ESCUELA DE NEGOCIOS

MAESTRÍA EN INTELIGENCIA DE NEGOCIOS Y CIENCIA DE DATOS

ESTIMACIÓN DEL COSTO MONETARIO DE UNA HOSPITALIZACIÓN EN UNA EMPRESA DE MEDICINA PREPAGADA MEDIANTE ALGORITMOS DE MACHINE LEARNING

Profesor

Manuel Eugenio Morocho

Autores

Ronny Santiago Benítez Quimbiulco

Félix Eduardo Jiménez Vásquez

2024

RESUMEN

El presente estudio tiene como objetivo desarrollar un modelo predictivo utilizando técnicas de Big Data y machine learning (ML) para estimar los costos de reclamaciones hospitalarias en el sector de medicina prepagada.

La precisión en la estimación de costos hospitalarios es esencial para una gestión eficaz en empresas de medicina prepagada. Los métodos tradicionales, basados en reglas fijas y datos históricos, suelen carecer de la flexibilidad necesaria para adaptarse a las complejidades y variaciones del entorno sanitario.

El enfoque de nuestro estudio se centra en la recopilación de datos históricos de reclamaciones y el preprocesamiento de estos para su análisis. Los métodos utilizados incluyen regresión lineal, Ridge y Lasso, además de un análisis de componentes principales (PCA) y una comparación con modelos avanzados de aprendizaje.

Tras evaluar diversos modelos, se determinó que la regresión lineal con selección de variables importantes y transformación logarítmica ofreció la mejor precisión en la predicción de costos. Los resultados obtenidos con PyCaret mostraron que modelos como LightGBM y Gradient Boosting proporcionaron métricas de error más bajas y mayor precisión predictiva, sugiriendo que el uso de técnicas de aprendizaje automático más avanzadas puede mejorar significativamente las predicciones. Se recomienda que la compañía continúe experimentando con estos modelos, capturando más datos relevantes para optimizar su capacidad predictiva y mejorar la gestión de recursos.

Keywords: Big Data, machine learning (ML), costos hospitalarios, medicina prepagada, regresión lineal, Ridge, Lasso, (PCA), PyCaret, LightGBM, Gradient Boosting.

ABSTRACT

The present study aims to develop a predictive model using Big Data and machine learning (ML) techniques to estimate hospital claim costs in the prepaid healthcare sector. Accurate estimation of hospital costs is essential for effective management in prepaid healthcare companies. Traditional methods, based on fixed rules and historical data, often lack the flexibility needed to adapt to the complexities and variations of the healthcare environment.

The focus of our study is on the collection of historical claim data and its preprocessing for analysis. The methods used include linear regression, Ridge, and Lasso, in addition to principal component analysis (PCA) and a comparison with advanced learning models.

After evaluating various models, it was determined that linear regression with important variable selection and logarithmic transformation offered the best accuracy in cost prediction. The results obtained with PyCaret showed that models such as LightGBM and Gradient Boosting provided lower error metrics and higher predictive accuracy, suggesting that the use of more advanced ML techniques can significantly improve predictions. It is recommended that the company continue experimenting with these models, capturing more relevant data to optimize predictive capabilities and improve resource management.

Keywords: Big Data, machine learning (ML), medical cost, medical insurance, linear regression, Ridge, Lasso, (PCA), PyCaret, LightGBM, Gradient Boosting.

ÍNDICE DEL CONTENIDO

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

INTRODUCCION

Estimar los costos hospitalarios de manera se ha vuelto una necesidad imperativa en las empresas de medicina prepaga, ya que esto tiene una incidencia directa en la planificación financiera y operativa de la organización, así como en la estimación de riesgos.

Por muchos años, se han utilizado métodos basados en reglas fijas y experiencias pasadas para calcular estos costos, pero estos enfoques pueden ser limitados y no siempre reflejan la realidad compleja y variable de los datos que a diario recibe la compañía por medio de sus reclamaciones de clientes.

Con el avance de la tecnología y el acceso a grandes volúmenes de datos, el machine learning (ML) ofrece una nueva forma de abordar este desafío. Los algoritmos de ML pueden analizar patrones y relaciones en los datos que los métodos tradicionales podrían pasar por alto, permitiendo estimaciones más precisas y adaptadas a las condiciones reales.

En el año 2024 la Auditora EY, considerada una de las 4 más importantes a nivel mundial, realiza una publicación donde se menciona que el ML es crucial en la industria aseguradora para mejorar la precisión en la evaluación de riesgos, agilizar el procesamiento de reclamaciones y ofrecer una experiencia personalizada al cliente, impulsando así la eficiencia operativa, económica y la innovación (Santenac et al., 2024).

Por otra parte, (Taloba et al., 2022) nos muestra que mediante la selección de variables principales como el índice de masa corporal (BMI), la edad, el género, si fuma y los costos individuales de atención médica, mediante un modelo de regresión lineal se puede estimar y predecir los costos de atención médica, así como en las hospitalizaciones.

Es por esta razón, que una precisión adecuada y efectiva en la estimación de costos no solo facilita una mejor planificación y asignación de recursos, sino que también ayuda a tomar decisiones más informadas y efectivas.

Este estudio explora cómo aplicar técnicas de ML para mejorar la estimación del costo de una hospitalización en una empresa de medicina prepaga. Al utilizar modelos avanzados y analizar datos históricos, buscamos superar las limitaciones de los métodos tradicionales y proporcionar una visión más clara y ajustada de los costos reales.

REVISIÓN DE LITERATURA

Panorama del Sector Asegurador Ecuatoriano

En los últimos años, el sector de seguros en Ecuador ha tenido un crecimiento notable. Según el Anuario 2023 publicado por Fedeseg (Federación Ecuatoriana de Empresas de Seguros), la demanda de pólizas de seguro, particularmente en los segmentos de vida y vehículos, ha tenido un crecimiento constante en los últimos 5 años. Durante el año anterior, esta tendencia alcista se manifestó en un aumento del 9,7% en las primas emitidas, que culminó con una suma total de 2.202,6 millones de dólares en diciembre de 2023 (Federación Ecuatoriana de Empresas de Seguro, 2024).

Dentro de los seguros de no vida, se encuentran los seguros de salud o planes de asistencia médica. Este tipo de seguro consiste en qué los usuarios por el pago de una prima periódica acceden de una red de prestadores médicos privados en donde reciben atención médica, quirúrgica y hospitalaria de la más alta calidad. El seguro se encargará de cubrir los gastos y asumir los riesgos.

Este segmento se enfrenta a un panorama complejo y en constante evolución, caracterizado por el aumento de los costos médicos, la prevalencia de enfermedades agudas o crónicas y la presión por ofrecer primas competitivas.

Según un estudio del Banco de la República de Colombia, se estima un aumento en la carga financiera en el sector salud debido al envejecimiento de la población y factores de riesgo relacionados con hábitos de vida poco saludables, con un incremento de costos del 40% entre 2022 y 2030 especialmente en enfermedades como cáncer, diabetes y enfermedades cardiovasculares (María et al., 2023).

En el Ecuador, podemos observar que con el pasar de los años, la población decide invertir en un seguro de salud privado, esto se ve reflejado en las cifras publicadas por la Federación Ecuatoriana de Empresas de Seguros, en donde se puede apreciar que la Variación de la Prima Neta en el ramo de Asistencia Médica en el año 2022 tuvo un incremento del 28,9% con relación al año 2021, y el año 2023 tuvo un incremento del 17,9% que significa una variación en más de USD 20,8 millones con relación al año 2022 (Federación Ecuatoriana de Empresas de Seguros, 2024).

TABLA DE VARIACIÓN NOMINAL (USD) Y RELATIVA (%)			
2019	Prima neta: 82,2M	Variación USD vs 2018 1,7M	Variación % vs 2018 2,2%
2020	Prima neta: 86,1M	Variación USD vs 2019 3,9M	Variación % vs 2019 4,8%
2021	Prima neta: 90,4M	Variación USD vs 2020 4,3M	Variación % vs 2020 5,0%
2022	Prima neta: 116,6M	Variación USD vs 2021 26,2M	Variación % vs 2021 28,9%
*** 2023**	Prima neta: 137,4M	Variación USD vs 2022 20,8M	Variación % vs 2022 17,9%

Ilustración 1 Variación Prima Neta Emitida.
Fuente: (Federación Ecuatoriana de Empresas de Seguros, 2024).

Está aceptación de la población por un seguro puede ser ocasionado por la saturación del sistema de salud público que a su vez refleja cada día una atención deficiente, esto acompañado al sentimiento de bienestar del individuo al contar con un respaldo económico en caso de una enfermedad, hacen que las empresas que ofertan dichos servicios tengan que tomar decisiones, diversifiquen el mercado y ofrezcan un servicio rentable y de calidad (Ortiz - Culcay et al., 2019).

Machine learning en Seguros

En su estudio, (Donado, 2022) define el ML como una rama de la Inteligencia Artificial que permite a las máquinas aprender de datos pasados sin ser explícitamente programadas para ello. Este enfoque permite a las máquinas analizar grandes cantidades de datos para identificar patrones y tomar decisiones sin intervención humana.

Hoy en día, los avances tecnológicos, como los macrodatos y el aprendizaje automático, han revolucionado las operaciones en varios sectores, incluido el

sector de los seguros, al mejorar la eficiencia y los procesos de toma de decisiones.

A continuación, se resaltará algunos de los usos más comunes del ML en el sector de seguros:

1. **Precios más precisos:** El ML permite a las aseguradoras evaluar riesgos de manera más precisa, lo que puede resultar en una fijación de precios más ajustada y personalizada.

   Según (Zhang & Walton, 2019) podemos utilizar modelos de aprendizaje automático, como los Modelos Lineales Generalizados (GLMs) y la Regresión de Procesos Gaussianos (GPR), en la estimación de precios en el contexto de seguros. Aspectos relevantes son:

   - Los GLMs son utilizados para modelar la relación entre primas, reclamaciones y otros factores en el seguro.
   - Al utilizar GLMs, las compañías de seguros pueden analizar datos históricos para predecir la demanda y las reclamaciones futuras, lo que lleva a estrategias de precios más precisas y optimización de ingresos.
   - GPR es una herramienta poderosa para modelar relaciones complejas entre variables y hacer predicciones basadas en datos limitados.
   - Al aplicar GPR, las compañías de seguros pueden optimizar estrategias de precios considerando incertidumbres en la demanda y las reclamaciones, lo que conduce a una gestión de ingresos más sólida.

2. **Mejora en la clasificación y retención de clientes:** Mediante el examen de los datos personales y de comportamiento de los clientes, los algoritmos de aprendizaje automático permiten a las organizaciones anticipar la pérdida de clientes y ejecutar tácticas de retención personalizadas, lo que genera una mayor rentabilidad.

Según el proyecto de investigación de (Andrea et al., 2023), los autores se enfocan en desarrollar una solución analítica para una compañía de seguros de automóviles que permita identificar a sus clientes vigentes según su probabilidad de retención y catalogarlos con la rentabilidad definida por las reglas de negocio utilizando algoritmos supervisados de clasificación como regresión logística, árboles de decisión, Random Forest, XGBoost, SVM, red neuronal y Catboost.

3. **Gestión de riesgos mejorada y prevención de fraude:** Al analizar datos en tiempo real, el ML puede ayudar a las aseguradoras a predecir costos futuros con la finalidad de reducir los riesgos económicos y a gestionar de forma oportuna el riesgo de fraude en las reclamaciones.

   Por ejemplo, la investigación de (Rodríguez, 2023) se centra en la identificación de actividades fraudulentas en la atención médica mediante la aplicación un modelo Isolation Forest y word2vec en la cual se analizan los reclamos de clientes presentados entre 2021 y 2023. El objetivo principal es identificar las irregularidades en la facturación de la atención médica y mejorar la eficacia en la identificación del fraude en el sector de los seguros de salud. En este sentido, el método Isolation Forest se aplica para separar los casos inusuales y descubrir posibles conductas fraudulentas, mientras que word2vec incorpora una agrupación del diagnóstico para mejorar el análisis.

Machine learning en la estimación de costos hospitalarios

Las metodologías tradicionales para la estimación de costos, como los modelos basados en reglas o la experiencia histórica, a menudo presentan limitaciones en cuanto a su precisión y flexibilidad. El ML ha emergido como una herramienta prometedora para abordar estas limitaciones, ofreciendo la capacidad de analizar grandes conjuntos de datos heterogéneos y descubrir patrones complejos que pueden ser utilizados para predecir con mayor precisión los costos hospitalarios.

En este contexto, la estimación precisa del costo de una hospitalización es fundamental para la toma de decisiones informadas en la gestión de riesgos, la tarificación de primas y la asignación de recursos.

En investigaciones académicas recientes, varios estudios se han centrado en estimar el costo monetario de la hospitalización en los seguros de salud utilizando metodologías de aprendizaje automático.

A continuación, se revisarán estudios similares, la metodología aplicada y los resultados obtenidos.

En su estudio, (Fan et al., 2024), para predecir los costos de hospitalización de pacientes con tuberculosis pulmonar utilizaron métodos de regresión múltiple y Perceptrón Multicapa (MLP). El objetivo del estudio fue mejorar la gestión de gastos y asignar recursos eficientemente. Durante el desarrollo se analizaron datos de 9570 pacientes en Kashgar entre 2020 y 2022 para estimar seis tipos de costos hospitalarios. El MLP mostró un mejor rendimiento predictivo, evaluado mediante R-cuadrado, error cuadrático medio (RMSE) y error absoluto medio (MAE), superando a la regresión múltiple. Factores, como la edad y la duración de la estancia, influyeron significativamente en los costos.

El estudio de (Donado, 2022) se centra en la predicción de la siniestralidad del asegurado en Seguros de Salud utilizando ML. Se utilizaron datos reales proporcionados por una entidad aseguradora, con 106,510 registros de asegurados activos en los últimos 12 meses. Se consideraron 14 variables independientes, incluyendo información del cliente y del producto. Aquí se implementaron varios algoritmos de ML, como el modelo lineal generalizado, SVM, y MARS. Se evaluaron las medidas de desempeño del modelo, como el error cuadrático medio (RSME) y el coeficiente de determinación R-cuadrado. Los resultados mostraron un RMSE de 32.9 y un R-cuadrado de 0.683, indicando un desempeño medio que sugiere la necesidad de explorar otros modelos para mejorar las predicciones.

En el estudio de (Casanova, 2022) sobre la predicción del costo de reclamaciones en el mercado asegurador, se aplicaron diferentes escenarios y modelos, incluyendo regresión lineal, árboles de regresión, Bagging y Gradient Boosting. Se identificaron variables significativas como edad, estado civil, número de hijos a cargo, y se encontró que los modelos de ensamble, especialmente Gradient Boosting, presentaron el mejor ajuste y menor error cuadrático medio en la predicción del costo de las reclamaciones.

El estudio de (Taloba et al., 2022) se centró en estimar y predecir los costos de hospitalización y atención médica utilizando modelos de regresión en ML. Se emplearon datos de registros de atención médica con información anónima sobre personas. Las variables incluidas fueron BMI, edad, género, hábito de fumar y número de hijos. Se implementó un modelo de regresión lineal y las medidas de evaluación incluyeron la comparación de los resultados con enfoques anteriores. Se obtuvo una precisión del 97.89% en la estimación de los costos de atención médica.

El estudio de (Rakshit et al., 2021) se centró en predecir los costos de atención médica de pacientes con cáncer de mama utilizando datos clínicos y económicos de 972 pacientes. Se emplearon 23 variables de actividades clínicas y se implementó el algoritmo KNN (K-Nearest Neighbors) para la predicción de costos. Las medidas de evaluación incluyeron el error absoluto porcentual medio (MAPE) y se compararon con otros métodos como Gradient Boosting, Redes Neuronales Artificiales y Elastic Net. Los resultados mostraron que el algoritmo propuesto superó a los métodos existentes, con un MAPE inferior al 6%.

Los Autores (ul Hassan et al., 2021), se enfocaron en predecir los costos de seguros médicos utilizando técnicas de inteligencia computacional. Se utilizaron datos de costos de seguros médicos obtenidos de KAGGLE. Las variables consideradas en el estudio incluyeron información sobre si el individuo es fumador, su índice de masa corporal (IMC) y su edad. Se aplicaron algoritmos como Regresión Lineal (LR), Stochastic Gradient Boosting (SGB), XGBoost (XGB), entre otros. Las medidas de evaluación utilizadas fueron el Error

Cuadrático Medio (RMSE) y la validación cruzada. Los resultados mostraron que el modelo SGB tuvo un AUC de 86%.

El estudio de (Maisog et al., 2019) demuestra la eficacia de algoritmos predictivos avanzados al identificar a las personas que solicitan reclamos de alto costos en el sistema de salud, logrando AUC-ROC superiores al 90%. Se utilizaron métodos como Bosques Aleatorios, Máquinas de Soporte Vectorial y Árboles Impulsados por Gradiente, destacando el rendimiento sobresaliente del algoritmo LightGBM. La identificación precisa de estos reclamantes permite implementar intervenciones efectivas y lograr ahorros significativos en costos de atención médica, abriendo nuevas oportunidades en el cuidado de la salud.

(Hanafy & Mahmoud, 2021) utilizaron técnicas de aprendizaje automático para predecir costos de seguros de salud, comparando varios modelos de regresión como regresión lineal múltiple, modelo aditivo generalizado, máquinas de vectores de soporte, bosques aleatorios, árbol de decisión, XGBoost, vecinos más cercanos, impulso gradual estocástico y redes neuronales profundas. Se evalúan métricas como MAE (0.17448), RMSE (0.38018) y un R-cuadrado del 85.8295 para el modelo de Impulso Estocástico. Los datos provienen de un conjunto de datos de Kaggle con siete atributos. El objetivo es mejorar la estimación de costos de seguros para atraer consumidores y optimizar la formulación de planes individuales.

IDENTIFICACIÓN DEL OBJETO DE ESTUDIO

La empresa de medicina prepagada objeto del estudio, gestiona una gran cantidad de datos relacionados con las reclamaciones hospitalarias presentadas por clientes individuales y empresariales. Miles de datos son procesados diariamente mediante los sistemas transaccionales, y esa información no es aprovechada en su totalidad.

En la actualidad, la compañía centra sus estimaciones y proyecciones en métodos estadísticos y actuariales tradicionales, aplicando únicamente las Reservas de Servicios Prestados y no Reportados (IBNR) que se refiere a las

reclamaciones futuras estimadas que una compañía de seguros espera pagar sobre pólizas que ya han sido emitidas pero que no han sido reportadas por los asegurados ni registradas en los libros de la compañía (Maribel et al., 2022).

Hasta el momento, no se ha implementado técnicas de ML que permitan anticipar los costos asociados a estas atenciones hospitalarias que puedan ocurrir dentro de la próxima vigencia de un cliente.

El no hacer estimaciones robustas y basadas en datos, genera problemas significativos en la planificación financiera y en la toma de decisiones estratégicas, afecta la eficiencia operativa y optimización de recursos.

El objetivo de este estudio es aplicar algoritmos de aprendizaje supervisado, en este caso modelos de regresión lineal para predecir con precisión los costos hospitalarios basados en datos de reclamaciones médicas, con el fin de mejorar la gestión económica y facilitar la toma de decisiones efectivas en cuanto a los costos hospitalarios.

Según (Yao et al., 2023), para una predicción exitosa, se requiere un proceso de entrenamiento del modelo de regresión el cual necesita utilizar una gran cantidad de datos de los individuos, como la edad, el sexo, la ocupación y demás datos que podemos encontrar en las reclamaciones. Mediante el análisis de estos datos, el algoritmo de regresión puede aprender la relación matemática entre variables, prediciendo así el costo del seguro de los clientes.

Los beneficios que podemos encontrar al tener al aplicar estas técnicas de ML en la predicción de costos hospitalarios son:

- La proyección óptima de los costos producto de hospitalizaciones de los usuarios permitirá una mejor planificación de los recursos financieros, estableciendo reservas adecuadas y ajustando las tarifas de los planes médicos para las empresas de manera más precisa.
- Con una estimación más acertada de los costos futuros hospitalarios, se podrán optimizar los recursos destinados a la gestión de reclamaciones,

mejorando la eficiencia operativa y reduciendo costos que no se requieren.

- El análisis de la siniestralidad de las empresas facilitará la identificación de patrones y tendencias, permitiendo la toma de decisiones estratégicas que reduzcan el riesgo.
- La proyección de los costos permitirá asignar otros recursos para la satisfacción de los clientes, como el servicio, mejorando así su satisfacción y fidelización con la empresa.
- Permitirá obtener una ventaja competitiva en el mercado de seguros de salud.

PLANTEAMIENTO DEL PROBLEMA

Naturaleza del Problema

La compañía ha centrado sus esfuerzos en la Analítica descriptiva, herramientas como Power BI y Qlik Sense son usadas para el análisis de costo y la toma de decisiones, sin embargo, este enfoque retrospectivo hace que las jefaturas y gerencias basen sus estrategias sobre lo que pasó, y no sobre lo que va a pasar.

En torno a la salud, cada usuario genera información muy valiosa y la enorme cantidad de datos acumulados sobre sus reclamaciones no se está utilizando de manera efectiva para generar descubrimientos valiosos y predecir tendencias futuras. La ausencia de un modelo predictivo que permita estimar cuánto puede costar una hospitalización, o estimar cuales pacientes pueden ser de alto costo crea una incertidumbre considerable en la planificación y gestión financiera de la empresa.

El estudio realizado por (Aguirre, 2022) indica que, tradicionalmente, las aseguradoras calculan las reservas mediante procedimientos basados en datos históricos de grupos de pólizas similares, lo cual puede resultar en estimaciones inexactas debido a la falta de consideración de las características individuales de los asegurados y los cambios en los montos de los siniestros a lo largo del tiempo. Esta inexactitud puede afectar negativamente la estabilidad financiera

de las compañías de seguros y sus decisiones de inversión. Además, existe una necesidad urgente de innovación en la industria aseguradora, impulsada por la competencia en el mercado y la demanda de personalización por parte de los clientes. Con el presente estudio, buscamos analizar y comparar diferentes modelos avanzados de ML que puedan predecir de manera más precisa las reservas o costo futuros, y dar la pauta para que estas técnicas puedan ser aplicadas en más procesos de negocio, facilitando decisiones más informadas y eficientes en la gestión del riesgo y los recursos financieros.

El estudio realizado por (Serrano et al., 2022) nos menciona la importancia de poder estimar de una forma adecuada los costos médicos en la industria de la salud, especialmente para la planificación, elaboración de presupuestos y toma de decisiones eficientes.

A pesar de la disponibilidad de datos, la predicción exacta de estos costos sigue siendo un desafío debido a la complejidad de las variables sociodemográficas y de salud de los pacientes, así como a la falta de herramientas analíticas específicas. Dentro de nuestra revisión literaria, la mayoría de los autores aplica varias técnicas de limpieza y segmentación de datos para obtener un resultado óptimo.

¿Por qué el Problema es Crítico para la Organización?

Este problema es crítico para la compañía por varias razones:

- La incapacidad para prever los costos de las reclamaciones lleva a una gestión financiera ineficiente, la cual dificulta la planificación adecuada de las reservas financieras y el ajuste de las tarifas en los planes médicos.
- La falta de previsión resulta en asignaciones ineficientes de recursos, lo cual incrementa los costos operativos y afecta la capacidad de la empresa para responder de manera efectiva.
- Una gestión deficiente en la proyección de los costos posibles y comportamiento del costo de la próxima vigencia del cliente podría causar grandes pérdidas monetarias para la compañía.

Justificación para Adoptar un Enfoque Analítico

Adoptar un enfoque analítico basado en técnicas de Big Data y ML se justifica por las siguientes razones:

- Las técnicas de ML permitirán analizar grandes volúmenes de datos históricos, identificando patrones no evidentes y mejorando la precisión en la predicción de costos futuros.
- Un modelo predictivo proporcionará una adecuada planificación y asignación de recursos, reduciendo costos operativos y mejorando la eficiencia en la gestión de reclamaciones.
- El análisis de datos facilitará datos valiosos que permitirán a la empresa tomar decisiones estratégicas informadas, mejorando la gestión de riesgos y la calidad del servicio

OBJETIVO GENERAL

Desarrollar e implementar un modelo predictivo basado en técnicas de Big Data y ML para estimar los costos de reclamaciones hospitalarias en el sector de seguros de salud, con el objetivo de mejorar la planificación financiera, optimizar recursos y tomar decisiones estratégicas informadas.

OBJETIVOS ESPECÍFICOS

1. **Recopilación y Preprocesamiento de Datos:** Recopilar datos históricos de reclamaciones hospitalarias, incluyendo detalles como tipo de reclamación hospitalaria, valor pagado, diagnostico, paciente, edad paciente, genero paciente, prestador, datos demográficos de los usuarios, y preprocesar estos datos para su análisis.
2. **Selección y Evaluación de Modelos Predictivos:** Evaluar y seleccionar algoritmos de ML adecuados para la predicción de costos, como regresión lineal.

3. **Entrenamiento y Validación de Modelos:** Entrenar los modelos seleccionados utilizando técnicas de validación cruzada para asegurar la robustez y precisión de las predicciones.
4. **Análisis de Resultados y Toma de Decisiones Estratégicas:** Identificar patrones y tendencias, y tomar decisiones estratégicas que mejoren la gestión de riesgos y optimicen la planificación financiera. Utilizar las predicciones del modelo para analizar las predicciones de los costos hospitalarios, clasificar posibles pacientes de alto costo, ajustar valores de primas y realizar alianzas estratégicas con hospitales que tengan los mejores costos en torno a enfermedades críticas.

JUSTIFICACIÓN Y APLICACIÓN DE LA METODOLOGÍA

Para nuestro estudio vamos a trabajar con una base de datos privada que contiene datos de las reclamaciones de casos hospitalarios que se dieron en el período enero 2023 / mayo 2024 en una empresa de medicina prepagada del Ecuador.

En vista de que la información médica puede contener datos sensibles, no se dispondrá de aquellos datos que puedan identificar a clientes, por su lado se incluyen códigos únicos de personas.

Recopilación de datos

La compañía cuenta con diversos sistemas transaccionales, en donde la información, en su mayoría, se almacenan en bases de datos estructuradas de tipo SQL.

Para la obtención de los datos se ha utilizado la herramienta Qlik Sense, la cual mediante un proceso de extracción, transformación y carga (ETL) nos permite relacionar datos de diferentes orígenes en un gran repositorio central para que posteriormente esos datos puedan ser consumidos por los usuarios.

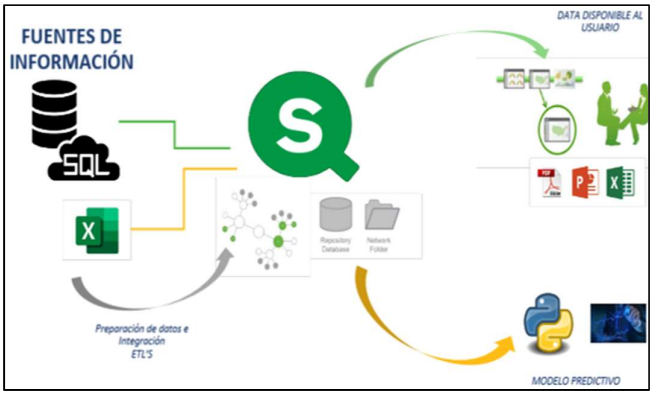


Ilustración 2 Ecosistema ETL.
Fuente: Los autores.

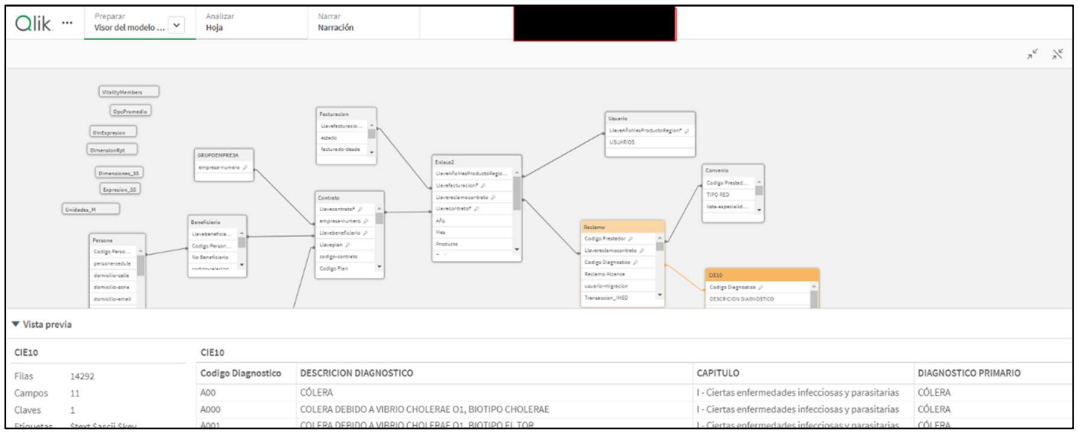


Ilustración 3 Ejemplo modelado de datos mediante Qlik Sense.
Fuente: Los autores.

Una vez transformada y cargado nuestros datos, procedemos a realizar la descarga de nuestra data set con los siguientes filtros de información:

a) Fecha incurrencia: Desde el 01-01-2023 al 31-05-2024
b) Tipo reclamo: “HOSPITALARIOS”
c) Lugar-atención: “Hospital”

Para toda la fase de ejecución se usará Phyton y las librerías que este entorno ofrece. Todos los pasos de nuestro estudio han sido cargados al repositorio de GitHub (Benítez & Jiménez, 2024).

Identificación y descripción de variables

Tabla 1 Variables

Nombre de la variable	Descripción	Tipo de variable	Tipo dato Phyton
IdCaso	Identificador único de caso hospitalario.	Cualitativa	object
Fecha Incurrencia	Fecha de la hospitalización.	N/A	object
Fecha inicio vigencia	Fecha desde que está afiliado el cliente.	N/A	object
Antigüedad cliente meses	Meses desde la afiliación hasta la hospitalización.	Cuantitativa	int64
Codigo paciente	Identificador único del cliente.	Cualitativa	int64
Edad Paciente	Edad del paciente a la fecha de hospitalización.	Cuantitativa	int64
Grupo Etario	Clasificación de rangos de edades.	Cualitativa	object
Genero	Genero Paciente.	Cualitativa	object
Provincia Cliente	Provincia de residencia del cliente.	Cualitativa	object
Producto	Tipo de producto contrato por el cliente.	Cualitativa	object
Codigo prestador principal	Identificador único del prestador de la hospitalización.	Cualitativa	int64
Tipo prestador principal	Categoría del prestador, si es hospital, centro médico, etc.	Cualitativa	object
Con convenio	Indica si el prestador tiene algún convenio de pago o financiamiento con la empresa.	Cualitativa	object
Provincia prestador principal	Provincia del establecimiento de la hospitalización.	Cualitativa	object
Tipo reclamo	Indica si es hospitalario o ambulatorio.	Cualitativa	object
Dias hospitalizacion	Cantidad de días que el cliente pasó hospitalizado.	Cuantitativa	int64
Nivel contrato	Indica el nivel del plan contratado, 3, 4, 5. Mientras más alto es el nivel, más cobertura dispone.	Cualitativa	int64
Canal Venta	Medio por el cual se vendió el contrato.	Cualitativa	object
Severidad	Indica si el caso tiene una severidad baja, media, alta, etc.	Cualitativa	object
Código Diagnostico	Código CIE10 que describe la enfermedad por la que se hospitalizó.	Cualitativa	object

Nombre de la variable	Descripción	Tipo de variable	Tipo dato Phyton
Código dx primario	Código de cabecera CIE10 que describe la enfermedad por la cual fue hospitalizado.	Cualitativa	object
Dx Primario	Descripción de la enfermedad general.	Cualitativa	object
Dx Final	Descripción de la enfermedad detallada.	Cualitativa	object
Familia dx relacionados	Indica si los diagnósticos están relacionados.	Cualitativa	object
Agrupacion Dx relacionados	Agrupa los diagnósticos en categorías similares.	Cualitativa	object
Terapia Intensiva	Indica si el cliente fue internado en terapia intensiva.	Cualitativa	object
Preexistencia	Indica si el diagnóstico es preexistente.	Cualitativa	object
Emergencia Vital	Indica si al momento de la hospitalización el cliente tenía en riesgo su vida.	Cualitativa	object
lugar-atencion	Indica el lugar donde se realiza la atención.	Cualitativa	object
Tipo procedimiento	Indica el procedimiento realizado, ejemplo Clínico o Quirúrgico.	Cualitativa	object
Tipo tratamiento	En base al diagnóstico indica si la enfermedad es aguda, crónica o crónica de tratamiento continuo	Cualitativa	object
Tipo transaccion	Si el caso se presentó por Crédito (financiado) o reembolso.	Cualitativa	object
Rubro facturado	Indica el desglose de los valores que se facturaron en la atención, ejemplo: Servicios hospitalarios, honorarios médicos, medicinas, etc.	Cualitativa	object
Valor facturado	Valor de la hospitalización antes de aplicar condiciones contractuales.	Cuantitativa	object
Valor cubierto	Valor que la empresa cubre del monto facturado.	Cuantitativa	object
Valor pagado	Valor que es pagado al prestador o cliente luego de aplicar las condiciones contractuales.	Cuantitativa	Object

Se obtiene un DataSet con 115118 registros y 36 columnas.

Limpieza, pre-procesamiento y/o transformación de datos.

Esta fase se centra en la corrección de los datos importados, para la obtención de un DataSet depurado que será utilizado en el análisis exploratorio preliminar y posteriormente será la fuente de los modelos de ML.

Dentro de las principales correcciones tenemos las siguientes:

- **a)** Corrección del tipo de dato, en las variables cualitativas y cuantitativas.
- **b)** Eliminación de registros con edades inconsistentes: En este punto se verifica que el DataSet contiene edades con valor -1, se procede a eliminar.
- **c)** Generación de DataSet agrupado por caso: En este punto se agrupa nuestra variable objetivo Valor facturado, y se obtiene columnas de valores por cada rubro.
- **d)** Se crea columnas adicionales para evitar el sobreajuste, indicando si dentro de la atención se incluye o no cualquiera de los servicios facturados.
- **e)** Verificación y corrección de datos pérdidas o duplicados: Después de la transformación no se encuentran datos de este tipo

Análisis exploratorio de los datos.

En esta fase realizaremos una visualización de las variables para posteriormente tratar los datos atípicos.

- **a) Descriptivos estadísticos:**

	Antigüedad cliente meses	Edad Paciente	Dias hospitalizacion	Nivel contrato	HONORARIOS MEDICOS	INSUMOS Y SUMINISTROS	LABORATORIOS	MEDICINAS	OTROS	PROTESIS	SERVICIOS HOSPITALARIOS	TERAPIA INTENSIVA	Valor facturado total	Valor cubierto total	Valor pagado total
count	15893.000000	15893.000000	15893.000000	15893.000000	15893.00000	15893.000000	15893.000000	15893.000000	15893.000000	15893.000000	15893.000000	15893.000000	15893.000000	15893.000000	15893.000000
mean	64.993393	34.373498	2.739823	4.221229	921.00452	368.809621	302.153985	585.716264	35.974894	153.061890	856.939937	56.552051	3280.213161	3062.982651	2878.211412
std	78.627108	24.430211	2.972112	1.396396	1432.04730	1400.482970	551.375334	1694.365019	260.657212	901.237278	1642.676018	456.689029	5145.101868	4689.336338	4512.352742
min	0.000000	0.000000	0.000000	3.000000	0.00000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	1.270000	0.000000	0.000000
25%	13.000000	9.000000	1.000000	3.000000	166.40000	0.000000	40.840000	58.470000	0.000000	0.000000	269.980000	0.000000	942.200000	881.260000	797.970000
50%	36.000000	34.000000	2.000000	4.000000	424.84000	65.700000	148.730000	220.750000	0.000000	0.000000	500.000000	0.000000	1916.500000	1784.380000	1672.110000
75%	77.000000	55.000000	3.000000	5.000000	1226.87000	281.070000	345.440000	568.040000	0.000000	0.000000	977.900000	0.000000	3881.720000	3611.610000	3399.630000
max	360.000000	98.000000	79.000000	8.000000	53193.53000	65617.220000	16805.380000	72701.160000	22200.730000	28746.530000	76806.470000	18450.000000	195617.220000	107617.270000	107617.270000

Ilustración 4 Descriptivos estadísticos.
Fuente: Los autores.

Observaciones:

- La mayoría de las hospitalizaciones son cortas, con una media de 2.74 días y una mediana de 2 días.
- Existe una alta variabilidad en los días de hospitalización, con un rango que va de 0 a 79 días
- Los honorarios médicos tienen un promedio de 921.00 y una alta desviación estándar de 1432.05.
- El 75% de los casos tiene costos por debajo de 3881.72, pero hay algunos casos con costos muy altos.
- El promedio facturado es de 3280.21, pero la mediana es de 1916.50, indicando la presencia de casos con valores extremadamente altos.

b) Visualización de variables

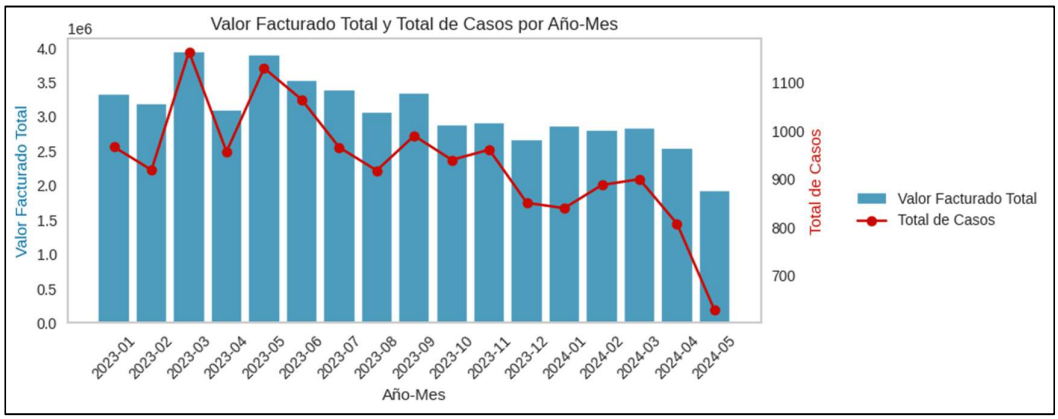


Ilustración 5 Evolución valor facturado y cantidad de casos por mes.
Fuente: Los autores.

Podemos observar que, a lo largo del tiempo, el valor facturado ha ido disminuyendo al igual que la cantidad de casos, sin embargo, como tenemos solamente 17 meses, no se aprecia una tendencia o algún tipo de estacionalidad.

Con las variables categóricas procedemos a realizar gráficos para entender la relación con nuestra variable objetivo Valor facturado total.

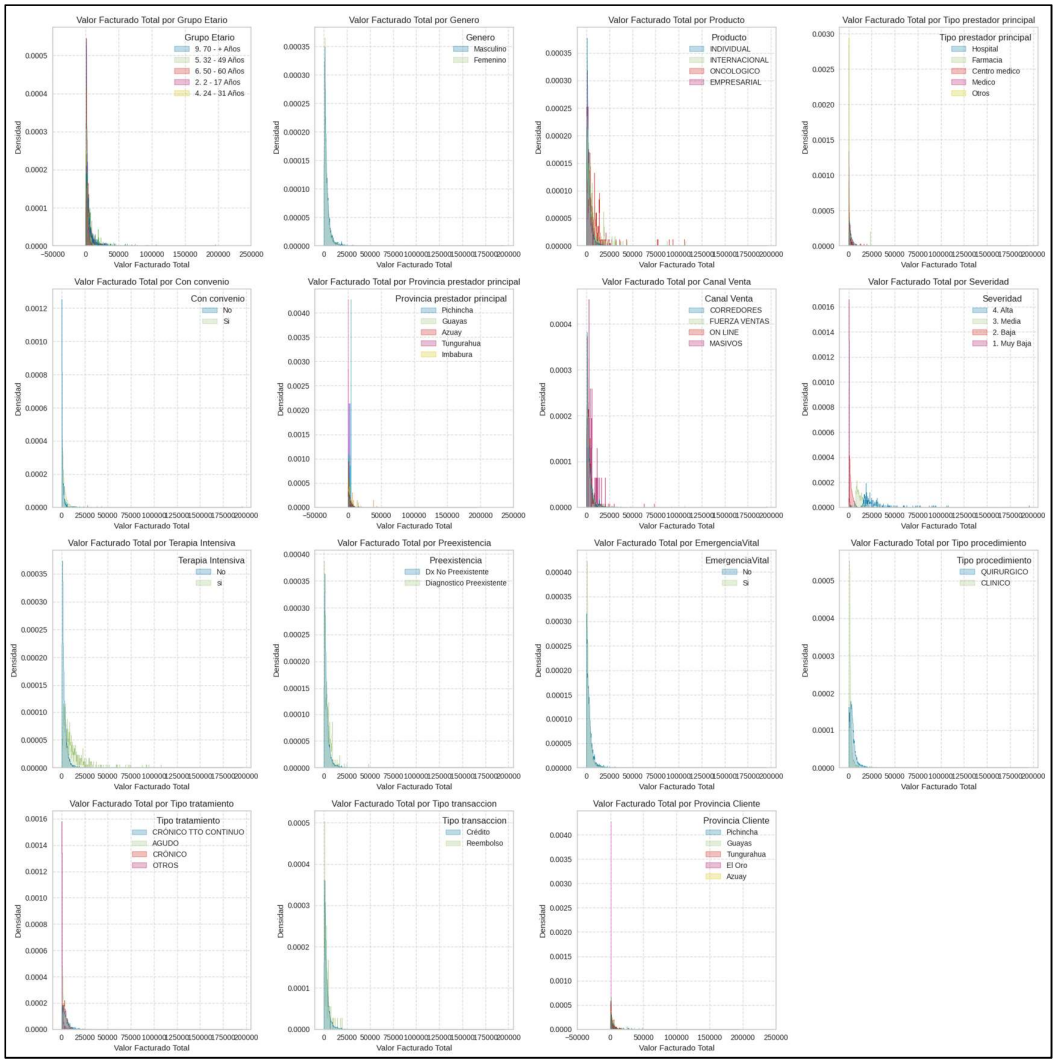


Ilustración 6 Gráficos de distribuciones categóricas.
Fuente: Los autores.

Podemos observar que hay clases dentro de las variables categóricas que tienen muy pocos datos, por ejemplo, Provincia Prestador y Provincia Cliente, por lo que se agruparán esos datos en una clase de tipo “Otros”.

En cuanto a las variables numéricas se establece la relación entre la variable objetivo mediante gráficos de dispersión.

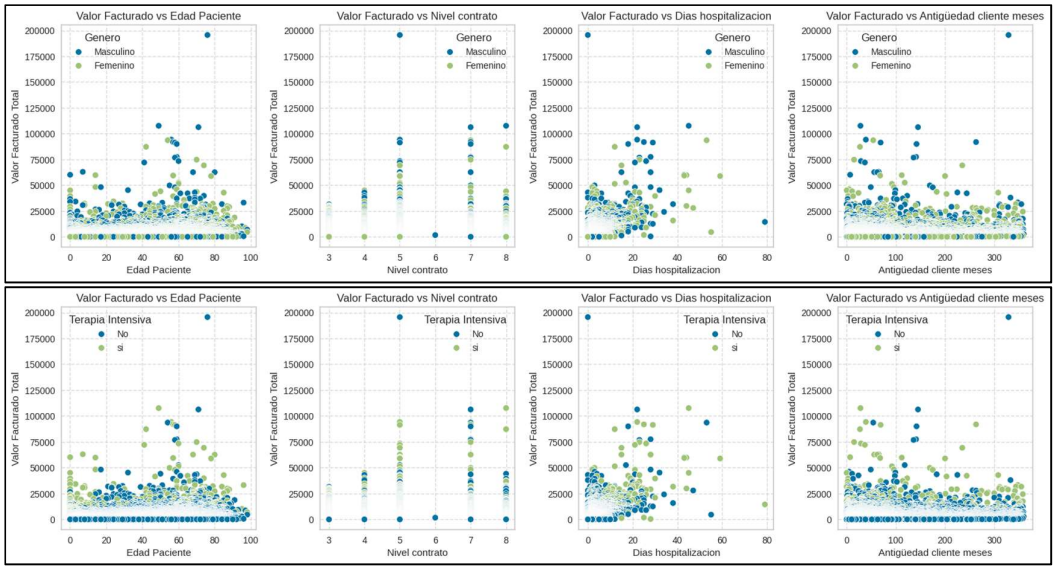


Ilustración 7 Relación variables numéricas por valor facturado total.
Fuente: Los autores.

Algunas observaciones que notamos se describen a continuación:

- El mayor monto facturado corresponde a edades comprendidas entre 32 y 49 años.
- Existe una tendencia positiva entre los días de hospitalización y el valor facturado total. A medida que aumentan los días de hospitalización, también tiende a aumentar el valor facturado total.
- La relación es más pronunciada para los pacientes que han estado en terapia intensiva.

Tratamiento de valores atípicos

- En base a la experiencia eliminamos todos los registros cuyo valor facturado sea menor a 300.00. (4.8 %)

  Se utiliza el método del rango intercuartílico (IQR) para identificar y eliminar los valores atípicos. Esto se logró calculando el IQR y filtrando los datos que se encontraban fuera de los límites definidos por: $Q1 - 1.5 \times IQR$ y $Q3 - 1.5 \times IQR$.

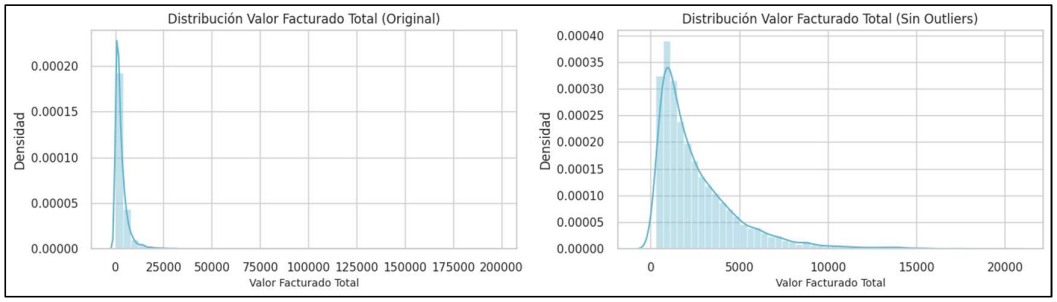


Ilustración 8 Distribución valor facturado. Fuente: Los autores.

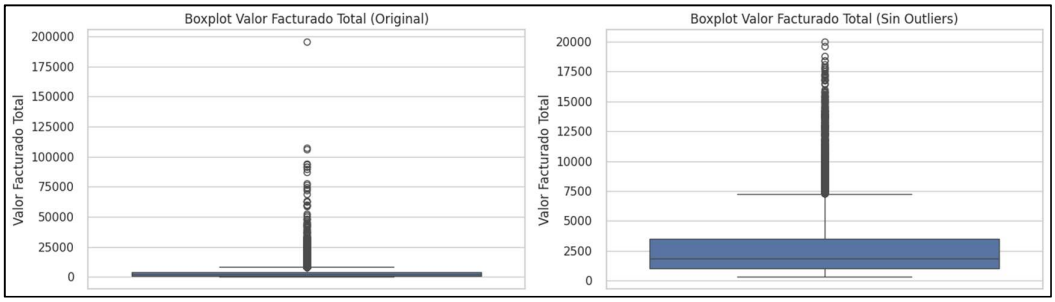


Ilustración 9 BoxPlot valor facturado. Fuente: Los autores.

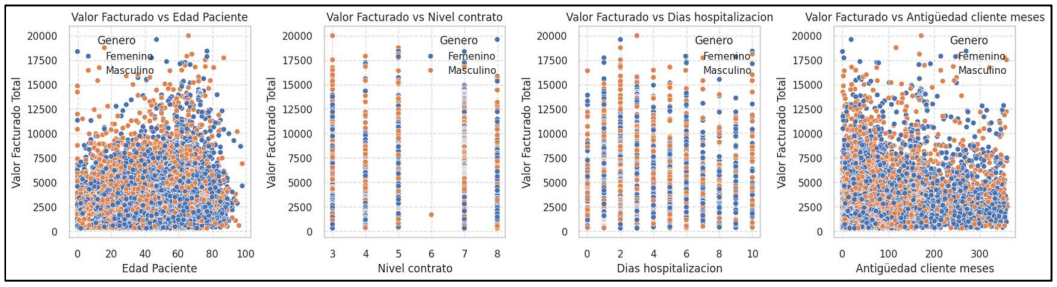


Ilustración 10 Gráficos de dispersión variables numéricas.
Fuente: Los autores.

c) Matriz de correlación.

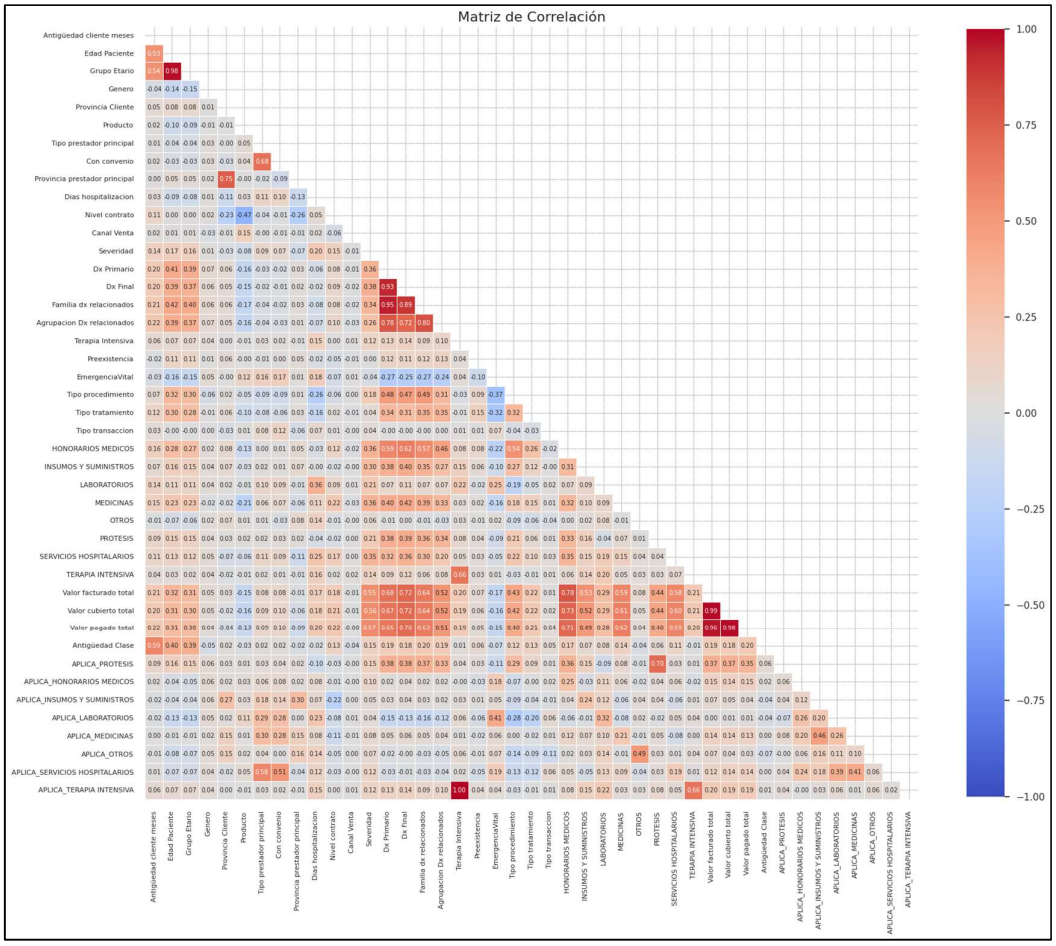


Ilustración 11 Matriz de correlación de Pearson.
Fuente: Los autores.

Observaciones:

La severidad de los casos influye significativamente en los costos médicos. Los procedimientos y tratamientos tienen un impacto notable en los honorarios médicos, medicinas y servicios hospitalarios. Además, la terapia intensiva y los diagnósticos son factores clave en la predicción de los costos totales de hospitalización.

Modelado de Datos:

Nuestro estudio busca probar diferentes métodos para obtener una adecuada precisión de los costos facturados.

Para las variables cualitativas se aplicará los siguientes métodos.

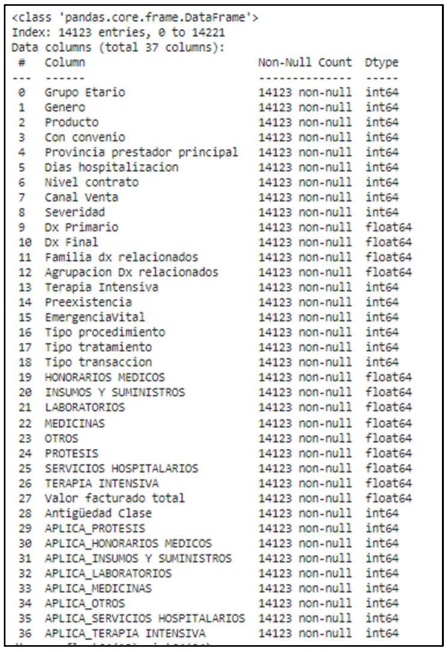
a) Para las variables con dos niveles se codifican con 0 y 1

b) Para las variables con más niveles se aplica Label Encoder

c) Para las variables con más de 50 clases como los Diagnósticos, e método One Hot Encoding no resulta efectivo ya que crea demasiadas columnas y su rendimiento se ve afectado, por lo tanto, aplicamos Target Encoding que transforma una variable categórica en una numérica utilizando la media de la variable objetivo para cada categoría. Primero, se calcula la media de la variable objetivo para cada categoría única y luego se sustituye cada categoría por su media correspondiente. Este método simplifica la integración de variables categóricas en modelos numéricos.

d) Para la ejecución de los modelos usamos las librerías de statsmodels y sklearn del Phyton.

A continuación, se muestran las variables que se utilizarán en la aplicación de los modelos, una vez tratadas las variables cualitativas.



```
<class 'pandas.core.frame.DataFrame'>
Index: 14123 entries, 0 to 14221
Data columns (total 37 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   Grupo Etario                    14123 non-null  int64
 1   Genero                          14123 non-null  int64
 2   Producto                        14123 non-null  int64
 3   Con convenio                    14123 non-null  int64
 4   Provincia prestador principal   14123 non-null  int64
 5   Dias hospitalizacion            14123 non-null  int64
 6   Nivel contrato                  14123 non-null  int64
 7   Canal Venta                     14123 non-null  int64
 8   Severidad                       14123 non-null  int64
 9   Dx Primario                     14123 non-null  float64
 10  Dx Final                        14123 non-null  float64
 11  Familia dx relacionados         14123 non-null  float64
 12  Agrupacion Dx relacionados      14123 non-null  float64
 13  Terapia Intensiva               14123 non-null  int64
 14  Preexistencia                   14123 non-null  int64
 15  EmergenciaVital                 14123 non-null  int64
 16  Tipo procedimiento              14123 non-null  int64
 17  Tipo tratamiento                14123 non-null  int64
 18  Tipo transaccion                14123 non-null  int64
 19  HONORARIOS MEDICOS              14123 non-null  float64
 20  INSUMOS Y SUMINISTROS           14123 non-null  float64
 21  LABORATORIOS                    14123 non-null  float64
 22  MEDICINAS                       14123 non-null  float64
 23  OTROS                           14123 non-null  float64
 24  PROTESIS                        14123 non-null  float64
 25  SERVICIOS HOSPITALARIOS         14123 non-null  float64
 26  TERAPIA INTENSIVA               14123 non-null  float64
 27  Valor facturado total           14123 non-null  float64
 28  Antigüedad Clase                14123 non-null  int64
 29  APLICA_PROTESIS                 14123 non-null  int64
 30  APLICA_HONORARIOS MEDICOS       14123 non-null  int64
 31  APLICA_INSUMOS Y SUMINISTROS    14123 non-null  int64
 32  APLICA_LABORATORIOS             14123 non-null  int64
 33  APLICA_MEDICINAS                14123 non-null  int64
 34  APLICA_OTROS                    14123 non-null  int64
 35  APLICA_SERVICIOS HOSPITALARIOS  14123 non-null  int64
 36  APLICA_TERAPIA INTENSIVA        14123 non-null  int64
```

Ilustración 12 Variables para modelo.
Fuente: Los autores.

1. Regresión Lineal Múltiple

La regresión lineal múltiple es una técnica que modela la relación entre una variable dependiente (y) y múltiples variables independientes $(x_1, x_2, \dots, x_n)$. El objetivo es encontrar los coeficientes $(\beta_0, \beta_1, \beta_2, \dots, \beta_n)$ que minimicen la suma de los errores cuadráticos entre las predicciones del modelo y los valores observados.

En el contexto de predicción de costos hospitalarios, estos modelos permiten identificar cómo factores como la edad, el género, la duración de la hospitalización y las condiciones médicas específicas influyen en el costo total. Al proporcionar una ecuación matemática basada en datos históricos, la regresión lineal múltiple es muy utilizada para este tipo de estudios.

Función Matemática:

$$[y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \epsilon]$$

Durante el presente programa pudimos observar diferentes enfoques, por ejemplo, en el enfoque econométrico, la regresión lineal se utiliza para analizar la relación entre variables económicas, enfocándose en la interpretación de coeficientes y validez de hipótesis económicas.

Por el contrario, en el enfoque de ML, la regresión lineal se emplea para predecir sin asumir una estructura específica, priorizando la precisión en la predicción sobre la interpretación de coeficientes.

2. Regresión Ridge

La regresión Ridge es una variante de la regresión lineal que añade una penalización a la magnitud de los coeficientes (β_j). Esta penalización es proporcional a la suma de los cuadrados de los coeficientes. El parámetro de regularización (α) controla la fuerza de la penalización. La regresión

Ridge ayuda a reducir el sobreajuste y a manejar la multicolinealidad al reducir los coeficientes grandes.

Función Matemática:

$$[\min_{\beta}\left(\sum_{i=1}^{m}\left(y_i-\beta_0-\sum_{j=1}^{n}\beta_j x_{ij}\right)^2+\alpha\sum_{j=1}^{n}\beta_j^2\right)]$$

3. Regresión Lasso

La regresión Lasso es otra variante de la regresión lineal que añade una penalización a la magnitud de los coeficientes (β_j). A diferencia de Ridge, la penalización es proporcional a la suma de los valores absolutos de los coeficientes. El parámetro de regularización (α) controla la fuerza de la penalización. La regresión Lasso no solo ayuda a reducir el sobreajuste, sino que también puede realizar selección de variables al reducir algunos coeficientes exactamente a cero, eliminando así algunas variables del modelo.

Función Matemática:

$$[\min_{\beta}\left(\sum_{i=1}^{m}\left(y_i-\beta_0-\sum_{j=1}^{n}\beta_j x_{ij}\right)^2+\alpha\sum_{j=1}^{n}|\beta_j|\right)]$$

En la predicción de costos, la selección del modelo adecuado es crucial para obtener predicciones precisas y evitar el sobreajuste. Los modelos Ridge y Lasso son particularmente útiles cuando se tienen muchas variables y existe el riesgo de multicolinealidad. Ridge es preferido cuando todas las variables son potencialmente importantes, mientras que Lasso es útil cuando se sospecha que solo un subconjunto de las variables tiene un impacto significativo en la variable dependiente.

4. Análisis de Componentes Principales (PCA)

El Análisis de Componentes Principales (PCA) es una herramienta crucial en la regresión lineal debido a su capacidad para abordar la multicolinealidad y reducir la dimensionalidad de los datos. Al transformar un conjunto de variables correlacionadas en un conjunto de componentes principales no correlacionados, PCA mejora la estabilidad y precisión del modelo de regresión. Esto no solo simplifica el modelo, sino que también facilita la interpretación de los resultados, permitiendo identificar las variables que más contribuyen a la variabilidad en los datos.

Además, la reducción de dimensionalidad lograda con PCA disminuye la complejidad computacional, haciendo el proceso de ajuste del modelo más eficiente y manejable, especialmente cuando se trabaja con grandes conjuntos de datos. Este enfoque permite a los investigadores y analistas concentrarse en las características más importantes de los datos, eliminando el ruido y mejorando la robustez del modelo. La combinación de estas ventajas hace que PCA sea una técnica valiosa y ampliamente utilizada en la construcción de modelos de regresión lineal efectivos y fiables.

5. Métricas de evaluación

Para la aplicación de nuestros modelos bajo el enfoque de ML se evaluarán las siguientes métricas:

Error Cuadrático Medio (MSE): Calcula el promedio de los cuadrados de los errores entre los valores predichos por el modelo y los valores reales. Un valor más bajo indica un mejor ajuste del modelo a los datos.

Raíz del Error Cuadrático Medio (RMSE): Es la raíz cuadrada del MSE y proporciona una medida del error en la misma escala que la variable dependiente. Es útil para interpretar la magnitud de los errores de predicción.

Coeficiente de Determinación (R-cuadrado): Indica la proporción de la variabilidad de la variable dependiente que es explicada por el modelo. Un valor cercano a 1 indica un buen ajuste del modelo a los datos, mientras que un valor cercano a 0 indica que el modelo no explica bien la variabilidad de la variable dependiente.

6. **Comparación de diferentes modelos de regresión avanzados usando PyCaret**

   PyCaret 3.0 es una biblioteca de ML en Python de código abierto y bajo código que automatiza los flujos de trabajo de aprendizaje automático, acelerando significativamente el ciclo de experimentación. Al sustituir cientos de líneas de código por unas pocas, PyCaret facilita el manejo de modelos y experimentos al integrar varias librerías y marcos de trabajo en una sola herramienta.

RESULTADOS

Con la finalidad de probar la eficacia de los modelos, se van a mostrar los resultados obtenidos aplicando diferentes formas de tratamiento de datos.

Realizamos las siguientes pruebas:

1. **Regresión lineal bajo enfoque econométrico**

   Evaluamos el rendimiento del modelo seleccionando diferentes variables para predecir el Valor facturado total.

   Por ejemplo, un primer modelo con todas las variables, posterior a ello se revisó las variables importantes aplicando el VIF, y al final el modelo se probó con las diferentes variables de diagnóstico.

Tabla 2 Resultados modelos con enfoque econométrico

Modelo	Método aplicado	R-squared:	Adj. R-squared:
lr	*Todas las variables*	1	1
lr	*Modelo Agrupación de Dx*	0.617	0.616
lr	***Mejor Modelo Dx Final***	**0.690**	**0.689**

Los resultados muestran que al usar todas las variables el modelo presenta problemas de sobreajuste, esto es de esperarse puesto que las variables de valores por rubros fueron creadas con el valor facturado. Estas variables por su fuerte relación no podrían servir como predictores, al contrario, vamos a ocupar las variables creadas en donde se tiene si una atención tiene o no un servicio en específico.

El modelo que mejor resultado generó es en el que se ocupa la variable Dx Final con un R-squared de 0.690.

De este modelo se realiza un análisis de los residuos

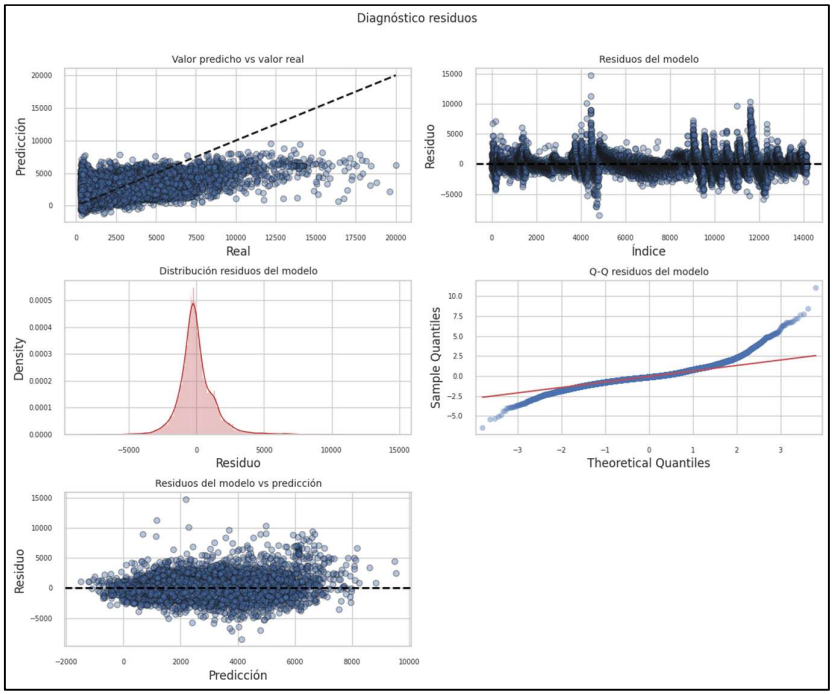


Ilustración 13 Análisis de residuos.
Fuente: Los autores.

Observaciones:

El análisis de los gráficos de residuos indica que el modelo de regresión se ajusta razonablemente bien a los datos. La mayoría de los supuestos de la regresión lineal parecen cumplirse, por otro lado, los errores se distribuyen de forma aproximadamente normal y no se observan patrones evidentes en los residuos que sugieran problemas graves con el modelo. Sin embargo, se ha detectado heterocedasticidad, lo que significa que la variabilidad de los errores no es constante a lo largo de todos los valores predichos.

Aunque esta limitación es importante, el modelo sigue siendo útil para entender la relación entre las variables estudiadas. Para mejorar la precisión de las predicciones y obtener resultados más confiables, sería beneficioso realizar ajustes adicionales, como explorar transformaciones de las variables o emplear técnicas de modelado que puedan manejar la heterocedasticidad.

2. Modelo de Regresión bajo el enfoque de machine learning

Para este enfoque vamos a realizar las predicciones con modelos de Regresión lineal, Lasso, Ridge y por último aplicaremos un PCA con una regresión lineal.

Algunos de los parámetros generales utilizados son:

- **test_size=0.20 y random_state=123**: Esta parte divide los datos en conjuntos de entrenamiento (80%) y prueba (20%), usando una semilla aleatoria fija (123) para asegurar que los resultados sean reproducibles.
- **Alpha:** El parámetro alpha controla la fuerza de regularización en Lasso y Ridge.

- Lasso: alpha = 0.1 lo que significa que se está aplicando una regularización moderada en el modelo Lasso. Esto permite que algunos coeficientes no sean reducidos completamente a cero, manteniendo así más características en el modelo.
- Ridge: alpha = 1.0 es un valor moderado para Ridge. Comparado con Lasso, Ridge tiende a conservar más características y solo reduce ligeramente los coeficientes. Un alpha más alto aumentaría la regularización, reduciendo aún más los coeficientes, pero potencialmente mejorando la generalización del modelo.

Al igual que en el enfoque econométrico, realizamos varias predicciones en torno a todas las variables y a la selección de variables importantes con diferentes variables de diagnóstico.

Resultados:

Tabla 3 Resumen de resultados

Métricas de entrenamiento

Modelo	Método aplicado	MSE	RMSE	MAE	R^2
lr	*Todas las variables*	7,837356E-18	0,000002799528	0,00000200408	1.0
lr	*Sin variables de rubros*	1.775249e+06	1332.384573	913.931034	0.692707
lr	*Sin variables de rubros + PCA*	1.899591e+06	1378.256457	965.449395	0.671183
ridge	*Sin variables de rubros*	1.775249e+06	1332.384606	913.930056	0.692707
lasoo	*Sin variables de rubros*	1.775249e+06	1332.384797	913.909549	0.692707
lr	*Variables importantes - Dx Final*	1.787877e+06	1337.115285	917.955623	0.690521
lr	*Variables importantes - Top Dx Final*	1.376865e+06	1173.398698	747.236783	0.656956
lr	*Variables importantes - Top Dx Final aplicado log*	1.655782e+06	1286.771798	733.313634	0.587465

Métricas de prueba

Modelo	Enfoque	MSE	RMSE	MAE	R^2
lr	*Todas las variables*	6,96753E-18	0,000002639608	0,000001897866	1.0
lr	*Sin variables de rubros*	1.732954e+06	1316.416917	910.894286	0.677826
lr	*Sin variables de rubros + PCA*	1.830529e+06	1352.970292	964.853864	0.659686
ridge	*Sin variables de rubros*	1.732916e+06	1316.402543	910.892321	0.677833
lasoo	*Sin variables de rubros*	1.732849e+06	1316.377313	910.853098	0.677846
lr	*Variables importantes - Dx Final*	1.746275e+06	1321.466835	915.040479	0.675350
lr	*Variables importantes - Top Dx Final*	1.444535e+06	1201.887943	753.277443	0.657271
lr	*Variables importantes - Top Dx Final aplicado log*	1.611212e+06	1269.335370	751.698914	0.617725

2.1. Regresión lineal con todas las variables

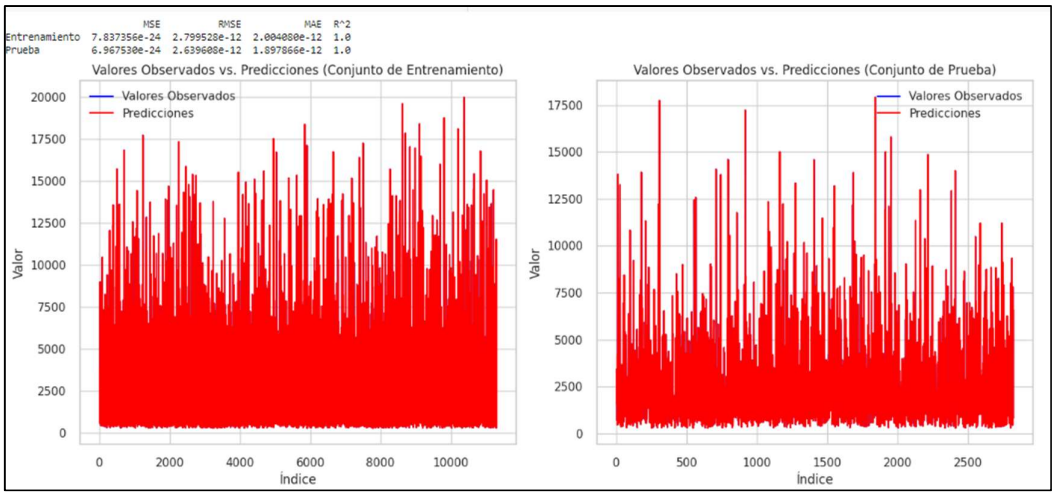


Ilustración 14 Valores observados vs predichos todas las variables.
Fuente: Los autores.

Observaciones:

Tal como lo vimos en la primera parte, el modelo se ajusta perfectamente esto porque las variables que tienen los valores por rubros están en base al valor facturado.

Por esta razón, se debe excluir esas variables en este enfoque.

2.2. Regresión lineal sin variables de rubros por valor facturado

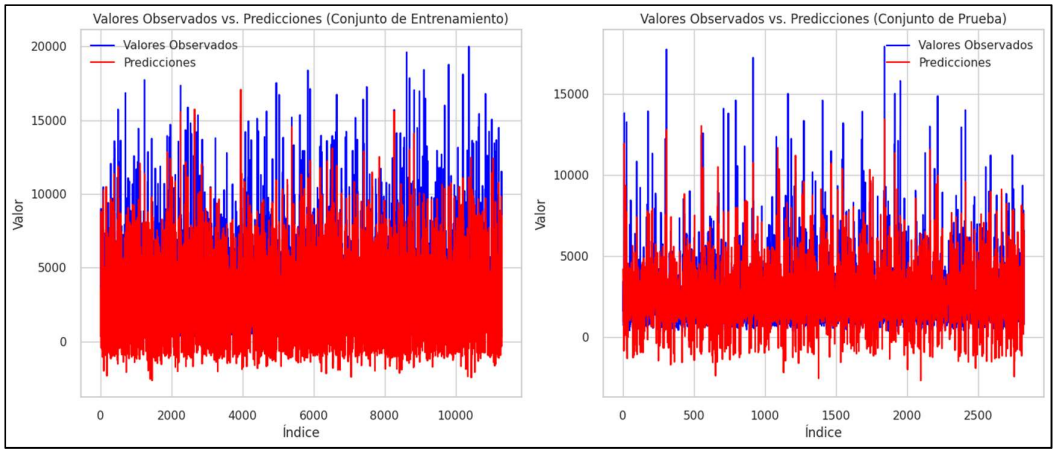


Ilustración 15 Valores observados vs predichos sin variables de rubros.
Fuente: Los autores.

Observaciones:

El modelo explica el 68% de la variabilidad en los datos con un R^2 de 0.6778. Sin embargo, el MAE de 910.89 sugiere una desviación promedio de 911 unidades, indicando que se necesita un refinamiento para mejorar la precisión en la estimación de costos hospitalarios.

2.3. Regresión Lasso y Ridge con todas las variables

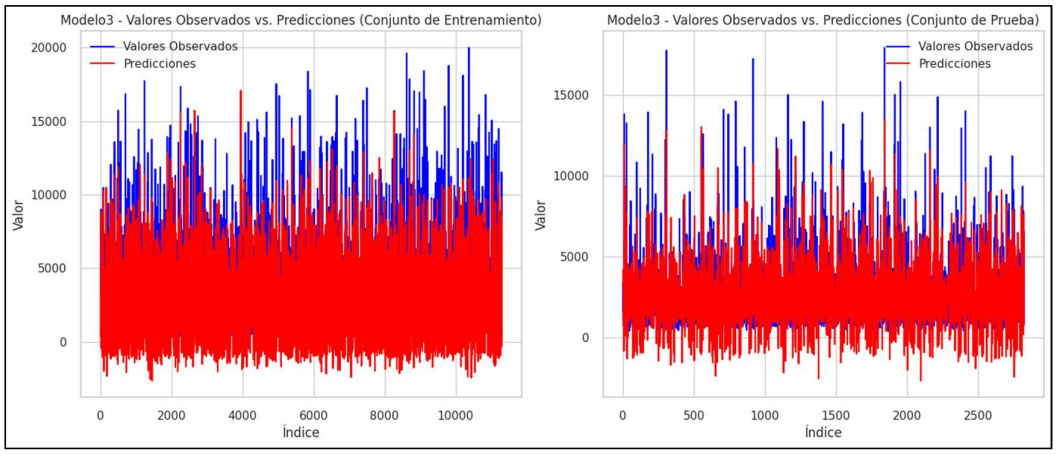


Ilustración 16 Valores observados vs predichos Lasso sin variables de rubros.
Fuente: Los autores.

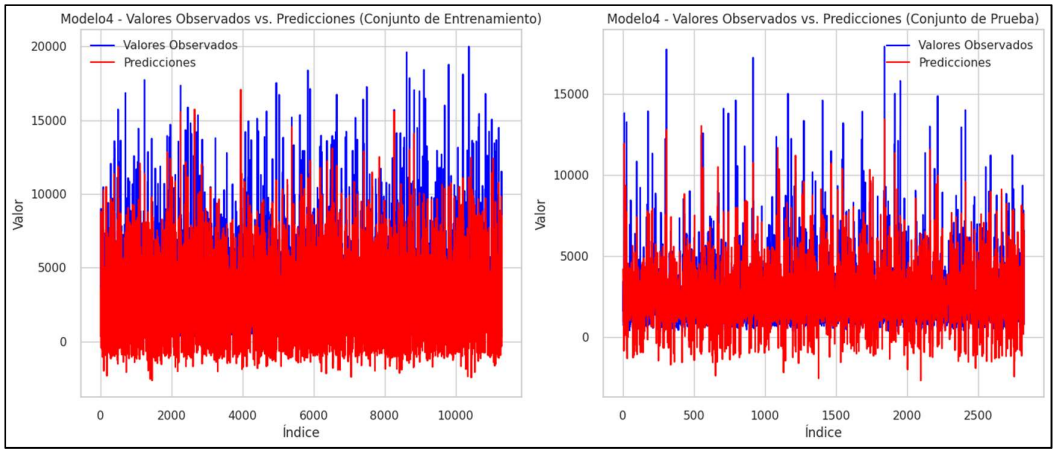


Ilustración 17 Valores observados vs predichos Ridge sin variables de rubros.
Fuente: Los autores.

Observaciones:

Los resultados de los modelos de regresión lineal, Ridge y Lasso, sin variables de rubros, son prácticamente idénticos en términos de MSE, RMSE, MAE y R^2. Esto sugiere que todos los modelos tienen un desempeño muy similar, con Lasso mostrando una ligera ventaja en precisión. Las diferencias mínimas en las métricas indican que los tres enfoques capturan las características de los datos de manera casi equivalente

2.4. Regresión lineal aplicado PCA

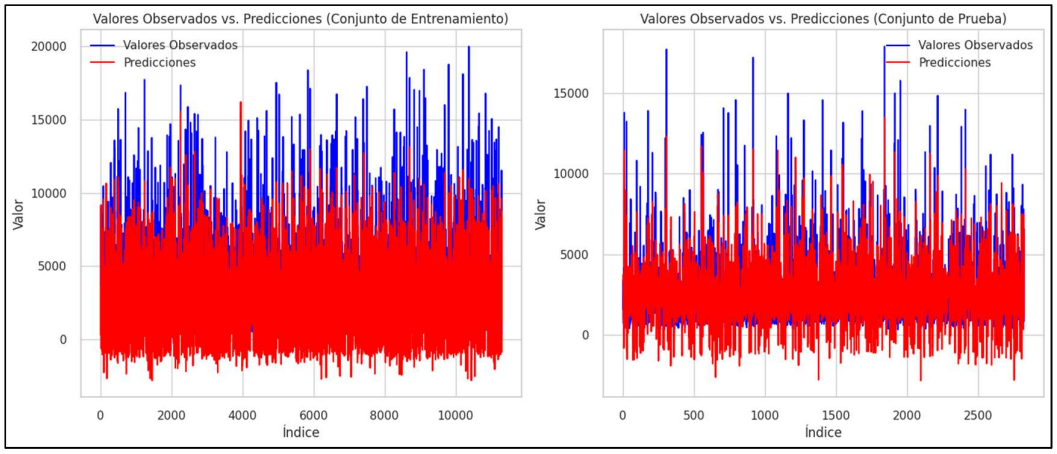


Ilustración 18 Valores observados vs predichos Regresión lineal con PCA sin variables de rubros. Fuente: Los autores.

Observaciones:

El modelo de regresión lineal sin variables de rubros más PCA muestra un MSE de 1.899591e+06, un RMSE de 1378.256457, un MAE de 965.449395 y un R^2 de 0.671183. Estos resultados indican un rendimiento ligeramente inferior en comparación con los modelos sin PCA, sugiriendo que la inclusión de PCA no mejora la precisión del modelo en este caso.

2.5. Regresión lineal con variables importantes.

En este punto se testearon diferentes variables de diagnóstico, y el mejor predictor fue la Variable Dx Final.

En esta categoría existen clases con muy pocos datos, por lo que nuestra propuesta es trabajar con los 50 diagnósticos mas frecuentes y aplicar una transformación logarítmica ajustado el modelo para que no prediga valores negativos.

En esta parte mostraremos este último resultado.

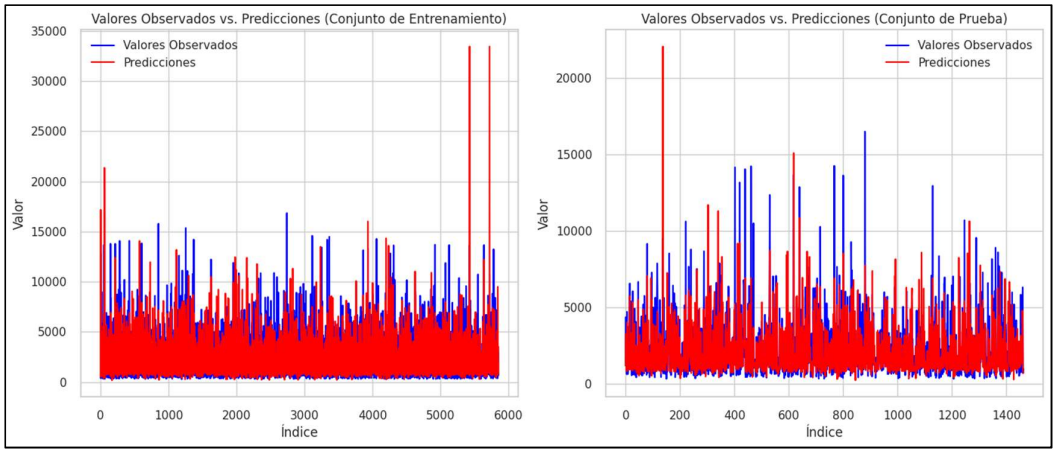


Ilustración 19 Valores observados vs predichos Regresión diagnósticos más frecuentes. Fuente: Los autores.

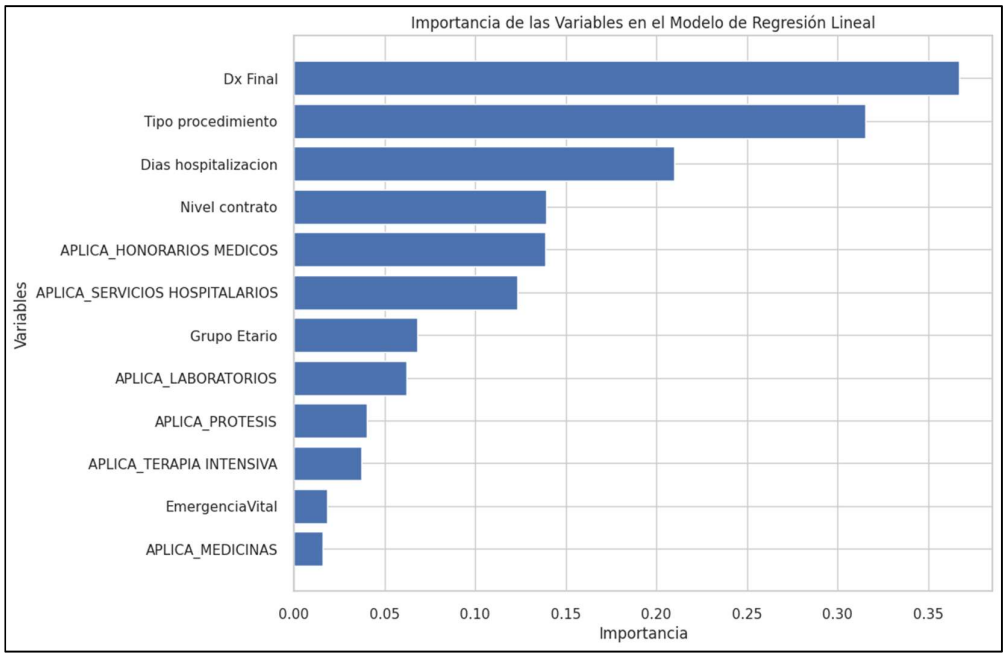


Ilustración 20 Importancia de las variables. Fuente: Los autores.

Observaciones:

El modelo parece capturar la tendencia general de los datos, pero lucha con los valores extremos (outliers). Esto podría ser una indicación de que el modelo no está completamente ajustado a estos valores más altos.

El comportamiento del modelo en el conjunto de prueba refleja el desempeño observado en el conjunto de entrenamiento, indicando que los problemas con la predicción de valores extremos no se deben al sobreajuste, sino que el modelo tiene dificultades para generalizar en estos casos.

DISCUSIÓN DE LOS RESULTADOS

Después de evaluar diversos modelos de regresión, hemos observado algunos problemas que pueden resumir en lo siguiente:

Sobreajuste

El sobreajuste sucede cuando un modelo se adapta demasiado a los datos de entrenamiento, capturando no solo las tendencias reales sino también el ruido aleatorio. Esto suele resultar en un excelente desempeño en los datos con los que se entrenó, pero en un rendimiento mucho peor al aplicarlo a datos nuevos o no vistos. Para evitar el sobreajuste, es esencial encontrar un equilibrio en la complejidad del modelo para que las predicciones sean más confiables y generalizables.

Para evitarlo es imprescindible realizar una correcta selección de variables y una transformación efectiva.

Presencia de valores negativos en las predicciones

Esto puede ser causado por:

Ausencia de Restricciones: Los modelos de regresión lineal estándar no tienen restricciones que impidan predecir valores negativos. Si el rango de los valores observados incluye cero o es muy cercano a él, es probable que el modelo haga predicciones negativas.

Distribución de Datos: Si la distribución de los datos de entrenamiento incluye valores cercanos a cero, el modelo puede extrapolar a valores negativos en la predicción.

Outliers y Sesgos: La presencia de outliers en los datos de entrenamiento puede sesgar el modelo, llevando a predicciones menos precisas y, en algunos casos, negativas.

Después de evaluar diversos modelos de regresión, la aplicación de regresiones Ridge y Lasso, a pesar de que se muestra una mejoría en los resultados, los mismos no tienen gran peso o variación en cuanto a un modelo de regresión lineal tradicional.

En cuanto a la aplicación de PCA, sus resultados no mostraron mejoría y podría entenderse a la variabiliad de los datos con los cuales se está trabajando.

De todos los testeos hemos observado que el modelo de regresión lineal en el cual se seleccionan las variables importantes de un top de diagnósticos con suficientes datos para predecir y aplicando una transformación logarítmica resulta ser la mejor opción para nuestras necesidades específicas. Aunque este modelo presenta un coeficiente de determinación (R^2) ligeramente inferior a otros modelos (0,6177), sus métricas de error, como el error cuadrático medio (MSE), la raíz del error cuadrático medio (RMSE) y el error absoluto medio (MAE), son favorables y significativamente mejoradas en comparación con otros enfoques.

Estos resultados muestran la gran variedad de datos y la necesidad de aplicar técnicas más avanzadas de aprendizaje para lograr una predicción más precisa. Bajo este criterio consideramos que un 61.77 % es considerado relativamente bueno por la magnitud y la variedad de datos que a los que nos enfrentamos durante el estudio.

Como una muestra de la aplicación de modelos de aprendizaje más avanzados que la compañía podría ir probando su eficacia, a continuación, se muestra una

tabla de comparación de resultados de varios modelos de ML aplicando la librería Pycaret.

Tabla 4 Resultados modelos de ML con Pycaret

Model	MAE	MSE	RMSE	R^2	RMSLE	MAPE	TT (Sec)
lightgbm	697.4893	1239695.2776	1110.2809	0.7859	0.4020	0.3453	1.4330
gbr	732.4538	1288338.5805	1133.2569	0.7768	0.4681	0.3899	0.6470
rf	731.2924	1315654.6687	1145.8514	0.7717	0.4148	0.3687	3.4110
xgboost	722.0533	1329003.8750	1150.4573	0.7698	0.4283	0.3598	0.1610
et	765.8413	1456208.2794	1205.4358	0.7472	0.4283	0.3748	2.7450
lar	916.2266	1790391.2625	1336.1504	0.6904	0.5413	0.6236	0.0390
br	915.8607	1790400.5875	1336.1418	0.6904	0.5405	0.6228	0.0650
llar	915.7110	1790562.4500	1336.1956	0.6904	0.5415	0.6228	0.0370
ridge	916.1345	1790385.4750	1336.1453	0.6904	0.5401	0.6234	0.0360
lasso	915.7104	1790562.5125	1336.1956	0.6904	0.5415	0.6228	0.0360
lr	916.2244	1790391.1625	1336.1504	0.6904	0.5413	0.6236	0.6310
en	931.8701	2228961.8750	1489.8137	0.6157	0.5717	0.5519	0.0390
dt	928.4314	2281326.7901	1507.2328	0.6037	0.5114	0.4347	0.0700
huber	956.7100	2340727.1295	1527.0154	0.5959	0.5781	0.5623	0.3720
knn	964.0736	2556730.0125	1595.2657	0.5588	0.5522	0.5542	0.0930
omp	1041.9799	2741006.2750	1652.7051	0.5270	0.6069	0.6600	0.0440
ada	1407.3777	2830800.2651	1681.4936	0.5066	0.8259	1.2118	0.7020
par	1360.0402	4115422.9297	1990.0083	0.2908	0.7018	0.8764	0.0890
dummy	1733.5062	5797422.3000	2404.6031	-0.0006	0.9003	1.2325	0.0320

Pycaret nos muestra que los modelos LightGBM, GBR y RF han mostrado un rendimiento sólido en la predicción, con LightGBM destacándose como el modelo más eficaz en términos de precisión y capacidad explicativa, dado su alto valor de R^2 de 0.7859 y su bajo MAE de 697.49. GBR sigue de cerca, con un R^2 \e 0.7768 y un MAE de 732.45, ofreciendo una precisión competitiva, pero con una ligera diferencia en comparación con LightGBM. Random Forest, aunque un poco menos preciso con un R^2 de 0.7717 y un MAE de 731.29, sigue siendo robusto y fiable.

Dado que estos modelos presentan características particulares, las mismas deben ser exploradas más a fondo para mejorar aún más las predicciones. Es crucial realizar una investigación adicional para afinar estos modelos, como la optimización de hiperparámetros, la inclusión de características adicionales, y la evaluación en diferentes subconjuntos de datos. Adicionalmente, explorar técnicas de ensamblaje o combinaciones de modelos podría proporcionar mejoras adicionales en la precisión y robustez de las predicciones. Este enfoque

permitirá maximizar el rendimiento y la aplicabilidad de los modelos en escenarios prácticos.

IMPLICACIONES SOBRE LA INNOVACION EMPRESARIAL

En un mundo cada vez más digital, la innovación empresarial se ha convertido en un elemento crucial para el éxito en el sector de seguros médicos. La incorporación de tecnologías avanzadas, como el Big Data, la computación en la nube (Cloud Computing), y la inteligencia artificial (AI) a través de ML, permite a las empresas no solo optimizar sus operaciones, sino también mejorar la experiencia del cliente y tomar decisiones más informadas.

A continuación, podemos observar como las nuevas tecnologías pueden aportar para que la compañía tome el rumbo de una empresa Data Driven.

Big Data y Almacenes de Datos (Data Warehouses - DWH)

Uno de los puntos altos en la innovación tecnológica en las empresas de seguros médicos es la gestión y explotación eficiente de grandes volúmenes de datos, comúnmente conocidos como Big Data.

Durante el estudio apreciamos como la compañía maneja gran cantidad de datos de los diversos sistemas transaccionales, estos datos son variados en su estructura, volátiles en su naturaleza, y de alta velocidad, lo que presenta tanto un desafío como una oportunidad.

Es en este punto donde los Data Warehouses son esenciales para centralizar y estructurar esta inmensa cantidad de información. Sirven como la columna vertebral para la integración de datos de diversas fuentes, permitiendo a las aseguradoras almacenar y acceder a información de manera coherente y organizada. Un DWH bien diseñado soporta la creación de Data Marts, que se especializan en áreas específicas como análisis de reclamaciones, evaluación de riesgos o estudios actuariales, facilitando un análisis más profundo y específico.

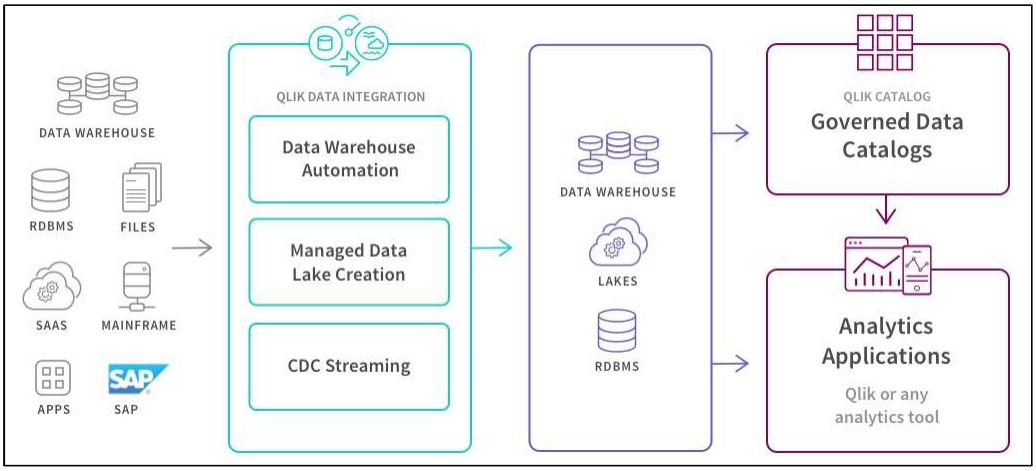


Ilustración 21 Automatización de Data warehouses.
Fuente: (Qlik, n.d.-b)

Computación en la Nube (Cloud Computing) para Aplicaciones de Machine learning e IA

Como siguiente paso en la cadena de innovación es la utilización de la Computación en la Nube. La nube no solo ofrece almacenamiento escalable, sino que también proporciona una infraestructura robusta para la ejecución de modelos de ML e inteligencia artificial, que son fundamentales para mejorar las operaciones en el sector de seguros médicos.

Según (Golightly et al., 2022), definen a la computación en la nube como una técnica que permite el acceso fácil y bajo demanda a una red de recursos informáticos compartidos, como servidores, aplicaciones y servicios. Este modelo se caracteriza por su escalabilidad, la capacidad de proporcionar recursos de manera rápida y eficiente, y la integración de diversas tecnologías de información

La adopción de esta tecnología permite a las aseguradoras enfocarse en la innovación y el mejoramiento continuo de sus soluciones sin la carga de gestionar complejas infraestructuras tecnológicas. Con la nube, el tiempo y los recursos que antes se destinaban a la instalación y mantenimiento de servidores ahora pueden dirigirse a proyectos más estratégicos, como el desarrollo de

nuevos modelos predictivos o la mejora de los servicios al cliente. Además, las actualizaciones automáticas y la capacidad de adaptarse a los cambios tecnológicos rápidamente garantizan que las aseguradoras siempre estén utilizando las herramientas más avanzadas, lo que les permite mantenerse competitivas en un entorno cada vez más digitalizado.

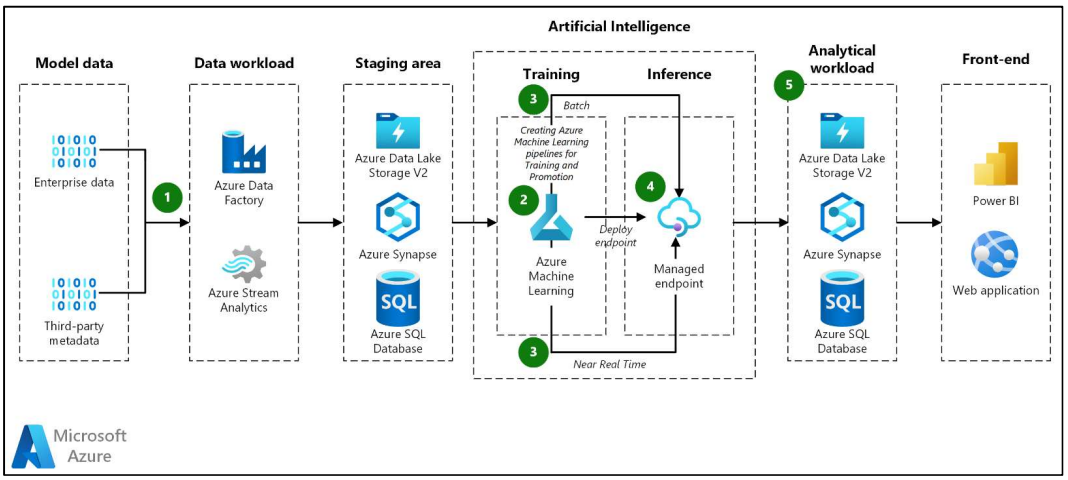


Ilustración 22 Aprendizaje automático de varios modelos con Azure ML - Azure Example Scenarios. Microsoft Learn Fuente: (Microsoft Azure, n.d.)

Business Intelligence y Analítica

Las herramientas de Business Intelligence transforman los datos en conocimiento accionable a través de visualizaciones y dashboards intuitivos. La analítica predictiva y prescriptiva, impulsada por ML, permite prever eventos como riesgos y fraudes, mientras que la integración con aplicaciones de negocio automatiza procesos clave, mejorando la experiencia del cliente y optimizando las operaciones diarias.

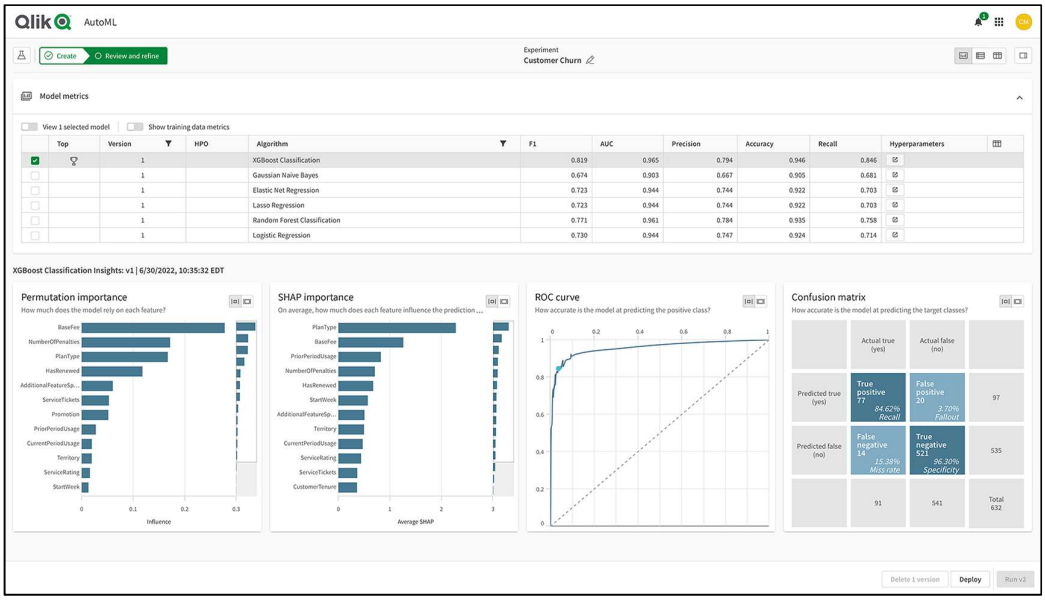


Ilustración 23 Capacidades de IA y aprendizaje automático.
Fuente: (Qlik, n.d.-a)

CONCLUSIONES Y RECOMENDACIONES

La irrupción del ML en el ámbito sanitario ha abierto un nuevo horizonte de posibilidades, particularmente en el terreno de la predicción de costos hospitalarios. Esta tecnología revolucionaria está transformando la forma en que se gestionan los recursos sanitarios, permitiendo análisis predictivos precisos y adaptativos que optimizan la eficiencia operativa y la toma de decisiones estratégicas

Conclusiones:

El ML, con su capacidad para procesar y analizar grandes volúmenes de datos complejos, como diagnósticos, tratamientos y duración de hospitalización, ofrece una visión óptima de los factores que determinan los costos de salud.

Al predecir costos hospitalarios, la compañía puede optimizar la asignación de recursos de manera más eficiente. Esto se traduce en una mejor gestión operativa, una distribución más adecuada del personal y una planificación

financiera más sólida, asegurando la sostenibilidad del sistema de prestaciones médicas a largo plazo.

Durante nuestro estudio, pudimos observar la gran cantidad de transacciones y costos hospitalarios que se generan mensualmente en la empresa, si bien la estimación de costos hospitalarios es un tema crítico puesto que las condiciones de salud dependen de varios factores que en la actualidad no se están analizando, el enfoque de nuestro estudio ofrece un punto de partida para poder visibilizar de mejor forma cómo se están asignando los recursos económicos.

También observamos la importancia de algunas variables, es por ello que para predecir de mejor forma los costos hospitalarios se debería aplicar submodelos o predicciones a cada variable importante, por ejemplo, predecir los costos hospitalarios por cada rubro.

Además, es necesario resaltar la importancia de haber iniciado el estudio aplicando regresiones lineales. Estas técnicas, aunque menos avanzadas, proporcionan una base sólida para entender la estructura de los datos y la relación entre las variables. La regresión lineal permite descomponer y analizar cómo cada variable afecta los costos hospitalarios, ofreciendo una perspectiva valiosa que puede complementar los hallazgos de los modelos más complejos. Esta fase inicial no solo facilita la interpretación de los resultados, sino que también ayuda a establecer un punto de referencia para evaluar el desempeño de los modelos avanzados.

Cabe destacar que la gran variabilidad observada en los datos puede presentar un desafío significativo para las predicciones. La presencia de esta variabilidad es una de las múltiples razones por las cuales estos modelos pueden tener dificultades en la precisión de los modelos y, por lo tanto, en la toma de decisiones basada en los resultados. Encontrar modelos que se ajusten adecuadamente a la realidad económica y reflejen de manera precisa esta variabilidad es crucial para mejorar las predicciones y lograr una mayor fiabilidad en las estimaciones de costos hospitalarios. Por lo tanto, la exploración y ajuste

continuo de los modelos de ML es esencial para abordar estos desafíos y optimizar la precisión de las predicciones.

Recomendaciones:

Como lo mencionamos anteriormente, predecir los costos hospitalarios es una tarea compleja que requiere de varios factores médicos como económicos.

Obtener y capturar la mayor cantidad de datos referente a una atención hospitalaria en base a las historias clínicas, evolución médica, condiciones de salud previas del paciente, etc., es crucial para enriquecer los modelos predictivos y aplicar algoritmos más sofisticados. En la actualidad, la compañía ha iniciado un proceso de transformación digital que permitirá capturar todos estos datos de documentos, fotografías o demás fuentes de información relacionados a la atención del paciente.

Con esos datos, a futuro se puede aplicar modelos de ML más avanzados sean de regresión o clasificación, que permitan tener una visión completa de la atención médica, por ejemplo, se puede crear modelos de clasificación o modelos de riesgo para los pacientes en donde se pueda predecir, por ejemplo:

- La probabilidad que un diagnóstico genere una complejidad.
- Predecir cuántos días un paciente puede permanecer hospitalizado por una patología.
- Que enfermedades preexistentes son más propensas a desarrollar patologías más graves o que requiera un procedimiento más costoso.

De esta forma la compañía podría anticiparse al costo monetario que se presenten a lo largo del tiempo por las atenciones hospitalarias que soliciten los clientes.

También, se recomienda testear el modelo en un entorno de pruebas con un diagnóstico que pueda ser controlado y no presente tanta variabilidad en sus

datos, a eso, se debe sumar un motor de reglas que, en base a las condiciones contractuales permita asignar o no el valor monetario.

A la par, mediante un equipo multidisciplinario, se recomienda probar al menos 3 de los modelos de aprendizaje automático avanzado, que mostraron los mejores resultados al aplicar Pycaret.

Mejora Continua y Adaptación

Crear un ciclo de retroalimentación que permita la mejora continua de los modelos a medida que se recopilan nuevos datos y se obtienen resultados de las predicciones. Esto asegurará que los modelos se mantengan actualizados y relevantes en un entorno de atención médica en constante cambio.

Optimización de Recursos y Planificación Financiera

Emplear modelos de ML con la finalidad de automatizar los procesos de Crédito que afectan directamente a la reserva de recursos monetarios por atenciones médicas.

Innovación

Aplicar una sólida estrategia de innovación empresarial marcando el camino hacia una empresa Data Driven que no solo integre nuevas tecnologías digitales, sino que enfoque sus esfuerzos en desarrollar e implementar una arquitectura de datos sólida que abarque toda la cadena de valor. Esto incluye desde la recopilación y gestión de datos en los sistemas operativos hasta su análisis y aplicación en la toma de decisiones estratégicas. En este enfoque, cada elemento, desde las herramientas de gestión hasta la cultura organizacional, debe alinearse para aprovechar al máximo el valor de los datos. Este cambio no solo impulsa la eficiencia operativa, sino que también permite una mayor agilidad y capacidad de respuesta en un entorno de negocios cada vez más complejo y competitivo.

ANEXOS

ANEXO 1

Tabla 5 Tabla comparativa de revisión bibliográfica sobre estudios similares recientes.

Autor	Nombre del estudio	Objeto de estudio	Metodologías aplicadas	Medidas de evaluación
(Fan et al., 2024)	Predicting Hospitalization Costs for Pulmonary Tuberculosis Patients Based on Machine Learning	Predicción de costos de hospitalización de pacientes con tuberculosis pulmonar	Regresión Múltiple, Perceptrón Multicapa (MLP)	R-cuadrado, RMSE, MAE
(Donado, 2022)	Machine Learning en modelos de predicción de la siniestralidad del asegurado en Seguros de Salud	Predicción de siniestralidad del asegurado en Seguros de Salud	Modelo Lineal Generalizado (GLM), Máquinas de Soporte Vectorial (SVM), Multivariate Adaptive Regression Splines (MARS)	RMSE, R-cuadrado
(Casanova, 2022)	Predicción Del Costo De Las Reclamaciones Con Machine Learning	Predicción del costo de reclamaciones en el mercado asegurador	Regresión Lineal, Árboles de Regresión, Bagging, Gradient Boosting	RMSE
(Taloba et al. ,2022)	Estimation and Prediction of Hospitalization and Medical Care Costs Using Regression in Machine Learning	Estimación de costos de hospitalización y atención médica	Regresión Lineal	Precisión
(Rakshit et al. ,2021)	A Machine Learning approach to predict healthcare cost of breast cancer patients	Predicción de costos de atención médica de pacientes con cáncer de mama	K-Nearest Neighbors (KNN), Gradient Boosting, Redes Neuronales Artificiales (ANN), Elastic Net	MAPE
(ul Hassan et al. ,2021)	A Computational Intelligence Approach for Predicting Medical Insurance Cost	Predicción de costos de seguros médicos	Regresión Lineal (LR), Stochastic Gradient Boosting (SGB), XGBoost (XGB)	RMSE, AUC

Autor	Nombre del estudio	Objeto de estudio	Metodologías aplicadas	Medidas de evaluación
(Maisog et al. ,2019)	Using massive health insurance claims data to predict very high-cost claimants: a Machine Learning approach	Identificación de solicitantes de reclamos de alto costo en el sistema de salud	Random Forest (RF), Máquinas de Soporte Vectorial (SVM), Gradient Boosted Trees (GBT), LightGBM	AUC-ROC
(Hanafy & Mahmoud 2021)	Predict Health Insurance Cost by using Machine Learning and DNN Regression Models	Predicción de costos de seguros de salud	Regresión Lineal Múltiple, Modelo Aditivo Generalizado (GAM), Máquinas de Soporte Vectorial (SVM), Random Forest (RF), Árbol de Decisión (DT), XGBoost (XGB), K-Nearest Neighbors (KNN), Stochastic Gradient Boosting (SGB), Redes Neuronales Profundas (DNN)	MAE, RMSE, R-cuadrado

REFERENCIAS

Aguirre, A. (2022). MODELOS DE PREDICCIÓN AVANZADOS PARA EL CÁLCULO DE RESERVAS EN LA INDUSTRIA ASEGURADORA. http://sedici.unlp.edu.ar/handle/10915/157814

Andrea, S., Salas, R., Fernanda, L., & Ortiz, C. (2023). Modelo Predictivo Probabilidad de Retención Seguro Automóviles.

Benítez, S., & Jiménez, F. (2024, August). PREDICCION COSTOS HOSPITALARIOS CON REGRESION LINEAL Y PCA. Https://Github.Com/Sbenitez87/CAPSTONE/Blob/Main/PREDICCION_COSTOS_HOSPITALARIOS_CON_REGRESION_LINEAL_Y_PCA_FINAL.Ipynb.

Casanova, Y. M. (2022). Predicción Del Costo De Las Reclamaciones Con Machine Learning. https://repositorio.escuelaing.edu.co/handle/001/2086

Donado, F. (2022). Machine Learning en modelos de predicción de la siniestralidad del asegurado en Seguros de Salud. http://hdl.handle.net/10017/54494

Fan, S., Abulizi, A., You, Y., Huang, C., Yimit, Y., Li, Q., & Zou, X. (2024). Predicting Hospitalization Costs for Pulmonary Tuberculosis Patients Based on Machine Learning. 1–17. https://doi.org/10.21203/rs.3.rs-4186975/v1

Federacion Ecuatoriana de Empresas de Seguro. (2024). Anuario FEDESEG 2023. Federacion Ecuatoriana de Empresas de Seguro. (06 de 07 de 2024). Obtenido de https://www.fedeseg.org/_files/ugd/f39f07_e42447c1463e48f5877626495b885fae.pdf

Federación Ecuatoriana de Empresas de Seguros. (2024, June 7). https://www.fedeseg.org/repprimanetaemitida. Https://Www.Fedeseg.Org/Repprimanetaemitida.

Golightly, L., Chang, V., Xu, Q. A., Gao, X., & Liu, B. S. C. (2022). Adoption of cloud computing as innovation in the organization. International Journal of Engineering Business Management, 14. https://doi.org/10.1177/18479790221093992

Hanafy, M., & Mahmoud, O. M. A. (2021). Predict Health Insurance Cost by using Machine Learning and DNN Regression Models. International Journal of Innovative Technology and Exploring Engineering, 10(3), 137–143. https://doi.org/10.35940/ijitee.C8364.0110321

Maisog, J. M., Li, W., Xu, Y., Hurley, B., Shah, H., Lemberg, R., Borden, T., Bandeian, S., Schline, M., Cross, R., Spiro, A., Michael, R., & Gutfraind, A. (2019). Using massive health insurance claims data to predict very high-cost claimants: a machine learning approach. https://www.researchgate.net/publication/338292264_Using_massive_health_insurance_claims_data_to_predict_very_high-cost_claimants_a_machine_learning_approach/comments

María, A., Ligia, I.-B., Melo-Becerra, A., Estefanía, D., María, P.-A., & Ramírez-Giraldo, T. (2023). Evolución y carga financiera de las Enfermedades Crónicas no Transmisibles en Colombia: 2010-2021.

https://repositorio.banrep.gov.co/server/api/core/bitstreams/01e2c642-b887-4e5a-aa4f-0c6f0b56a690/content

Maribel, A., Acosta, M., Carrillo, E. X., & Quito, L. (2022). Aplicación de los métodos Chain-Ladder y Link Ratio para la estimación de reservas IBNR para siniestros ocurridos y no reportados en una empresa de seguros.

Microsoft Azure. (n.d.). Aprendizaje automático de varios modelos con Azure Machine Learning - Azure Example Scenarios. Retrieved August 8, 2024, from https://learn.microsoft.com/es-es/azure/architecture/ai-ml/idea/many-models-machine-learning-azure-machine-learning

Ortiz -Culcay, O., Fernández -García, C., & Pérez -Rico, C. (2019). Análisis de cobertura de medicina prepagada en Pichincha (2019-2020). https://doi.org/10.29076/issn.2528-7737vol15iss38.2022pp1-13p

Qlik. (n.d.-a). Aprendizaje automático automatizado para equipos de analítica. Retrieved August 8, 2024, from https://www.qlik.com/es-es/products/qlik-sense/ai

Qlik. (n.d.-b). Automatización de Data Warehouses. Retrieved August 8, 2024, from https://www.qlik.com/es-es/data-warehouse-automation

Rakshit, P., Zaballa, O., Pérez, A., Gómez-Inhiesto, E., Acaiturri-Ayesta, M. T., & Lozano, J. A. (2021). A machine learning approach to predict healthcare cost of breast cancer patients. Scientific Reports, 11(1). https://doi.org/10.1038/s41598-021-91580-x

Rodríguez, M. B. (2023). Aplicación de modelo de detección de anomalías Isolation Forest para la detección de reclamos fraudulentos en una compañía del sector de medicina prepagada. https://dspace.udla.edu.ec/handle/33000/15219

Santenac, I., Majkowski, E., Vermeulen, P., Sun-Young, A., Fattibene, L., & Hurynovich, A. (2024). 2024 Global Insurance Outlook Strengthening trust to unlock innovation and growth. https://www.ey.com/en_gl/insights/insurance/global-insurance-industry-trends

Serrano, D. R. S., Rincón, J. C., Mejía-Restrepo, J., Núñez-Valdez, E. R., & García-Díaz, V. (2022). Forecast of Medical Costs in Health Companies Using Models Based on Advanced Analytics. Algorithms, 15(4). https://doi.org/10.3390/a15040106

Taloba, A. I., Abd El-Aziz, R. M., Alshanbari, H. M., & El-Bagoury, A. A. H. (2022). Estimation and Prediction of Hospitalization and Medical Care Costs Using Regression in Machine Learning. Journal of Healthcare Engineering, 2022. https://doi.org/10.1155/2022/7969220

ul Hassan, C. A., Iqbal, J., Hussain, S., AlSalman, H., Mosleh, M. A. A., & Sajid Ullah, S. (2021). A Computational Intelligence Approach for Predicting Medical Insurance Cost. Mathematical Problems in Engineering, 2021. https://doi.org/10.1155/2021/1162553

Yao, L., Lin, Y., Mo, Y., & Wang, F. (2023). Performance Evaluation of Financial Industry Related Expense Forecasting Using Various Regression Algorithms

for Machine Learning. Highlights in Science, Engineering and Technology, 57. https://doi.org/10.54097/hset.v57i.10007

Zhang, Y., & Walton, N. (2019). Adaptive Pricing in Insurance: Generalized Linear Models and Gaussian Process Regression Approaches. http://arxiv.org/abs/1907.05381