JUGADOR 1

PUNTUACIÓN MÁS ALTA 2500

JUGADOR 2

DETECCION DE FRAUDES EN TARJETAS DE CREDITO

START MENU SIGN IN

AÑADE UNA BREVE DESCRIPCIÓN

MENU 01 07 12

PROBLEMA

- TODOS LOS DIAS SE LLEVAN A ACABO MILLONES DE TRANSACCIONES POR MEDIO DE TARJETAS DE CREDITO. LA GRAN MAYORIA SON LLEVADAS A ACABO POR EL TITULAR DE ESTA. SIN EMBARGO, EXISTEN CASOS DE ROBO DE IDENTIDAD Y CARGOS NO RECONOCIDOS QUE ES IMPORTANTE IDENTIFICAR

VUELVE A LA PÁGINA AGENDA

SIGN IN

BACK TO AGENDA PAGE

EL SET DE DATOS VIENE DE KAGGLE CREDIT CARD FRAUD DETECTION (HTTPS://WWW.KAGGLE.COM/DATASETS/MLG-ULB/CREDITCARDFRAUD)

	Time	V1	V2	V3	V4	V5	V6	V7	V8	V9	...	V21	V22	V23	V24	V2
0	0.0	-1.359807	-0.072781	2.536347	1.378155	-0.338321	0.462388	0.239599	0.098698	0.363787	...	-0.018307	0.277838	-0.110474	0.066928	0.12853
1	0.0	1.191857	0.266151	0.166480	0.448154	0.060018	-0.082361	-0.078803	0.085102	-0.255425	...	-0.225775	-0.638672	0.101288	-0.339846	0.16717
2	1.0	-1.358354	-1.340163	1.773209	0.379780	-0.503198	1.800499	0.791461	0.247676	-1.514654	...	0.247998	0.771679	0.909412	-0.689281	-0.32764
3	1.0	-0.966272	-0.185226	1.792993	-0.863291	-0.010309	1.247203	0.237609	0.377436	-1.387024	...	-0.108300	0.005274	-0.190321	-1.175575	0.64737
4	2.0	-1.158233	0.877737	1.548718	0.403034	-0.407193	0.095921	0.592941	-0.270533	0.817739	...	-0.009431	0.798278	-0.137458	0.141267	-0.20601

5 rows × 31 columns

- **28 variables Anonimas**
- **Time**
- **Amount**
- **284807 registros**

OBJETIVOS

Se utilizara el set de datos de kaggle para poder desarrollar un modelo de clasificacion, y de esta manera montar una aplicacion que nos permita predecir si una transaccion es Fraudulenta o no

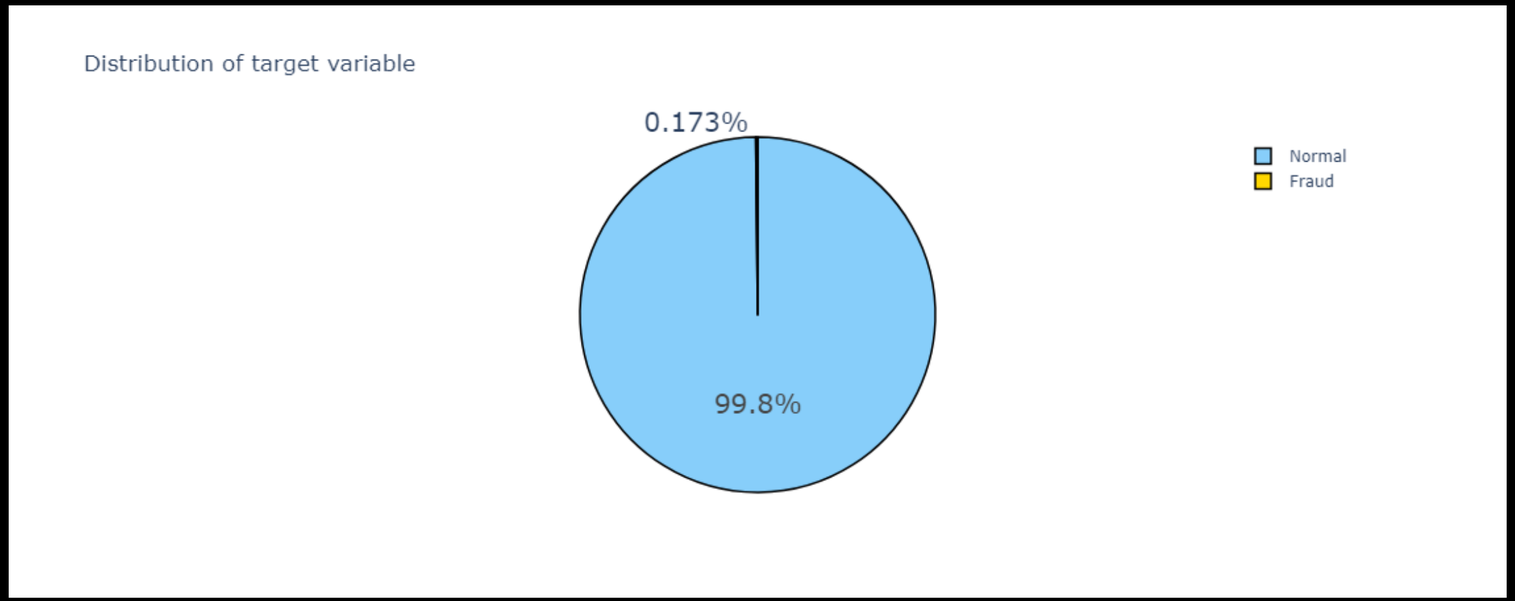
SIGN IN

VUELVE A LA PÁGINA AGENDA

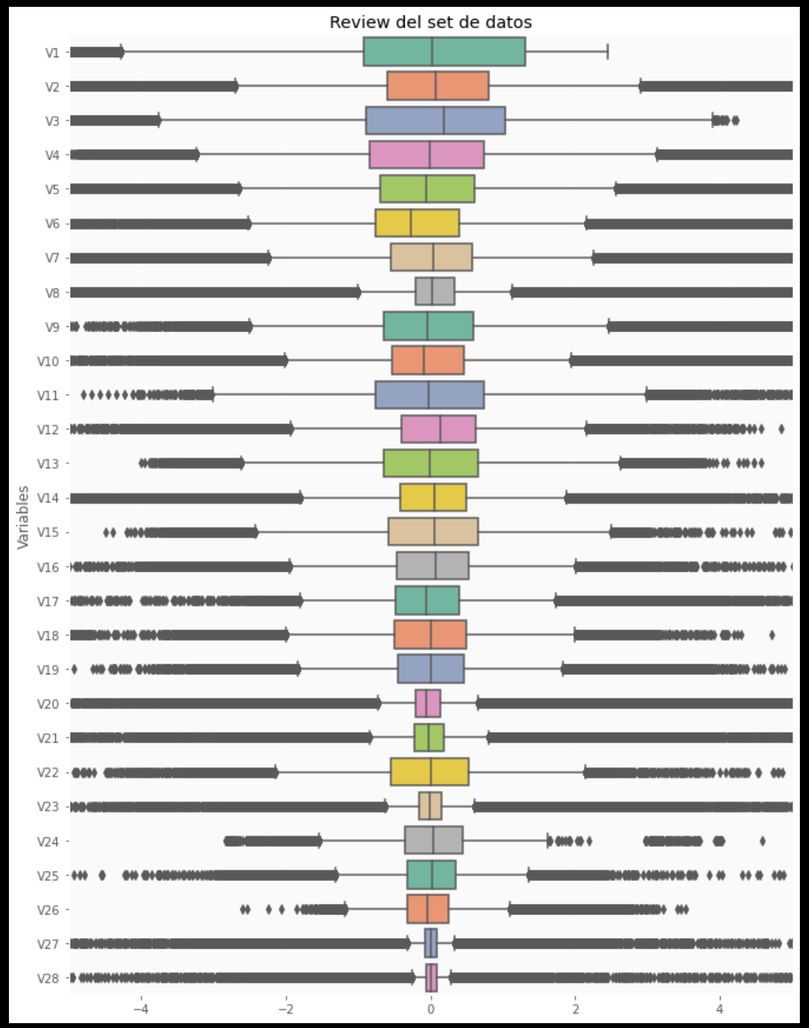
ANALISIS EXPLORATORIO

EDA Y ANALISIS

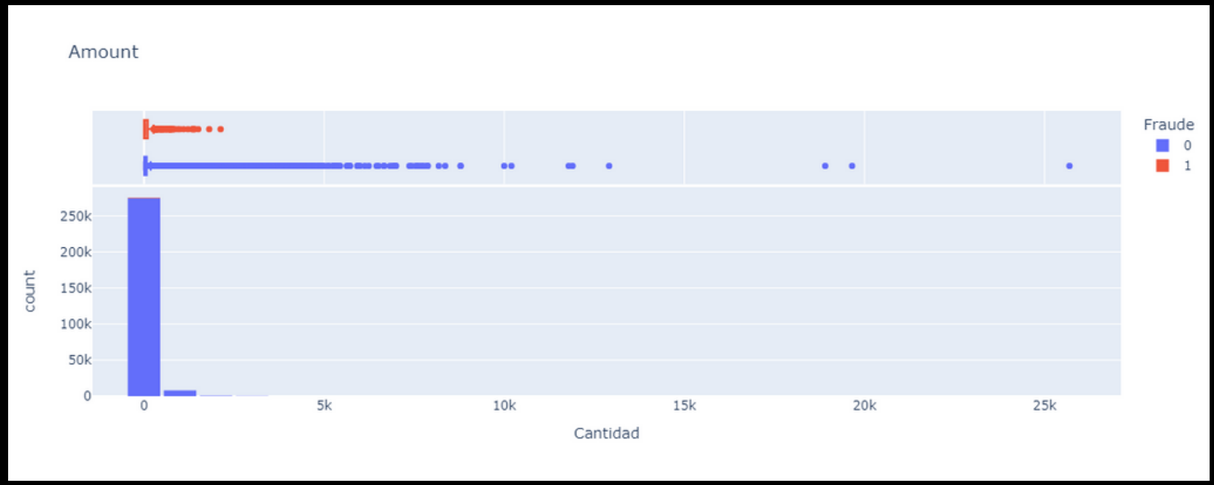
Dataset sumamente desbalanceado



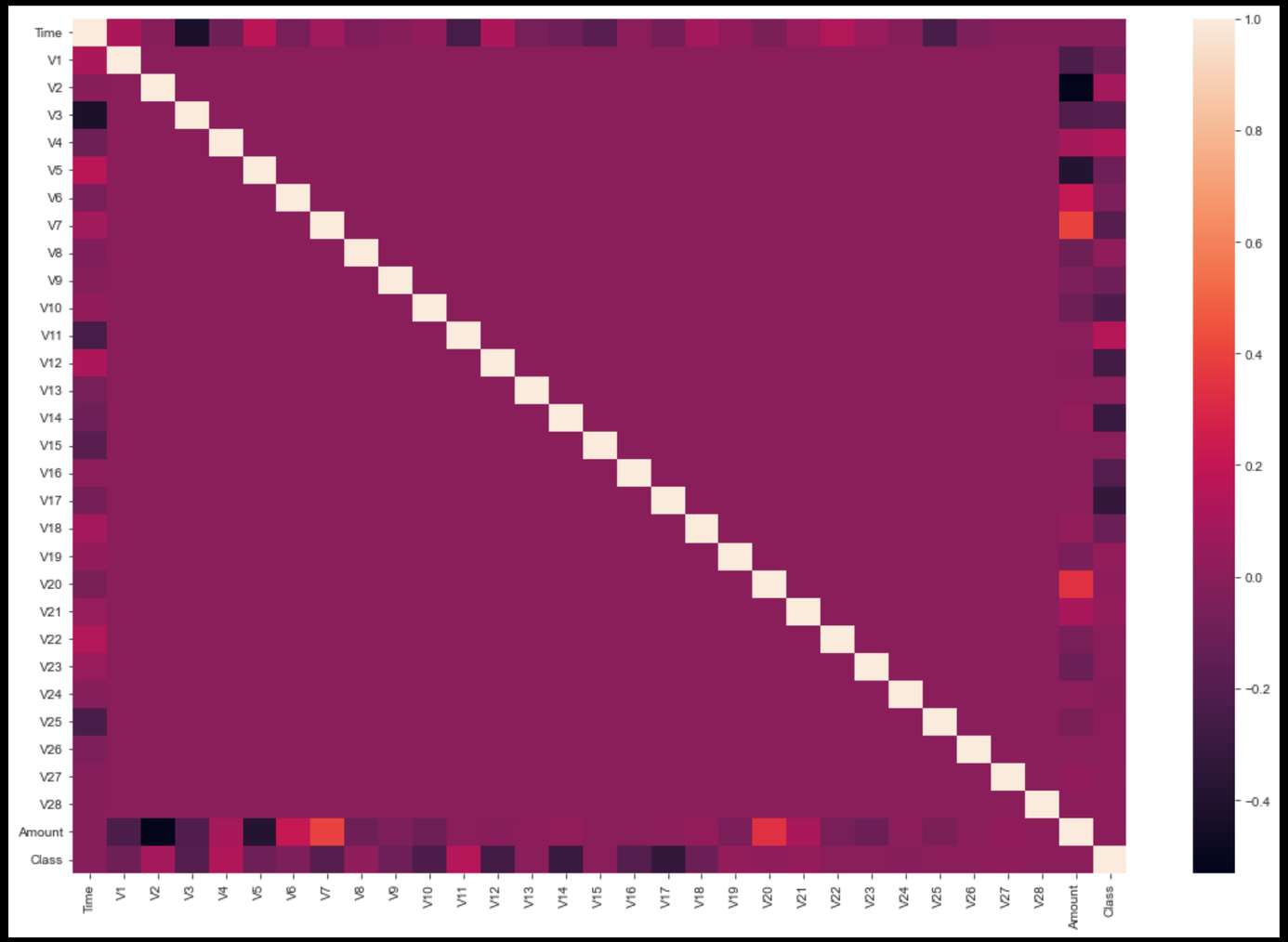
Variables Anonimas e importe variables con muchos outliers



Las variables anonimas y el importe se encuentran distribuidos con muchos outliers en especial la variable importe o cantidad de la transaccion



Correlacion de variables



Del mapa de calor, se puede observar que no hay fuertes correlaciones positivas o negativas entre ningún par de variables. De igual forma la variable tiempo no es significativa para el modelo

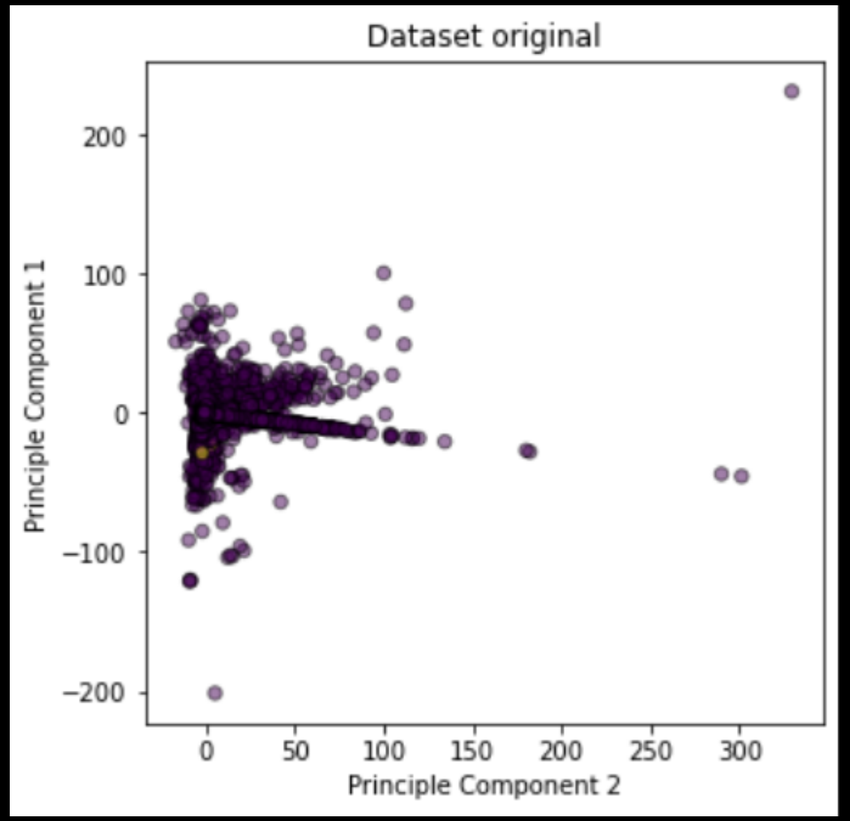
SIGN IN

VUELVE A LA PÁGINA AGENDA

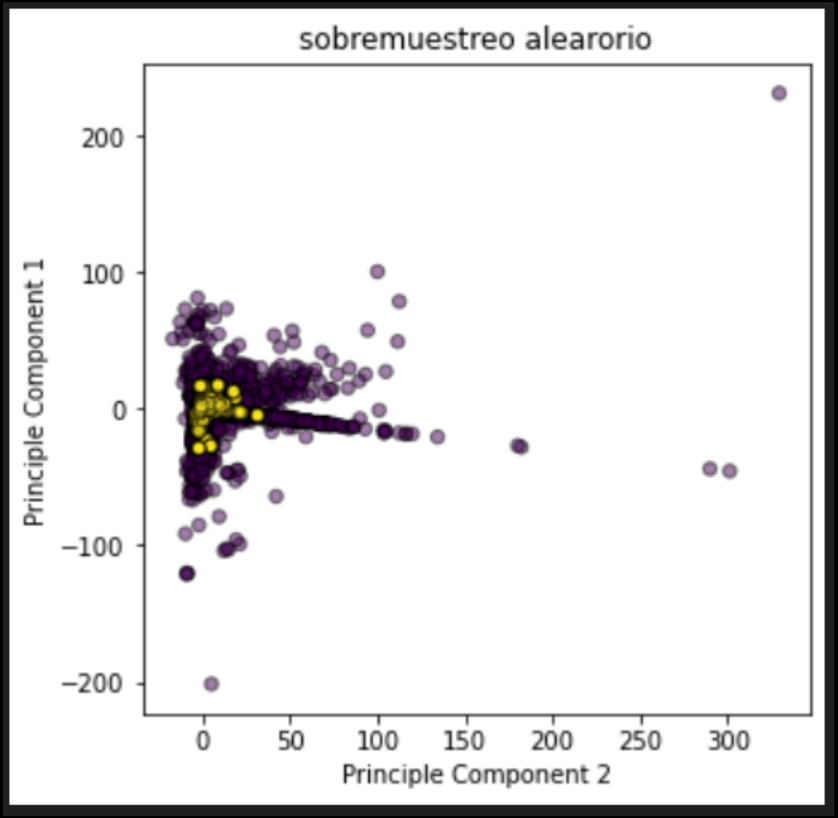
MODELADO Y EVALUACION

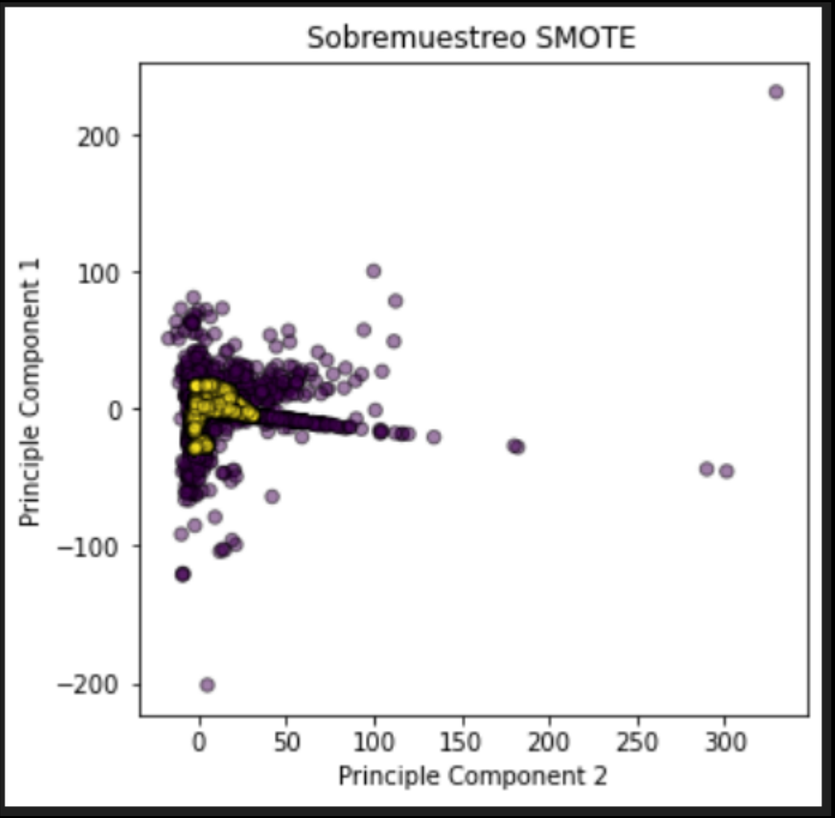
EL PROBLEMA DEL DESBALANCE

Tecnicas para lidiar con los dataset desbalanceados

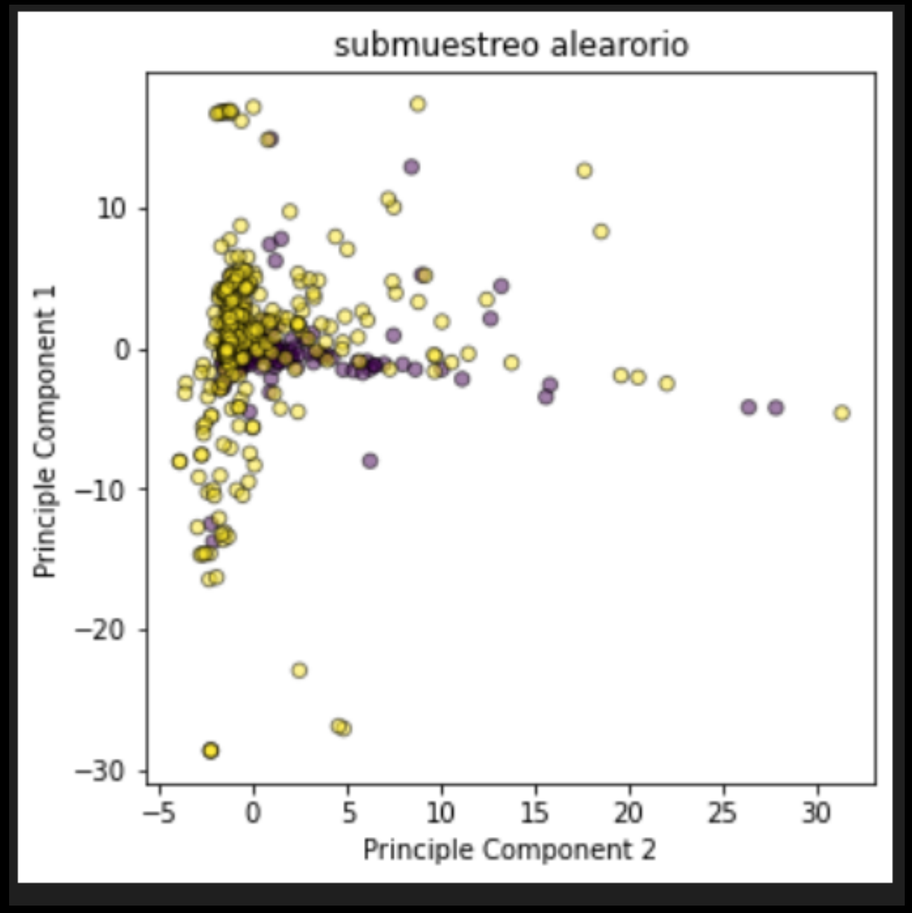


Tecnias de Sobremuestreo





Tecnica de Submuestreo

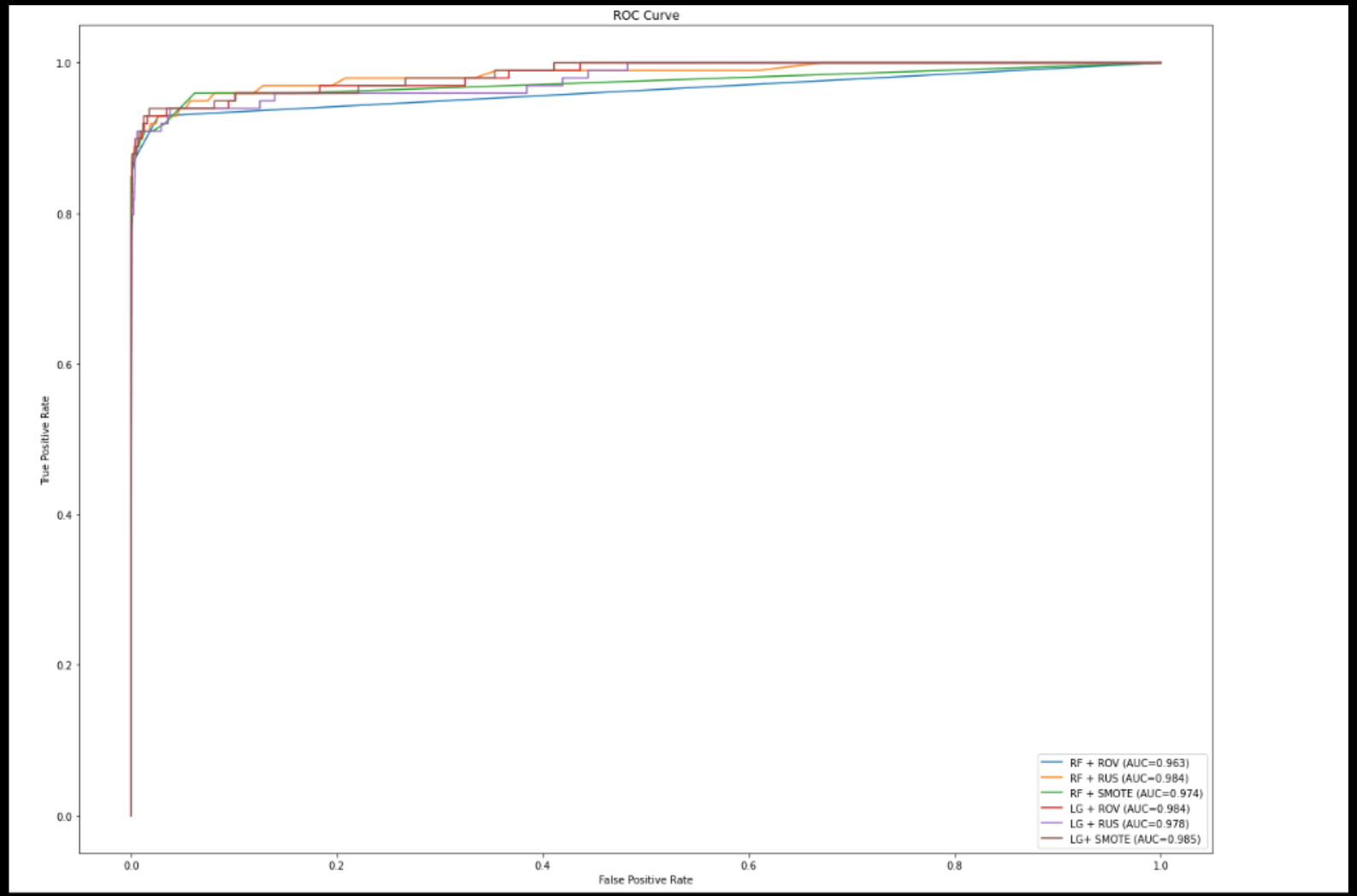


Comparacion de Modelos

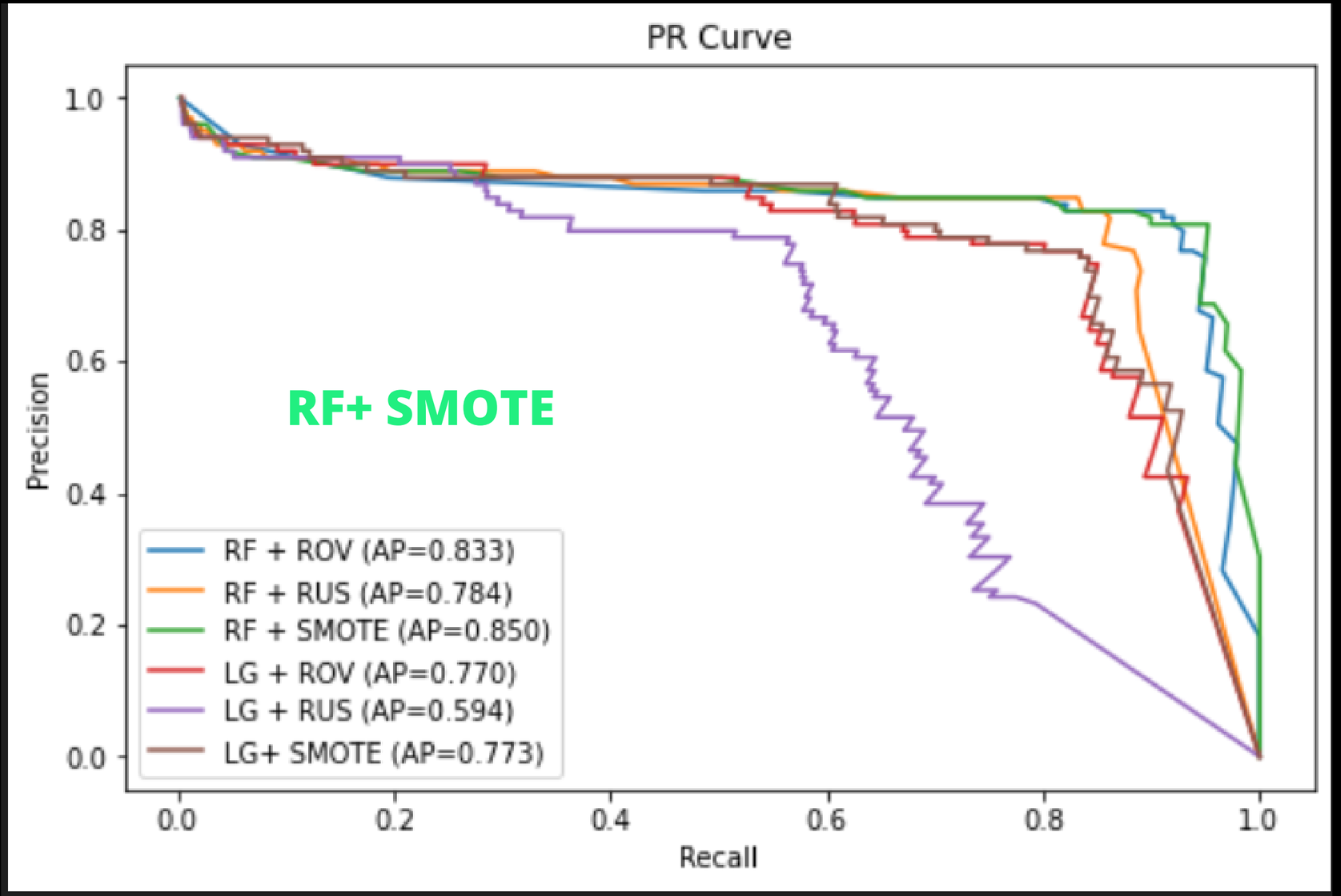
Se evaluaran los modelos con la adicion de las tecnicas de balanceo de dataset previamente mencionadas. Las confifuraciones seran las siguientes:

- **RF+ROV (Random Forest con sobremuestreo aleatorio)**
- **RF+RUS (Random Forest con submuestreo aleatorio)**
- **RF+SMOTE (Random Forest con metodo SMOTE)**
- **LR+ROV (Regresion Logistica con sobremuestreo aleatorio)**
- **LR+RUS (Regresion Logistica con submuestreo aleatorio)**
- **LR+SMOTE (Regresion Logistica con metodo SMOTE)**

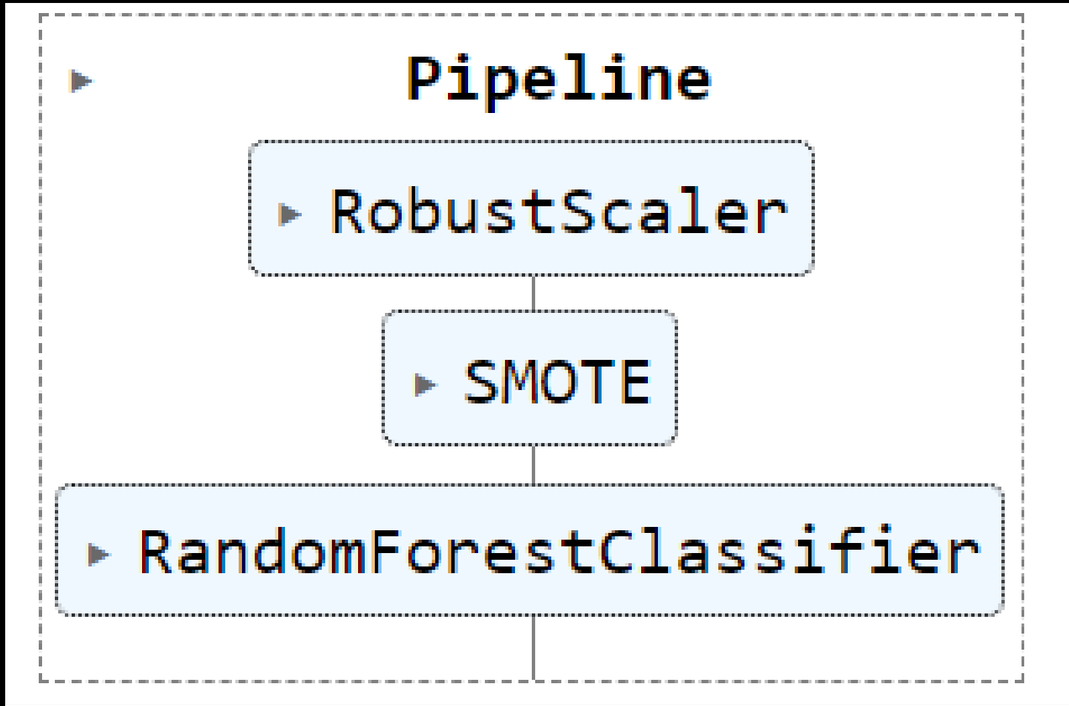
ROC Curve


ROC Curve
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
RF + ROV (AUC=0.963)
RF + RUS (AUC=0.984)
RF + SMOTE (AUC=0.974)
LG + ROV (AUC=0.984)
LG + RUS (AUC=0.978)
LG+ SMOTE (AUC=0.985)

PR Curve


PR Curve
Precision
Recall
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
RF+ SMOTE
RF + ROV (AP=0.833)
RF + RUS (AP=0.784)
RF + SMOTE (AP=0.850)
LG + ROV (AP=0.770)
LG + RUS (AP=0.594)
LG+ SMOTE (AP=0.773)

Modelo escogido



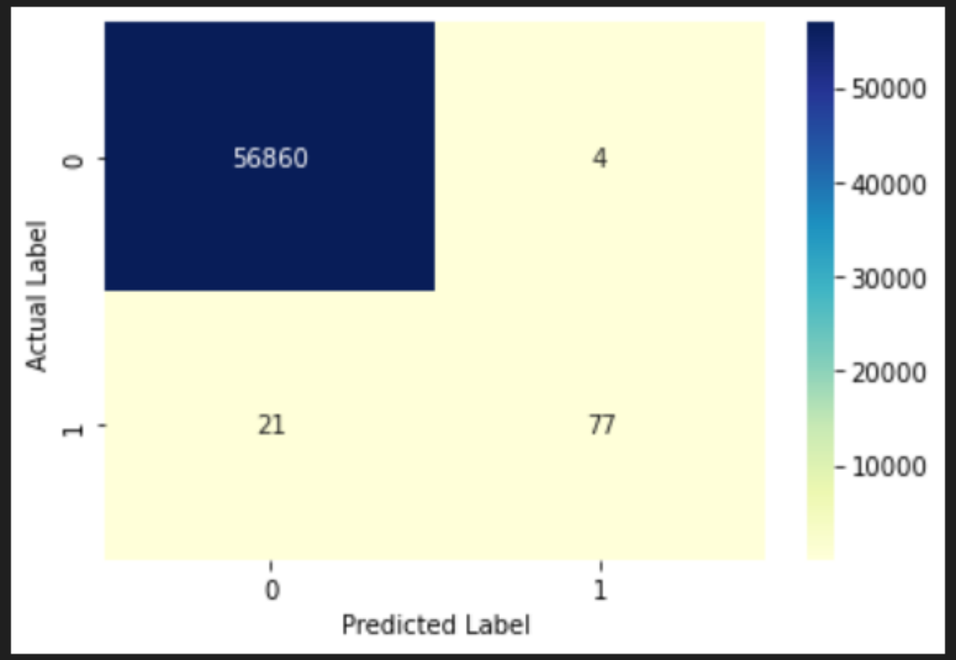
Modelo Elegido Despues del GridSearch:

- **Robustscaler para lidiar con los outliers**
- **SMOTE para balancear el dataser**
- **Random Forest con los siguientes parametros {'max_depth':16, 'n_estimators': 150}**

Resultados

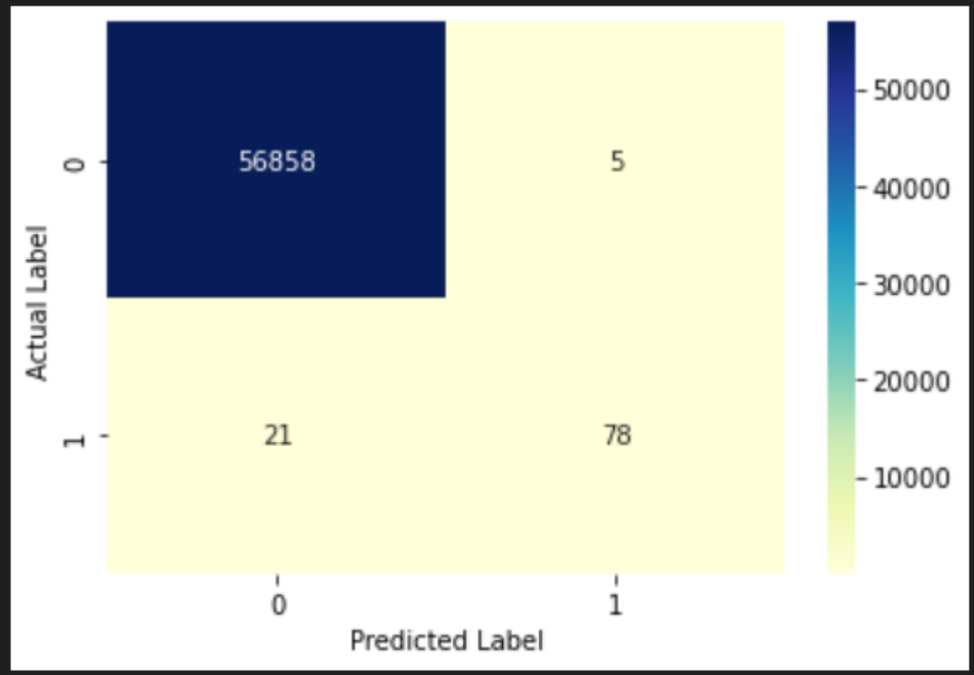
Datos de prueba

Recall = 78.57% AP = 74.73% Precision = 95.06%



Datos de validacion

Recall = 78.79% AP = 74.08% Precision = 93.98%



Imortancia de las Variables predictoras

	feature	importance
16	V17	0.169251
11	V12	0.142064
13	V14	0.116059
9	V10	0.081721
15	V16	0.078381
10	V11	0.065602
8	V9	0.034805
17	V18	0.033838
6	V7	0.027143
3	V4	0.025921

SIGN IN

VUELVE A LA PÁGINA AGENDA

DESPLIEGUE

- **Frontend: esta creado con Streamlite**
- **Backend: esta creado con FastApi**

Conclusiones

- **Las tecnicas de balanceo de dataser mostraron ser una buena herramienta para la creacion de un clasificador en un data set desbalanceado**
- **El Robust scaler es una buena tecnica par amanejar los outliers. Sin embargo una alternativa seria cambiar la distribuicon de los datos o quitar los outliers**
- **Las variables Anonimas son las que tienen mas peso en el modelo, sin embargo estas no son accesibles a saber la informacion que tenian en un inicio**

- En un inicio se creyo que el importe o la cantidad de la transaccion nos podia indicar su naturaleza. Sin embargo, parece ser que esta variable no guarda alguna relacion alguna.
- El sobremuestreo por lo general mostro mejores resultados y es recomendable su uso en dataset pequeños ya que no nos deshacemos de datos. Sin embargo, tecnicas como SMOTE que generan datos sinteticos nos llevan a entrenar nuestro modelo con datos Falsos o artificiales.
- Por el contrario para set de datos grandes donde quitar datos sea una opcion el submuestreo es una buena opcion