Caso de estudio: Tarifación de seguro de salud.

Daniel Manco, Esteban Arcila, Astrid Daniela Giraldo.

Departamento de
Ingeniería Industrial

Universidad de Antioquia

Medellín, Colombia

2024

Introducción

En el ámbito de los seguros de salud, la determinación precisa del costo de las tarifas de las pólizas es crucial tanto para las aseguradoras como para los asegurados. Este caso de estudio se centra en el desarrollo de un modelo predictivo para estimar el costo de las tarifas de las pólizas de salud. Utilizando datos históricos que incluyen características demográficas, historial médico y hábitos de estilo de vida de los asegurados, el modelo busca identificar patrones y relaciones que influyen en los costos de las pólizas.

La implementación de un modelo de predicción preciso puede ofrecer múltiples beneficios, tales como:

- Optimización de Precios: Permite a las aseguradoras establecer tarifas más justas y competitivas.
- Gestión de Riesgos: Ayuda a las aseguradoras a evaluar y mitigar riesgos, ofreciendo planes personalizados.
- Mejora en la Satisfacción del Cliente: Contribuye a la transparencia y equidad en los precios, mejorando la relación con los asegurados.

Este estudio emplea técnicas de regresión y modelos avanzados de aprendizaje automático para predecir el costo de las tarifas de las pólizas, con el objetivo de mejorar la precisión y eficiencia en la fijación de precios dentro del sector de seguros de salud.

Diseño de la solución.

Teniendo en cuenta la problemática, información disponible y el proceso analítico y administrativo a realizar, se plantea un diagrama de procesos que contiene el esquema que guiará la implementación de la solución propuesta.

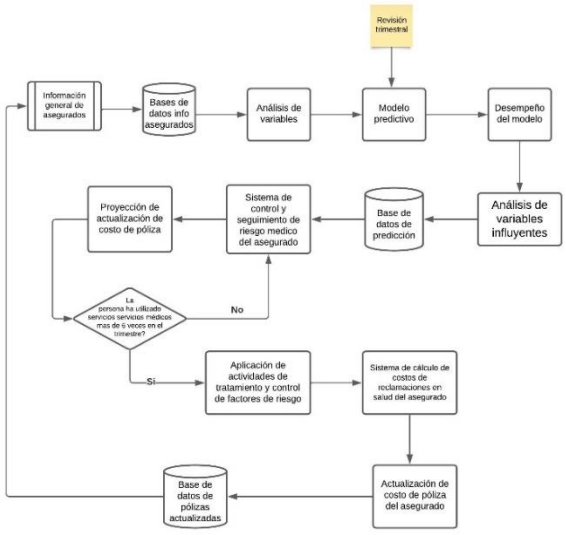

Imagen 1.Diseño de solución propuesto – Elaboración propia

Desarrollo del caso de estudio.

A. Limpieza y transformación.

En esta parte del proyecto, se analizaron las siguientes bases de datos, pertenecientes a los periodos de enero de 2018 a diciembre de 2019:

BD_Expuestos: Base de datos con la población expuesta

BD_Siniestros: Base de datos del costo y conteo de las reclamaciones.

BD_Sociodemograficas: Base de datos con las características sociodemográficas.

Al hacer una revisión inicial de la información, en estos no se encuentran mayor irregularidad respecto a datos, datos repetidos, etc. Por tanto, se procede con la concatenación de estos.

B.Análisis exploratorio

Después de completar el proceso de limpieza y transformación de los datos, se llevó a cabo un análisis exploratorio que fue fundamental para comprender la naturaleza de los datos antes de realizar cualquier suposición. Este análisis nos permitió identificar patrones y detectar valores atípicos. Para lograrlo, se procedió a graficar tanto las variables categóricas como las numéricas, lo que nos permitió observar y comprender su comportamiento de manera más clara y detallada.

Dentro de este exploratorio, se rescata como información importante las siguientes gráficas y su respectivo análisis.

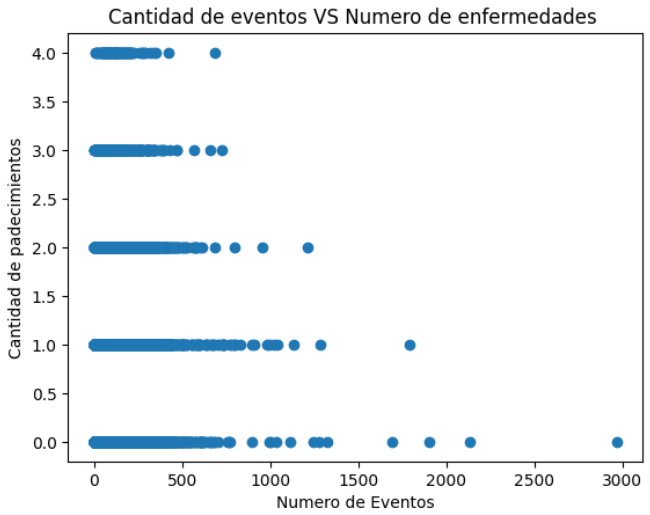

Imagen 2. Diagrama de dispersión entre variables – Elaboración propia

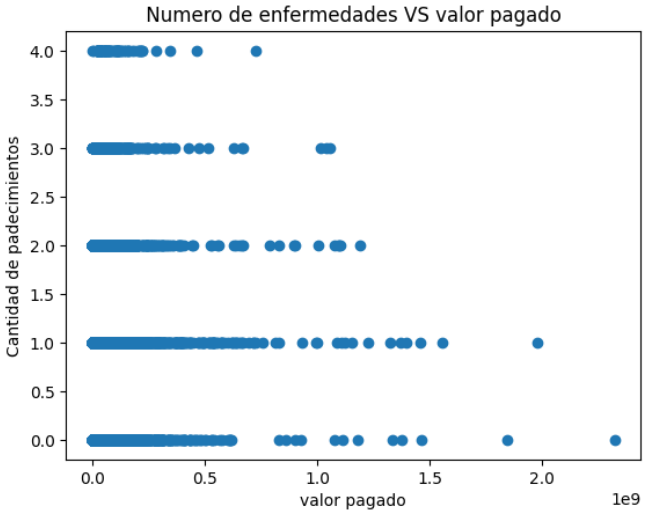

Imagen 3. Diagrama de dispersión entre variables – Elaboración propia

De las gráficas se puede notar que:

- El número de eventos no es mayor a 35 o 40.
- Un porcentaje muy bajo de los afiliados presentan enfermedades como cáncer, EPOC, enfermedad cardiovascular e hipertensión; la diabetes es la que más se presenta en los afiliados. Sin embargo, el porcentaje de afiliados que presentan enfermedades es muy bajo.
- La proporción de mujeres es mayor a la de hombres.

Se decide no eliminar ninguna variable ya que la cantidad que se tiene de ellas es baja y podría sesgar demasiado los resultados de los modelos posteriores.

También, se procede a analizar las ciudades y sus valores de pago,

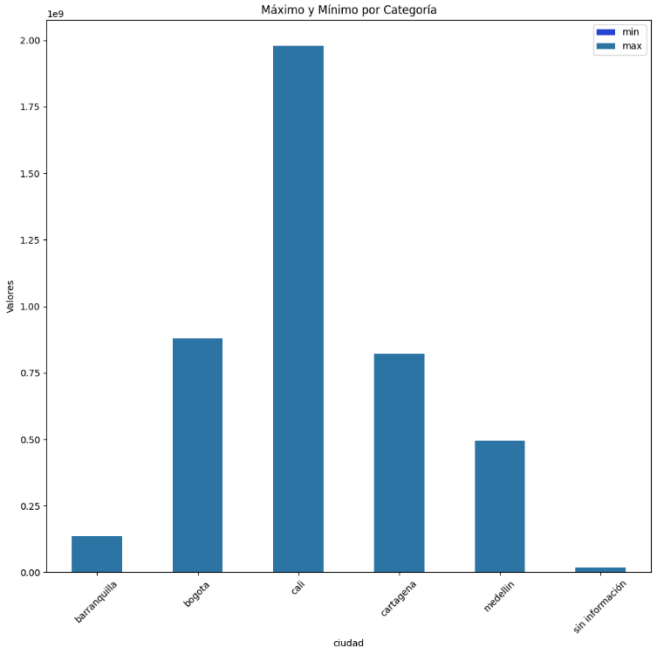

Imagen 4. Gráfico de barras entre ciudades y valores pagados – Elaboración propia

Se observa que la ciudad de Cali tiene el valor pagado más alto respecto a otras ciudades, esto puede darse por posibles valores atipicos dentro de esta, ya que su dieferencia es importante.

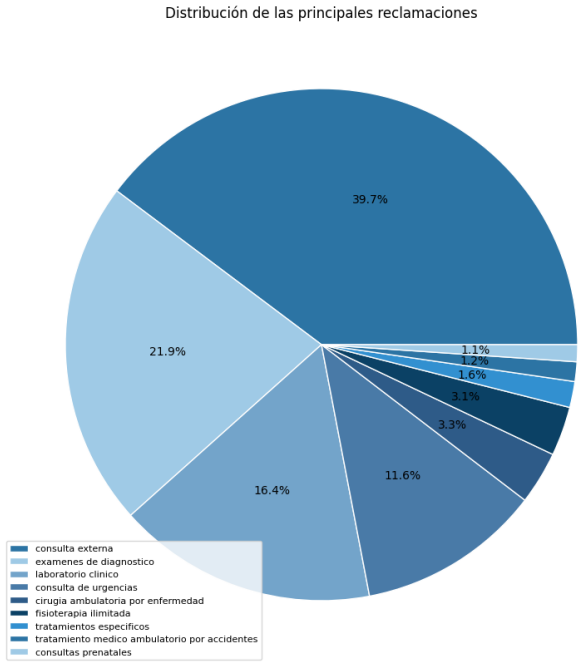

Imagen 5. Gráfica de torta para reclamaciones – Elaboración propia

Se puede notar claramente que algunas variables tienen una cantidad de datos diferente que impide analizar su distribución. De las que sí se logra mostrar sus valores se identifica que:

- La consulta externa, exámenes de diagnóstico, laboratorio clínico y consulta de urgencias son las reclamaciones más comunes.

Por último, se muestra la distribución de la edad con respecto al número de usuarios.

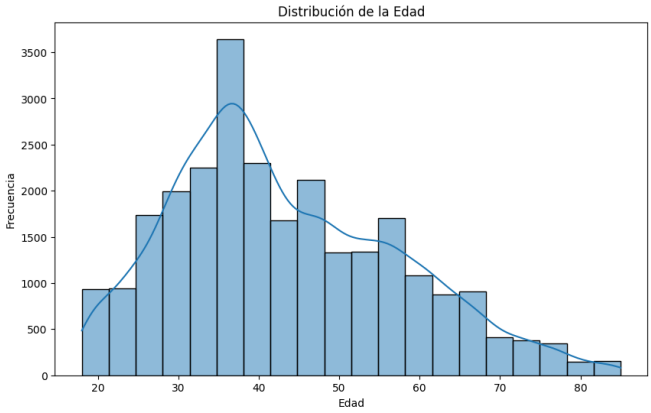

Imagen 6. Histograma para edades – Elaboración propia

En este gráfico, se observa que la edad de los usuarios oscila entre los 35 y 40 años, esto puede darse a que las personas en esta edad suelen empezar a presentar sus primeros problemas de salud, especialmente los crónicos.

C.Selección de algoritmos y técnicas de modelado.

Para la selección de los modelos esperados a implementar, se decide implementar diferentes técnicas de modelado, ya que esto permitirá seleccionar un modelo en especifico que tenga una mejor adaptación a los datos.

D. Selección de variables

Antes de la aplicación de selectores de variables, se decide generar una base de datos, la cual contiene las siguientes variables:

Eventos, valor_pagado, ciudad, cancer, epoc, diabetes, hipertensión, enf_cardiovascular, genero, edad, cantidad_enfermedades, clasifica_diagn.

Además, se hace una matriz de correlación.

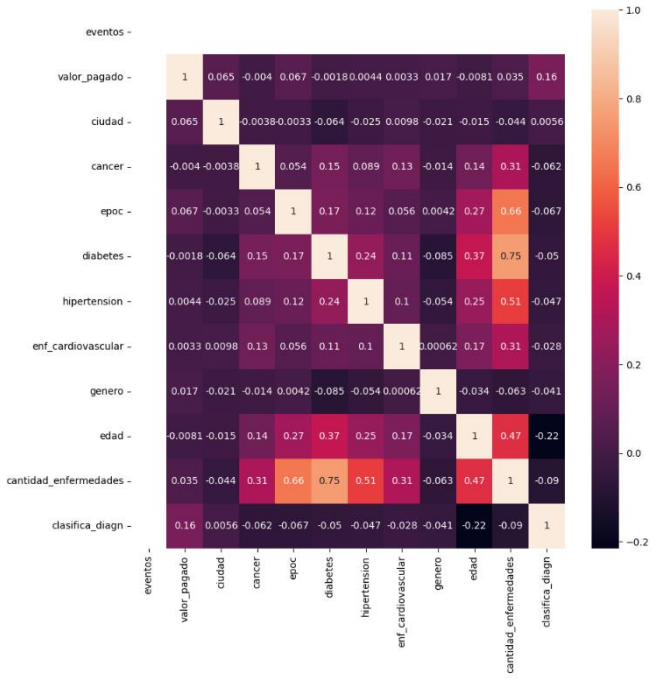

Imagen 7. Matriz de correlación entre variables – Elaboración propia

Para observar la correlación entre estas, se muestra una salida comparativa entre variables.

epoc	edad	0.274590
cantidad_enfermedades	cancer	0.306266
cancer	cantidad_enfermedades	0.306266
enf_cardiovascular	cantidad_enfermedades	0.307979
cantidad_enfermedades	enf_cardiovascular	0.307979
diabetes	edad	0.373857
edad	diabetes	0.373857
cantidad_enfermedades	edad	0.469697
edad	cantidad_enfermedades	0.469697
cantidad_enfermedades	hipertension	0.508123
hipertension	cantidad_enfermedades	0.508123
cantidad_enfermedades	epoc	0.655265
epoc	cantidad_enfermedades	0.655265
diabetes	cantidad_enfermedades	0.753796
cantidad_enfermedades	diabetes	0.753796

Imagen 8. comparación de correlación entre variables – Elaboración propia

- Existe una correlación positiva moderada (0.33) entre la edad de una persona y la clasificación de diagnóstico de neoplasia. Esto sugiere que, en general, a medida que la edad aumenta, también puede aumentar ligeramente la probabilidad de recibir un diagnóstico de neoplasia.
- Hay una correlación positiva moderada (0.38) entre la edad y la presencia de diabetes. Esto sugiere que a medida que la edad aumenta, también aumenta la probabilidad de tener diabetes.
- Existe una correlación positiva moderada a fuerte (0.46) entre la edad de una persona y la cantidad de enfermedades que tiene. Esto indica que a medida que una persona envejece, es más probable que desarrolle más enfermedades.
- Hay una correlación positiva moderada a fuerte (0.51) entre la cantidad de enfermedades que tiene una persona y la presencia de hipertensión. Esto sugiere que las personas con más enfermedades tienen una mayor probabilidad de tener hipertensión y viceversa.
- Existe una correlación positiva fuerte (0.66) entre la cantidad de enfermedades que tiene una persona y la presencia de EPOC (Enfermedad Pulmonar Obstructiva Crónica). Esto indica que las personas con más enfermedades tienen una mayor probabilidad de tener EPOC.
- Hay una correlación positiva fuerte (0.76) entre la presencia de diabetes y la cantidad total de enfermedades que tiene una persona. Esto sugiere que las personas con diabetes tienden a tener más enfermedades en comparación con aquellas que no tienen diabetes.

Selector de variables Lasso: Al implementar este selector de variables, selecciona las siguientes características:

	Feature	Coefficient
10	clasifica_diagn	407841.269622
3	epoc	185178.257494
1	ciudad	156375.662108
7	genero	59402.998263
8	edad	18715.435114
5	hipertension	9496.883154
2	cancer	1886.363644
6	enf_cardiovascular	1349.328952
4	diabetes	-6787.915270

Imagen 9. Coeficiente de características Lasso – Elaboración propia

- Los coeficientes positivos indican una asociación positiva entre la característica y la variable de respuesta, mientras que los coeficientes negativos indican una asociación negativa.
- Epoc (Enfermedad Pulmonar Obstructiva Crónica): Tiene el coeficiente más alto positivo, lo que significa que esta característica tiene la mayor influencia positiva en la variable de respuesta.
- Edad: Es la segunda característica más importante, con un coeficiente positivo alto. Esto sugiere que la edad también tiene una fuerte influencia positiva en la variable de respuesta.
- Eventos: También tiene un coeficiente positivo alto, lo que indica que esta característica tiene una influencia significativa en la variable de respuesta.
- Género: Tiene un coeficiente positivo, lo que significa que el género también está relacionado positivamente con la variable de respuesta, aunque en menor medida que las características anteriores.
- Hipertensión, Ciudad de Medellín, Ciudad de Bogotá, Clasificación de diagnóstico de enfermedades endocrinas, nutricionales y metabólicas, y Cáncer: Todas estas características

tienen coeficientes positivos, lo que indica que también tienen una influencia positiva en la variable de respuesta, pero menos que las características anteriores.

SelectKbest: Al implementar este selector de variables, selecciona las siguientes características: **'ciudad', 'cancer', 'epoc', 'diabetes', 'hipertension', 'enf_cardiovascular', 'genero', 'edad', 'cantidad_enfermedades', 'clasifica_diagn'.**

Para este caso de estudio, se decide no implementar ningún selector de variables, ya que son pocas variables y estos seleccionar casi la mayoría de estas.

F. Comparación y selección de técnicas

RESULTADOS Y ANÁLISIS DE MODELOS

A continuación, se mostrarán los resultados de los modelos realizados. Dentro de estos modelos se evalúan métricas de desempeño como: MAE, MSE, RMSE y R-SQUARED.

Modelo de regresión Lineal

Las métricas de desempeño para este modelo fueron las siguientes:

```
MSE: 5914485824411.961
R2: 0.035348596356535866
RMSE: 2433292.9189685117
MAE: 1747132.0037556915
```

Imagen 10. Métricas de desempeño del modelo de Regresión lineal – Elaboración propia

Modelo de regresión Ridge y Lasso

Las métricas de desempeño para este modelo fueron las siguientes:

```
Ridge MSE: 5857336061336.568
Lasso MSE: 5857340699454.616
R2: 0.035348596356535866
RMSE: 2433292.9189685117
MAE: 1747132.0037556915
```

Imagen 11. Métricas de desempeño del modelo de Regresión Ridge y Lasso – Elaboración propia

Modelo Random Forest

Las métricas de desempeño para este modelo fueron las siguientes:

```
MSE: 5312521786734.141
RMSE: 2304890.840524588
MAE: 1562733.0373864206
R-squared: 0.13352880527546263
```

Imagen 12. Métricas de desempeño de Modelo Random Forest – Elaboración propia

Modelo XGBoost

Las métricas de desempeño para este modelo fueron las siguientes:

```
MSE: 5305282655550.235
RMSE: 2304890.840524588
MAE: 1560600.5554112373
R-squared: 0.13470950606079368
```

Imagen 13. Métricas de desempeño de Modelo XGBoost – Elaboración propia

Modelo Decision Tree Classifier

Las métricas de desempeño para este modelo fueron las siguientes:

```
MSE: 6028680218050.424
RMSE: 2455337.0884769415
MAE: 1613284.9545046464
R-squared: 0.016723514585795463
```

Imagen 14. Métricas de desempeño de Modelo Decision Tree Classifier – Elaboración propia

Modelo SGD Regressor

Las métricas de desempeño para este modelo fueron las siguientes:

```
MSE: 5920914429502.3
RMSE: 2433292.9189685117
MAE: 1748598.7363645518
R-squared: 0.0343000922079133
```

Imagen 15. Métricas de desempeño de Modelo SGD Regressor – Elaboración propia

El desempeño general de los modelos se puede considerar bajo, pero esto tiene dos explicaciones principales:

- La Variable objetivo tiene una escala mayor a las variables predictoras.
- La variable objetivo no se puede escalar respecto a las variables predictoras, ya que esto afectaría a los valores predichos por los modelos.

Por tanto, para mejorar las métricas de desempeño del modelo, se necesitaría de más variables relacionadas con la variable objetivo, para así, otorgar características a los modelos que les permita, mejor desempeño.

G. Despliegue de modelos para ejecución del reto

Para el despliegue, se decidió escoger el modelo de *Random Forest*, ya que este mostró mejor desempeño respecto a los demás modelos.

Por ultimo se hace el despliegue ingresando la siguiente información al modelo con el fin de predecir el valor de la póliza según las características del usuario:

```
data_pred={
    'edad': 40,
    'epoc': 1,
    'diabetes':0,
    'clasifica_diagn':3,
    'ciudad': 2,
    'hipertension':0,
    'cancer':0,
    'enf_cardiovascular': 1
```

Imagen 16. Datos para despliegue del modelo – Elaboración propia

Por último, se obtiene el valor a pagar por el usuario según las características de este.

```
El costo del seguro es:  2871310.323932727
```

Imagen 17. Resultado del despliegue del modelo – Elaboración propia

Esta estimación del valor de la póliza podría no acercarse al valor mas cercano a la realidad teniendo en cuenta el desempeño de los modelos expuestos y al contexto dato anteriormente.

Conclusiones

A continuación, se postula la conclusión general de este estudio:

- Es necesario obtener información adicional en cuanto a variables se trata para obtener un mejor desempeño en las predicciones de los modelos.
- Se debe tener en cuenta la relación que tiene la variable objetivo con respecto a las demás variables, en este caso, fue difícil encontrar una relación entre estas, haciendo que el desempeño se viera afecto negativamente.
- Teniendo en cuenta que estos datos no son didácticos, sino reales, se debe buscar otras técnicas para el tratamiento de estas distintas a las convencionales, o en su defecto, revisar variable por variable, encontrando relación entre estas y por ende, acotar las variables a un grupo en las cuales se encuentre una predicción acertada al momento de implementar modelos de predicción.

Referencias

[1] Amazon, «AWS Amazon,» Amazon Web Services, 1 Enero 2022. [En línea]. Available: https://aws.amazon.com/es/what-is/python/. [Último acceso: 25 Septiembre 2022].

[2] A. S. Alberca, «Aprende con Alf,» Aprende con Alf, 14 Junio 2022. [En línea]. Available: https://aprendeconalf.es/docencia/python/manual/pandas/. [Último acceso: 25 Septiembre 2022]

[3] scikit-learn, « scikit-learn Machine Learning in Python» Diciembre 2022. [En línea]. Available: https://scikit-learn.org/stable/