Analítica III - Caso de estudio de tarifación para seguros de salud

Cesar Iván Ávila Díaz. cesar.avila@udea.edu.co
Verónica Andrea Morales González. veronica.moralesg@udea.edu.co

Repositorio GitHub: https://github.com/veronica1908/FINANZAS_E4_AIII

Resumen: El caso de estudio aborda el desafío de "UdeA Insurance" en predecir costos y demandas futuras de servicios de salud para determinar precios adecuados para sus seguros. Se elaboró un modelo predictivo utilizando datos históricos de reclamaciones, y se diseñaron herramientas para ayudar al equipo de ventas. El proceso incluyó limpieza y transformación de datos, análisis exploratorio y selección de variables, donde se identificaron factores clave. Se emplearon técnicas de modelado avanzadas, incluyendo árboles de decisión, Random Forest y Gradient Boosting, optimizando hiperparámetros con GridSearchCV. Finalmente, el modelo fue evaluado y ajustado para su despliegue, demostrando su efectividad en mejorar la precisión de la tarifación de seguros de salud.

Palabras Clave: Modelo predictivo, Seguros de salud, Tarifación.

1. Descripción del caso problema

En el sector de los seguros, el riesgo se refiere a la posibilidad de que se produzca una pérdida o daño cuantificable en términos monetarios. En los seguros de salud, este riesgo se manifiesta a través de eventos que requieren intervención médica, como enfermedades, accidentes, exámenes diagnósticos o partos. El departamento de actuaria de "UdeA Insurance" ha estado recopilando datos históricos sobre los costos asociados a reclamaciones de servicios de salud y enfrenta el desafío de prever futuros costos y demandas de servicios de salud para determinar el precio correcto de los seguros. La tarea principal es desarrollar un modelo predictivo que permita estimar con precisión las futuras utilizaciones y costos de servicios de salud, y utilizar estas predicciones para fijar precios adecuados para los seguros. Además, se deben proporcionar herramientas que asistan al equipo de ventas en la comunicación de tarifas y beneficios a los clientes.

2. Diseño de la solución para el caso de estudio

La solución propuesta se muestra gráficamente a continuación:

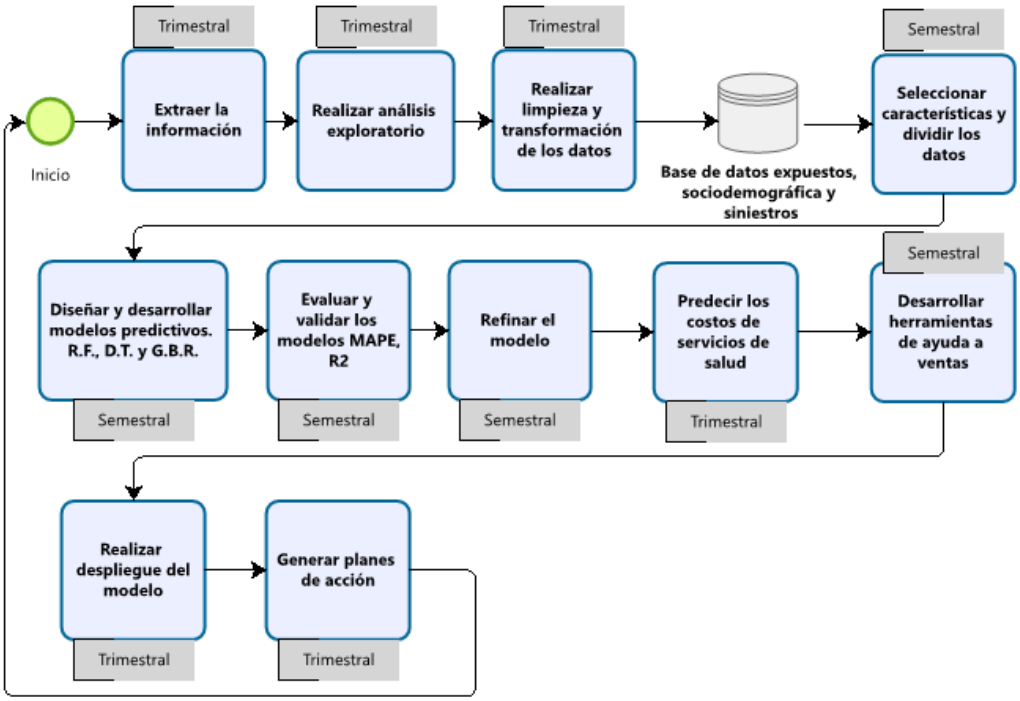


Ilustración 1. Solución teórica propuesta. Elaboración propia.

3. Desarrollo del caso de estudio

a. *Limpieza y transformación*

Durante esta etapa, se realizaron operaciones de limpieza y transformación de datos para asegurar la coherencia y calidad, esto incluyó manejar valores nulos y ajustar el formato de algunas de las variables.
Para este caso, se tomó primero la base de expuestos en la que se muestra que hay 300900 registros. Se muestra la fecha de inicio y fecha de fin de vigencia de la póliza y también hay fecha de cancelación para algunas de las pólizas, las que no han sido canceladas se encuentran vacías, por lo que se agrega la fecha de finalización a las que están vacías, pero también se crea una variable binaria que indica si la póliza fue cancelada (1) o no (0). En la base sociodemográfica se eliminó los valores nulos de la variable mujer (-1), los valores nulos de la variable ciudad se agregaron a la categoría “sin información” que ya existía, se creó la variable edad a partir de la fecha de nacimiento con respecto al 31 de diciembre de 2019 y se procedió a unir las tablas expuestos y sociodemográfica con el filtro de los asegurados registrados en la segunda base y esta se unió a la tercera base que corresponde a siniestros.
Se crea otra variable a partir de la edad con grupos de 10 años para visualizar mejor la distribución de los datos. Se extrae el mes y año de pago a partir de la fecha, para ver la distribución de los datos en el tiempo. Se calcula también el tiempo de cobertura de la póliza a partir de las fechas de inicio y fin. Se realiza también la verificación de que no haya fechas de inicio mayores a las fechas de cancelación o fin, se identifican 65 registros con esta novedad, los cuales son eliminados.

b. Análisis exploratorio

En el análisis exploratorio, se realiza una revisión rápida para la base sociodemográfica, obteniéndose lo siguiente: Se tiene datos de 267233 personas, el 54% de los asegurados son mujeres y 46% hombres. En cuanto a la ciudad en la que se encuentran los asegurados, se tiene que el 48% está en Cali, 19% en Bogotá, 17% en Cartagena, 13% en Medellín, 3% en Barranquilla y se tiene menos del 1% sin información. Más del 99% de los asegurados no tienen cáncer, solo 925 de ellos tienen esta enfermedad. El 2% tiene EPOC. El 6% tiene diabetes. Solo el 1% sufre de hipertensión. Menos del 1% sufre alguna enfermedad cardiovascular.

Para la base de siniestros, se tiene 3308480 registros distribuidos en 24 meses, se tiene la variable de reclamación que muestra el tipo de servicio solicitado por el asegurado, con 41 tipos diferentes de reclamación donde el 38% de los casos son por consulta externa, seguido de exámenes de diagnóstico con el 21%. Se muestra el código del diagnóstico, en total 5830 códigos diferentes, donde el 83% es "9", que pertenece a "Diagnóstico pendiente" de acuerdo con la siguiente variable que es el nombre como tal del diagnóstico. La variable “Eventos”, muestra la cantidad de eventos por asegurado, la mayoría de ellos solo tienen un evento registrado, representando el 70.4% de los casos. Se muestra un máximo de eventos por asegurado, de 279 con menos del 1% de los casos y un promedio de 2 eventos. El valor pagado se muestra en billones, la gran mayoría está alrededor de 0,1 billones, es decir, 100 millones. Hay un 3% de los datos que está en 97,88 millones.

Se elabora una matriz de correlación, observándose que la correlación positiva más alta es entre la edad y la diabetes con 0.42. También se da entre la edad y la hipertensión (0.25), EPOC (0.21), enfermedad cardiovascular (0.16) y cáncer (0.14). También entre hipertensión y diabetes (0.29), entre cáncer y diabetes (0.18). En cuanto a correlaciones negativas, se tiene como se espera entre duración y cancelación de la póliza.

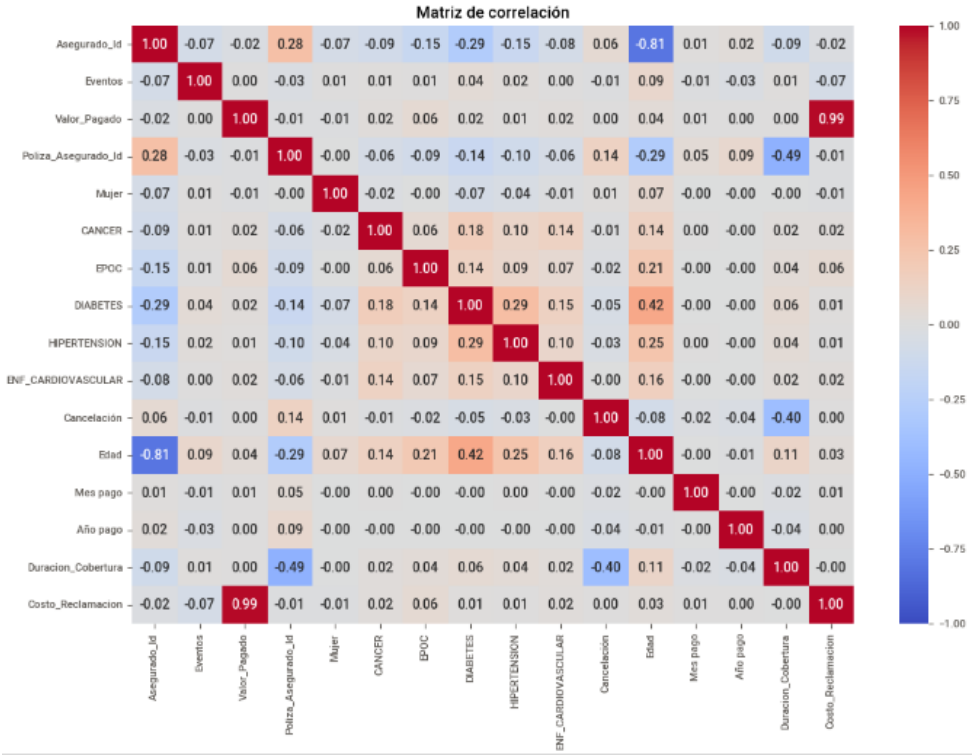


Ilustración 2. Matriz de correlación. Elaboración propia.

El análisis exploratorio con respecto a la variable objetivo arrojó lo siguiente:

```
Estadísticas descriptivas de Valor_Pagado:
count    3.778234e+06
mean     8.017543e+05
std      2.910020e+06
min      1.020000e+05
25%      4.346523e+05
50%      4.844292e+05
75%      6.315640e+05
max      1.977986e+09
```

En cuanto al valor pagado por género, se tiene que hay datos más altos para el género hombre, sin embargo, se muestra un valor atípico muy alto en la categoría de mujer, siendo el valor pagado de 1,9 billones de pesos.

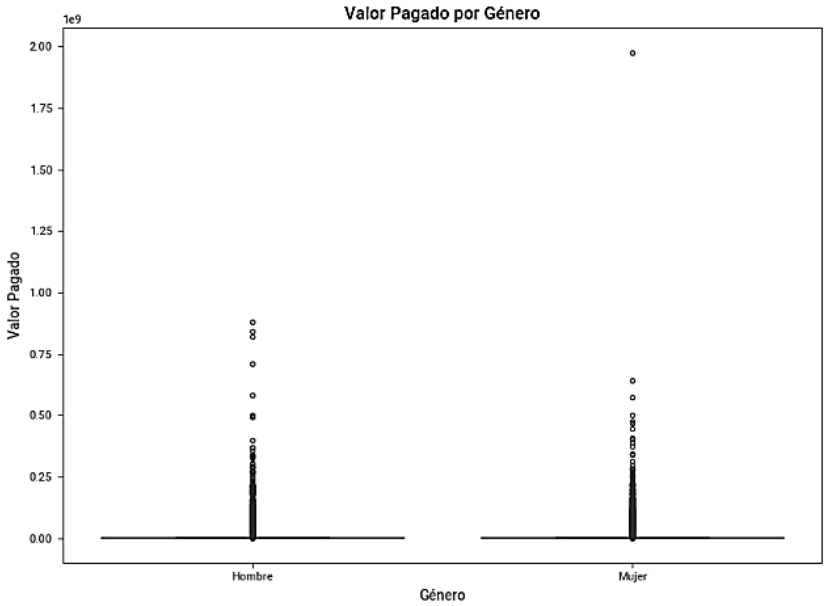


Ilustración 3. Valor pagado por género. Elaboración propia.

Se grafica la variable objetivo con respecto a la edad utilizando los grupos definidos y allí se observa que hay una relación positiva en general, donde a mayor edad, mayor es el valor pagado por el seguro.

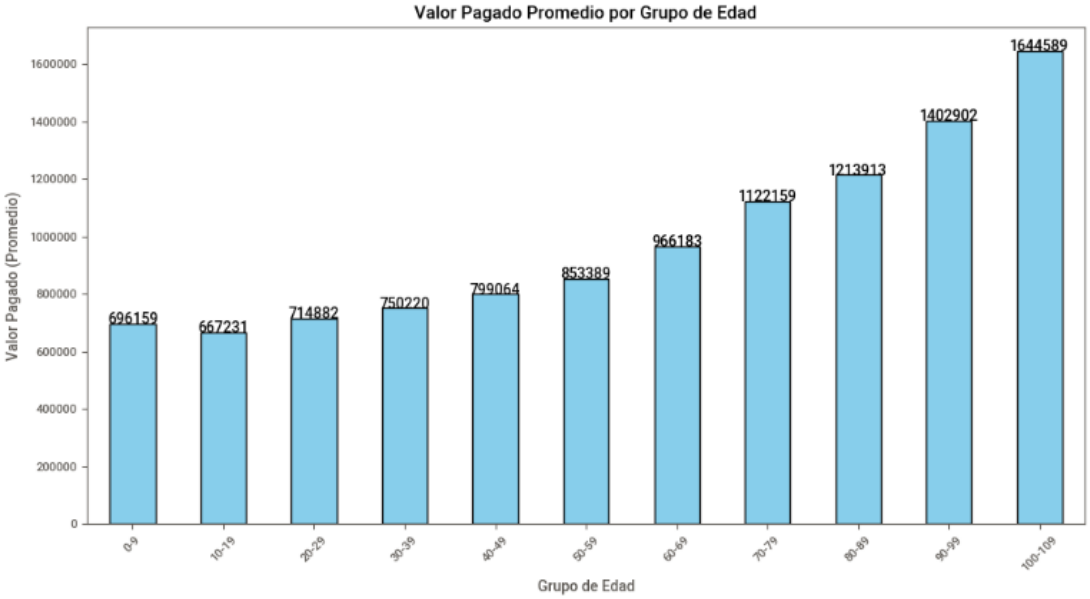


Ilustración 4. Valor pagado por grupo de edad. Elaboración propia.

Con respecto a la ciudad, se grafica el promedio del valor pagado en cada una de ellas y se tiene un promedio más alto en la ciudad de Medellín, seguido por la ciudad de Cartagena y Bogotá. Las diferencias son relativamente bajas, pero pueden representar grandes diferencias en masa.

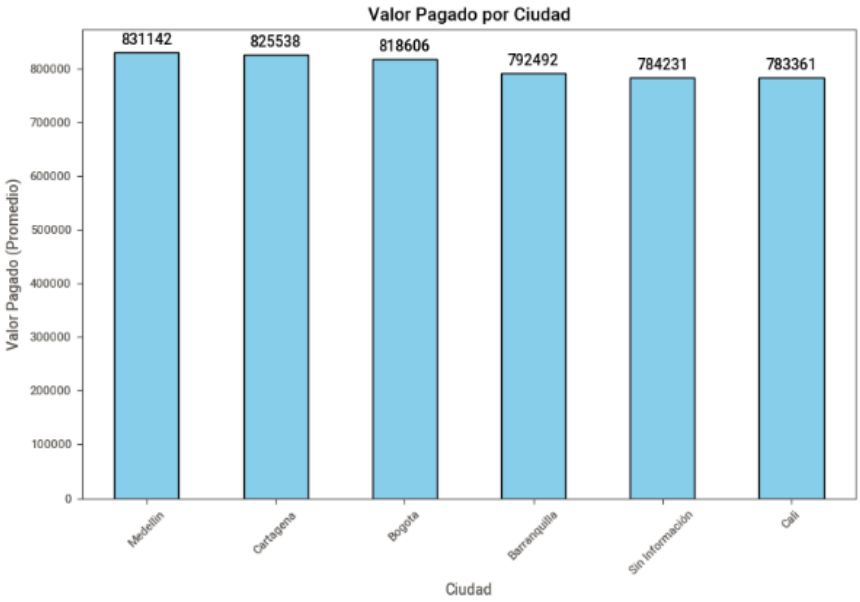


Ilustración 5. Valor pagado por ciudad. Elaboración propia.

Al revisar el promedio del valor pagado por enfermedad, se tiene que es la diabetes la que genera valores más altos de pago y el cáncer el que tiene menor valor promedio.

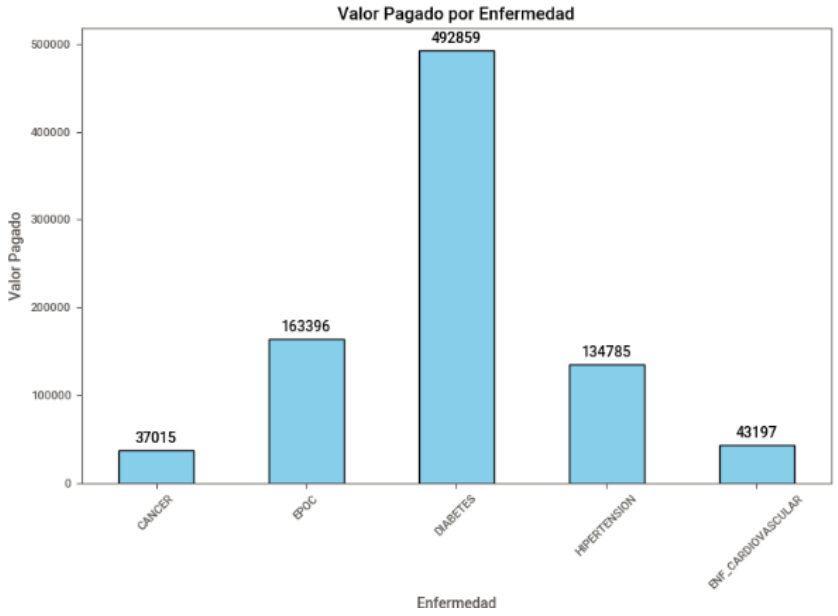


Ilustración 6. Valor pagado por enfermedad. Elaboración propia.

Finalmente, al revisar la tendencia del total del valor pagado mes a mes durante los dos años, se ve un aumento en el tiempo, con una caída de valor en enero de 2019 y continúa aumentando en los siguientes meses.

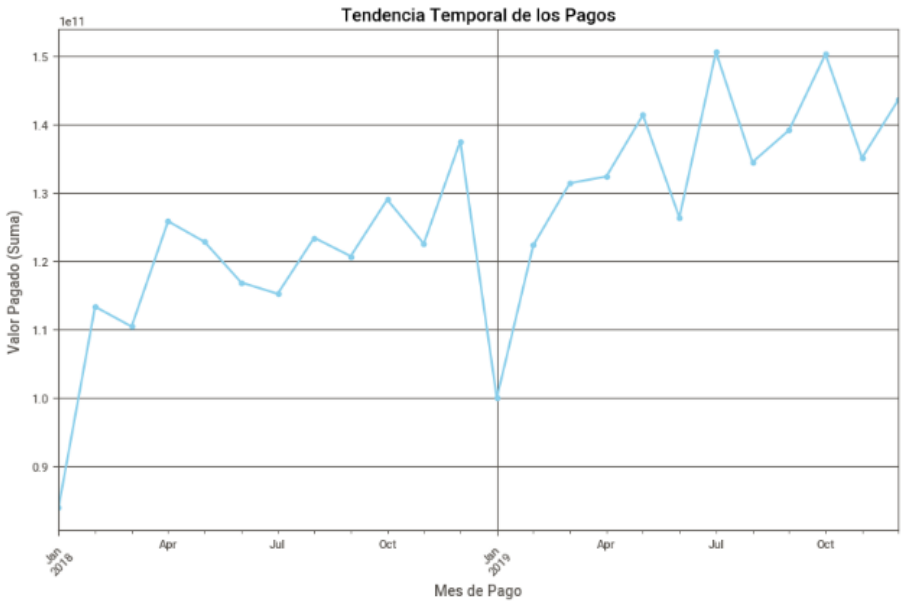


Ilustración 7. Tendencia temporal de los pagos. Elaboración propia.

c. Selección de variables

En el proceso de selección de variables, se llevó a cabo una serie de pasos para identificar y determinar qué características eran más relevantes para el modelo predictivo. Se tuvo limitaciones computacionales tanto para la selección como para medir los modelos iniciales, por lo cual se realizó los siguientes ajustes a la base final:

Se decidió trabajar solo con los datos de 2019, que son los más recientes y representan el 52% de los datos, se creó una nueva variable para reclasificar la edad en un número menor de grupos y esto es de acuerdo con la clasificación de grupos etáreos definidos por el ministerio de salud en Colombia, estos son Primera Infancia (0-5 años, Infancia (6-11 años), Adolescencia (12-18 años), Juventud (14-26 años), Adultez (27-59 años) y Persona Mayor (60 años o más), de estos grupos elegimos solo los dos últimos para trabajar los modelos [1].
Como la variable "Reclamación" tiene 41 valores diferentes, se realiza una reclasificación para disminuir las categorías a un total de tres: Atención ambulatoria y consultas, tratamientos y cirugías hospitalarios y la tercera categoría como otros servicios. Esta reclasificación se realiza en una nueva variable nombrada "Cat_reclamacion" y se realiza con ayuda de inteligencia artificial para definir dónde ubicar cada tipo de reclamación.

Una vez realizados estos ajustes, se procede a eliminar las variables que no aportan al modelo: Mes_Pago, Asegurado_Id, Diagnostico_Codigo, Diagnostico_Desc, Eventos, Poliza_Asegurado_Id, FECHA_INICIO, FECHA_CANCELACION, FECHA_FIN, FechaNacimiento, Edad, Mes pago, Año pago, Costo_Reclamacion, Duracion_Cobertura, Reclamacion, Grupo_Edad y Cancelación.
Se verifican los valores únicos para las variables que serán utilizadas en los modelos y se muestran a continuación:

```
Valor_Pagado          288521
Mujer                      2
Ciudad                     6
CANCER                     2
EPOC                       2
DIABETES                   2
HIPERTENSION               2
ENF_CARDIOVASCULAR         2
Grupo etareo               2
Cat_reclamacion            3
```

Se realiza la conversión a dummies y para ello se crea una instancia de LabelEncoder que se utilizará para transformar las variables categóricas en valores numéricos.
Luego se realiza el escalado y estandarizado de los datos con el 15% de los datos definidos, ya que con un porcentaje mayor, el tiempo de corrida para la evaluación de métricas en los modelos superaba las dos horas. Se utiliza el árbol de decisión en primera instancia para ver el comportamiento en la selección de variables:

```
|--- Cat_reclamacion_Tratamientos y Cirugías Hospitalarios <= 3.64
|   |--- Cat_reclamacion_Otros Servicios <= 6.31
|   |   |--- EPOC <= 1.93
|   |   |   |--- value: [685282.90]
|   |   |--- EPOC >  1.93
|   |   |   |--- value: [1367521.53]
|   |--- Cat_reclamacion_Otros Servicios >  6.31
|   |   |--- Mujer <= -0.35
|   |   |   |--- value: [698401.67]
|   |   |--- Mujer >  -0.35
|   |   |   |--- value: [3157322.11]
|--- Cat_reclamacion_Tratamientos y Cirugías Hospitalarios >  3.64
|   |--- EPOC <= 1.93
|   |   |--- Grupo etareo_Adultez (27-59 años) <= -0.66
|   |   |   |--- value: [9191167.70]
|   |   |--- Grupo etareo_Adultez (27-59 años) >  -0.66
|   |   |   |--- value: [5657822.62]
|   |--- EPOC >  1.93
|   |   |--- Ciudad_Bogota <= 0.92
|   |   |   |--- value: [11745379.40]
|   |   |--- Ciudad_Bogota >  0.92
|   |   |   |--- value: [17947743.43]
```

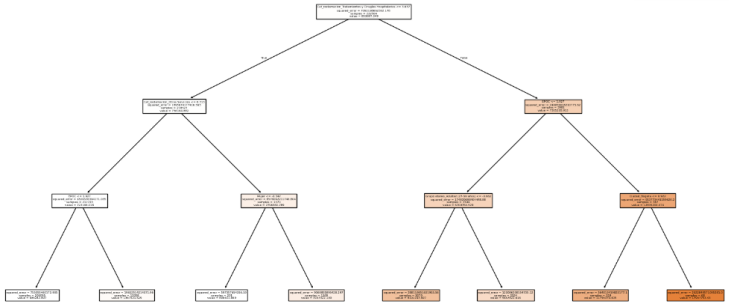


Ilustración 8. Árbol de decisión. Elaboración propia.

El árbol de decisión muestra que las características clave que se utilizan para las divisiones son Cat_reclamacion_Tratamientos y Cirugías Hospitalarios, Cat_reclamacion_Otros Servicios, EPOC, Mujer, Grupo etareo_Adultez (27-59 años) y Ciudad Bogotá.

El siguiente paso fue la selección oficial de variables, que se realizó utilizando varios modelos de regresión, incluyendo Decision Tree Regressor, Random Forest Regressor y Gradient Boosting Regressor con threshold="0.6*mean", se obtuvo la siguiente selección de variables:

```
Data columns (total 8 columns):
 #   Column                                                Non-Null Count   Dtype
---  ------                                                --------------   -----
 0   Cat_reclamacion_Tratamientos y Cirugías Hospitalarios  222504 non-null  float64
 1   DIABETES                                              222504 non-null  float64
 2   ENF_CARDIOVASCULAR                                    222504 non-null  float64
 3   EPOC                                                  222504 non-null  float64
 4   Grupo etareo_Adultez (27-59 años)                     222504 non-null  float64
 5   Grupo etareo_Persona Mayor (60 años o más)            222504 non-null  float64
 6   HIPERTENSION                                          222504 non-null  float64
 7   Mujer                                                 222504 non-null  float64
```

d. Selección de algoritmos y técnicas de modelado

Una vez definidos los modelos y seleccionadas las variables, se realizó la evaluación de modelos, para ello se evaluó el rendimiento de cada modelo utilizando la métrica de R2 y de Error Cuadrático Medio (MSE). Esta evaluación proporcionó una medida cuantitativa del desempeño predictivo de cada modelo en el conjunto de datos. Se compararon los resultados de los diferentes modelos utilizando tanto todas las variables disponibles como un conjunto seleccionado de variables. Esto permitió determinar qué modelos funcionaban mejor con qué conjunto de características.

```
R2 para todos los modelos con todas las variables:
DecisionTreeRegressor: 0.050973328395073667
RandomForestRegressor: 0.07176886611736874
GradientBoostingRegressor: 0.12967780211497343

R2 para todos los modelos con variables seleccionadas:
DecisionTreeRegressor: 0.1104159744043787
RandomForestRegressor: 0.11178003570784394
GradientBoostingRegressor: 0.121197126688727
```

El modelo Decision Tree muestra una mejora significativa en el R2 cuando se utilizan variables seleccionadas en lugar de todas las variables. Aunque el valor de R2 es más alto con las variables seleccionadas, es el de puntaje más bajo, lo que sugiere que el modelo no explica completamente la variabilidad en los datos observados.

Al igual que el modelo de árbol de decisión, Random Forest también muestra una mejora en el R2 cuando se utilizan variables seleccionadas.

El modelo Gradient Boosting tiene el valor más alto de R2 entre los tres modelos cuando se utilizan todas las variables. Sin embargo, este valor disminuye ligeramente cuando se usan variables seleccionadas. Aunque proporciona el R2 más alto, aún puede no ser suficiente para explicar toda la variabilidad en los datos.

```
MSE para todos los modelos con todas las variables:
DecisionTreeRegressor: 6529255902518.11
RandomForestRegressor: 6466827711828.786
GradientBoostingRegressor: 6171389402699.112

MSE para todos los modelos con variables seleccionadas:
DecisionTreeRegressor: 6309041898061.868
RandomForestRegressor: 6304130532528.377
GradientBoostingRegressor: 6244897581517.448
```

Para el DecisionTreeRegressor, se observa que el MSE disminuye cuando se utilizan variables seleccionadas en comparación con todas las variables. Esto podría indicar que algunas variables no aportan mucha información adicional al modelo y podrían estar introduciendo ruido.

Para el RandomForestRegressor, también se muestra una disminución en la selección de las variables.

Para el GradientBoostingRegressor, se ve uu aumento en el MSE cuando se utilizan variables seleccionadas en lugar de todas las variables. Es posible que el modelo inicialmente sobreajustara los datos con todas las variables disponibles. La selección de variables podría haber ayudado a mitigar este sobreajuste al eliminar el ruido y la redundancia en los datos.

Observando los resultados proporcionados, vemos que el GradientBoostingRegressor tiene el R2 más alto tanto con todas las variables como con las variables seleccionadas. Además, presenta el MSE más bajo en ambos casos.

e. Afinamiento de hiperparámetros

El afinamiento de hiperparámetros se realizó utilizando la técnica RandomizedSearchCV, esta técnica selecciona un número fijo de combinaciones aleatorias de hiperparámetros para probar, ya que esto es más eficiente en términos de tiempo de cómputo, ya que en este caso tenemos espacios de hiperparámetros muy grandes. Se realiza el afinamiento para el GradientBoostingRegressor y para el RandomForestRegressor.

Después de realizar un proceso de búsqueda aleatoria para ajustar los hiperparámetros del modelo de Gradient Boosting, se encontraron los siguientes mejores valores: un learning_rate de aproximadamente 0.1396, una profundidad máxima (max_depth) de 3, un mínimo de muestras por hoja (min_samples_leaf) de 3, un mínimo de muestras para dividir un nodo (min_samples_split) de 4, un número de estimadores (n_estimators) de 269 y un factor de muestreo (subsample) de alrededor de 0.7140. Con estos parámetros, se obtuvo un coeficiente de determinación (R2) de aproximadamente 0.1213, lo que sugiere que el modelo es capaz de explicar alrededor del 12.13% de la variabilidad en los datos de manera satisfactoria. Aunque el rendimiento del modelo no es el mejor, estos resultados proporcionan una base para continuar con el análisis y la evaluación del modelo.

El proceso de afinamiento del modelo Random Forest logró encontrar los siguientes mejores hiperparámetros:

Se utilizó el bootstraping (con bootstrap establecido en True), una profundidad máxima de 16 (max_depth), se seleccionaron características con max_features establecido en 'log2', un mínimo de muestras por hoja de 3 (min_samples_leaf), un mínimo de muestras para dividir un nodo de 8 (min_samples_split) y 174 estimadores (n_estimators). Con estos hiperparámetros, el modelo alcanzó un coeficiente de determinación (R2) de aproximadamente 0.1170, lo que sugiere que es capaz de explicar alrededor del 11.70% de la variabilidad en los datos de manera satisfactoria.

f. Evaluación y selección del modelo

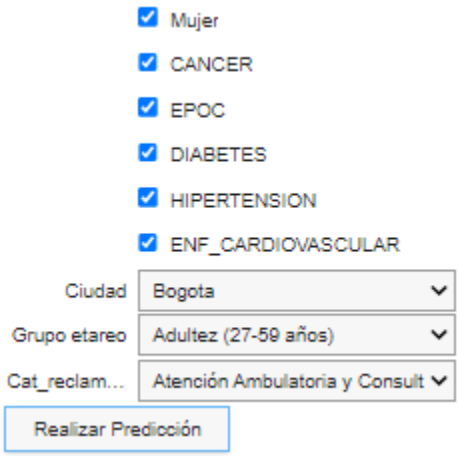
Primero se dividió el conjunto de datos en entrenamiento y prueba usando train_test_split. Luego, se entrenaron los modelos (GradientBoostingRegressor y RandomForestRegressor) con los datos de entrenamiento y se realizó predicciones en el conjunto de prueba. Finalmente, se calculó el R2 en el conjunto de prueba para cada modelo y se seleccionaó el de mejor rendimiento.

Para este caso, el mejor modelo es GradientBoostingRegressor con un R2 en conjunto de prueba de 0.1374. Esto significa que el modelo puede explicar aproximadamente el 13.74% de la variabilidad en los datos de prueba. Es importante tener en cuenta que, aunque este valor de R2 parece bajo, puede ser significativo para este caso de estudio con un conjunto de datos tan grande.

g. Despliegue del modelo

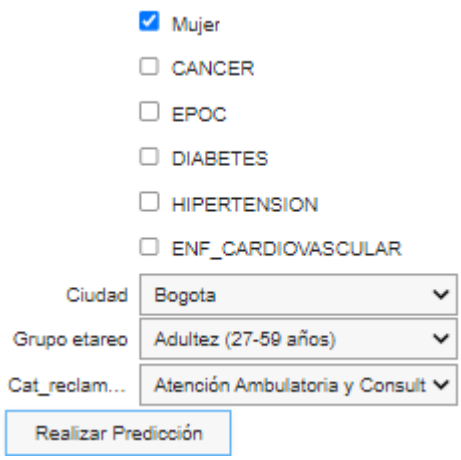
Para el despliegue, se utiliza las características del entrenamiento del modelo (X_train) y se elabora un ayuda ventas en Python que ayude a predecir el valor del seguro de acuerdo con las características seleccionadas. Algunos de los resultados son los siguientes:

1)



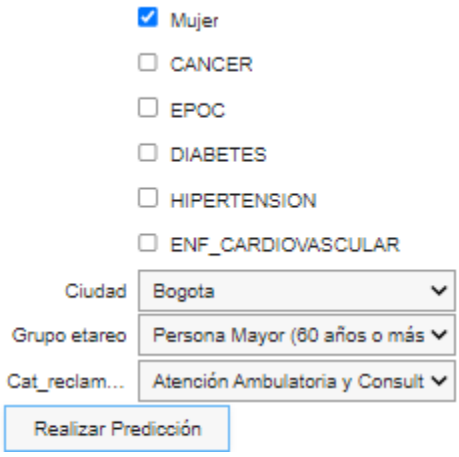
Valor estimado del seguro: $681.367

2)



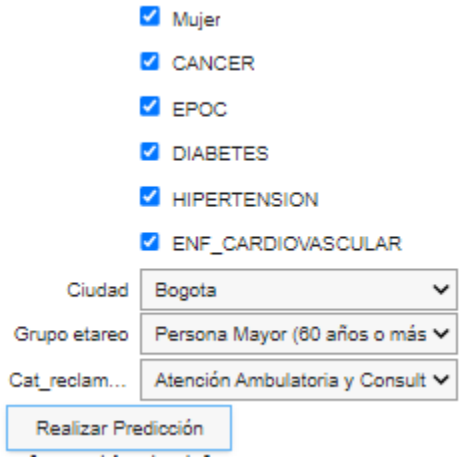
Valor estimado del seguro: $630.281

3)



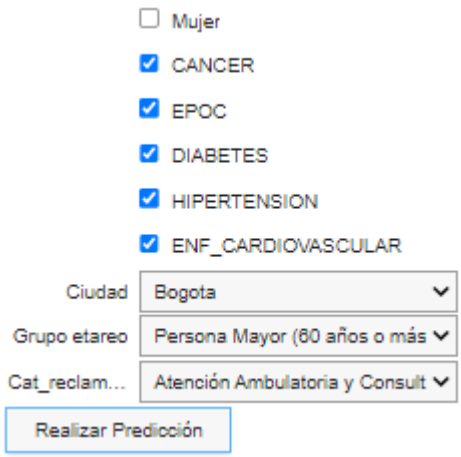
Valor estimado del seguro: $723.218

4)



Valor estimado del seguro: $749.288

5)



Valor estimado del seguro: $749.288

Se observa que los valores son lógicos, ya que, en la primera corrida, se consideró el grupo etáreo de Adultéz con todas las enfermedades ($681.367) y al compararlo con la segunda corrida en la que se mantienen las características excepto las enfermedades, el valor para el seguro es más bajo ($630.281), tal como se espera. En la tercera corrida se mantuvo sin enfermedades pero se cambió el grupo etáreo a Persona mayor y se obtuvo un costo más elevado que las dos primeras corridas ($723.218), luego en la cuarta corrida, se agregan las enfermedades y se encuentra un valor superior a la tercera corrida ($749.288), finalmente se verifica si hay algún cambio al desmarcar la variable mujer y no se obtiene variación alguna ($749.288, por lo que el género no es tan representativo en este caso.

4. Conclusiones y recomendaciones

Basándose en el caso de estudio de tarifación para seguros de salud, se concluye que la limpieza exhaustiva y la transformación de datos son pasos críticos para la construcción de modelos predictivos precisos. La selección cuidadosa de variables y el análisis exploratorio profundo son fundamentales para identificar patrones y relaciones relevantes en los datos. Además, la evaluación y el ajuste de los modelos, junto con el desarrollo de herramientas interactivas para el despliegue, son aspectos clave para proporcionar soluciones efectivas y prácticas en el sector de seguros de salud. Se encontró que el modelo Gradient Boosting presentó el mejor rendimiento, con un coeficiente de determinación (R2) del 13.74% en el conjunto de prueba.

En cuanto a recomendaciones, se sugiere un seguimiento continuo del rendimiento del modelo, la exploración de otras técnicas de modelado y la incorporación de más características para mejorar aún más la precisión del modelo y adaptarse a los cambios en el entorno de datos y del negocio.

5. Referencias

[1] Ministerio de salud, «Ciclo de vida,» [En línea]. Available: https://www.minsalud.gov.co/proteccionsocial/Paginas/cicloVida.aspx#:~:text=La%20siguiente%20clasificaci%C3%B3n%20es%20un,(60%20a%C3%B1os%20y%20m%C3%A1s).. [Último acceso: Mayo 2024].