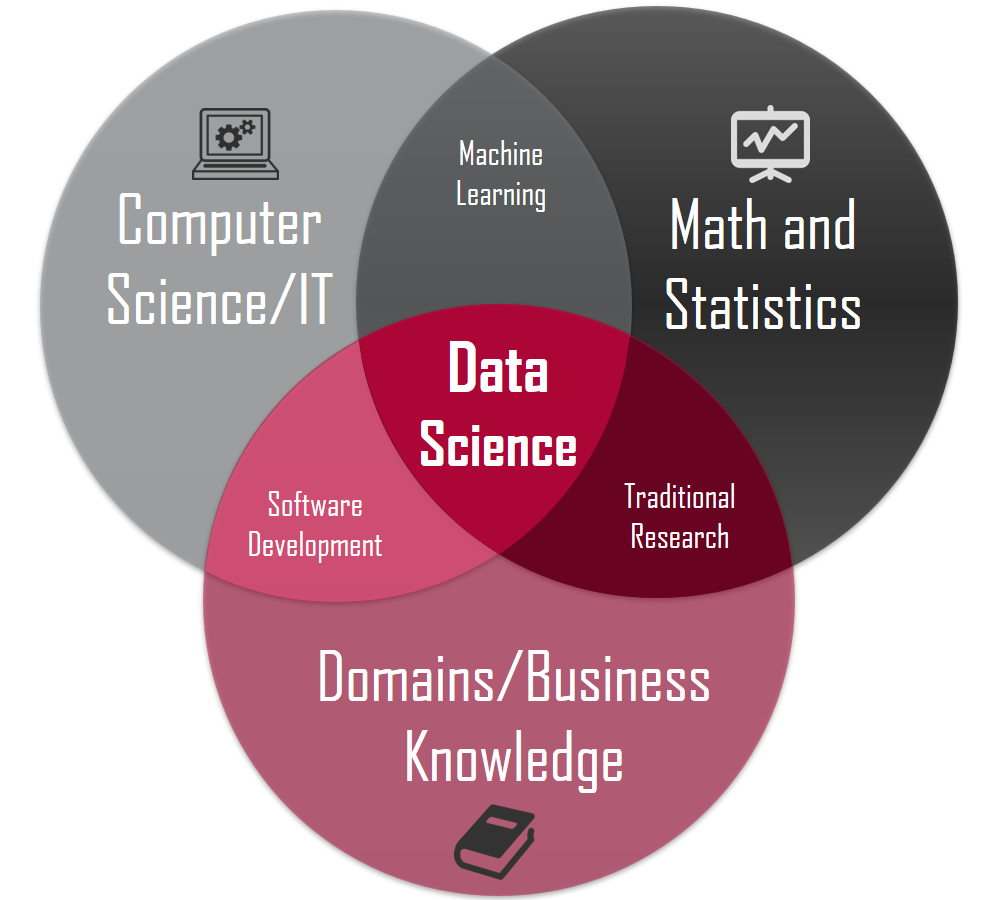
**Documentación**

**Proyecto Car Insurance Data**

**CURSO 19130 - Data Sciense**



02 JULIO

Creado por: Rodrigo Gonzalez

**Contenido**

Descripción del caso de Negocio...................................................................................3

Tabla de Versionado....................................................................................................3

Objetivos del modelo.................................................................................................3 Descripción de los datos .........................................................................................4 Hallazgos encontrados en el EDA .................................................................................5

Algoritmo Elegido.........................................................................................................8 Métricas de desempeño del modelo............................................................................9 Método de Ensamble...............................................................................................10

Método de HiperParametro........................................................................................11

Futuras Lineas.........................................................................................................11

Conclusiones..........................................................................................................11

Proyecto BI Sciense

**Descripción del caso de negocio**

El DS seleccionado trata de una compañía de seguros que quiere determinar en el último año si los clientes asegurados tuvieron un siniestro.

Mediante este indicador podrán determinar que costos tendrán a futuro, como ser más rentable y estimar la prima que deben cobrar.

**Tabla de Versionado**

|  |  |  |
| --- | --- | --- |
| Fecha | Versión | Referencia de cambios |
| 24/04/2022 | 1.0 | Se realizó la primer entrega donde realizamos DS+Analisis de los datos (Data Wrangling, Data Adquisicion y EDA). |
| 17/05/2022 | 2.0 | Se realizó la segunda entrega donde elegimos nuestra variable Target y seleccionamos algoritmo para generar modelo de Machine Learning. |
| 05/06/2022 | 3.0 | Se realizó tercera entrega donde incluimos readecuaciones detectadas en los últimos controles. |
| 02/07/2022 | 3.0 | Se realizó entrega final donde incluimos modelo de boosting y mejoras de los modelos. |

**Objetivos del modelo**

Al tener cada vez más clientes, quiere aprovechar los beneficios del “Machine Learning” para predecir si los clientes que ingresen nuevos, si generara un reclamo por un siniestro, teniendo en cuenta los datos históricos de la Base de Datos.

Como se trata de un problema de Clasificación, utilizaremos los algoritmos de este tipo para generar el modelo.

Nuestro DS proviene de nuestra organización (First Party) y es de tipo Estructura.

Además se pretende estimar que costos tendrá en el último año financiero.

**Descripción de los datos**

Los datos refieren a una compañía de seguros la cual quiere determinar en el último año si sus clientes asegurados realizaron siniestros. Los tipos de campos son Int, float y Object.

**Origen de DS Kaggle.com (Link** [**https://www.kaggle.com/datasets/sagnik1511/car-insurance-data**](https://www.kaggle.com/datasets/sagnik1511/car-insurance-data)**).**

**Variables**:

* ID
* RANGO-EDAD
* SEXO
* RACE
* EXPERIENCIA\_MANEJO
* EDUCACION
* INGRESO
* SCORE
* DUEÑO\_AUTO
* AÑO\_AUTO
* CASADO
* HIJOS
* POSTAL\_CODE
* KILOMETRAJE
* VEHICLE-TYPE
* FALLAS\_VELOCIDAD
* DUIS
* PAST\_ACCIDENTS
* OUTCOME

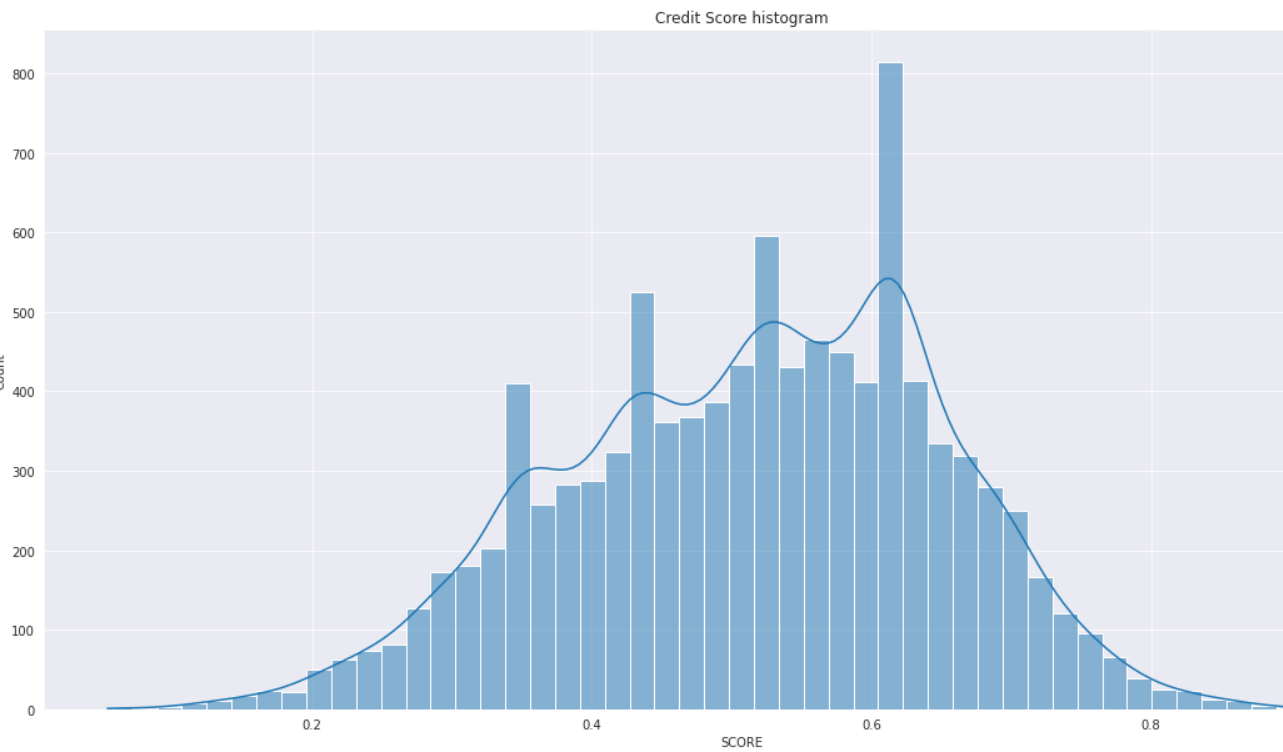
**Acciones realizadas al DS:**

* Identificamos 982 valores nulos en el Campo Score, para completarlos, tomamos en cuenta que tipo de ingreso tienen y calculamos el promedio de Score de cada tipo. El motivo, es que esta muy asociada al Score. Para completarlos, tomamos en cuenta que tipo de ingreso tienen y calcularemos el promedio de Score de cada tipo.
* La otra variable con nulos es el Kilometraje, se usara para completar el promedio.
* No eliminamos ninguna variable.

**Hallazgos encontrados en el EDA**

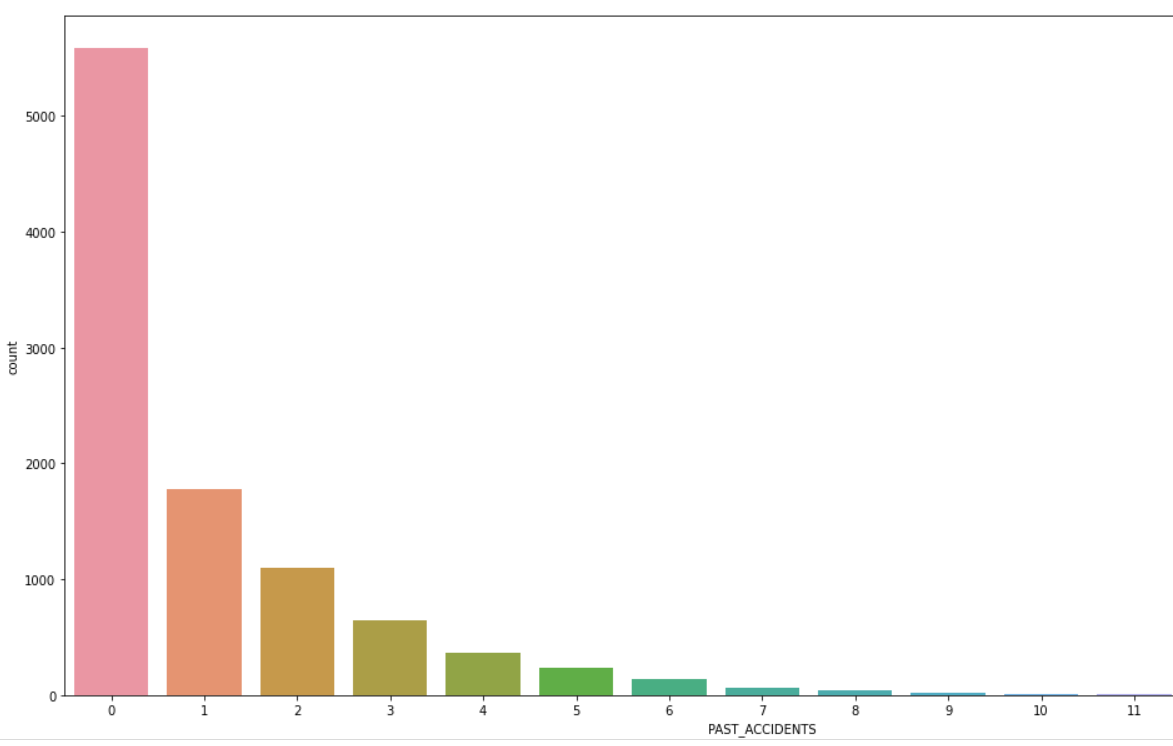
**Grafico 1:**

Identificamos que los valores obtenidos del Score se pueden validar como correctos, ya que el mayor % de clientes se agrupan en la mitad del gráfico.



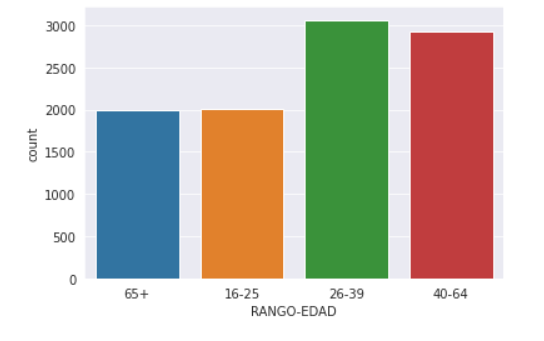
**Grafico 2:**

Podemos chequear que la mayoría de los clientes no tienen accidentes en su historial. Esto permitirá presumir que tiene alta probabilidad nuestra base que no usara nuestro seguro, generando menos costos.



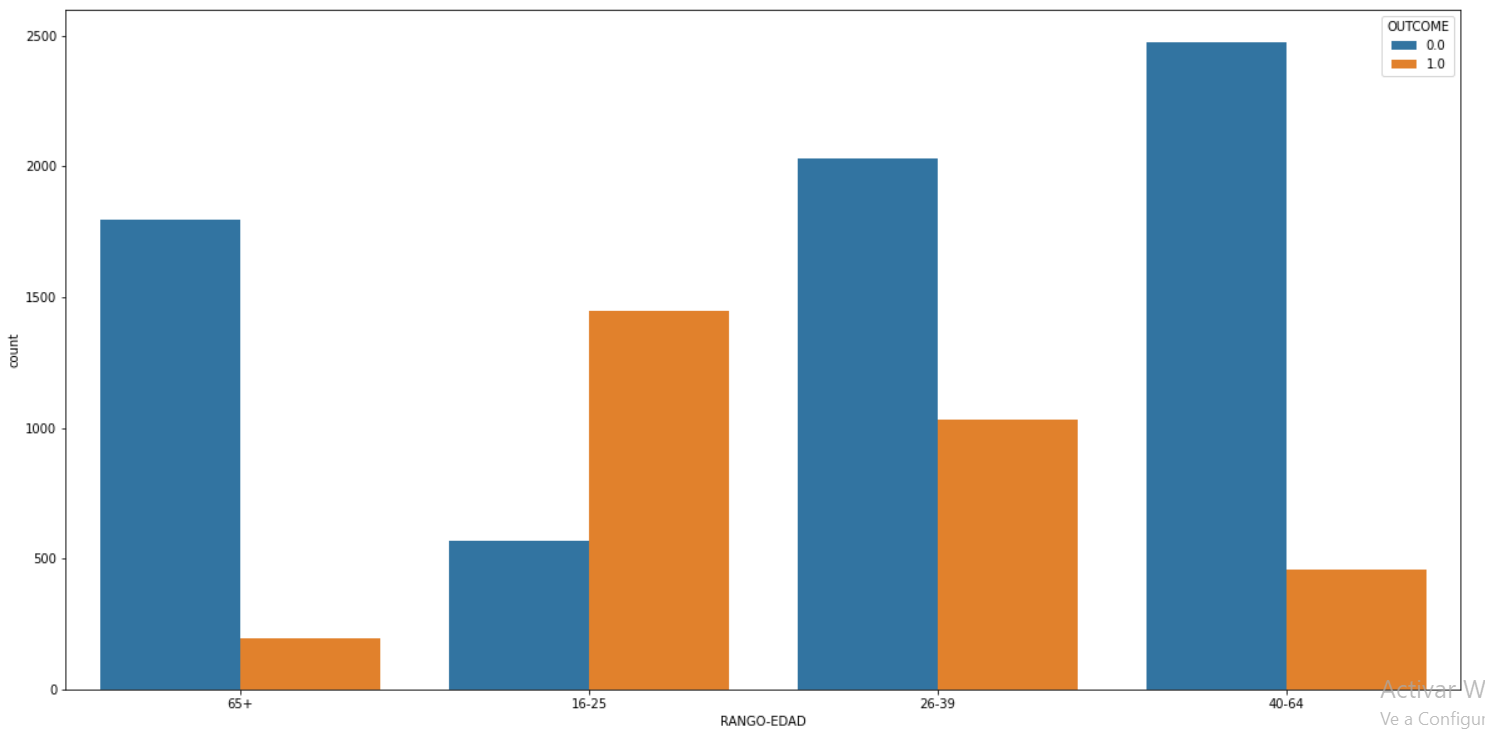
**Grafico 3:**

Grafico por distribución de edad, observamos que tenemos un gran % de mayores de 26 años en nuestro universo, tiene sentido ya que es una edad promedio de compra de vehículo.

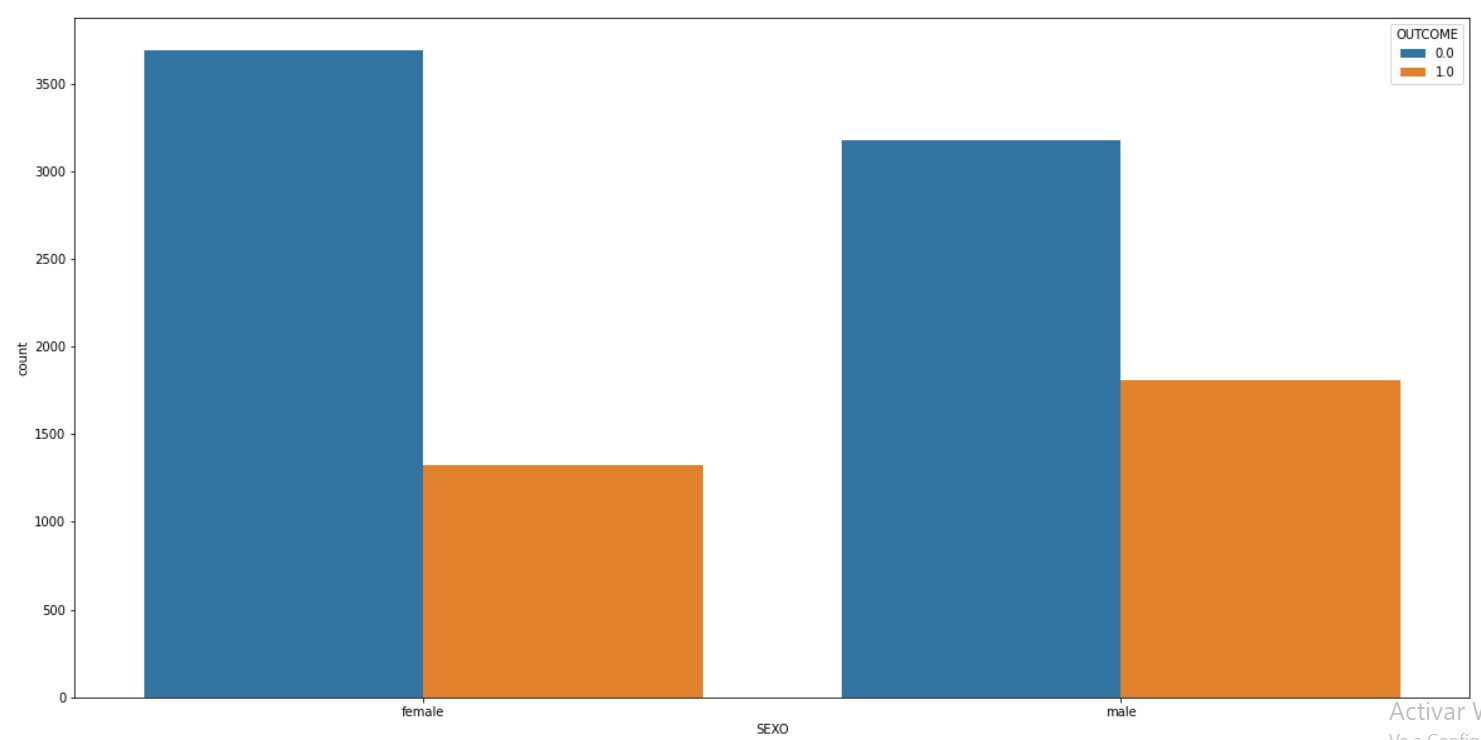


**Grafico 4:**

Chequeamos como se distribuye nuestra variable dependiente con respecto a la edad de los clientes. Comprobamos que a menor edad, más probabilidad de realizar un reclamo en nuestra compañía.

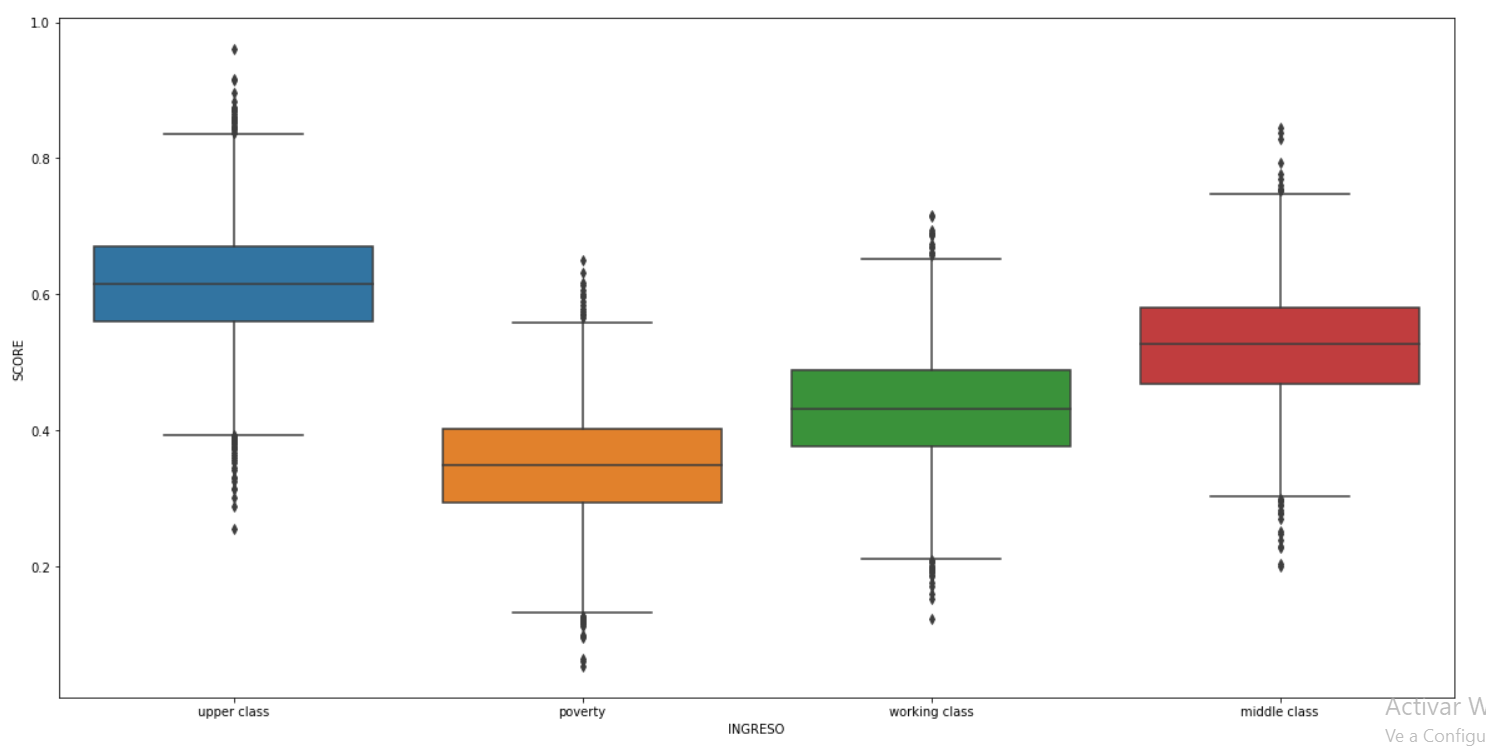
**Grafico 5:**

Chequeamos como se distribuye nuestra variable dependiente con respecto a la edad de los clientes. Comprobamos que a menor edad, más probabilidad de realizar un reclamo en nuestra compañía.



**Grafico 6:**

En este grafico demostramos que cuanto más ingreso tiene el cliente, más probabilidades de tener un Score alto.



**Algoritmo Elegido**

**Arboles de decisión:**

Utilizamos el modelo con dos alturas máximas diferentes (3 y 5) dos debido al tamaño del DS.

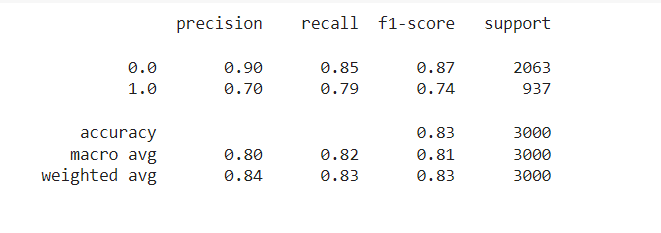
**Random Forest:**

* Número de árboles 100.

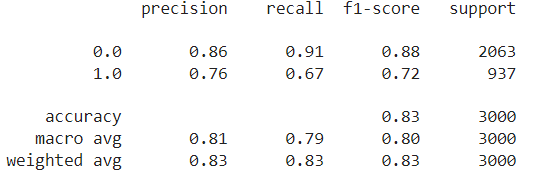
En ambos algoritmos, utilizamos 30% del DS para test y 70% para train.

**Métricas de desempeño del modelo**

**Arboles de decisión- Primer modelo:**



**Arboles de decisión- Segundo modelo:**

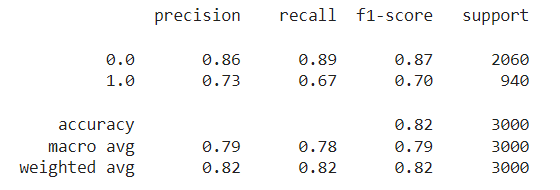


Elegimos la métrica de evaluación de Accuracy ya que investigue que es una buena medida cuando las clases de variables de destino en los datos están casi equilibradas.

Chequeamos el primer árbol y comprobamos que tenemos un 70% de acierto cuando la variable de salida es 1 (Cliente que generaría un siniestro) y la salida 0, un 90% de acierto.

Teniendo en cuenta el 2 arbol mejoramos los % de acierto de la variable de salida 1, generando un mejor modelo.

**Random Forest:**



Tenemos un % de acierto casi idéntico que el calculado en el anterior algoritmo con el arbol de decisión con 100 árboles (82%).

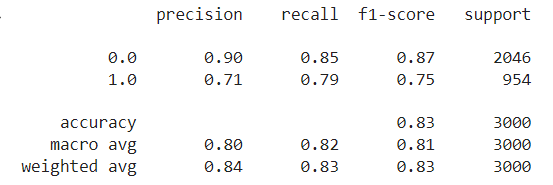
Si achicamos a 50 árboles nos da un % de 81%.

**Conclusiones de ambos modelos:**

No encontramos diferencias significativas entre el modelo de Arboles de Decisión y Random Forest en el DS elegido, ambos no dan % elevados de acierto con la variable de salida 1 (Si tendrán un siniestro).

**Metodo de Ensamble**

Utilizamos el algoritmo "XG Boost" para implementar optimización en nuestro modelo, dando como resultado un % mayor de eficacia pero sigue muy bajo la predicción de la variable de salida 1 (SI generara un siniestro).



**Método de Hiperparametro**

Utilizamos el metodo "Halving Grid Search CV" para implementar optimización en nuestro modelo, dando como resultado un % mayor de eficacia.



**Futuras Líneas**

Considero que para mejorar los resultados obtenidos deberíamos tener más cantidad de registros para hacer más robusto el DS. Por otro lado es necesario poder implementar otros algoritmos para determinar si los elegidos fueron los correctos.

**Conclusiones**

Los resultados obtenidos con la variable dependiente de casos con reclamo, no fue el % esperado. De igual manera con los modelos de Boosting y HiperParametro se mejoran los resultados.