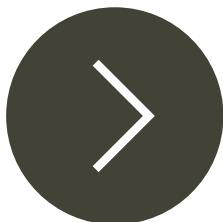


Análisis de Auditoría en búsqueda de Empresas Fraudulentas



Ignacio Sabatell





Introducción



Nuestro proyecto convierte al auditor en un verdadero “detective financiero”.

Utilizando técnicas de Machine Learning, analizamos múltiples indicadores de riesgo –financieros, operativos e históricos– para predecir qué empresas presentan mayor probabilidad de fraude.

Datos





Datos

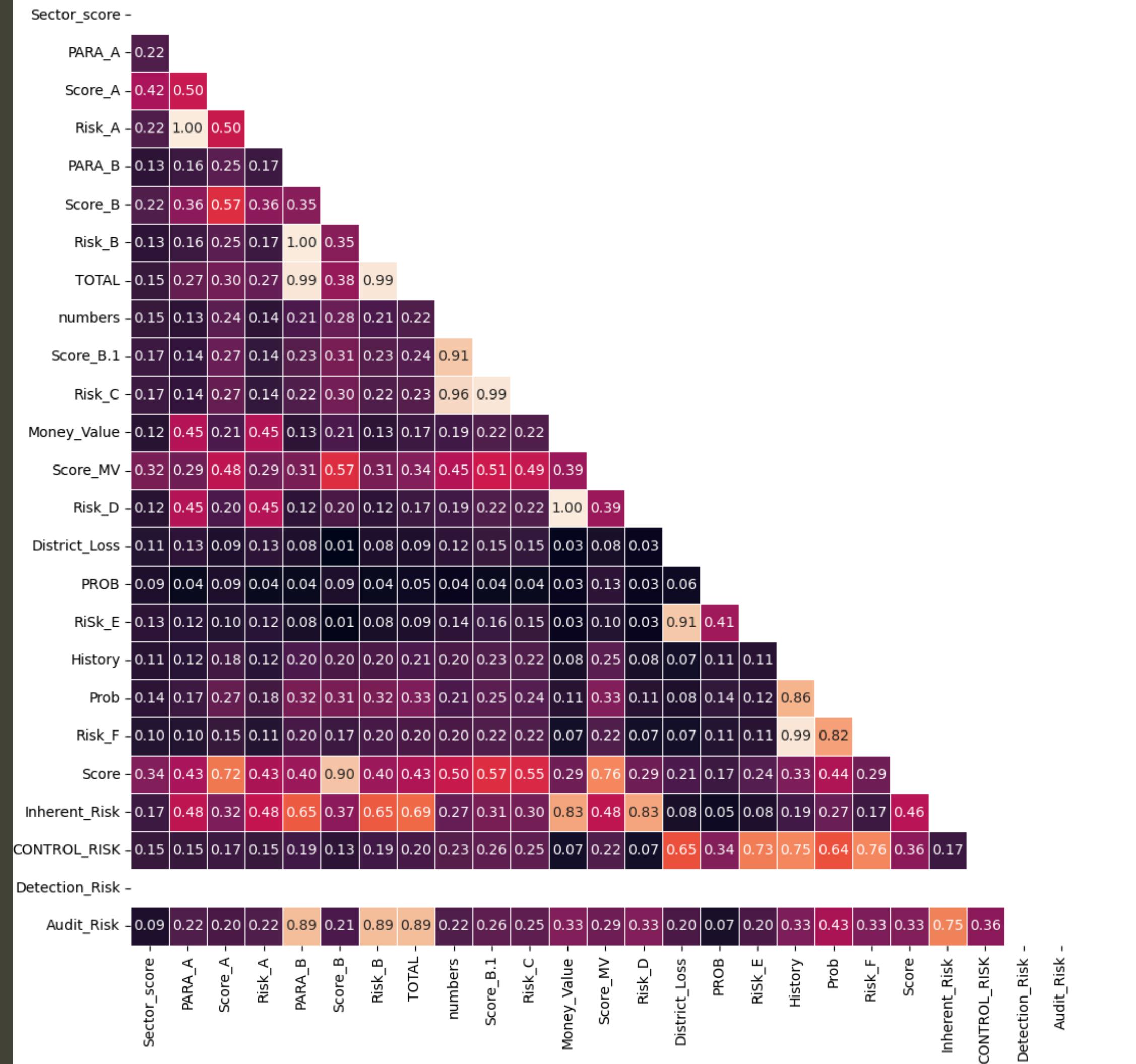
- 1 Score_A, Score_B, Score_MV: indicadores derivados de distintos factores de riesgo.
- 2 Inherent Risk, Control Risk, Detection Risk: miden el riesgo inherente y la efectividad de los controles existentes.
- 3 Probabilidades de fraude: la probabilidad calculada de que la empresa esté incurriendo en fraude.



Matriz Correlación

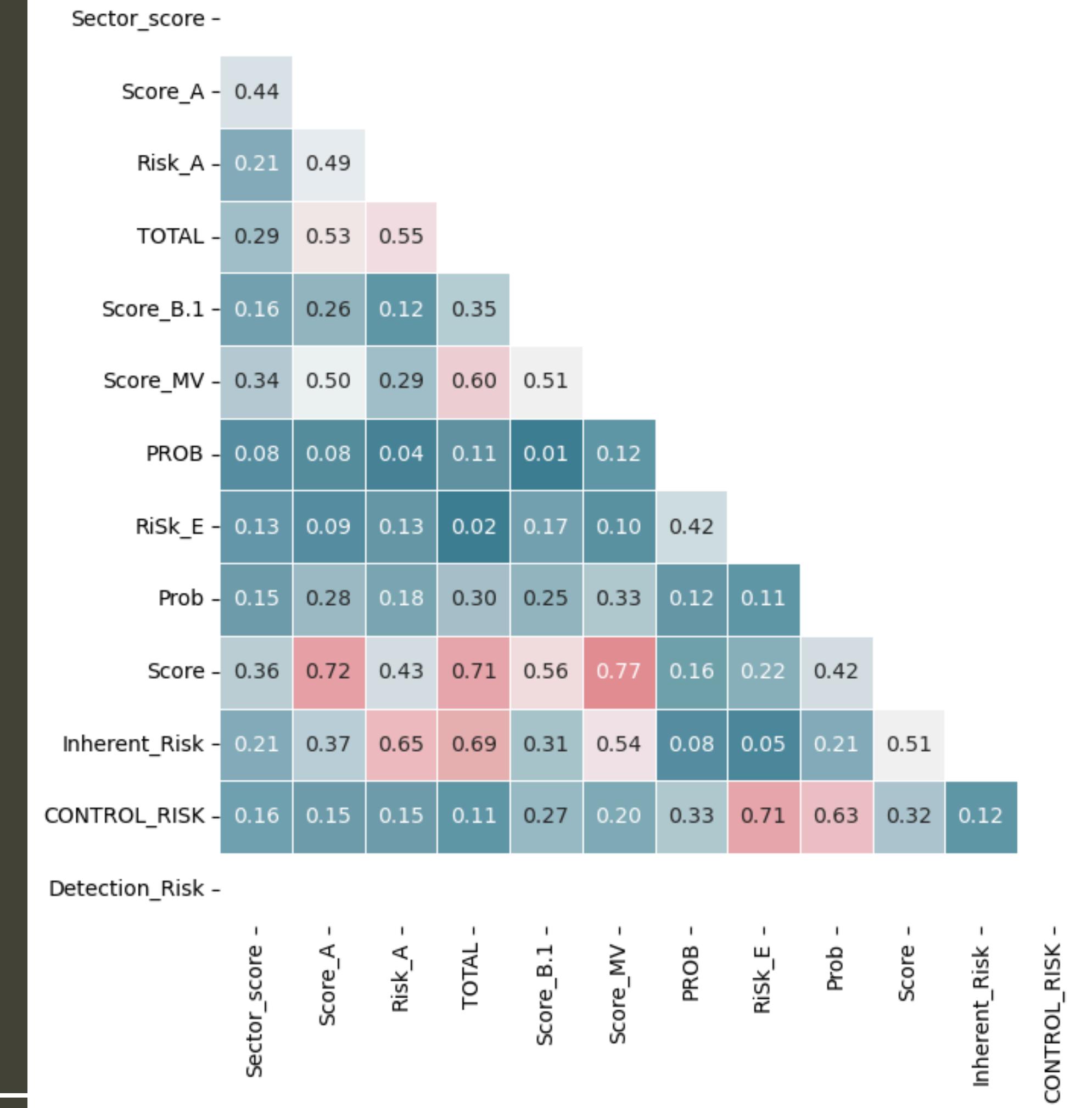
Si la Correlación > 0.8
Eliminamos columnas

- Reduce redundancia de información.
- Mejora la interpretabilidad del modelo.
- Evita sobreajuste (overfitting) al eliminar features que aportan lo mismo.
- Hace el modelo más eficiente y rápido de entrenar.



Matriz de features reducida

Usamos estas columnas para entrenar el modelo.





Modelos utilizados



- Bagging
- Random Forest
- Gradient Boosting
- Adaboost





Random forest

En los Random Forests, cada árbol solo aprenderá a partir de un subconjunto de los puntos de datos y de las características.

Por lo tanto, cada estimador aprende con datos de entrenamiento diferentes, asegurando independencia.

Random Forest Metrics

Accuracy: 0.9806451612903225

F1-score: 0.9767441860465116

Confusion Matrix:

```
[[89  0]
 [ 3 63]]
```

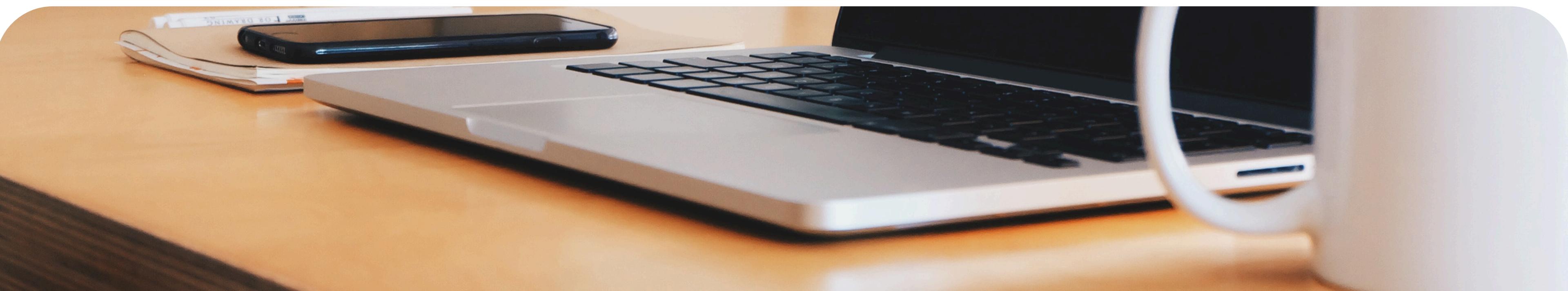
Classification Report:

	precision	recall	f1-score	support
0	0.97	1.00	0.98	89
1	1.00	0.95	0.98	66
accuracy			0.98	155
macro avg	0.98	0.98	0.98	155
weighted avg	0.98	0.98	0.98	155



Aplicación del modelo

Streamlit





Datos a tener en cuenta

- El dataset al ser pequeño las predicciones son muy altas.
- Debería haber una leyenda más concreta de los datos para su interpretación