estas métricas para las principales variables del conjunto de datos relacionado con los gastos médicos.

El promedio y la mediana ayudan a identificar cuánto se gasta comúnmente en atención médica, mientras que la desviación estándar y el rango muestran qué tan variados son esos montos. Además, la asimetría y la curtosis aportan información sobre la forma de la distribución y si existen gastos inusualmente altos o bajos.

Tabla 3 Estadísticas descriptivas de las variables a emplear

Variable	*N*	*Media*	*Desv. Est.*	*Mínimo*	*P25*	*Mediana*	*P75*	*Máximo*
aseguradora	71249	1.893	0.315	0.000	2.000	2.000	2.000	2.000
proveedor	71249	501.426	285.829	0.000	199.000	489.000	695.000	1171.000
tiporeclamo	71249	0.138	0.345	0.000	0.000	0.000	0.000	1.000
edad	71249	30.819	15.871	0.000	19.000	33.000	42.000	122.000
cie10	71249	1032.553	631.782	0.000	442.000	962.000	1562.000	2239.000
diagnostico	71249	1030.129	658.688	0.000	479.000	1002.000	1596.000	2384.000
presentado	71249	15.055	11.881	0.000	4.850	12.730	22.000	69.480
pagado	71249	9.588	7.559	-0.400	2.960	8.000	13.780	33.000
copago	71249	4.289	15.794	0.000	0.680	2.100	5.510	871.820
deducible	71249	0.659	5.561	0.000	0.000	0.000	0.000	260.000
tipoatencion	71249	0.006	0.074	0.000	0.000	0.000	0.000	1.000
tipoafiliado	71249	0.468	0.499	0.000	0.000	0.000	1.000	1.000

sexo	71249	0.517	0.500	0.000	0.000	1.000	1.000	1.000

Esta tabla muestra las estadísticas de las columnas numéricas de los datos, como la cantidad de registros, el valor promedio, la desviación estándar, los valores mínimo y máximo, y los percentiles (25%, 50%, y 75%).

Acorde al análisis de tabla anterior se determina las siguientes observaciones:

Aseguradora: Aunque es una columna categórica, ha sido convertida a valores numéricos. Tiene dos valores posibles (representados como 0 y 2), lo que sugiere que hay tres aseguradoras en los datos.

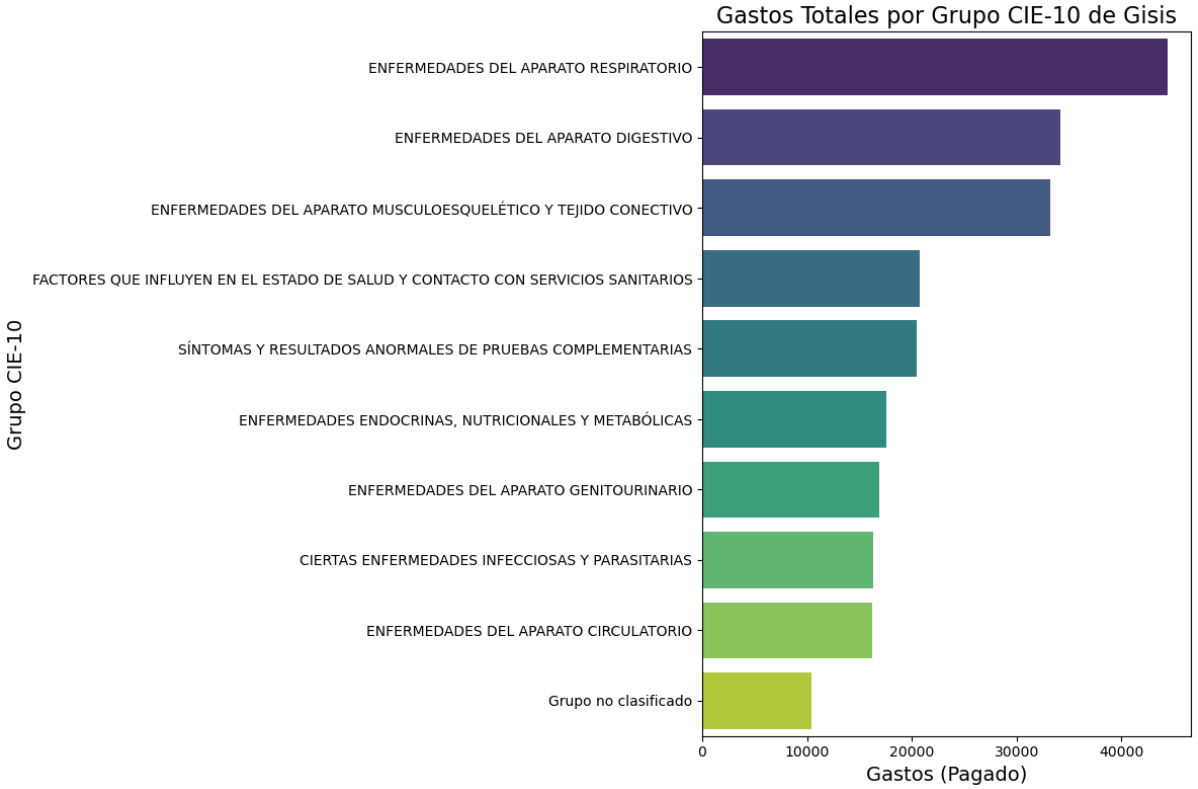
- **Proveedor**: La columna "Proveedor" también se codifica numéricamente, con un amplio rango de valores, indicando muchos proveedores en los datos.
- **Tipo Reclamo**: Mayormente con valor 0, lo que indica que la mayoría de los reclamos podrían tender a ser de un solo tipo.
- **Edad**: La edad promedio es de 30.82 años, con un rango que va de 0 a 122 años, lo que sugiere la presencia de asegurados de todas las edades.
- **CIE-10:** Al igual que otras columnas categóricas, ha sido convertida en números. Representa el código diagnóstico, con un rango de 0 a 2239.
- **Diagnóstico**: Representa el diagnóstico y su estadística es similar a la del CIE-10.
- **Presentado**: Indica el monto presentado por los asegurados. El promedio es de 15.05 unidades monetarias, con un rango que va de 0 a 69.48.
- **Pagado**: El monto promedio pagado es de 9.59 unidades, con un rango desde -0.4 (posiblemente reembolsos con errores) hasta 33 unidades.
- **Copago**: Promedio de 4.29 unidades, con un máximo muy alto de 871.82, lo que podría representar casos excepcionales.
- **Deducible**: El deducible es relativamente bajo, con un promedio de 0.65 unidades y un máximo de 260.
- **Tipo Atención**: Esta columna tiene valores muy pequeños, lo que sugiere que la mayoría de los registros pertenecen a un tipo de atención mayoritario.
- **Tipo Afiliado**: Tiene un valor promedio de 0.47, lo que indica una distribución cercana al 50% entre los distintos tipos de afiliados.
- **Sexo**: Similar al tipo de afiliado, el promedio de 0.51 indica que la distribución entre hombres y mujeres está balanceada.

4.2 Puesta en marcha y funcionamiento

Para poder proceder con la elaboración de modelos predictivos fue necesario agrupar inicialmente los diagnósticos acordes a su categoría CIE 10, las cuales se pueden observar en el gráfico 8. Se establecieron características que afectan a similares órganos o con factores comunes como su origen.

Para poder entrenar al modelo inicial se escogió a una empresa en particular y se observó su comportamiento. A partir de esta empresa se procederá a elaborar una metodología base que podrá ser replicable para el resto de las empresas, permitiendo escoger el mejor modelo. En el gráfico se puede visualizar el comportamiento de las observaciones por diagnostico en la empresa seleccionada.

Figura 8 Gastos totales por Grupo CIE-10



Para mejorar el análisis y no tener un fuerte impacto de datos atípicos se eliminaron los outliers como parte de la depuración de datos.

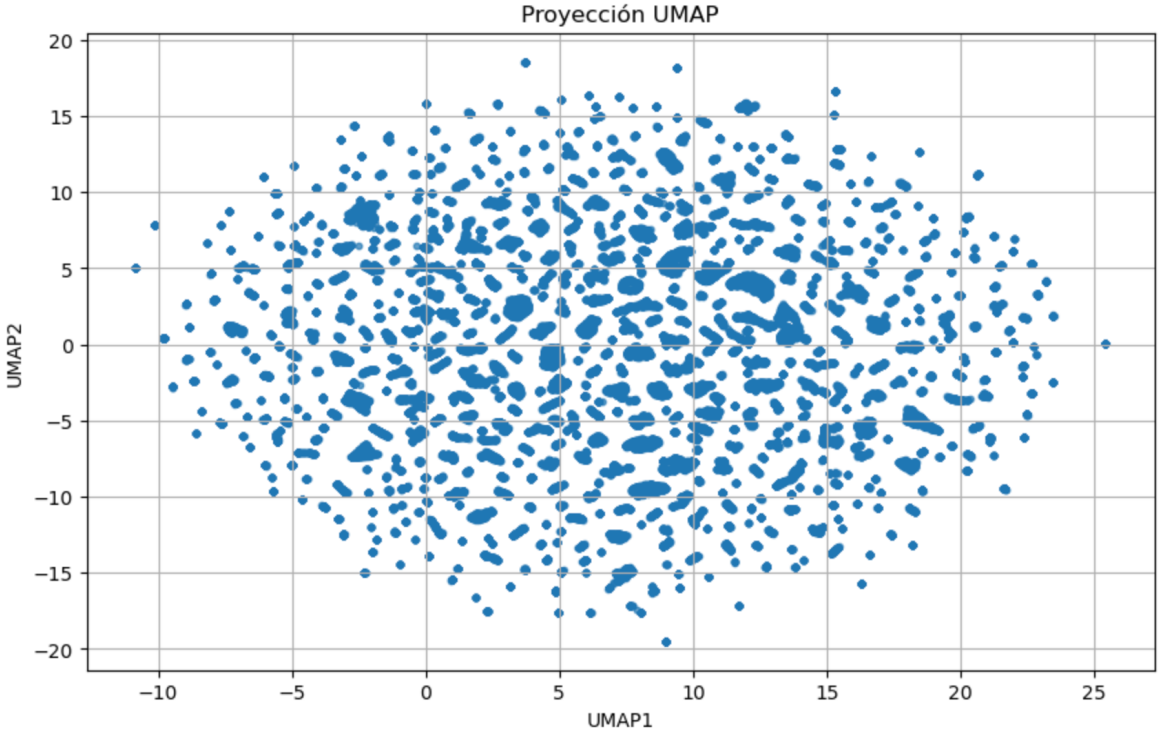
A partir de las variables edad, sexo, valores pagados, deducibles, copago, tipo de atención, afiliado, temporada del año y grupo de diagnóstico se procede a crear clústeres que

faciliten el análisis de los datos al contar con información variada. Seguido de esto, se procede a aplicar la reducción de dimensionalidad de los variables en base a la técnica no lineal UMAP.

Tal como se puede apreciar en la figura 9, hasta el tercer componente se ha capturado una parte importante de la varianza explicada. Es decir, los tres primeros componentes son los que realmente aportan información relevante para el modelo, y los demás componentes no agregan tanto valor.

Este método ayuda a reducir la dimensionalidad de los datos sin perder un gran nivel de información, permitiendo que el modelo sea más eficiente, al reducir la cantidad de características a considerar sin tener un impacto negativo en la capacidad de predicción y favorece a la interpretabilidad.

Figura 9 Varianza explicada por componente

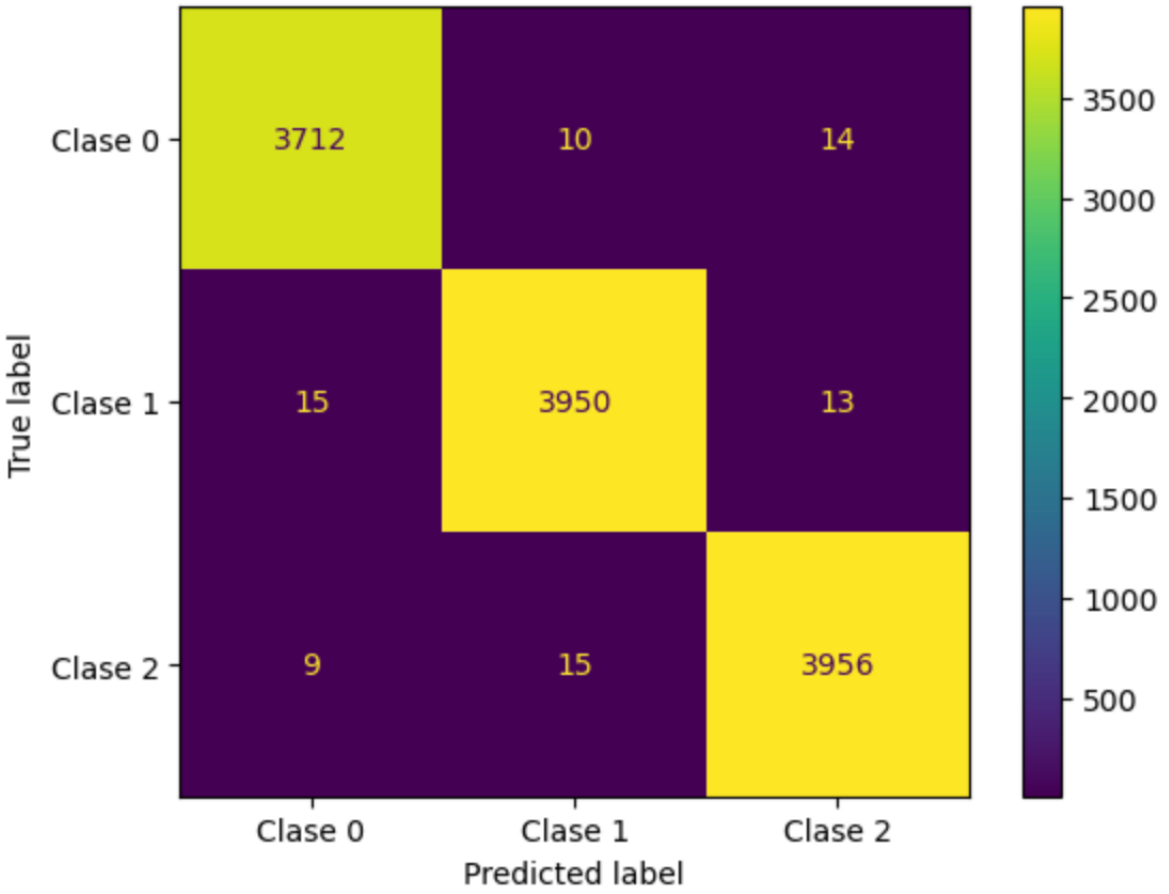


Adicionalmente, se aplicó KNN para poder determinar la clasificación a la que pertenecen los puntos ruidosos. Este método permite asignar los puntos de ruido a una categoría acorde a la cercanía que exista con otros puntos que sí cuentan con una clasificación clara. Es decir, dependiendo de sus vecinos más cercanos, lo que permite que

los puntos ya no estén dispersos, sino que puedan ser parte de un clúster para el agrupamiento final. En este caso, los clústers dependen de categorías de los reembolsos, por ejemplo, datos sobre quién solicitó el mismo y sus características demográficas.

Luego se procedió con un árbol de decisión para clasificar a qué clúster pertenece cada grupo. En total, existieron 2,508 observaciones, de las cuales acertó en 2,497 casos. Es decir, las etiquetas fueron correctas para el 99.56% de las observaciones acorde a lo observado en la figura 10.

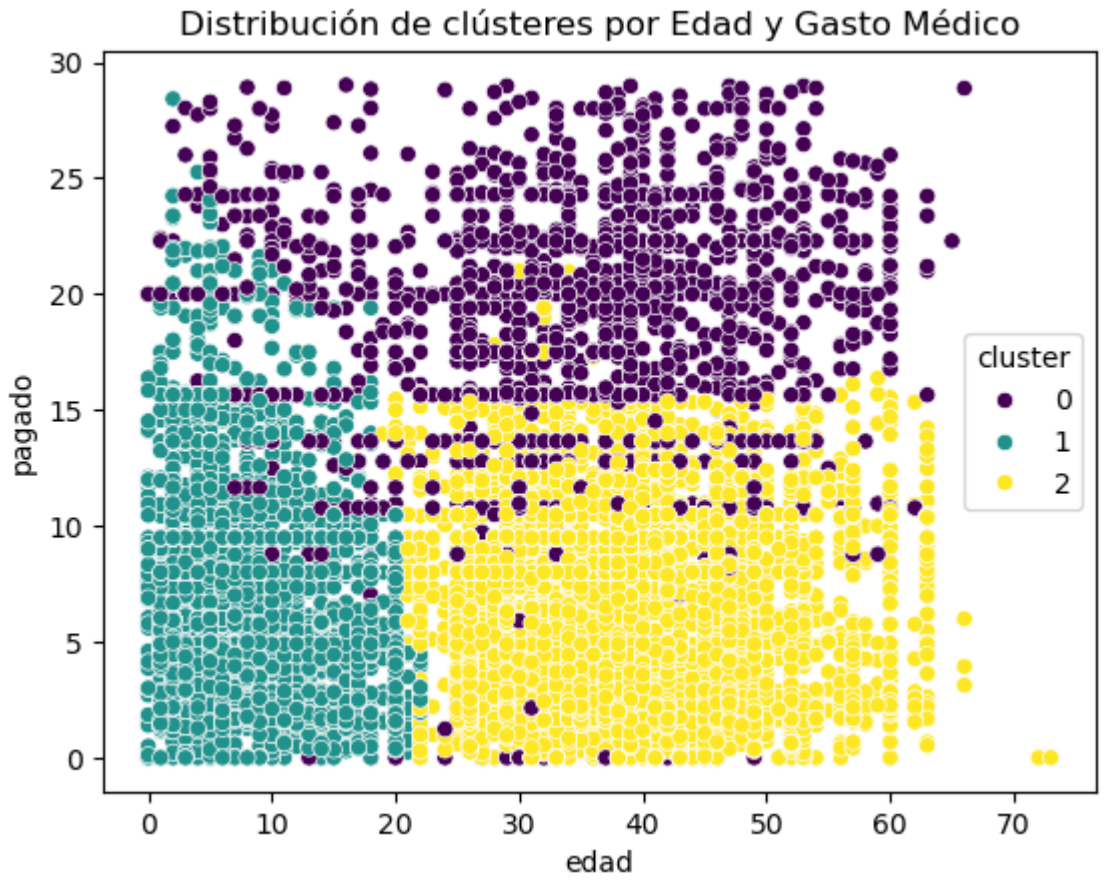
Figura 10 Matriz de confusión para la clasificación de clústeres



Como se puede apreciar en la figura 11, la clase 0 está conformada por personas que cuentan con un gasto elevado, sin tomar en consideración como factor principal la edad, ya que está se muestra en un amplio rango. Por otra parte, el grupo 1 está conformado por personas jóvenes, es decir, desde recién nacidos hasta adultos jóvenes que cuentan con un gasto médico entre medio y bajo. Finalmente, el grupo 2 son personas con gasto médico que

también es considerado como medio y bajo, pero de edad más avanzada, esto implica adultos y adultos mayores.

Figura 11 Distribución de clústeres



Para poder continuar con la predicción, se realiza un proceso de optimización de hiperparámetros para tres modelos de regresión: Random Forest, Gradient Boosting y XGBoost, utilizando GridSearchCV para encontrar la mejor configuración de parámetros.

Ajustar los hiperparámetros es una estrategia que busca mejorar la precisión de los modelos al modificar variables esenciales como la cantidad de estimadores, la velocidad de aprendizaje o la profundidad de los árboles. En este proceso, el código establece distintos valores posibles para cada uno de estos parámetros y los analiza de manera sistemática para encontrar la mejor combinación.

Para realizar esta optimización de manera organizada y eficiente, se utiliza un pipeline de scikit-learn. El pipeline incluye dos pasos principales: un preprocesador que convierte las características categóricas en variables numéricas utilizando OneHotEncoder y el modelo de

regresión correspondiente. Este enfoque garantiza que las transformaciones sean aplicadas de manera consistente en los datos de entrenamiento y prueba, lo que facilita la evaluación del rendimiento de los modelos.

En cada ciclo del proceso, se utiliza GridSearchCV para explorar todas las combinaciones posibles de parámetros y evaluar el desempeño de los modelos mediante validación cruzada. Al emplear cinco divisiones o folds, se logra una estimación más confiable del comportamiento del modelo. Además, se adopta el valor negativo del error cuadrático medio como métrica de evaluación, lo que facilita la elección de la configuración que genere el menor margen de error.

Al finalizar el proceso, el código almacena el mejor modelo entrenado para cada uno de los algoritmos en un diccionario llamado best_models. Además, el modelo con el mejor rendimiento general se guarda en la variable best_model_trained. Esto asegura que, al final del proceso, se haya identificado el modelo más adecuado, con los mejores hiperparámetros, para el conjunto de datos en cuestión.

Luego de esto, se procedió a elaborar un Voting Regresor el cual se comparó con los resultados previamente obtenidos.

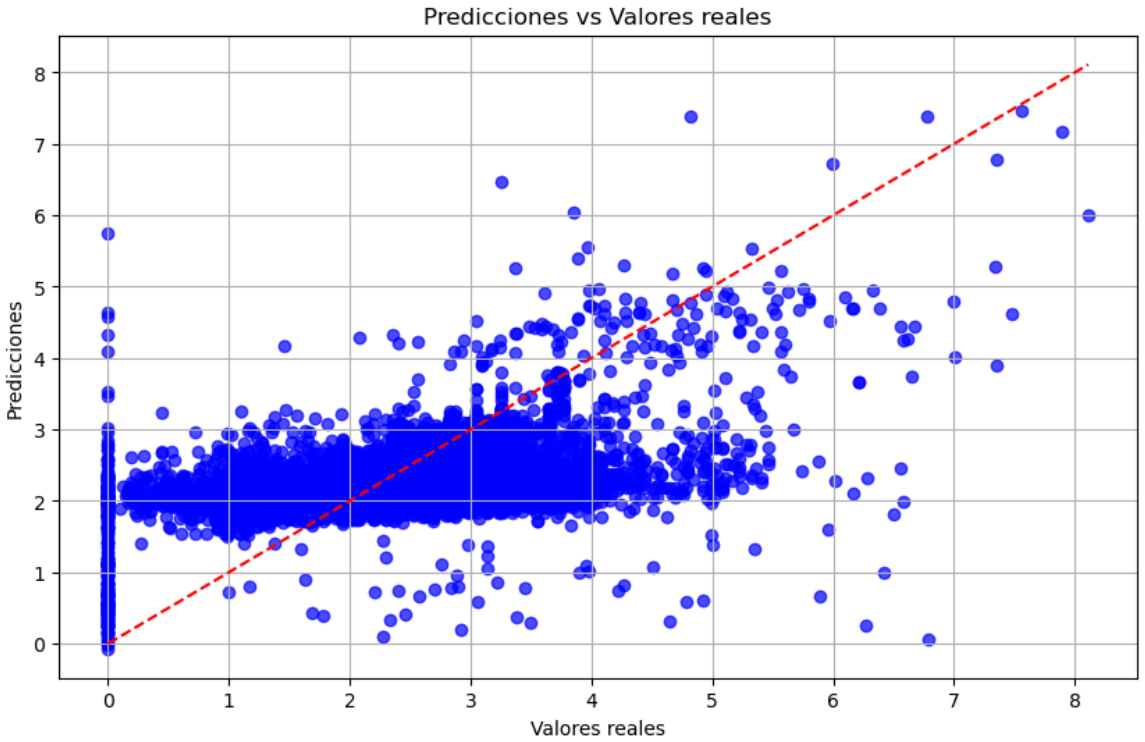
Los resultados fueron los presentados en la tabla 3.

Tabla 4 Métricas de evaluación de modelos

	Random Forest	***Gradient Boosting***	***XGBoost***	***Voting Regressor***
MSE	0.7825	0.8174	0.8174	0.7844
MAE	0.6654	0.6933	0.6935	0.6758
R2	0.2457	0.8174	0.2120	0.2438

En resumen, Random Forest es el modelo que muestra el mejor rendimiento en términos de precisión (menor MSE y MAE), aunque Gradient Boosting tiene el mejor R2, lo que indica que logra capturar la variabilidad de los datos con mejor precisión. Por su parte, XGBoost tiene un rendimiento globalmente más bajo en comparación con los otros modelos.

Figura 12 Valores reales vs predicciones



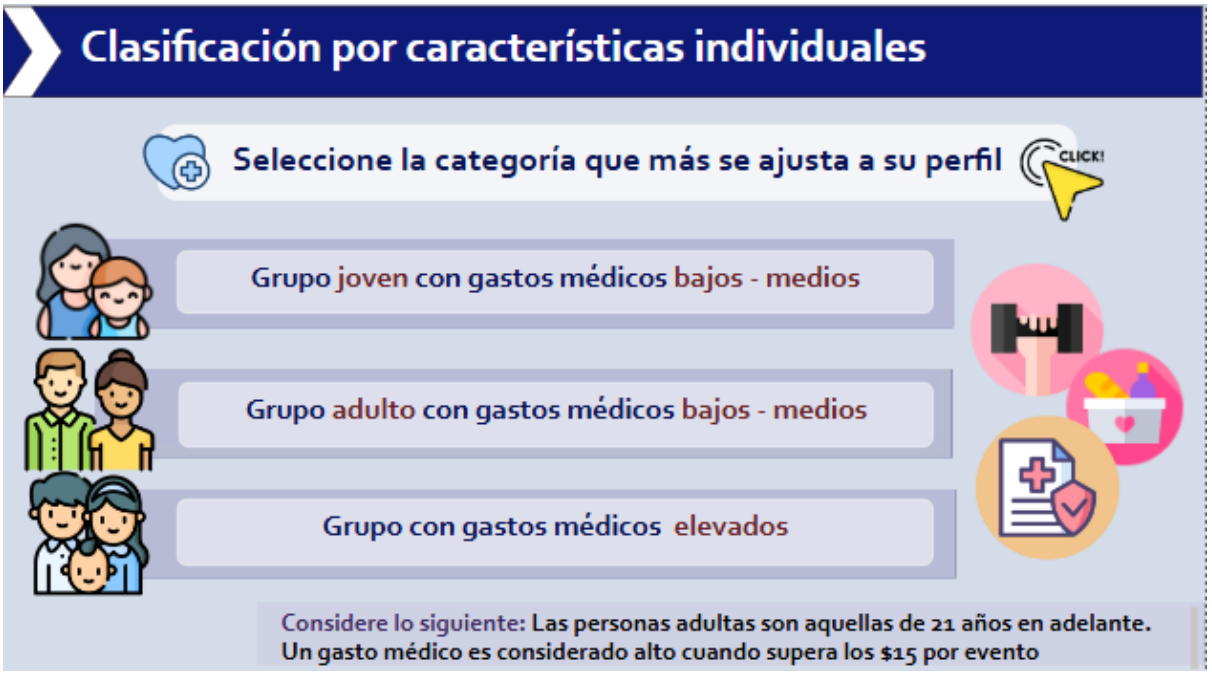
Actualmente, el código empleado fue entrenado para predecir cuál sería el mejor modelo para una empresa en específico. Sin embargo, permite que se pueda aplicar a otra empresa y se almacene cuál sería el mejor modelo acorde a los resultados y métricas observadas. Esta facilidad de replicabilidad permite que el código sea escalable, de tal forma que se optimice el tiempo y los recursos, además de que permite que, en caso de ser necesario, se puedan realizar ajustes de forma rápida y sencilla.

Adicionalmente, gracias a esta replicabilidad se puede comparar de manera objetiva qué modelo funciona mejor en diferentes contextos, lo que proporciona una visión más clara sobre las fortalezas y debilidades de cada modelo en distintos entornos. Este enfoque permite tomar decisiones más informadas reduciendo la asimetría de información, así como facilitar la mejora continua.

4.2.1. Elaboración del dashboard

Para la elaboración del dashboard se consideraron distintas características que podrían ser interesantes para los usuarios.

Figura 13 Página de inicio del dashboard



En la primera sección del dashboard, el usuario debe identificarse proporcionando información clave sobre su perfil, específicamente su edad y el nivel de gasto promedio por siniestro registrado. Esta autoidentificación permite personalizar la visualización y contextualizar los resultados del modelo predictivo.

Para efectos del análisis, si el usuario tiene menos de 21 años, se clasifica como una persona joven. Caso contrario, se considera una persona adulta. Posterior a esta distinción etaria se procede a segmentar mejor los patrones de gasto y atención médica.

En cuanto al nivel de gasto, se establecen dos categorías principales: bajo-medio y elevado. Se considera que un gasto es elevado si el promedio por siniestro supera los $15. La segmentación permite al sistema ajustar dinámicamente los resultados y recomendaciones, facilitando así una experiencia más personalizada y útil para el usuario final.

Además, esta sección introductoria actúa como un filtro inicial para las visualizaciones posteriores, asegurando que los indicadores mostrados reflejen con mayor precisión el perfil del usuario y permitiendo comparar su situación con la de otros grupos similares dentro del conjunto de datos.

Figura 14 Categorización en el dashboard

Una vez seleccionada la categoría correspondiente, se presenta al usuario el dashboard personalizado. Este panel es interactivo y puede filtrarse por variables clave como año, sexo, filial y aseguradora. Esto permite adaptar la visualización a diferentes contextos y necesidades de análisis.

En esta sección, el usuario podrá consultar indicadores relevantes como el ahorro promedio por evento médico, el copago cubierto por la aseguradora y el pago realizado por el asegurado. Además, los datos pueden desglosarse según el tipo de atención médica, ya sea hospitalaria o ambulatoria, lo que facilita una comprensión más detallada del comportamiento del gasto.

Asimismo, el dashboard incluye predicciones específicas por tipo de enfermedad, lo que permite anticipar posibles costos asociados a condiciones médicas frecuentes. Esta funcionalidad resulta especialmente útil tanto para la aseguradora como para el asegurado,

ya que contribuye a reducir la brecha de información característica de este sector, promoviendo una toma de decisiones más informada y eficiente.

En conjunto, esta herramienta no solo facilita el análisis del gasto médico, sino que también permite evaluar de manera objetiva los beneficios de contar con un seguro, aportando claridad sobre su impacto económico en diferentes escenarios.

4.2.2 Limitaciones y riesgos del modelo

El presente modelo predictivo constituye una herramienta valiosa para estimar los gastos médicos de los empleados de clientes corporativos, es importante reconocer ciertas limitaciones y riesgos inherentes a su implementación.

Las principales limitaciones radican en la calidad y representatividad de los datos utilizados para el entrenamiento. Debido a que las fuentes provienen de aseguradoras con estructuras y criterios de codificación distintos, existe el riesgo de que ciertos grupos poblacionales estén subrepresentados; especialmente en variables categóricas como diagnósticos, tipo de servicio, tipo de atención, entre otros, lo cual podría afectar la precisión del modelo para esos segmentos. Asimismo, los datos históricos pueden contener sesgos preexistentes, lo que podría perpetuar desigualdades si no se controlan adecuadamente.

Otro riesgo relevante está relacionado con la posibilidad de que el modelo sea utilizado para tomar decisiones que afecten directamente a individuos, como el acceso a beneficios médicos o la contratación de pólizas. Aunque el enfoque propuesto trabaja con datos agregados y no busca identificar casos individuales, es necesario establecer límites claros sobre su uso, a fin de evitar prácticas discriminatorias o interpretaciones erróneas de los resultados.

En cuanto a los aspectos técnicos, el modelo podría experimentar una disminución en su desempeño predictivo si no se actualiza regularmente con datos recientes, en especial en un contexto donde las condiciones médicas y las políticas de cobertura hospitalaria y ambulatoria pueden cambiar rápidamente.

Para mitigar estos riesgos, se recomienda la implementación de auditorías periódicas del modelo, el monitoreo de métricas de equidad y la participación de un comité de

supervisión multidisciplinario que evalúe su impacto de manera continua por parte de la agencia de seguros. Estos mecanismos permitirán detectar desviaciones tempranas, ajustar los parámetros del modelo según sea necesario y garantizar un uso ético y transparente de la herramienta.

5. Conclusiones y recomendaciones

5.1. Conclusiones

A lo largo del desarrollo de este trabajo de titulación se logró diseñar un modelo predictivo de gastos médicos orientado a empleados de clientes corporativos, alcanzando así el objetivo general planteado. El modelo permite anticipar el valor estimado que una compañía de seguros deberá cubrir, basándose en reclamos históricos de los empleados. Esto no solo representa un avance técnico significativo, sino que además constituye una herramienta estratégica para la optimización de la gestión de recursos en el campo asegurador, al mejorar la planificación financiera y contribuir a una mayor satisfacción de los clientes.

Un componente clave en la construcción del modelo fue la segmentación previa de los datos, la cual permitió entender mejor los patrones subyacentes en los gastos médicos. Debido a la alta dimensionalidad y al carácter categórico de muchas variables presentes en el conjunto de datos, fue necesario aplicar una técnica de reducción de dimensionalidad como UMAP. Esta técnica facilitó la representación visual y matemática de los datos en un espacio reducido, permitiendo que el algoritmo KMeans pudiera identificar de manera eficiente tres grupos diferenciados de gastos médicos. Estos grupos fueron posteriormente validados con expertos del negocio, quienes confirmaron que los mismos estaban directamente relacionados con tipos específicos de diagnósticos médicos, lo que ratifica la solidez del enfoque adoptado.

Aunque el modelo se desarrolló inicialmente con los datos de una sola empresa que cuenta con un alto volumen de empleados, se demostró que la estrategia metodológica empleada y el pipeline de procesamiento son replicables en diferentes filiales y empresas, adaptándose a sus características particulares. Para esto se adoptó un enfoque agnóstico al origen de los datos, priorizando la estructura de estos sobre la procedencia, lo que permitió construir un modelo orientado a la filial sin perder generalidad ni robustez. Esta flexibilidad resulta esencial dada la heterogeneidad presente entre empresas del mismo grupo asegurador.

En términos de validación, se utilizaron múltiples métricas de rendimiento que permitieron evaluar la precisión y capacidad de generalización del modelo predictivo. Este análisis exhaustivo facilitó la selección del mejor algoritmo de aprendizaje automático y proporcionó una base objetiva para justificar su implementación. La solidez de los resultados refuerza la confiabilidad del sistema propuesto en contextos reales de toma de decisiones.

Además, como parte del entregable final, se desarrolló una herramienta de visualización interactiva que integra el modelo predictivo más eficiente. Esta herramienta no solo facilita la interpretación de los resultados por parte del usuario final, sino que también fortalece la capacidad analítica de la agencia de seguros y sus clientes corporativos, permitiéndoles explorar diferentes escenarios y planificar con mayor precisión sus recursos.

Finalmente, los resultados obtenidos reflejan que la metodología propuesta es efectiva para abordar el problema planteado. No obstante, se identifican oportunidades claras de mejora en futuras versiones del modelo. Estas incluyen la incorporación de nuevas variables que puedan aportar mayor poder explicativo como datos relacionados con estilo de vida, frecuencia de atención médica o tipo de cobertura, así como la exploración de técnicas de modelado más sofisticadas, como redes neuronales profundas o modelos basados en árboles optimizados, que podrían aumentar significativamente el rendimiento del sistema.

5.2. Recomendaciones

A partir de los hallazgos alcanzados en este proyecto, se recomienda considerar la ampliación de la aplicación del modelo predictivo hacia otras empresas clientes de la agencia de seguros. Si bien el presente trabajo se enfocó en una empresa con una cantidad considerable de empleados, el enfoque desarrollado ha demostrado ser flexible y escalable, permitiendo su adaptación a diferentes estructuras empresariales. Para lograrlo de manera efectiva, será necesario ajustar los hiperparámetros del modelo según las particularidades de cada filial, teniendo en cuenta la heterogeneidad de los datos y las diferencias en comportamiento entre regiones o segmentos.

Asimismo, se sugiere establecer un proceso continuo de actualización y mantenimiento del modelo. Dado que los patrones de atención médica, políticas internas y perfiles de los empleados pueden evolucionar con el tiempo, es fundamental reentrenar el

modelo de forma periódica utilizando datos actualizados. Esto permitirá que las predicciones se mantengan actualizadas y, al mismo tiempo, que el modelo se adapte con mayor eficacia a posibles variaciones en el entorno o en los patrones de uso de servicios médicos.

Una recomendación clave para futuras actualizaciones del modelo es la incorporación de nuevas variables explicativas. Variables relacionadas con el estilo de vida, historial clínico, nivel de cobertura del seguro, frecuencia de atención médica, e incluso indicadores socioeconómicos, podrían aportar mayor valor explicativo y enriquecer la precisión de las predicciones. Esta ampliación del espectro de datos también puede contribuir a una segmentación más fina y contextualizada de los empleados.

En cuanto a los algoritmos empleados, si bien el modelo actual ha mostrado un desempeño aceptable, se recomienda experimentar con modelos más avanzados de aprendizaje automático, tales como CatBoost o incluso redes neuronales profundas. Estas técnicas pueden ser especialmente útiles en contextos con alta complejidad o en situaciones donde existan relaciones no lineales entre las variables, permitiendo así un ajuste más fino del comportamiento del gasto médico.

Por otro lado, se considera estratégico el desarrollo de herramientas de visualización interactivas que permitan a los clientes corporativos explorar los resultados del modelo de manera sencilla e intuitiva. Dashboards personalizados por cliente pueden facilitar la comprensión del comportamiento de los gastos médicos, ayudando a identificar tendencias, prevenir situaciones de alto riesgo financiero y tomar decisiones informadas sobre estrategias de cobertura o prevención.

También es recomendable fortalecer el trabajo colaborativo entre los equipos técnicos encargados del desarrollo del modelo y los expertos del negocio. Esta sinergia garantiza que los resultados obtenidos sean no solo técnicamente sólidos, sino también relevantes y útiles para la toma dentro del campo asegurador.

Por último, se recomienda establecer e incorporar métricas concretas que permitan evaluar cómo el modelo influye en la satisfacción del cliente y en la optimización de los procesos dentro de la agencia de seguros. Analizar estos indicadores brindará una retroalimentación constante que impulse mejoras progresivas y refuerce el valor entregado tanto a la aseguradora como a sus clientes empresariales.

6. Referencias

Amazon Web Services. (s.f.). *Amazon Athena*. Obtenido de https://aws.amazon.com/athena/

Amazon Web Services. (s.f.). *Amazon Simple Storage Service (Amazon S3)*. Obtenido de https://aws.amazon.com/s3/

Amin, M., Liu, O. R., & Kensaku, S. A. (2020). Learning hidden patterns from patient multivariate time series data using convolutional neural networks: A case study of healthcare cost prediction. *Journal of Biomedical Informatics*.

Andika, T., Putra, J., Citra Lesmana, D., & Purnaba, G. (2021). *Prediction of Future Insurance Premiums When the Model is Uncertain.*

Asamblea Nacional del Ecuador. (2021). Obtenido de Ley Orgánica de Protección de Datos Personales (LOPDP): https://www.registroficial.gob.ec/

Burgos, M., & Manterola, C. (2010). Cómo interpretar un artículo sobre pruebas diagnósticas. *Rev. Chilena de Cirugía*, pp. 301-308.

Chan, N.-W., Lee, A.-H., & Zainol, Z. (2021). Predicting Employee Health Risks using Classification Ensemble Model. *International Conference on Information Retrieval and Knowledge Management*.

(2008). *Constitución de la República del Ecuador.* Montecristi / Quito: Registro Oficial.

Dutta, S., Bose, P., & Bandyopadhyay, S. (2023). Forecasting health insurance premium using machine learning approaches. *Asia-Pacific Journal of Science and Technology*.

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., ... & Vayena, E. (2018). AI4People—An ethical framework for a good AI society. *Minds and Machines*, 689–707.

González, A. (2022). Protección de datos personales y su aplicación en el sector salud en Ecuador. *Revista Jurídica Digital Andina*, 89–105.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys (CSUR)*, 1–42.

Haiyan, Y., Saeed, P., Hang, Q., Renying, X., & Hongxia, M. (2024). Personalized Algorithmic Pricing Decision Support Tool for Health Insurance: The Case of Stratifying Gestational Diabetes Mellitus into Two Groups. *Information & Management* .

Hassan, C., Iqbal, J., Hussain, S., AlSalman, H., Mosleh, M., & Sajid, S. (2021). A Computational Intelligence Approach for Predicting Medical Insurance Cost. *Mathematical Problems in Engineering*.

INEC. (2024). *Índice de precios al consumidor*. Obtenido de Ecuador en cifras: https://www.ecuadorencifras.gob.ec/indice-de-precios-al-consumidor/

Ismail, P., Stam, F., Portrait, A., & Van Witteloostuijn, X. K. (2024). Addressing unanticipated interactions in risk equalization: A machine learning approach to modeling medical expenditure risk. *Economic Modelling*.

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 389–399.

Kafuria, A. (2022). Predictive Model for Computing Health Insurance Premium Rates Using Machine Learning Algorithms. *International Journal of Computer (IJC)* , pp 21-38.

Kandula, A., Kalyanapu, S., Rayapalli, S. R., Modugumudi, V., & Kanikella, S. (2024). Medical Insurance Predictive Modelling: An Analysis of Machine Learning Methods. *2nd IEEE International Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation*.

Kemboi, N., Kasozi, J., & Nkurunziza, J. (2021). A Comparative Analysis of Machine Learning Models for the Prediction of Insurance Uptake in Kenya. *Data*, pp. 1 - 17.

Kuo, C.-Y., Yu, L.-C., Chen, H.-C., & Chan, C.-L. (2018). Comparison of models for the prediction of medical costs of spinal fusion in Taiwan diagnosis-related groups by machine learning algorithms. *Healthcare Informatics Research*, pp. 29 - 37.

Kwon, H., Park, S., Park, Y., Park, D., & Baik, S. (2024). Development of blood demand prediction model using artificial intelligence based on national public big data. *Digital Health*.

Lurie, P. (2007). *Actuarial Methods in Health Insurance Provisioning, Pricing and Forecasting Prepared by Peter Lurie.*

Manchester, S. (2014). An Examination of the Rising Costs of Employer-Sponsored Health Insurance in the United States: What Has Caused This Increase and What Can Be Done to Remedy the Problem? *Seminar Research Paper Series.*, Paper 25.

Mathauer, I., & Oranje, M. (2024). Machine learning in health financing: benefits. *Bulletin of the World Health Organization*, pp. 216 - 224.

Matloob, I., Khan, S., Hussain, F., Haider, W., Rukaiya, R., & Khalique, F. (2021). Need-based and optimized health insurance package using clustering algorithm. *Applied Sciences*.

Mendenhall, W. (2002). *Introducción a la probabilidad y estadística.* Paraninfo.

Mladenovic, S., Milovancevic, M., Mladenovic, I., Petrovic, J., Milovanovic, D., Petković, B. R., & Barjaktarović, M. (2020). Identification of the important

variables for prediction of individual medical costs billed by health insurance. *Technology in society*.

Molano, N., Rodríguez, K., Valera, D., & Vanegas, O. (2017). Calidad de vida en pacientes incidentes vs. prevalentes. ¿Hay diferencia en la calidad de vida? *Rev. Colomb. Nefrol.*, pp. 141 - 148.

Orji, U., & Ukwandu, E. (2024). Machine learning for an explainable cost prediction of medical insurance. *Machine Learning with Applications*.

Panda, S., Purkayastha, B., Das, D., Chakraborty, M., & Biswas, S. (2022). Health Insurance Cost Prediction Using Regression Models. *International Conference on Machine Learning, Big Data, Cloud and Parallel Computing,* .

Ridzuan, A., Azman, A., Marzuki, F., Faudzi, W., Aziz, S., & Bakar, N. (2024). Health Insurance Premium Pricing Using Machine Learning Methods. *Journal of Advanced Research in Applied Sciences and Engineering Technology*, pp. 134 - 141.

Sandra, J., Joshi, S., Ravi, A., Kodipalli, A., Rao, T., & Kamal, S. (2024). Prediction of Cost for Medical Care Insurance by Using Regression Models. *Lecture Notes in Electrical Engineering*.

Santos, A., Leal, G., & Balancieri, R. (2024). Identification of high-risk beneficiaries in private healthcare insurance. *Health Informatics Journal*.

Sharma, A., & Jeya, R. (2024). Prediction of Insurance Cost through ML Structured Algorithm. *International Conference on Computing, Power, and Communication Technologies*.

Stock, J., & Watson, M. (2002). *Introducción a la econometría.* Pearson.

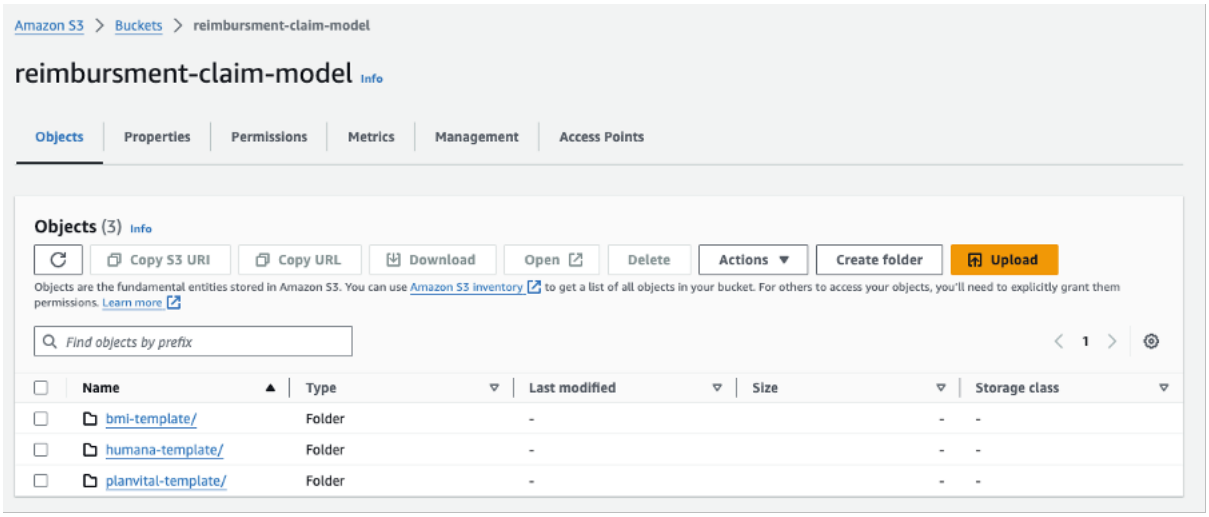
Syarifah, D., & Herdianto, K. (2023). Prediction of Health Insurance Claims Using Logistic Regression and XGBoost Methods. *Procedia Computer Science*, pp. 1012-1019.

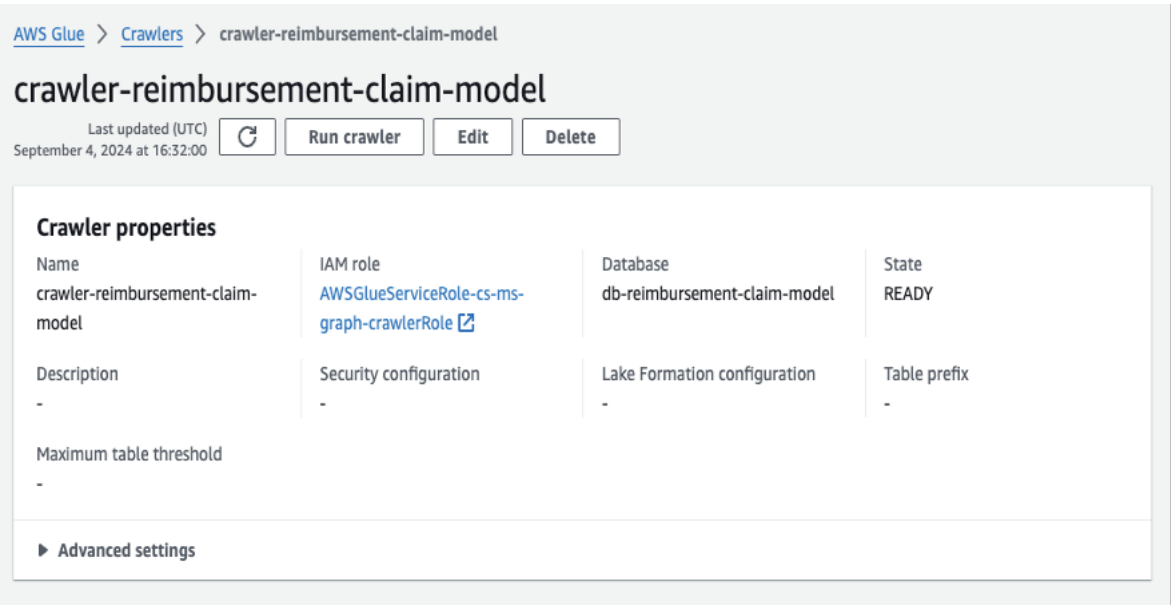
Van Rossum, G., & Drake, F. L. (2009). *Python 3 Reference Manual.* CreateSpace.

7. Anexos

Anexo 1.

Integración de datos en AWS





Como se puede obsevar en las imágenes se procedió a cargar la data en AWS para poder generar los modelos correspondientes y almacenar los resultados. De esta forma, se pueden comparar los distintos modelos y evaluar cuál posee un mejor rendimiento. Además, previo a la ejecución de los modelos, se realizó un análisis exploratorio que permitía comprender mejor los datos a emplear.

	cie10	total_pago	promedio_pago	cantidad_filas	diagnostico
1015	I10	288532367.0	8.905320e+05	324	HIPERTENSION ESENCIAL (PRIMARIA)
55	A09	137517748.0	1.049754e+06	131	DIARREA Y GASTROENTERITIS DE PRESUNTO ORIGEN I...
1130	J00	88077807.0	6.935260e+05	127	RINOFARINGITIS AGUDA (RESFRIADO COMUN)
1379	K21	72874524.0	3.643726e+06	20	ENFERMEDAD DEL REFLUJO GASTROESOFAGICO
1227	J22	69470886.0	2.105178e+06	33	INFECCION AGUDA NO ESPECIFICADA DE LAS VIAS RE...
1425	K30	66440828.0	8.004919e+05	83	DISPEPSIA
1307	J45	56887219.0	4.375940e+06	13	ASMA
1995	M511	55143556.0	2.757178e+06	20	TRASTORNOS DE DISCO LUMBAR Y OTROS CON RADICUL...
1181	J069	48195575.0	5.238649e+05	92	INFECCION AGUDA DE LAS VIAS RESPIRATORIAS SUPE...
1213	J20	47972184.0	1.142195e+06	42	BRONQUITIS AGUDA

La tabla anterior muestra los 10 principales diagnósticos en la que se incurren en mayor gasto presentado para la compañía de seguro; de manera general se puede interpretar que las mismas pertenece al tipo de atención *"Ambulatorio" y al grupo de enfermedades ocasionadas en las vías respiratorias.*

Anexo 2.

Balance de datos

	aseguradora	filial	proveedor	tiporeclamo	cie10	diagnostico	tipoatencion	tipoafiliado	sexo
count	84559	84559	84559	84559	84559	84559	84559	84559	84559
unique	3	20	2025	2	2727	2674	2	2	2
top	PLANVITAL	GISIS S.A.	CENTRO MEDICO VERIS	PAGO PROVEEDOR	A09	DIARREA Y GASTROENTERITIS DE PRESUNTO ORIGEN I...	AMBULATORIO	DEPENDIENTE	MUJER
freq	70750	46775	24998	66320	2877	3345	82884	44688	43575

En la tabla anterior, se puede deducir lo siguiente; hay un desbalance en los registros por aseguradora; la predomina es Plan Vital con unos 70750 registros, y hay unos 66320 reclamos asociados a pago a proveedores, por lo que se busca un balance entre estas observaciones.