APLICACIONES DE LA ANALÍTICA: FINANZAS

POR:
DANIELA LÓPEZ ANAYA
VALENTINA MÚNERA PULGARÍN

DOCENTE:
ANDRES MAURICIO GOMEZ ARCILA

INGENIERÍA INDUSTRIAL
FACULTAD DE INGENIERÍA
UNIVERSIDAD DE ANTIOQUIA
2023

A. DISEÑO DE LA SOLUCIÓN

El diseño de la solución (observar Ilustración 1), consiste en un modelo de aprendizaje supervisado de regresión, el cual va a predecir el precio promedio de las reclamaciones que pueda realizar el usuario en dos años, el asesor del seguro ingresa los datos necesarios al modelo y descarga la predicción, a partir del resultado calcula el precio del seguro y se lo informa al cliente, si el cliente decide adquirir el seguro, la información del usuario se ingresa a la base de datos, la cual entrena al modelo mensualmente.

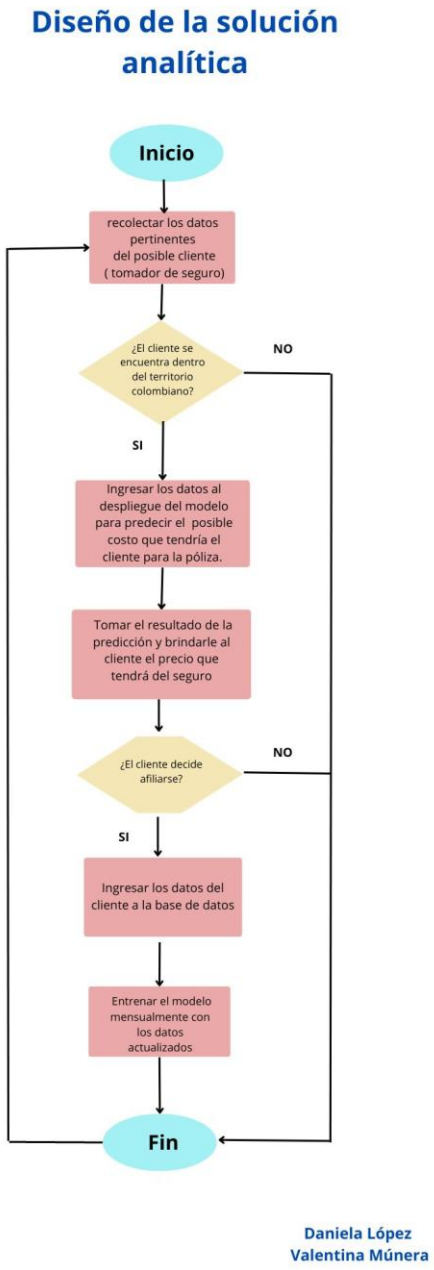

Ilustración 1. Diseño de solución

B. ANALISIS EXPLORATORIO Y TRANSFORMACIÓN

Inicialmente, fue suministrado una base de datos SQL "BD_insurance2019dataset.db" para examinar, esta cuenta con seis tablas.

La tabla "SocioDemograficas" cuenta con la descripción de los códigos de cada una de las regiones de los usuarios, contiene los siguientes campos:

1. **Regional_id:** código de la regional
2. **Regional_Desc:** nombre de la regional

La tabla "Reclamaciones" cuenta con la descripción de los códigos de cada una de las reclamaciones que tiene el seguro, contiene los siguientes campos:

1. **Reclamacion_Id:** código de reclamación
2. **Reclamacion_Desc:** nombre de reclamación

La tabla "Diagnostico" cuenta con la descripción de los códigos de los diagnósticos existentes, cuenta con los siguientes campos:

1. **Diagnostico_Codigo:** código de diagnostico
2. **Diagnostico_Desc:** información de diagnostico

La tabla "Utilizaciones" la cual es tipo transaccional, ya que en ella se encuentran las reclamaciones realizadas por los usuarios. Cuenta con la información de las reclamaciones y usuarios que la realizaron, las fechas en las que la realizaron, el diagnostico, el precio y la cantidad.

La tabla "Regional" la cual es tipo maestra, ya que contiene un solo registro por paciente. Cuenta con información pertinente del usuario, como lo es el código ID, sexo, fecha de nacimiento, código regional y patologías.

La tabla "Genero" cuenta con la descripción de los códigos de los géneros, contiene los siguientes campos:

1. **Sexo_Cd:** código de sexo
2. **Sexo_desc:** genero

Con las tablas mencionadas anteriormente, se crea una tabla por medio de SQL, la cual quedo con los siguientes campos:

1. **afiliado_id:** ID del usuario
2. **sexo_desc:** código de sexo
3. **fechanacimiento:** fecha de nacimiento del usuario
4. **regional_desc:** código de la regional del usuario
5. **cancer:** si el usuario padece de cáncer
6. **epoc:** si el usuario padece de epoc
7. **diabetes:** si el usuario padece de diabetes
8. **hipertensión:** si el usuario padece de hipertensión
9. **enf_cardiovascular:** si el usuario padece de ineficiencia cardiovascular

10. **fecha_reclamacion:** fecha de reclamación
11. **reclamacion_desc:** código de reclamación
12. **diagnostico_desc:** código de diagnostico
13. **cantidad:** cantidad de reclamación
14. **precio:** precio de reclamación
15. **prom_precio:** precio promedio de reclamación

Este procedimiento se realiza con el fin de facilitar la transformación de los datos y optimizar espacio y tiempo de ejecución:

Con la tabla final creada se realizó un análisis de los nulos encontrados en cada uno de los campos, en donde se encontró que el porcentaje de nulos no superaba el 2%, por lo cual se decide eliminar estos registros para no afectar el rendimiento del modelo.

Ahora bien, con la tabla transformada y limpia se realizó la visualización de los datos, en donde se encontró lo siguiente:

En la Ilustración 2, observamos que las reclamaciones con mayor precio promedio son tratamiento quirúrgico hospitalario por cáncer y complicaciones con un precio promedio de 27.123 millones de pesos, tratamiento quirúrgico hospitalario por enfermedad congénita con un precio promedio de 13.4138 millones de pesos y tratamiento médico hospitalario por cáncer y complicaciones con un precio promedio de 13.4038 millones de pesos, en los demás tratamientos se observa un precio promedio menor a 10.8348 millones de pesos.

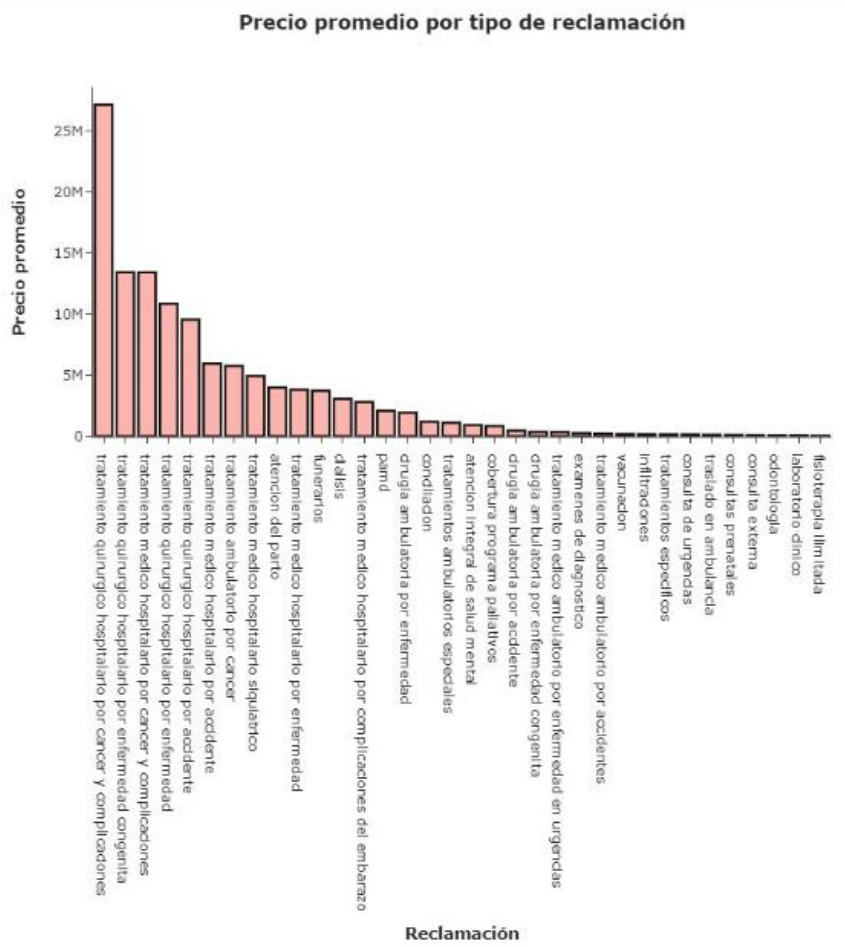

Ilustración 2. Precio promedio por tipo de reclamación

En la ilustración 3, se observa la evolución del precio promedio de las reclamaciones por el año de nacimiento del paciente, en donde los pacientes con el mayor precio promedio de reclamación tienen entre – y – años de edad, así mismo se puede observar que entre menos edad menos precio promedio de reclamación tienen, esto se debe a que han realizado un menor número de reclamaciones en el transcurso de los años.

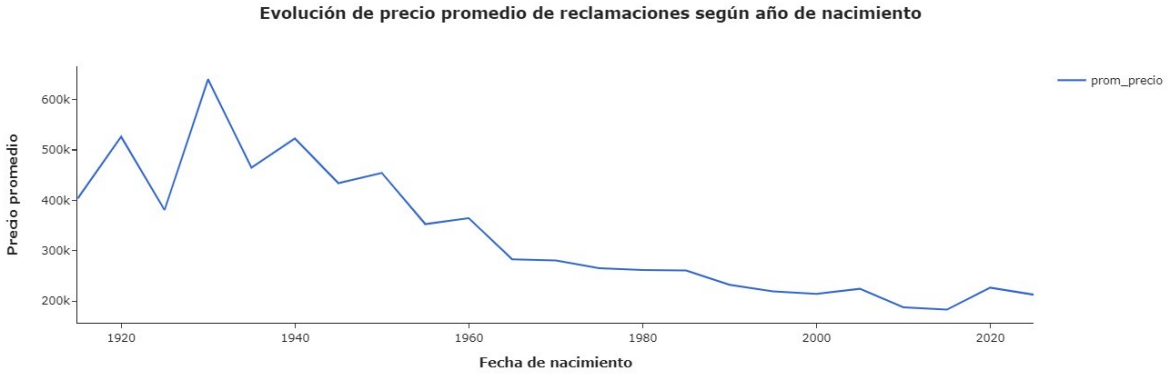

Ilustración 3.Evolución de precio promedio de reclamaciones según año de nacimiento.

En la ilustración 4, se observa el porcentaje de reclamaciones según el sexo del paciente, donde se encontró que los pacientes de sexo femenino tienen el mayor porcentaje de reclamaciones, siendo este el 62.9% y el sexo masculino con un porcentaje de 37.1%.

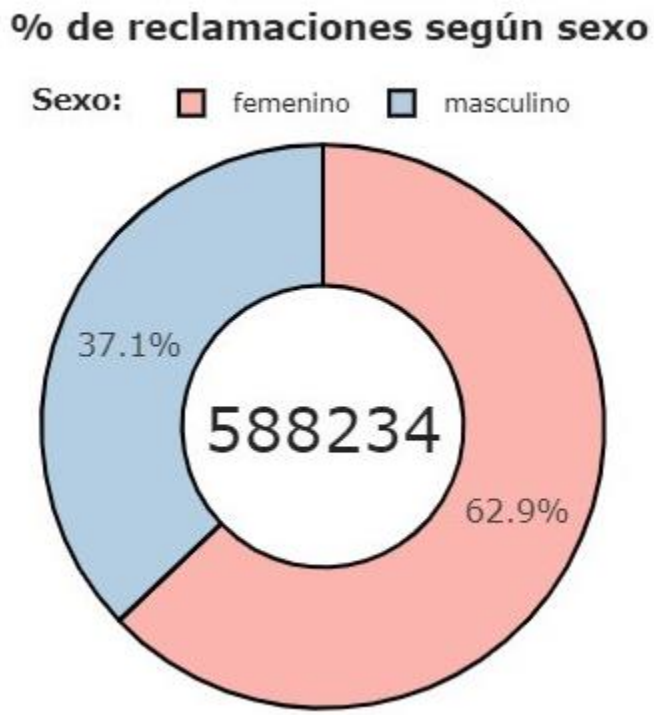

Ilustración 4. Porcentaje de reclamaciones según sexo

En la Ilustración 5, se observa el precio promedio por la presencia de patologías, donde se encontró que los pacientes que sufren de alguna patología tienen mayor precio promedio por reclamaciones que un paciente que no sufre de ella, así mismo se puede observar que los pacientes con cáncer son los pacientes con el mayor precio promedio en comparación de los demás pacientes que sufren de otro tipo de patología.

Adicionalmente, observando la Tabla 1 se evidencia que el sexo femenino al ser el genero con mayor número de reclamaciones no registra con el mayor precio promedio, en comparación con el sexo masculino, que a pesar de contar con menos número de reclamaciones obtiene un mayor numero de precio promedio.

sexo_desc	prom_precio
femenino	257128.99
masculino	289.617.288

Tabla 1. Precio promedio por sexo

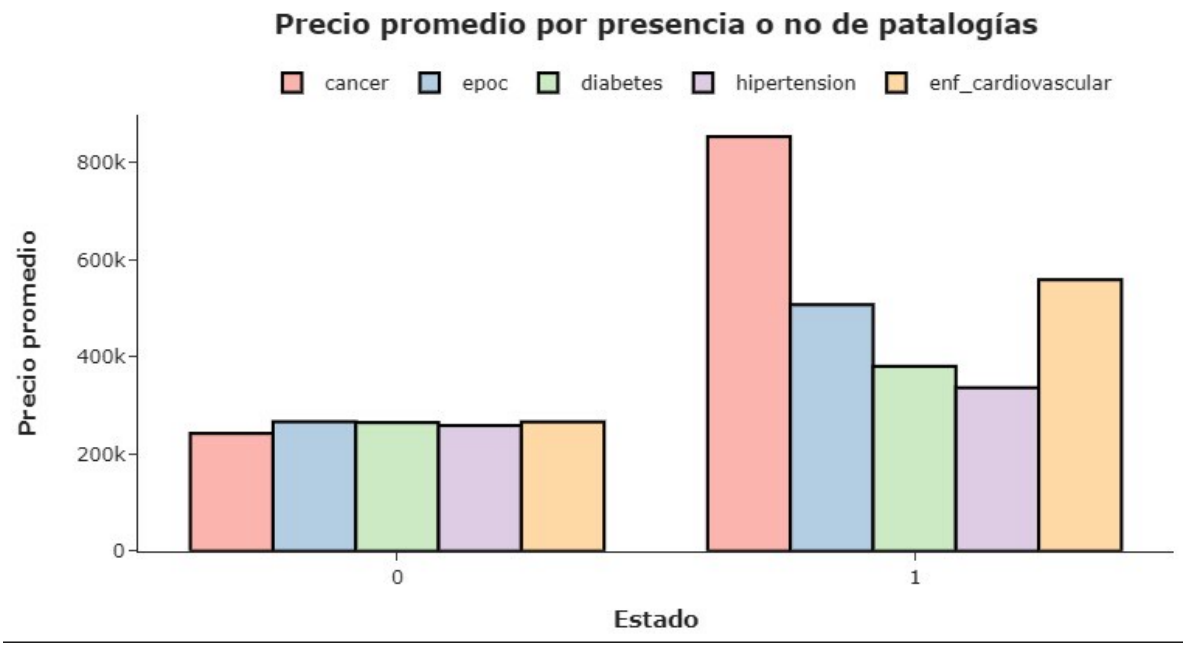

Ilustración 5.Precio promedio por presencia o no de patologías.

C. SELECCIÓN DE VARIABLES

Tras realizar la limpieza de los datos, 2 variables fueron eliminadas, quedando en total 13 variables de entrada y 1 de salida. Sin embargo, se observó la necesidad de aplicar métodos de selección de variables para así determinar cuales tenían mayor influencia con la variable de salida.

El método seleccionado fue KBest, seleccionando 9 variables, las cuales fueron:

- diabetes
- enf_cardiovascular
- cancer
- epoc
- hipertension
- regional antioquia
- regional norte
- regional centro
- edad

El resultado obtenido por el método de selección no selecciona el sexo del usuario, el cual es una variable que se considera importante para este modelo, por lo cual se decide realizar una grafica para visualizar de mejor manera los resultados.

En la Ilustración 6, podemos observar que los resultados obtenidos no tienen una diferencia significativa, por lo cual se cree pertinente utilizar todas las variables de entrada (13) que se tenían inicialmente.

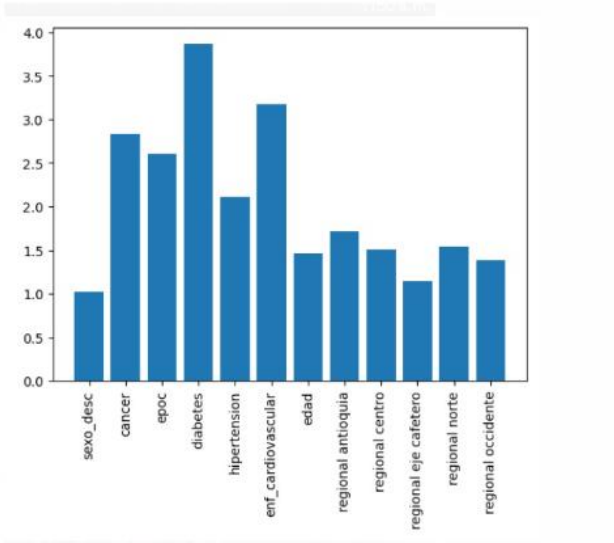

Ilustración 6. Features seleccionado por método K best.

D. SELECCIÓN DEL MODELO

Teniendo en cuenta las características del problema y solución planteada de un modelo de predicción de aprendizaje supervisado de regresión y siguiendo la guía de Scikit Learn (Scikit Learn, 2023), se seleccionó un tipo de modelo para aplicar: SGD Regression. La ruta seguida para seleccionar los anteriores modelos se puede visualizar en la Ilustración 7.

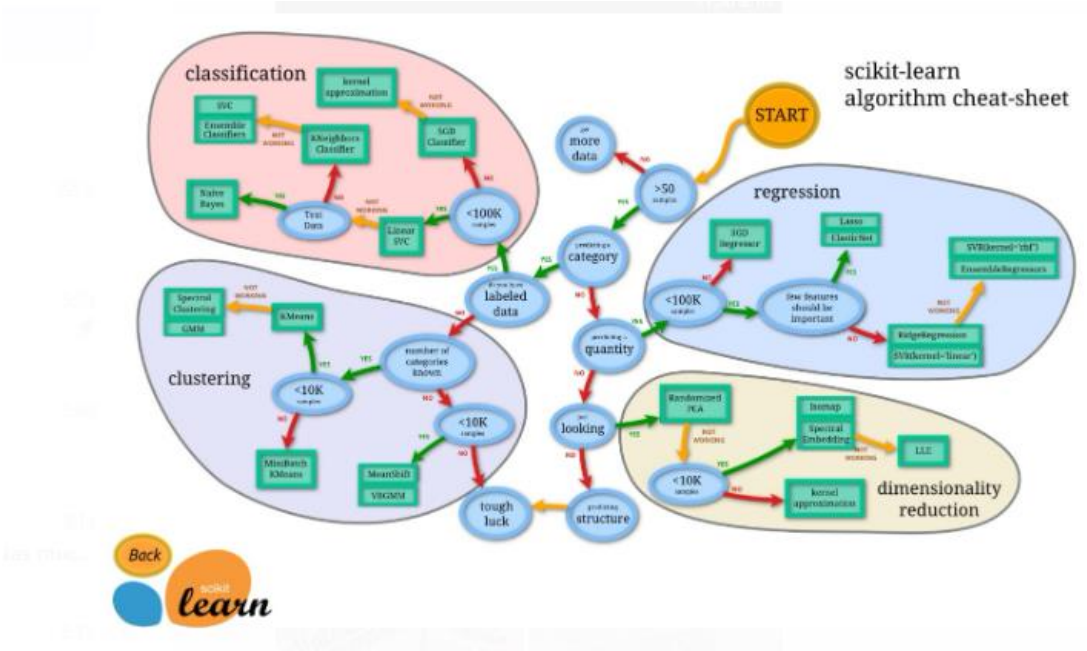

Ilustración 6. Ruta Scikit Learn.

El algoritmo de descenso de gradiente estocástico es un método de optimización iterativo que busca encontrar los coeficientes de regresión que minimizan la función de costo del modelo. En cada iteración, el algoritmo selecciona aleatoriamente un subconjunto de datos de entrenamiento (un solo ejemplo o un lote pequeño) y ajusta los coeficientes utilizando el gradiente de la función de costo calculado en ese subconjunto.

Adicionalmente se seleccionaron otros modelos de regresión para aplicar, los cuales son:

- Ridge: La regresión Ridge encuentra el conjunto de coeficientes que minimiza la suma de los cuadrados de las diferencias entre los valores observados y los valores predichos, teniendo en cuenta tanto el ajuste a los datos como la penalización de la regularización. Es importante destacar que el término de intercepción no se penaliza en Ridge.
- Regresión Lineal: El modelo de regresión lineal asume que existe una relación lineal entre la variable dependiente y las variables independientes

E. COMPARACIÓN DE MODELOS

Para la comparación de modelos y la selección de uno de ellos se tuvieron en cuenta tres métricas:

MSE: Error cuadrático medio es una medida comúnmente utilizada para evaluar el rendimiento de un modelo de regresión. El MSE mide la magnitud promedio de los errores cuadráticos entre los valores reales y los valores predichos por el modelo.

RMSE: Error cuadrático medio raíz es una medida también comúnmente utilizada para evaluar el rendimiento de un modelo de regresión de manera similar al MSE (Mean Squared Error). El RMSE mide la raíz cuadrada de la magnitud promedio de los errores cuadráticos entre los valores reales y los valores predichos por el modelo.

MAE: Error absoluto medio a diferencia del MSE y el RMSE, que miden los errores cuadráticos, el MAE mide la magnitud promedio de los errores absolutos entre los valores reales y los valores predichos por el modelo.

Aplicando a cada modelo las métricas mencionadas anteriormente, se obtuvo el siguiente resultado (Ver tabla 2):

Modelo	MAE	MSE	RMSE
SGD Regression	6.712799	0.00819	0.00191
SGD Regression	6.714385	0.00819	0.00197704
Random Forest	7.4710911	0.008643	0.0019800
Ridge	6.711570	0.008192	0.00188097
Regresión lineal	6.71153346326	-	-

Tabla 2. Resultado metricas

Al observar los resultados obtenidos de cada métrica por modelo, se selección el segundo modelo SGD Regresión el cual tiene los siguientes hiperparametros (Ver tabla 3):

max_iter	tol
50	1e-3

Tabla 3. hiperparametros

F. DESPLIEGUE DEL MODELO

El despliegue del modelo se realizó por medio de la librería Pickle, la cual permite un proceso por de una jerarquía de objetos Python los cuales se convierten en un flujo de bytes.

G. REFERENCIAS

Scikit Learn. (7 de Marzo de 2023). *scikit-learn.* Obtenido de scikit-learn: https://scikit-learn.org/stable/tutorial/machine_learning_map/index.html