UNIVERSIDAD
DE ANTIOQUIA

TRABAJO ANALÍTICA DE DATOS APLICADA EN FINANZAS

SANTIAGO GÓMEZ BERRÍO
DIEGO ANDRÉS LUNA PATERNINA
MARIA CLARA SALAZAR DUQUE

ANDRÉS MAURICIO GÓMEZ ARDILA
APLICACIONES EN ANALÍTICA

UNIVERSIDAD DE ANTIOQUIA
FACULTAD DE INGENIERÍA
DEPARTAMENTO DE INGENIERÍA INDUSTRIAL
MEDELLÍN

2023

Analítica de Datos Aplicada en Finanzas

1. **Diseño de solución propuesto:**

   La solución propuesta va encaminada a la predicción de precios de la prima de tarificación en los diferentes seguros que ofrezca la aseguradora "UdeA Insurance", dado que si todos los clientes de la cartera tienen perfiles de riesgo idénticos, la aseguradora simplemente cobra la misma prima para todos los asegurados porque tienen la misma pérdida esperada. Sin embargo, en este caso, los asegurados no son necesariamente homogéneos; el conocimiento de estas características o variables, como edad, sexo, estilo de vida, entre otras, pueden mejorar la capacidad de calcular primas justas para los asegurados individualmente, ya que pueden usarse para estimar o predecir las pérdidas esperadas con mayor precisión (Frees, E. W., 2021).

   De acuerdo a lo anteriormente dicho, se presenta el diseño de la solución en la *figura 1*.

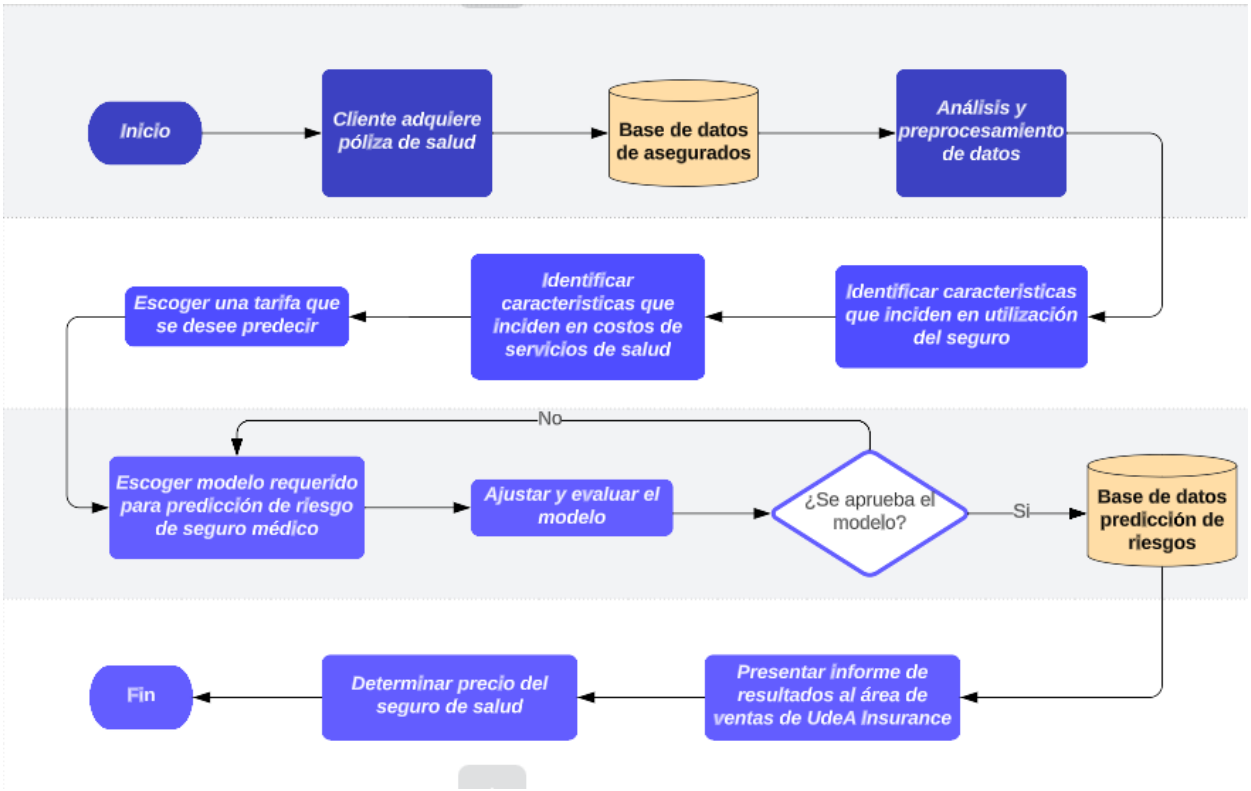

Figura 1. *Diagrama de solución propuesta*

2. **Limpieza y transformación:**

   Se observó detenidamente el contenido de cada una de las tablas de la base de datos, revisando la presencia de nulos, y se cambió los tipos de datos de algunas variables como el formato de las fechas, para realizar series de tiempo, y proceder con el análisis exploratorio. En la variable sexo se arregló la distribución de la categoria masculina y femenina, lo mismo que para la variable regional, donde habían datos categorizados como N/D y se puso como regional 6, ya que este solo era un dato. Además, se obtuvo una columna de edad, obtenida a partir de la variable fecha de

nacimiento, la cual fue eliminada posteriormente para trabajar solo con la edad. Finalmente, se realizó un merge de la tabla utilizaciones y sociodemográfico, ya que la edad se considera una variable clave en el tema de los seguros, en muchos casos es un factor determinante para definir el precio que un asegurado debe pagar por el seguro.

3. **Análisis exploratorio:**

Al analizar los gráficas de series de tiempo de la variable de fecha de reclamación, se identificó la necesidad de delimitar los datos pertenecientes al año 2019, específicamente en los meses de septiembre a diciembre, para un mejor modelamiento, ya que allí se presentaba mayor concentración de las reclamaciones realizadas por los asegurados como se muestra en la *figura 2*.

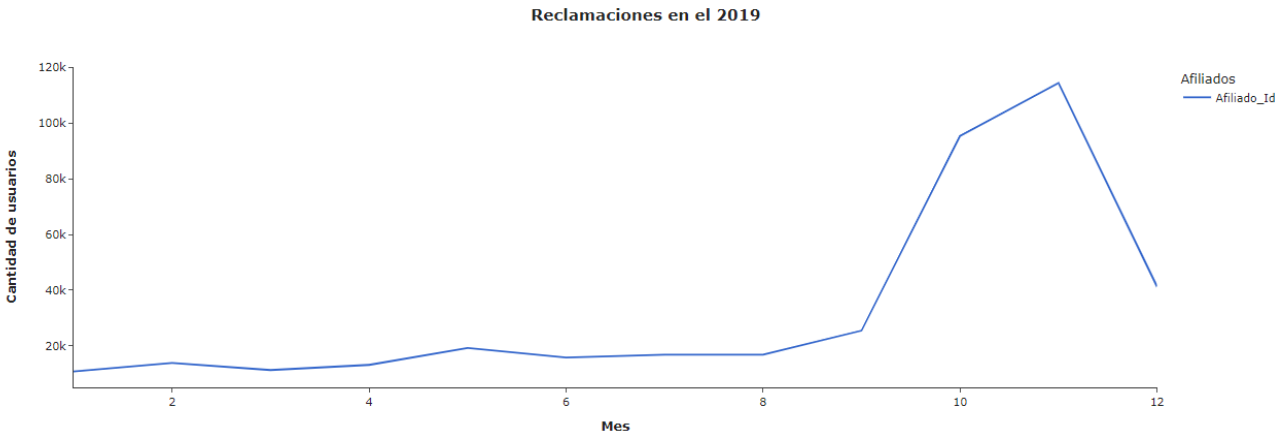

Figura 2. *Reclamaciones mensuales en el 2019*

Adicionalmente, se notó lo siguiente:

- Las causas de reclamaciones se generaron principalmente por consultas externas, exámenes de diagnóstico y laboratorio clínico (*figura 3*).

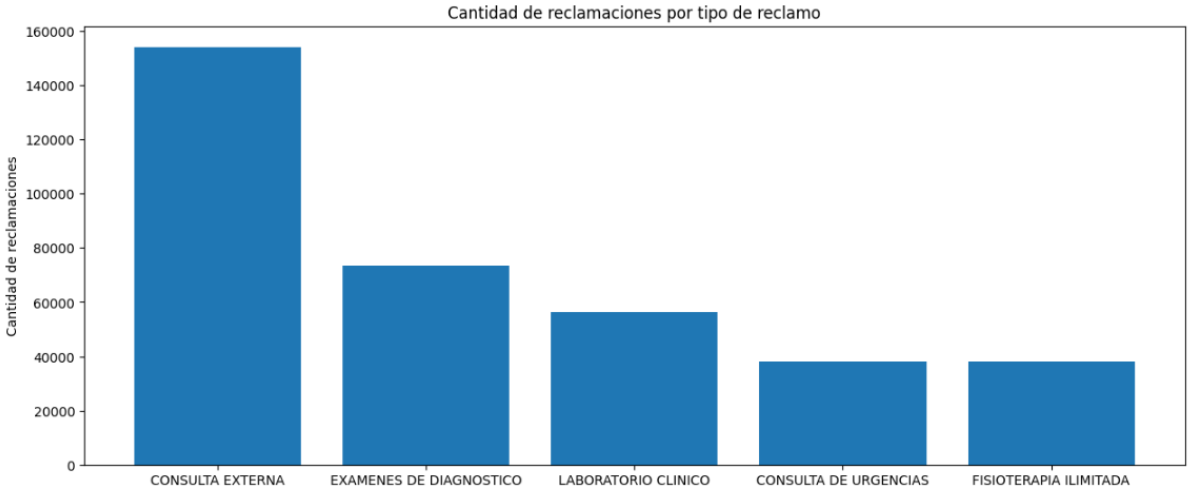

Figura 3. *Cantidad de reclamaciones por tipo de reclamo*

- Los diagnósticos más frecuentes al reclamar los seguros suelen ser aquellos que se encuentran pendientes, otros controles generales de salud de rutina, exámenes de laboratorio, hipotiroidismo no especificado y diabetes mellitus especificada.
- Las regiones más comunes en las que se encuentran los afiliados son centro, antioquia, occidente y norte. Como se muestra en la *Tabla 1*.

	Regional_Desc	Cantidad
0	REGIONAL CENTRO	106907
1	REGIONAL ANTIOQUIA	39582
2	REGIONAL OCCIDENTE	36030
3	REGIONAL NORTE	28259
4	REGIONAL EJE CAFETERO	7377
5	Sin Información	49

Tabla 1. *Regiones*

- La patología mayormente identificada antes de hacer la delimitación de fechas era la hipertensión, pero a partir de septiembre de 2019 los clientes presentan mayormente enfermedad cardiovascular, seguida de EPOC y diabetes.
- El género de los afiliados es femenino en mayor medida.
- Los precios de tarificación se encuentran entre 0 y 200 mil, como se observa en la *Figura 4*.

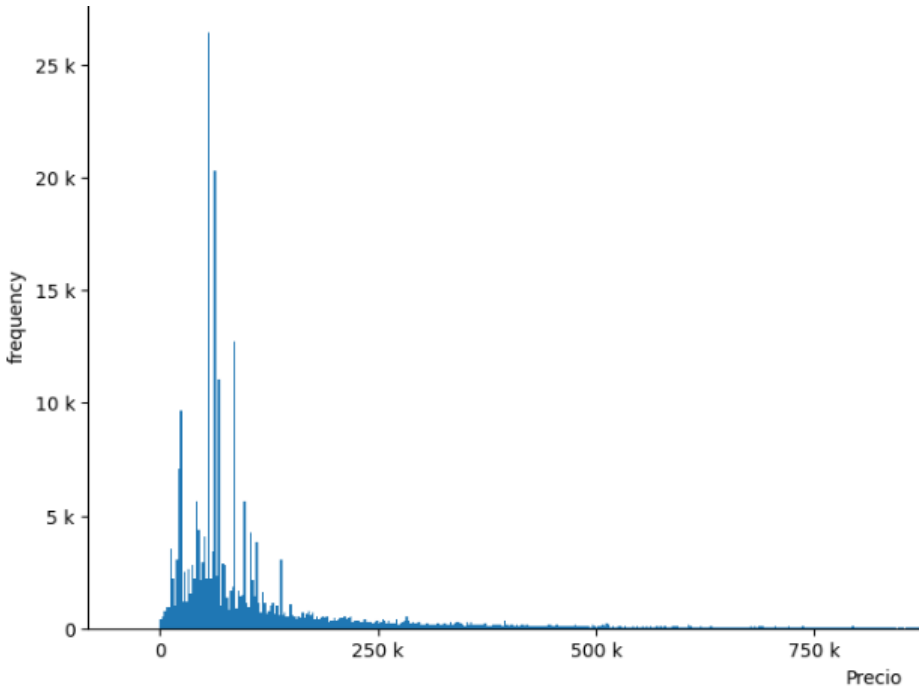

Figura 4. *Precio de tarificación*

- A partir de los 40 años de edad sube más el precio de tarificación, lo cual tiene sentido porque ya se van presentando más complicaciones de salud.

4. **Selección de variables:**
   Para la selección de variables se realizó una matriz de correlación para observar el comportamiento de las variables que más aportan al modelo y se utilizó el método K-best; las variables con mayor puntaje fueron cantidad con un total de 38,57 y reclamación de diagnóstico con un total de 30,95. Sin embargo, se optó por continuar trabajando con todas, para tener más datos que lograran explicar el modelo.

5. **Selección de algoritmos y técnicas de modelado:**
   De acuerdo a una malla obtenida de la página principal de scikit-learn (*Figura 5)*, se siguió la ruta hasta llegar al modelo adecuado. En este caso, se optó por el modelo SGD Regressor.

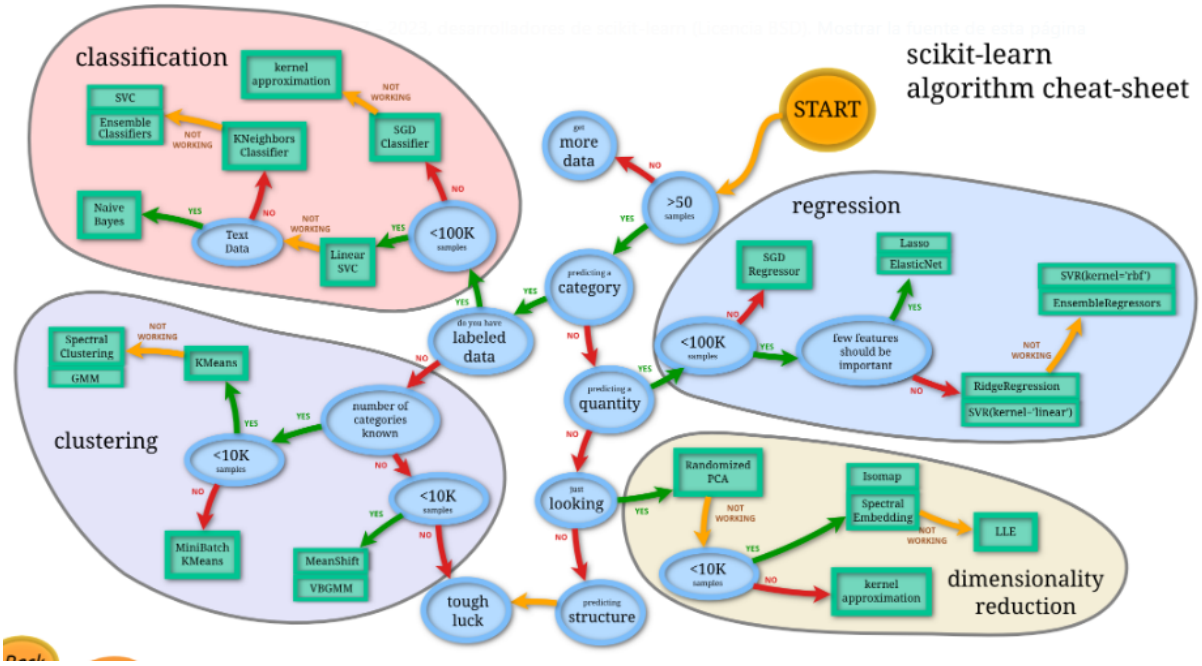

Figura 5. *Malla de modelos.* ***Fuente:*** *scikit-learn*

Sin embargo, el modelo SGD Regressor no arrojó los resultados esperados. Debido a esto, se realizaron varios modelos como Linear Regression, Decision Tree Regressor y finalmente, Randomforest regressor.

6. **Evaluación y selección del modelo:**
   Para seleccionar el modelo que mejor se adapte y represente los datos se tuvieron en cuenta dos métricas de rendimiento, el R-cuadrado y un MAE, en cada uno de los cuatro modelos implementados. A continuación se muestran los resultados de cada modelo.

Modelos	R-cuadrado	MAE
Random Forest Regressor	0,1689	91382,613
Linear Tree Regressor	0,03363	96690,379
Linear regression	0,06214	109359.67
SGD Regressor	0,0594	109882,73

Tabla 2. *Métricas de evaluación de los modelos*

Como se puede observar en la *Tabla 2,* el modelo que mejor se ajusta a los datos es el Random Forest Regressor porque tiene un MAE más bajo que los demás modelos indicando un mejor ajuste, lo cual se puede reafirmar con el valor del R-cuadrado, ya que este se mide en una escala de entre 0 y 1, siendo los valores más cercanos a 1 los que mejor se ajustan al modelo. Como se puede ver el valor del R-cuadrado del modelo Random Forest Regressor es el más cercano a 1 de todos los modelos, por esto se eligió este modelo por encima de los demás. A pesar de que estas métricas siguen dando un poco bajas se considera normal, ya que se están trabajando con datos reales y con estos no se logra obtener precisiones tan altas.

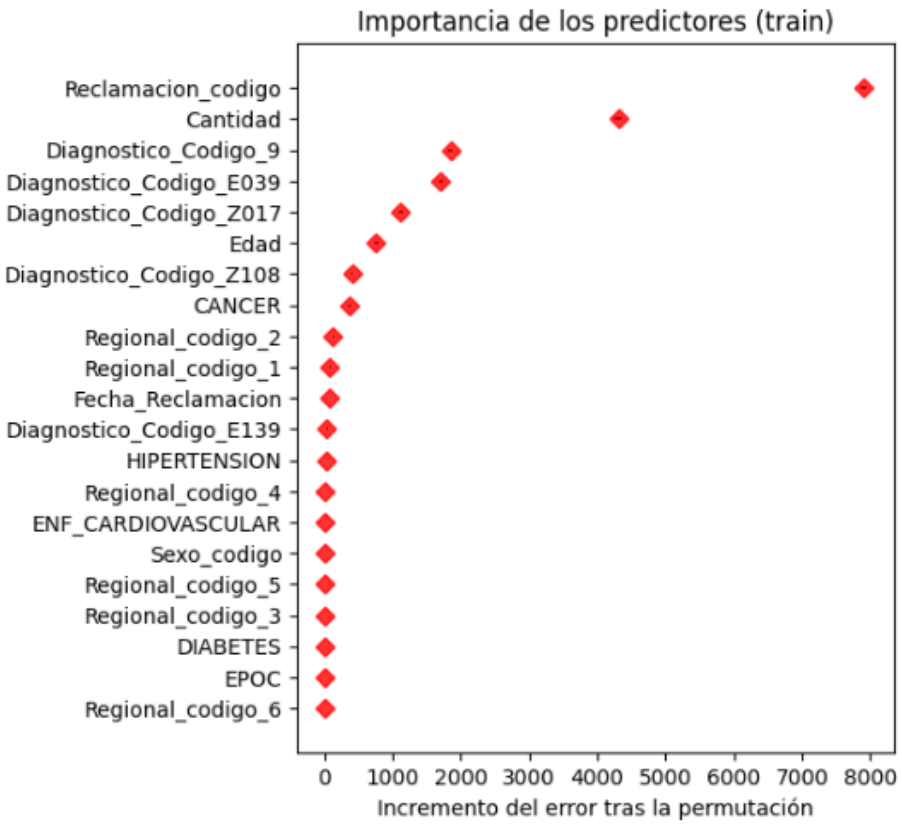

Figura 6. *Importancia de los predictores*

En la *Figura 6* se logra observar la importancia de las variables a la hora de predecir en el modelo, donde claramente se destacan el código de la reclamación, el diagnóstico y la edad, mientras que la región y las enfermedades que presenta el asegurado se observan cómo variables no tan significativas en su nivel de importancia respecto al modelo.

7. **Despliegue del modelo:**

Todos los asegurados que comparten factores de riesgo idénticos se asignan a la misma clase de riesgo y se consideran homogéneos desde el punto de vista de los precios; la aseguradora "UdeA Insurance" les cobra la misma tarifa.

Por lo tanto, incorporar y diferenciar características de riesgo importantes de las personas en el proceso de fijación de precios del seguro es un componente relevante tanto para la determinación de la prima justa para los clientes como para la sostenibilidad a largo plazo de la aseguradora.

De acuerdo a esto, se utilizó la librería pickle la cual hizo que se guardará en un archivo para evitar que cambie los resultados si se realizan diferentes corridas del código. Sumado a esto, se creó un diccionario con diferentes datos que permiten realizar la predicción objetivo. En este caso se tuvieron en cuenta los datos presentados en la *Figura 7*.

```
##Se crea un Dic con datos para predecir el precio
data_pred = {
    "Sexo_codigo": 0,
    "CANCER": 1,
    "EPOC": 0,
    "DIABETES": 0,
    "HIPERTENSION":0,
    "ENF_CARDIOVASCULAR":0,
    "EDAD":48,
    "Fecha_Reclamacion":11,
    "Reclamacion_codigo":9,
    "Cantidad":1,
    "Regional_codigo_1":0,
    "Regional_codigo_2":1,
    "Regional_codigo_3":0,
    "Regional_codigo_4":0,
    "Regional_codigo_5":0,
    "Regional_codigo_6":0,
    "Diagnostico_Codigo_9":1,
    "Diagnostico_Codigo_E039":0,
    "Diagnostico_Codigo_E139":0,
    "Diagnostico_Codigo_Z017":0,
    "Diagnostico_Codigo_Z108":0,
```

Figura 7. *Diccionario de datos para predecir tarifa*

A partir de estos datos entregados al archivo pickle, se obtuvo una predicción del seguro de $61.217,66.

Referencia

Frees, E. W. (2021). Chapter 8 Clasificación de Riesgos | Loss Data Analytics. github. https://ewfrees.github.io/Loss-Data-Analytics-Spanish/C-RiskClass.html