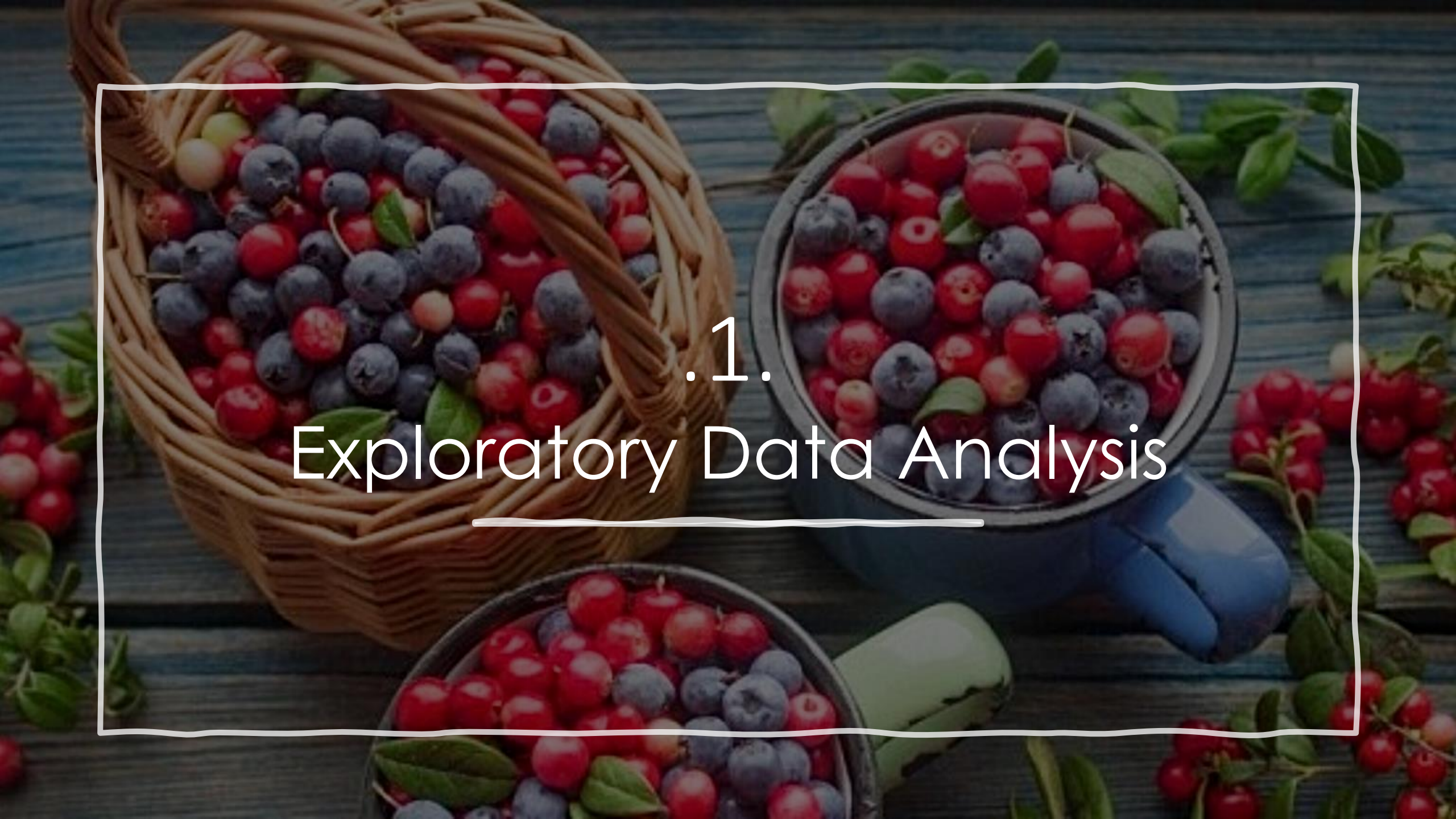
Datamecum

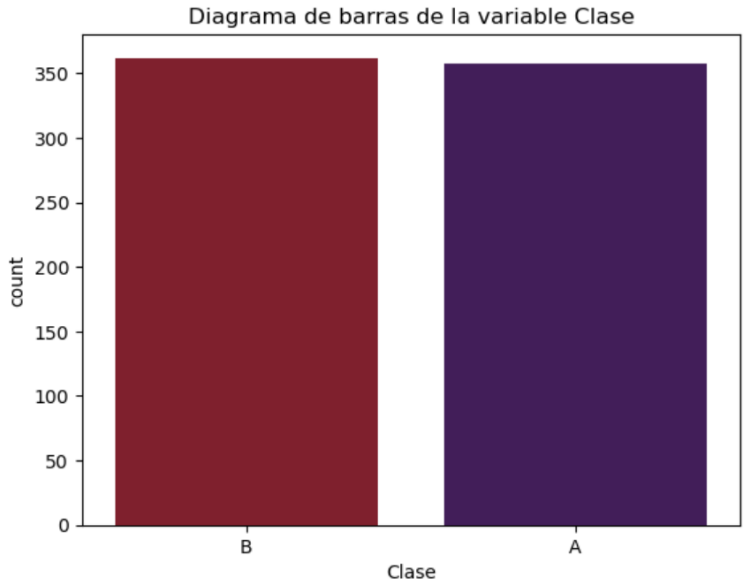
Datathon 2023

Pilar Madariaga



.1.
Exploratory Data Analysis

Raw Data I

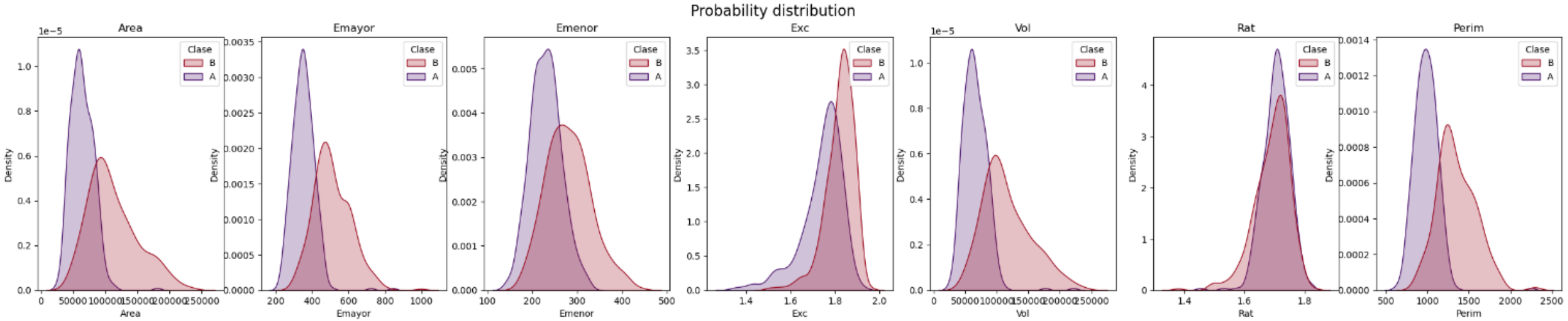


Variable Respuesta : Clase de arándano

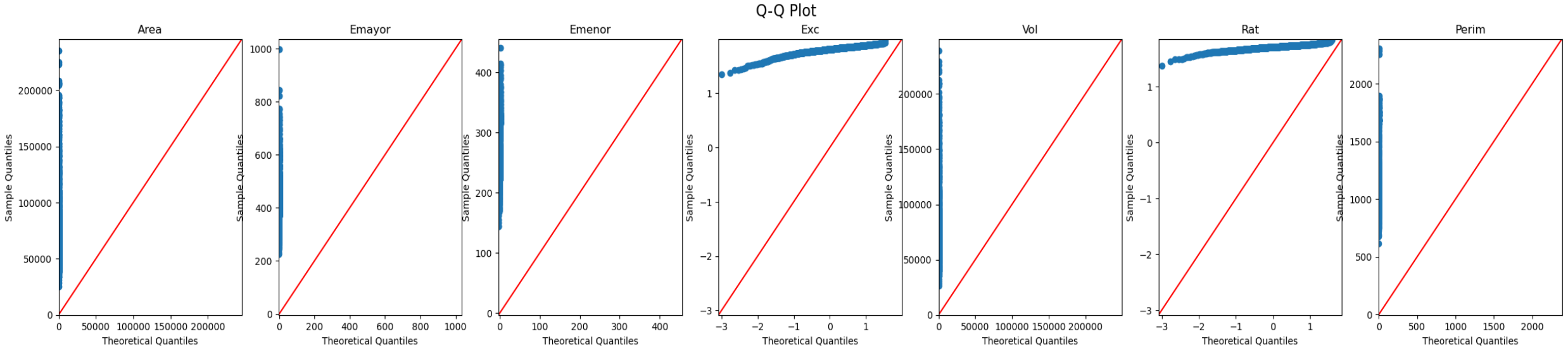
	count	unique	top	freq	mean	std	min	25%	50%	75%	max
Area	681.0	NaN	NaN	NaN	88262.286344	38589.31613	25387.0	60199.0	79735.0	105053.0	235047.0
Emayor	686.0	NaN	NaN	NaN	431.197846	115.843789	225.629541	345.601388	409.21384	496.804625	997.291941
Emenor	683.0	NaN	NaN	NaN	255.521046	49.624764	143.710872	219.586405	248.606869	286.957802	440.497127
Exc	676.0	NaN	NaN	NaN	1.781422	0.090054	1.34873	1.744768	1.798546	1.841674	1.962124
Vol	677.0	NaN	NaN	NaN	91020.830133	39822.725399	26139.0	61496.0	82555.0	108296.0	239093.0
Rat	680.0	NaN	NaN	NaN	1.698053	0.05306	1.379856	1.668171	1.704924	1.733535	1.830632
Perim	673.0	NaN	NaN	NaN	1167.043938	267.437031	619.074	970.754	1129.072	1313.092	2303.69
Clase	720	2	B	362	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Variables Predictoras: Area, Eje Mayor, Eje menor, Excentricidad, Volumen Ratio, Perímetro

Raw Data II


Probability distribution
Area
Emayor
Emenor
Exc
Vol
Rat
Perim
Clase
B
A
Density
1e−5

Distribución No Normal

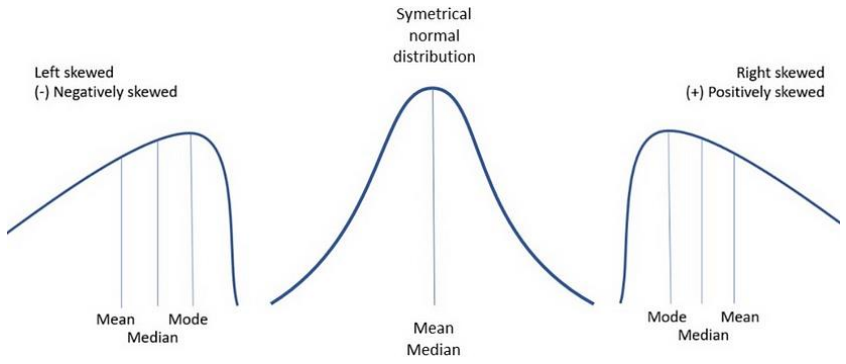


Missing Values

	Total	%
Area	39	5.416667
Emayor	34	4.722222
Emenor	37	5.138889
Exc	44	6.111111
Vol	43	5.972222
Rat	40	5.555556
Perim	47	6.527778
Clase	0	0.000000

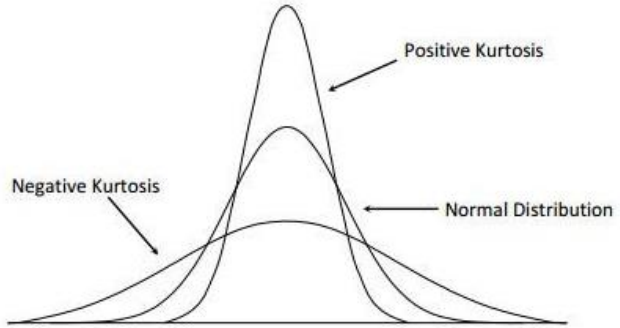
- Se observa un porcentaje pequeño de valores nulos distribuidos de manera no uniforme entre las diferentes columnas del dataset.
- Estrategia de tratamiento de valores faltantes: **Eliminación**

Skew and Kurtosis



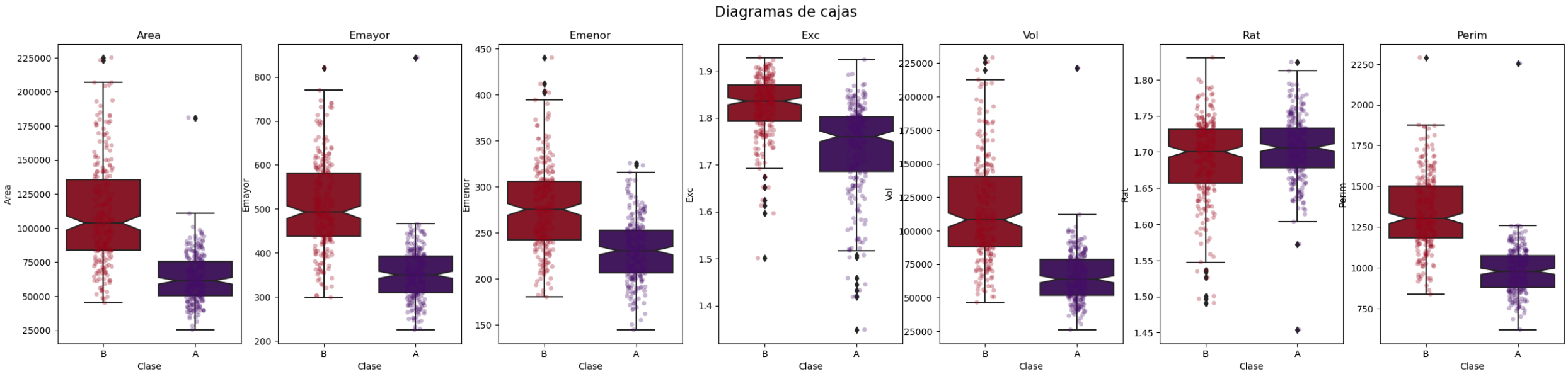
Skweness es una medida que determina la **asimetría** de las colas.

	skew	kurtosis
Area	1.157159	1.059180
Emayor	0.870055	0.789054
Emenor	0.652404	0.423626
Exc	-1.390112	2.841782
Vol	1.127781	0.979255
Rat	-1.101462	3.244730
Perim	0.812233	0.693695

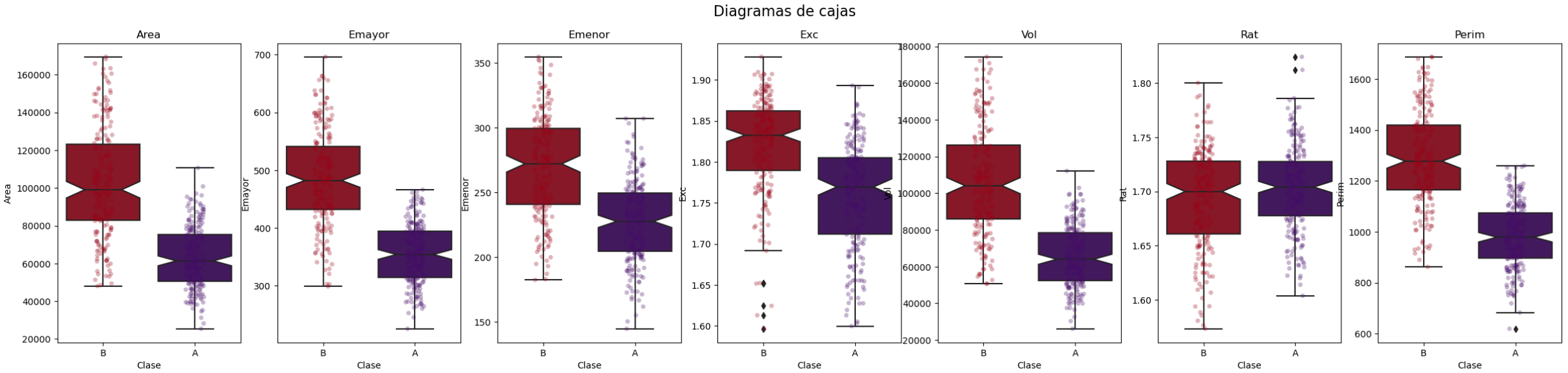


Curtosis es una medida que determina la forma de la distribución, observando el pico y la pesadez de las colas (***tailedness***).

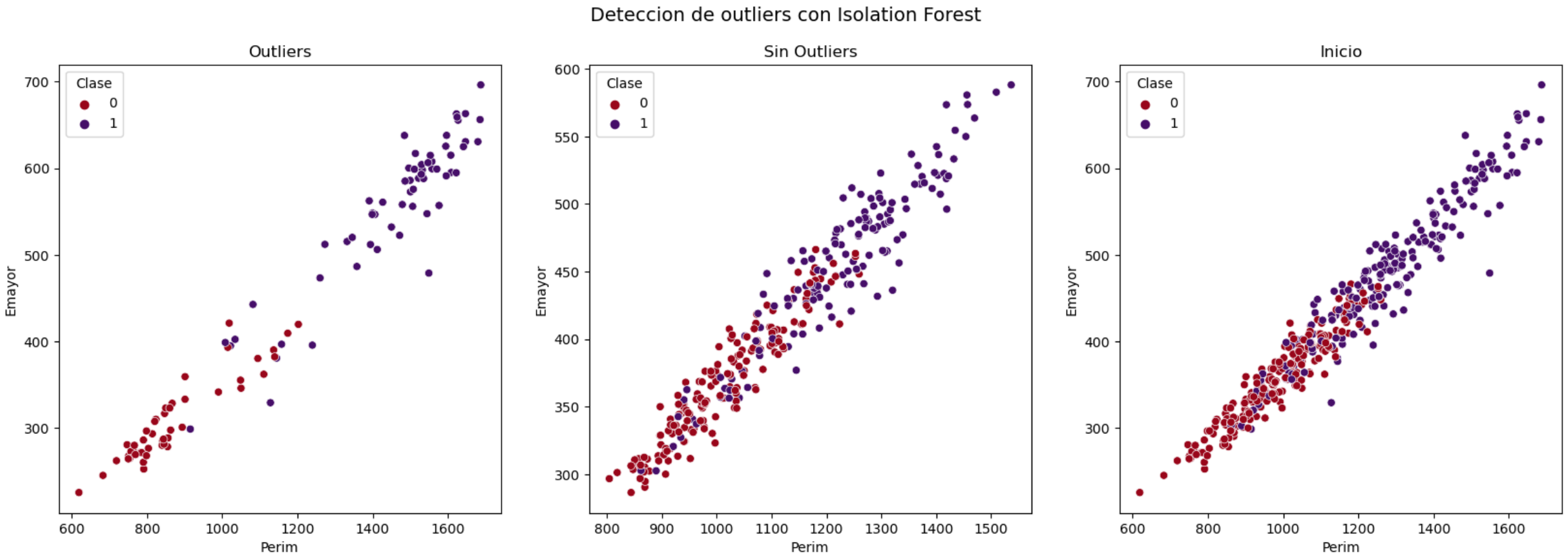
Outliers: IQR Method



Without Outliers I

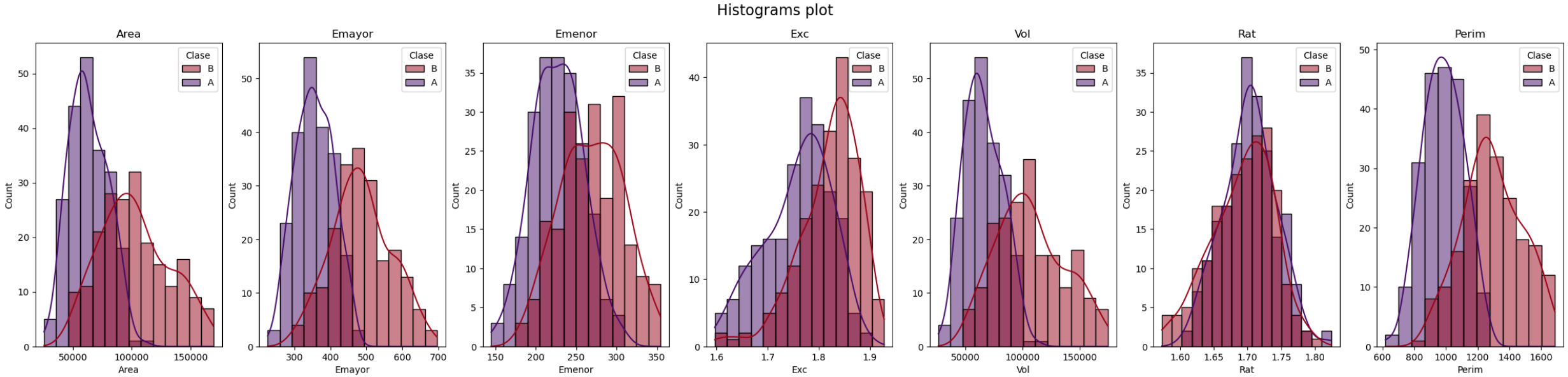


Anomalies: Isolation Forest

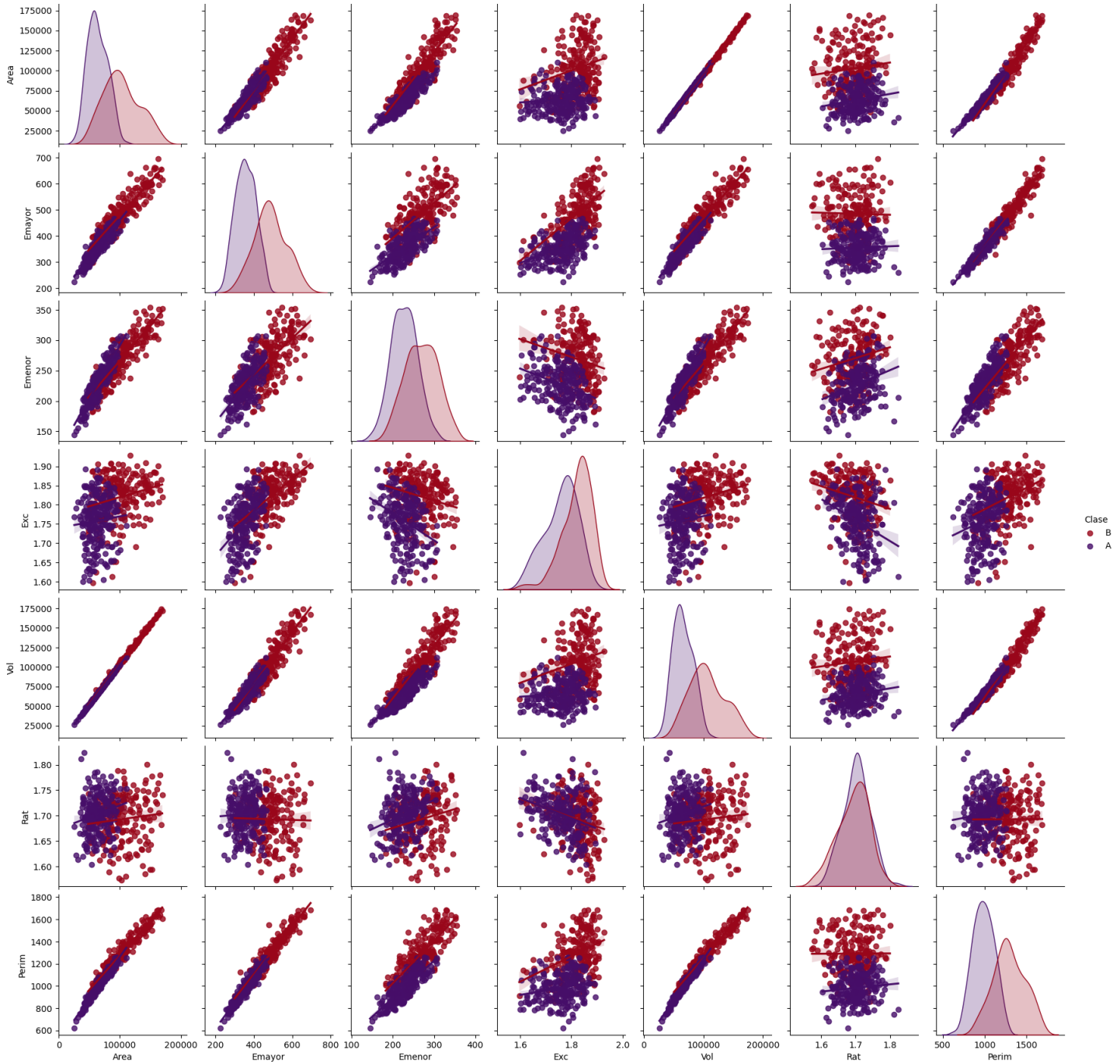


After cleaning

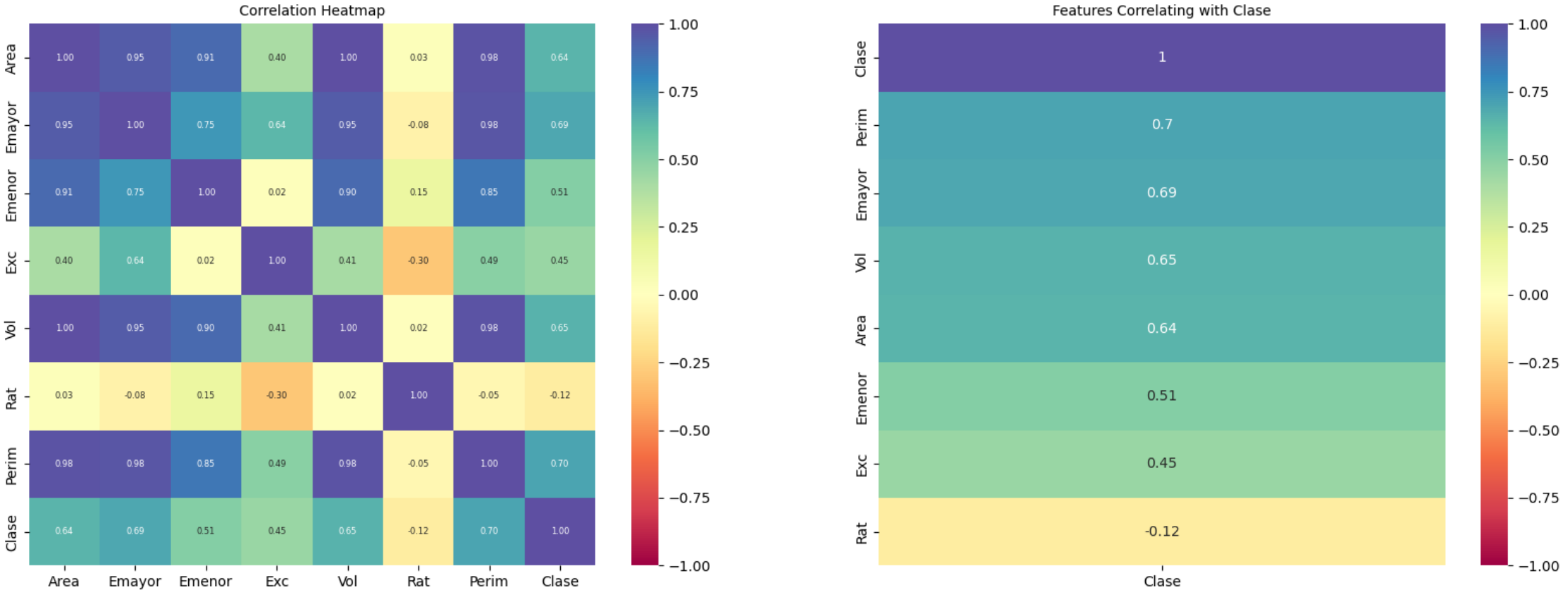
	skew	kurtosis
Area	0.769034	0.009354
Emayor	0.539407	-0.293369
Emenor	0.251378	-0.392472
Exc	-0.603576	-0.154965
Vol	0.754136	-0.035302
Rat	-0.263829	-0.011196
Perim	0.428334	-0.456457



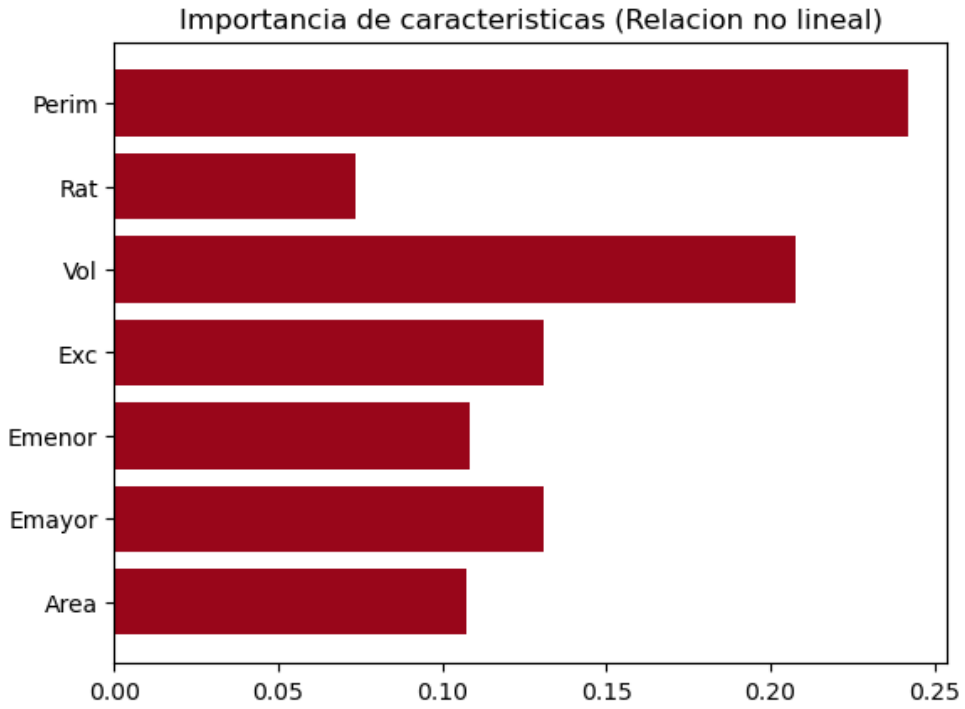
Multicollinearity



Correlation Matrix



Feature Importance



- Análisis de la importancia de las variables predictoras.
- Se decide no descartar ninguna variable predictora.

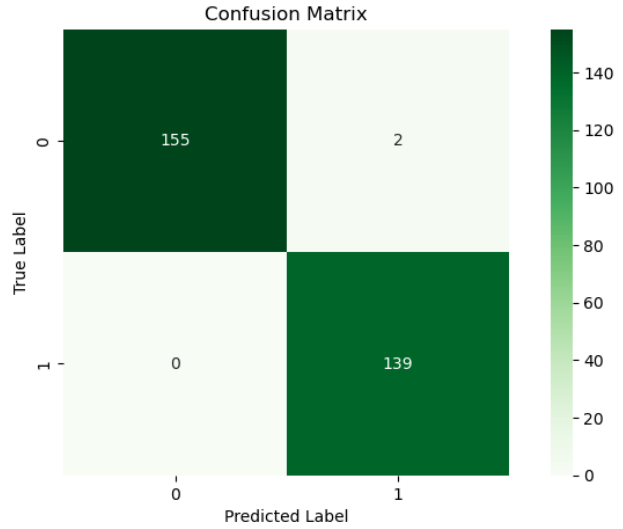
Datamecum

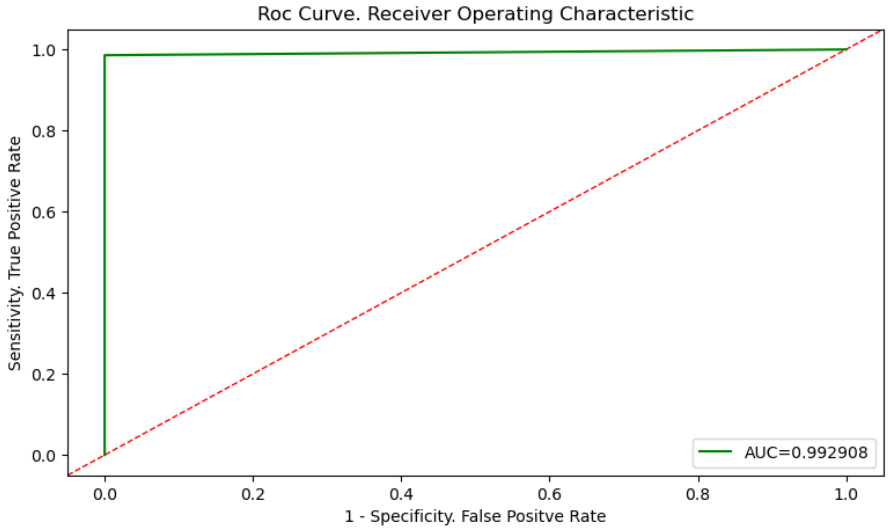
.2.
Fit and Evaluation

XGBClassifier

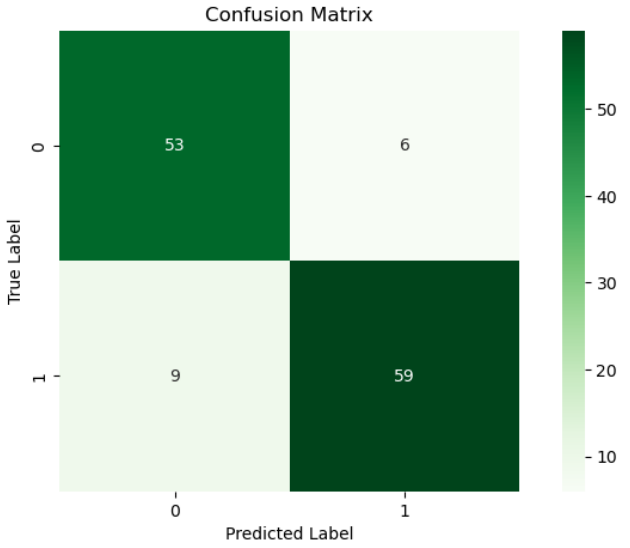
AUC: 0.88 F1:88

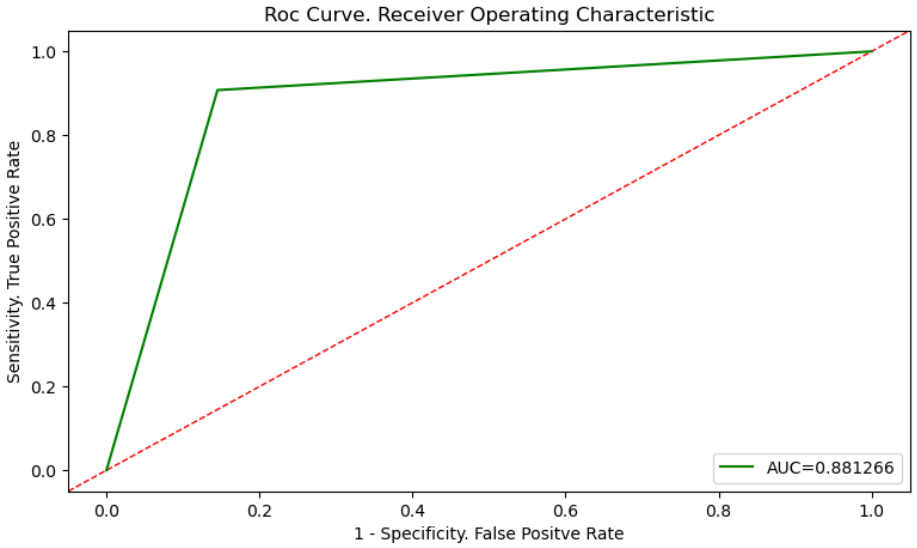
XGBClassifier with estimators (TRAIN)





XGBClassifier with estimators (TEST)

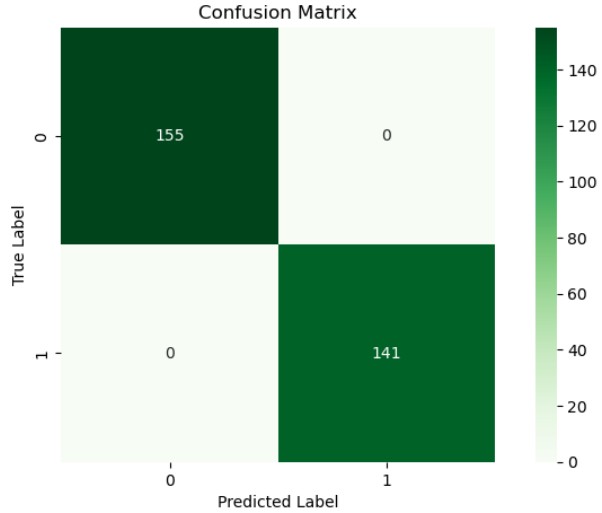


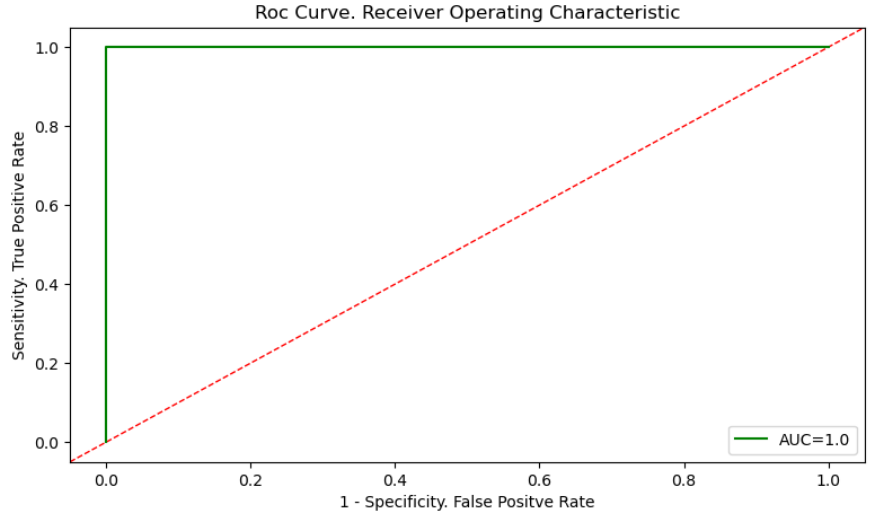


HistGradientBoostingClassifier

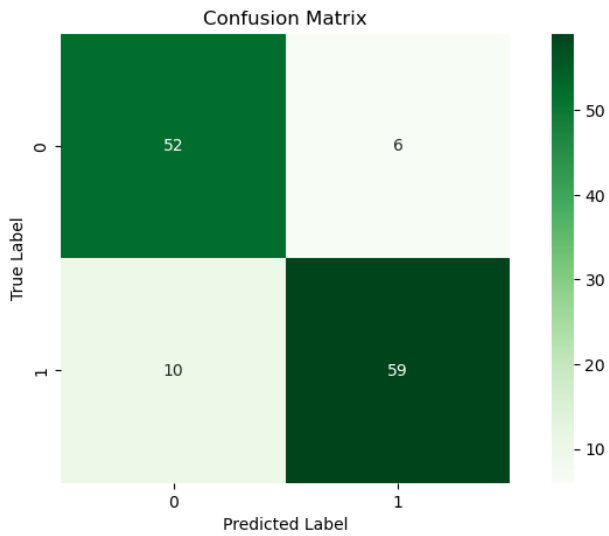
AUC: 0.87 F1:88

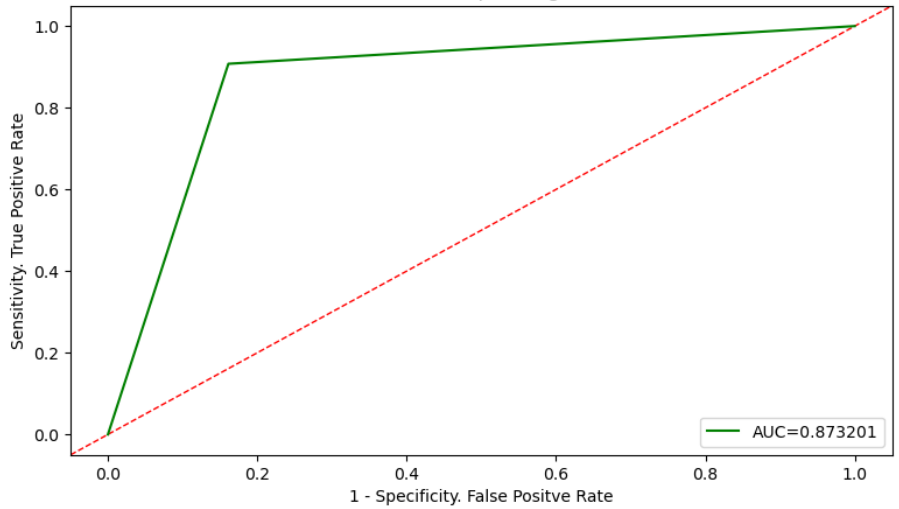
HistGradientBoostingClassifier (TRAIN)





HistGradientBoostingClassifier (TEST)



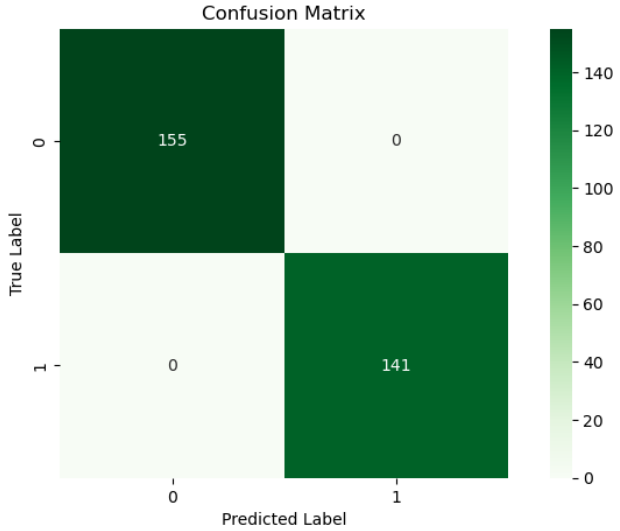


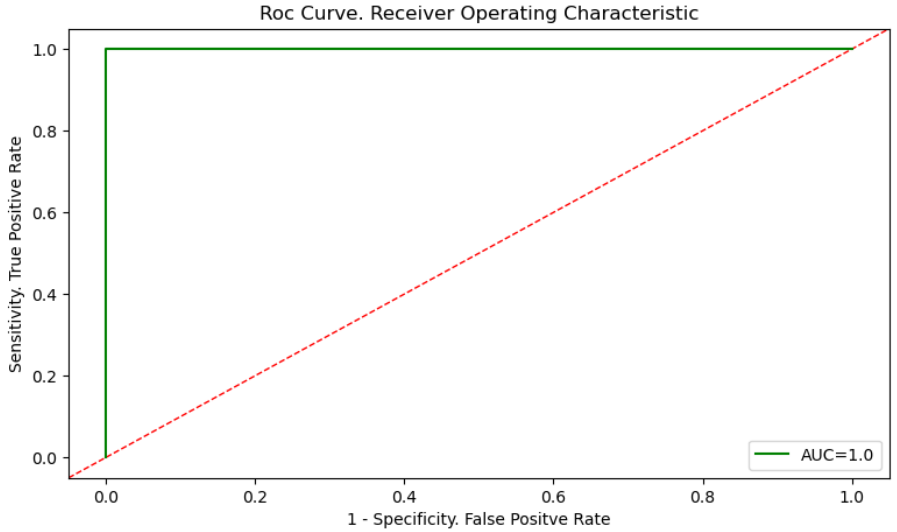
Random Forest

AUC: 0.87 F1:88

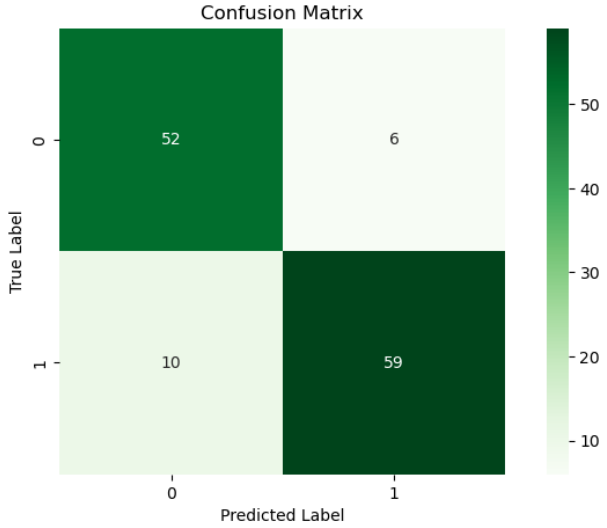
```
RandomForestClassifier(max_depth=3, n_estimators=10, random_state=0)
```

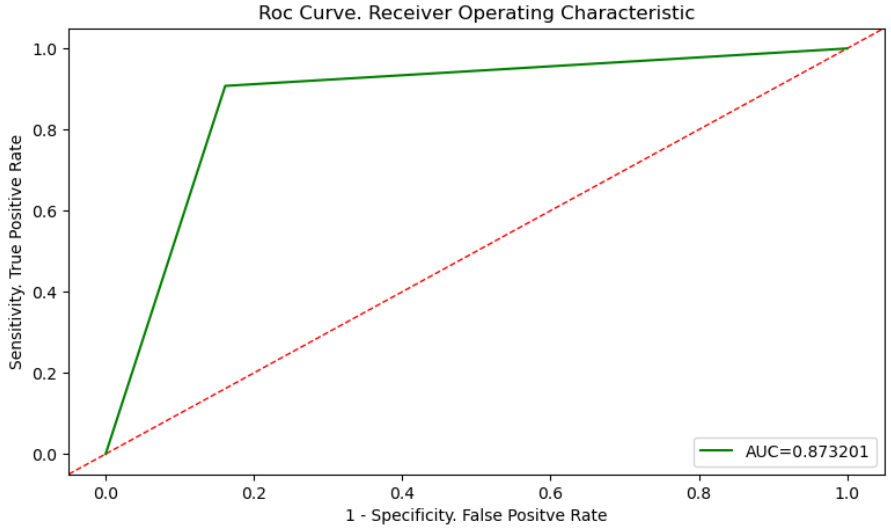
Random Forest optimizado(TRAIN)





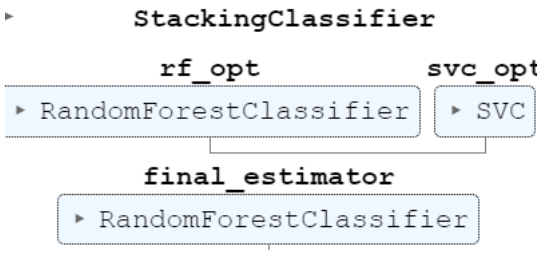
Random Forest optimizado (TEST)



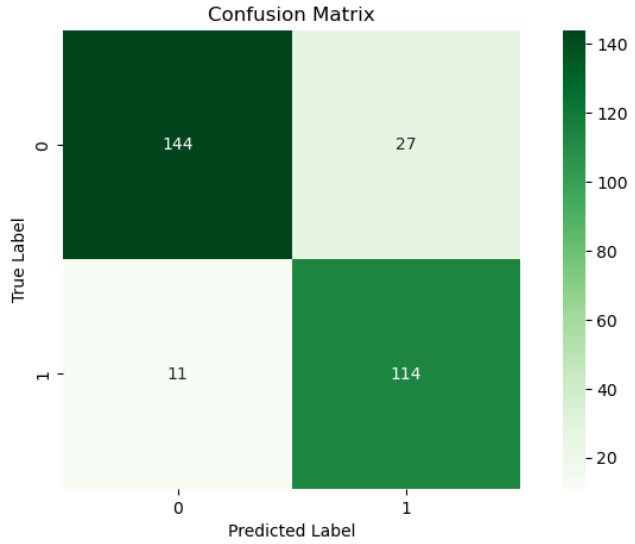


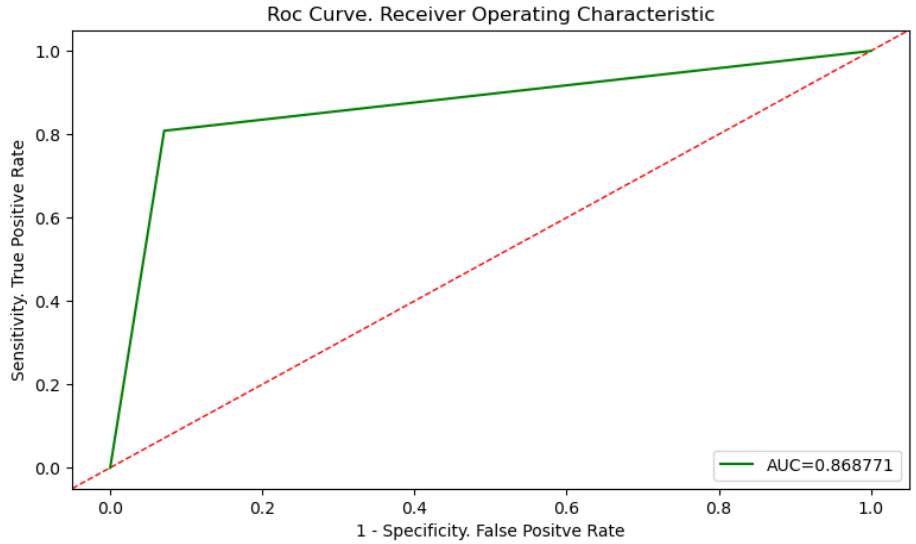
StackingClassifier

AUC: 0.85 F1:85

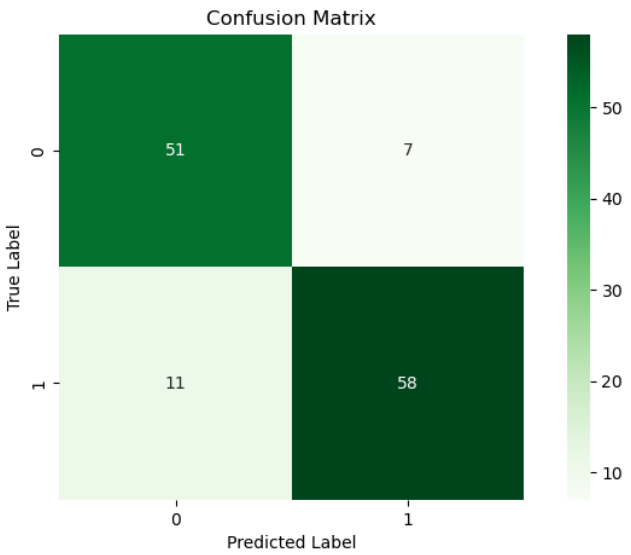


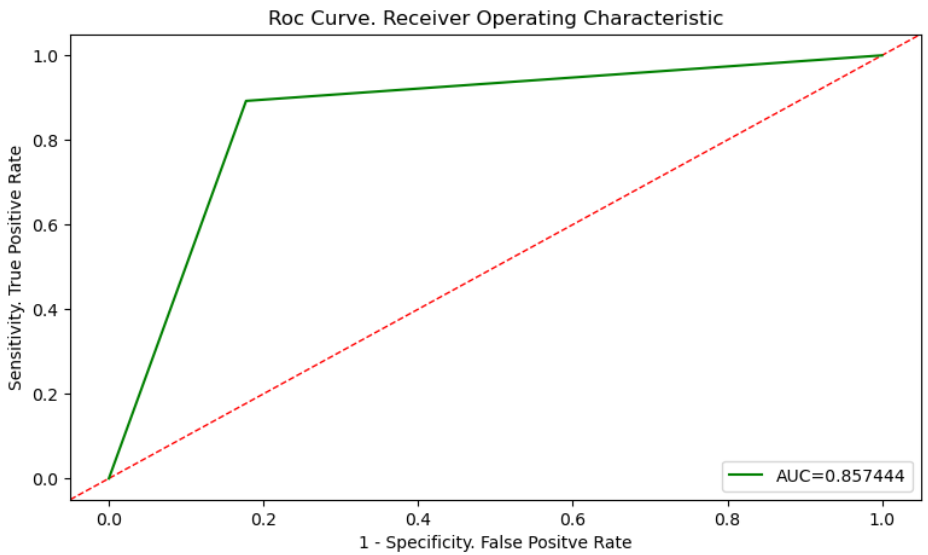
StackingClassifier (TRAIN)





StackingClassifier (TEST)





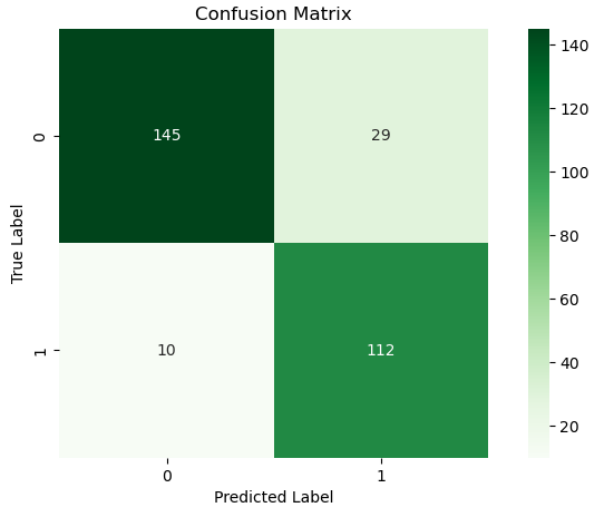
SVM

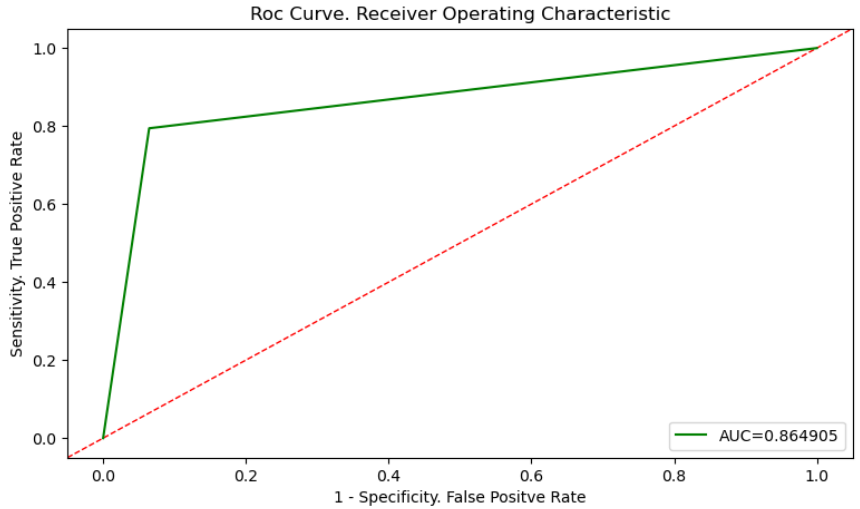
AUC: 0.86 F1:85

SVC

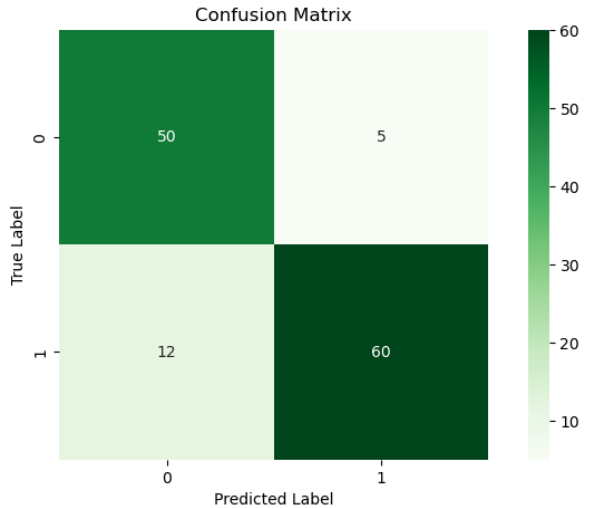
```
SVC(C=10, gamma=0.01, kernel='linear', probability=True, random_state=0)
```

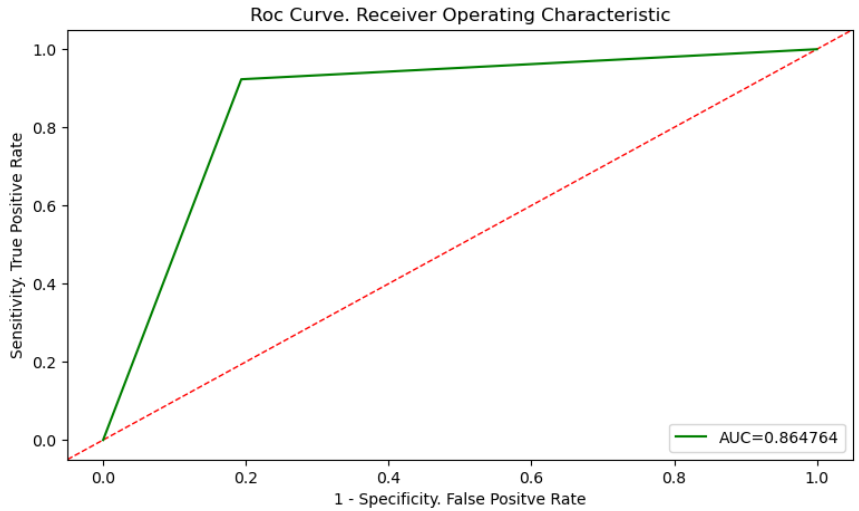
SVM optimizado (TRAIN)





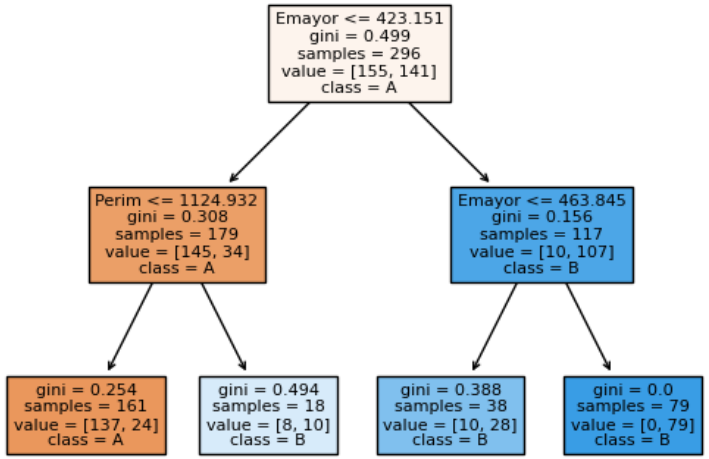
SVM optimizado(TEST)



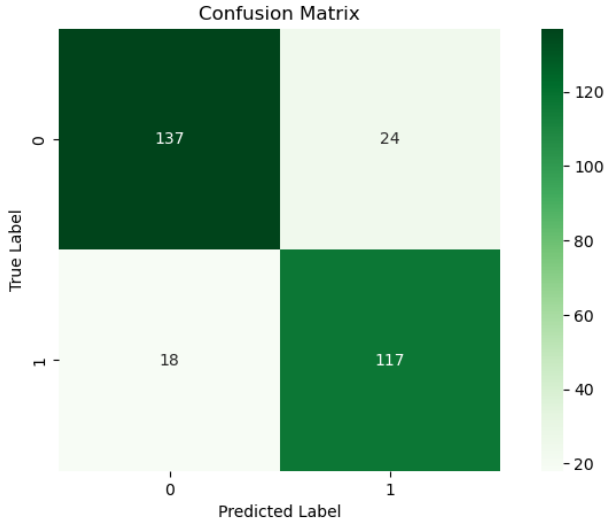


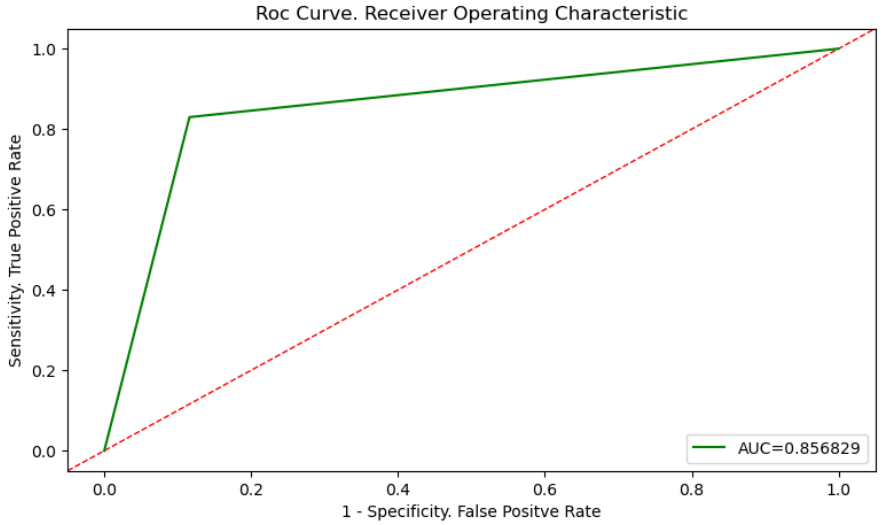
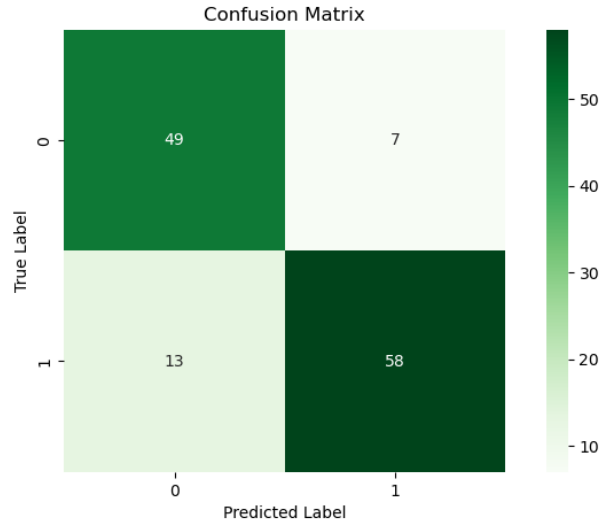
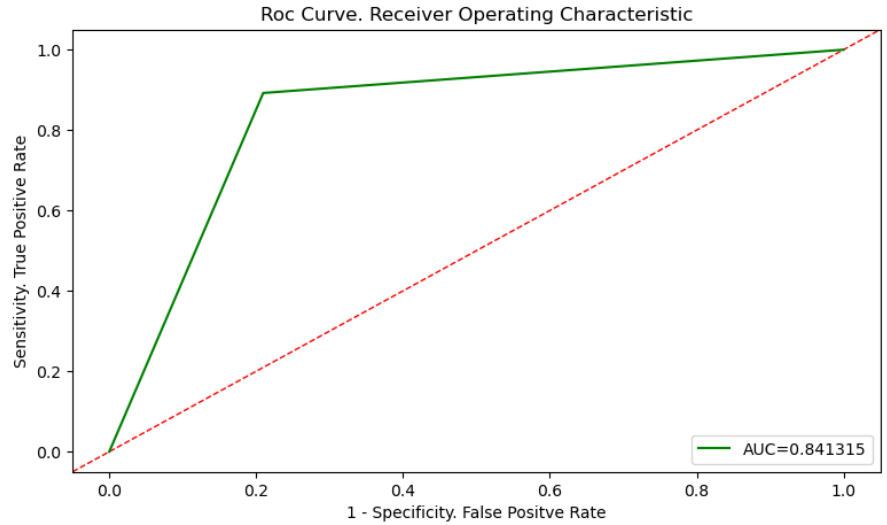
DecisionTreeClassifier

AUC: 0.84 F1:84



Decission Tree sin optimizar (TRAIN)





Decission Tree sin optimizar(TEST)