Saturdays.AI
Guayaquil

UNIVERSIDAD DEL PACÍFICO
ESCUELA DE NEGOCIOS
ECUADOR

Fraude Bancario

Proyecto Inteligencia Artificial

https://github.com/soriaster/Fraude_Bancario

fppt.com

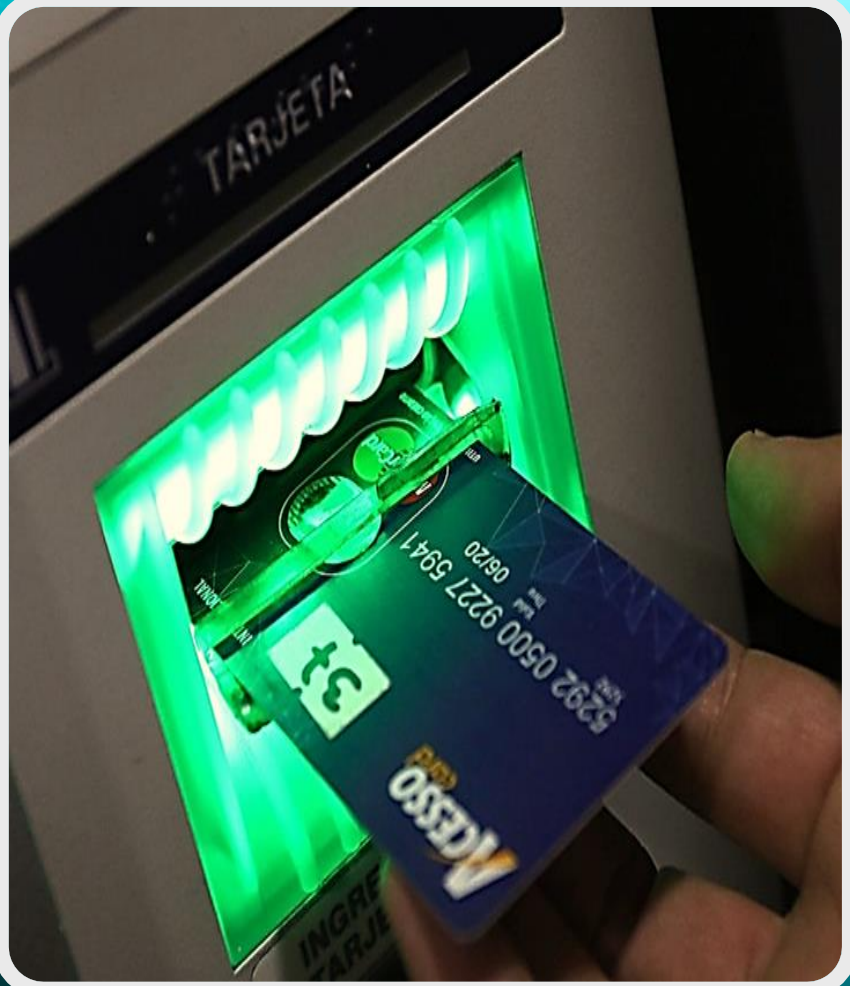
Integrantes

					
Michelle Nieto	Elisa Cisneros	Andrés Soriano	Alex Samaniego	Alex Vizuete	Daniel Quinde

fppt.com

LA PROBLEMATICA

- Los ciberdelincuentes aprovechan los avances tecnológicos para descubrir nuevas vulnerabilidades en los sistemas.
- Ellos se fijan en los patrones de consumo de las personas. Y utilizan técnicas de clonación de tarjetas en cajeros automáticos en lugares concurridos para pasar inadvertidos.



fppt.com

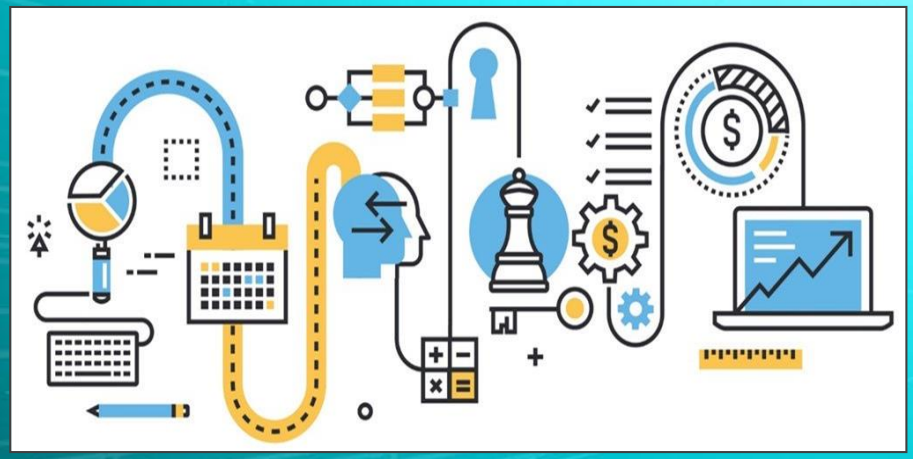
EL MERCADO OBJETIVO

- Estas vulnerabilidades existen en todas las industrias que ofrecen a sus clientes las tarjetas de crédito y de débito como método de pago.
- Los negocios más comunes son:
  - ❑ Comercios / Retail,
  - ❑ Entidades Financieras y
  - ❑ Gasolineras

fppt.com

LA ESTRATEGIA

- Se ha desarrollado un demo, donde se utilizó un set de datos de un Banco Europeo, con transacciones de tarjetas de crédito y débito, en un período de dos días, y con información de transacciones fraudulentas y no fraudulentas.
- Con este historial nos fue posible realizar el entrenamiento de un algoritmo para poder clasificar los datos fraudulentos.



fppt.com

LOS DATOS: DATA SET

- Data Set Base: 284807
- Separamos los datos identificados como fraude Campo 1, de los que no son Fraude Campo 0. Y se guardaron en dos DataFrame
- Fraudes 492 registros -> clase minoritaria
- No fraudes 284315 registros -> clase mayoritaria
- Los datos fraudulentos equivalen al 0.17% del Dataset total

```
[ ]    1 pd.value_counts(ftc.Class, sort = True)
```

```
0    284315
1       492
```

fppt.com

EL DATA SET

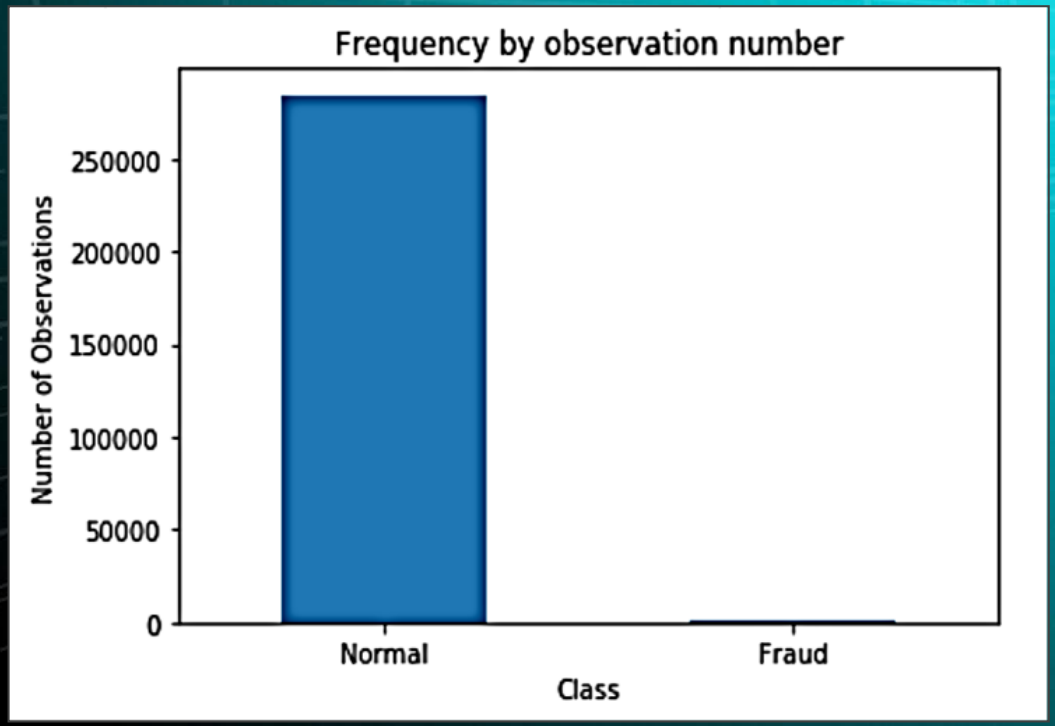
- Por seguridad y confidencialidad de los datos, tenemos 28 campos / variables sin etiqueta

	Time	V1	V2	V3	V4	V27	V28	Amount	Class
0	0.0	-1.359807	-0.072781	2.536347	1.378155	0.133558	-0.021053	149.62	0
1	0.0	1.191857	0.266151	0.166480	0.448154	-0.008983	0.014724	2.69	0
2	1.0	-1.358354	-1.340163	1.773209	0.379780	-0.055353	-0.059752	378.66	0
3	1.0	-0.966272	-0.185226	1.792993	-0.863291	0.062723	0.061458	123.50	0
4	2.0	-1.158233	0.877737	1.548718	0.403034	0.219422	0.215153	69.99	0

fppt.com

EL BALANCEO DE DATASET

- Para el balanceo de datos con el modelo de regresión logística utilizamos el metodo de penalización Class_weight="balanced", el cual se encargará de equilibrar la clase mayoritaria a minoritaria



fppt.com

QUE HICIMOS

- Por motivos de eficiencia y alto rendimiento, utilizamos tres modelos de inteligencia artificial:
- Regresión logística y
- XGBOOST
- Random Forest
- Y determinamos cual de las tres tiene mejor desempeño en la clasificación.

fppt.com

EL ENTRENAMIENTO Y LA PREDICCIÓN

- Entrenar el data set balanceado y comparar con diferentes modelos de entrenamiento.
- Modelos a Probar: Regresión logística, Random Forrest y XGBoost
- Obtener:
- TP (True Positive) TN (True Negative)
- FP (False Positive) FN (False Negative)

.

fppt.com

Saturdays.AI
Guayaquil

Universidad del Pacífico
Escuela de Negocios
Ecuador

Regresión Logística Con Penalización

Regresión Logística con Penalización

	precision	recall	f1-score	support
0	1.00	0.98	0.99	71087
1	0.07	0.91	0.13	115

```
roc_auc_score(y_test, y_pred_reglog)
```

```
0.9468012557790485
```

Random Forest

Random Forest con Balanceo Oversampling

	precision	recall	f1-score	support
0	1.00	1.00	1.00	71090
1	0.93	0.73	0.82	112

```
roc_auc_score(y_valid, y_pred_rfos)
```

```
0.8660292285432952
```

XGBoost

XBOOST con Método Ensambling

	precision	recall	f1-score	support
0	1.00	0.99	0.99	71090
1	0.08	0.86	0.15	112

```
roc_auc_score(y_valid, y_pred_exgb)
```

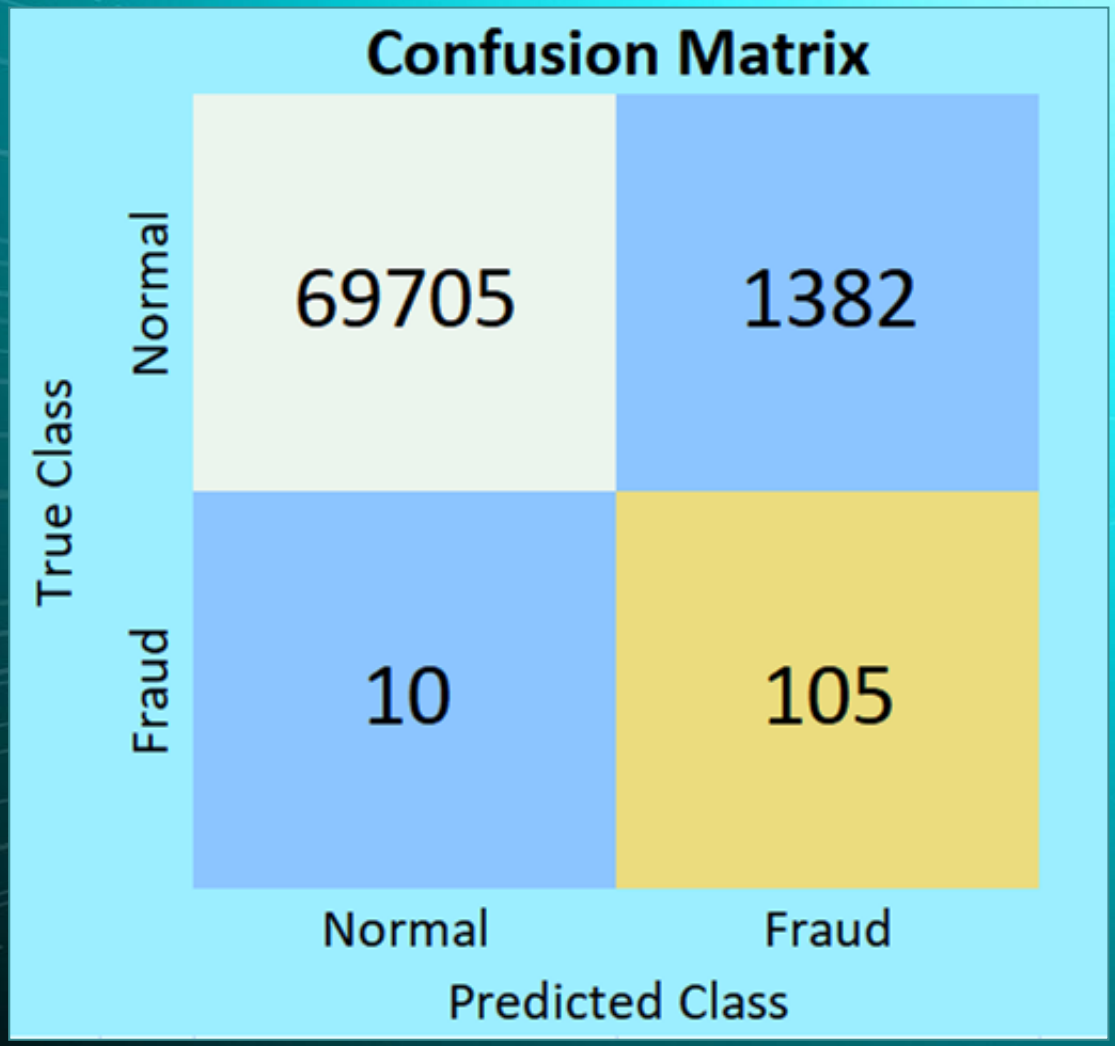
```
0.9210879569157808
```

fppt.com

LA MATRIZ DE CONFUSIÓN

Regresión Logística Con Penalización

AUC que tan buen clasificador es
Recall capacidad para poder identificar fraude y no fraude
Presición Éxito en predecir nuestro data set desbalanceado



fppt.com

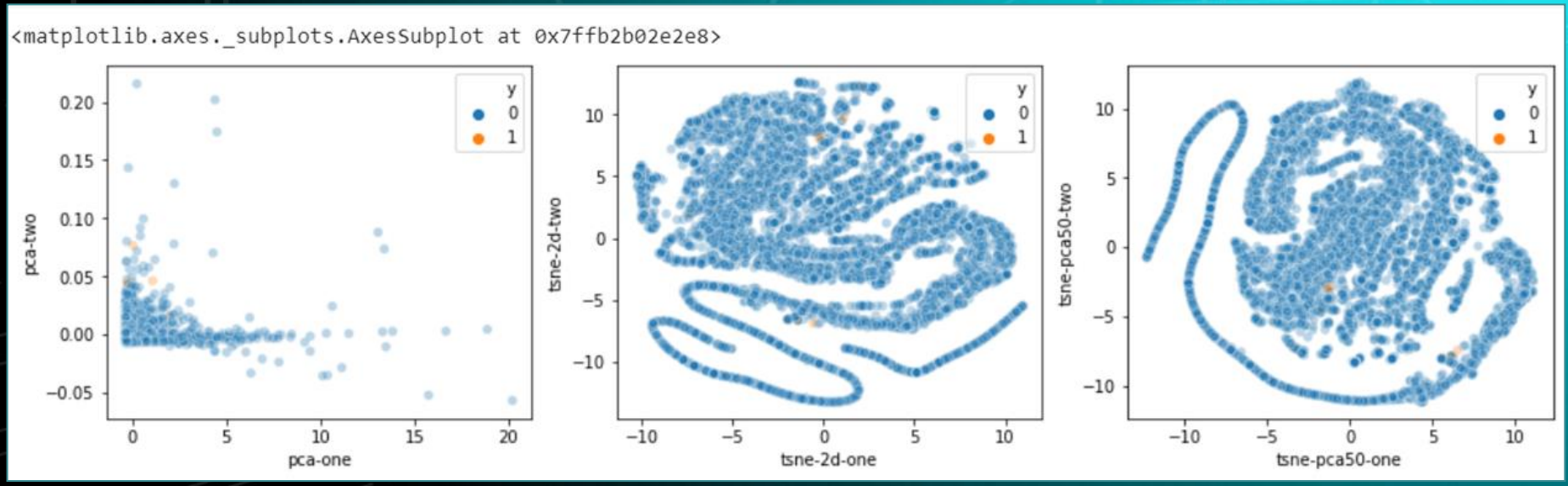
ANÁLISIS COMPARATIVO

MODELO	XGBOOST	REGRESIÓN LOGÍSTICA CON PENALIZACION	RANDOM FOREST
RECALL	0,86	**0,91**	0,73
AUC	0,92	**0,94**	0,86

fppt.com

MODELOS DE VISUALIZACIÓN T-SNE

- **PCA. –** Establecemos la reducción de dimensionalidad con el análisis de componentes principales
- **TSNE. -** El cual consiste en que los puntos cercanos se atraen y los distantes se repelen, con la finalidad de reducir las dimensiones.



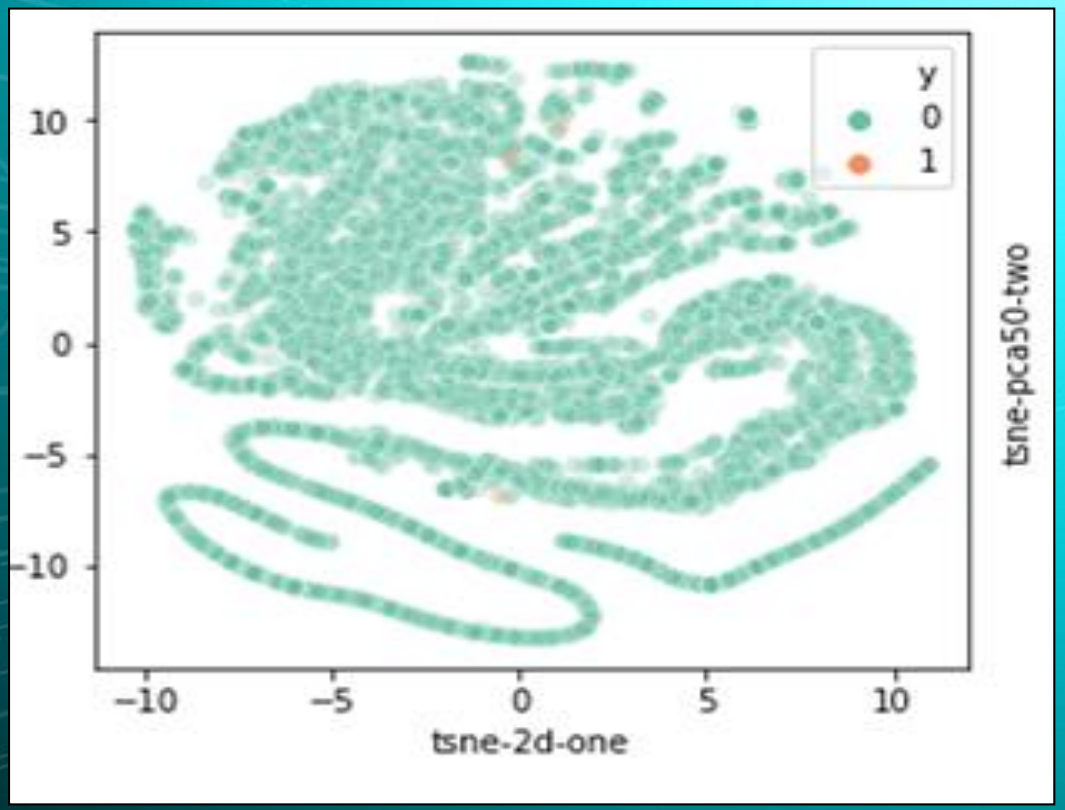
fppt.com

INTERPRETACIÓN T-SNE

►Es un algoritmo de aprendizaje automático para visualización desarrollado por L. Van Der Maaten.

►Es una técnica no lineal para la reducción de dimensionalidad, que permite visualizar DATA-SET de hasta 30M.

►t-SNE se utiliza en áreas como seguridad informática, investigación del cáncer, bioinformática.



fppt.com

RESULTADOS

- Regresión Logística con penalización fue el modelo que mejor se comportó.
- La característica de penalización class_weight="balanced" es la encargada de balancear internamente el dataset utilizando los valores dependientes y_train, ponderando las clases para asegurar una combinación equilibrada.
- El recall fue de 0,91; el ROC_AUC de 0,94; la precisión del modelo seleccionado es de 7%, sin embargo es preferible que el modelo entrenado sea asertivo en los datos fraudulentos.

fppt.com

CONCLUSIONES

- El data set era altamente desbalanceado y con variables restringidas por la privacidad del banco.
- Elaboramos el proyecto con varios modelos de inteligencia artificial para analizar los resultados con criterio estadístico de RECALL y AREA BAJO LA CURVA.
- Hemos comprobado que la tecnología está tan abierta, que los modelos están mejorándose constantemente. Y esto nos permite obtener mejores resultados, basados en los criterios de tiempos de respuesta y porcentaje de éxito en los resultados en las predicciones.

fppt.com

GRACIAS!

Fraude Bancario

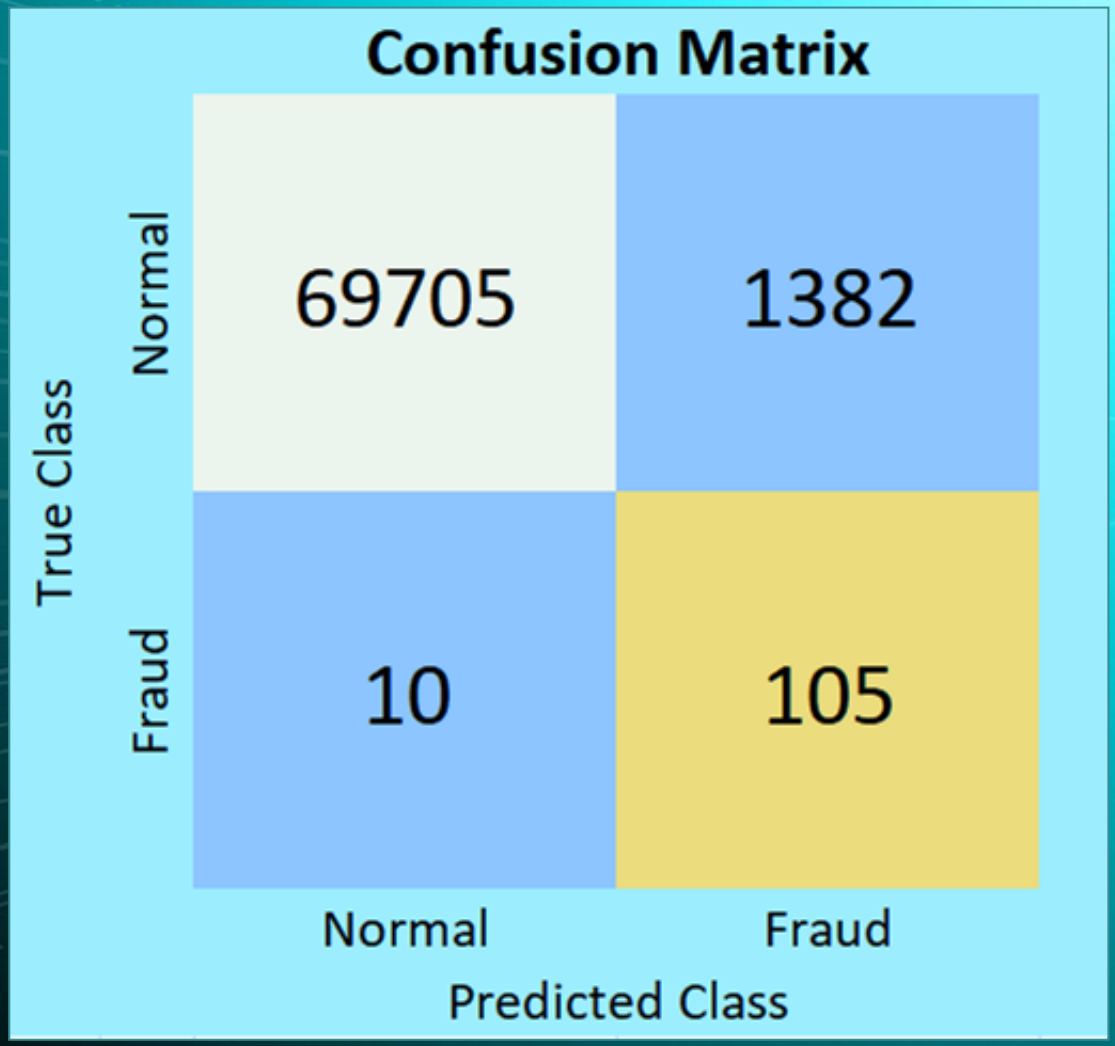
Proyecto Inteligencia Artificial

https://github.com/soriaster/Fraude_Bancario

fppt.com

LA MATRIZ DE CONFUSIÓN

Regresión Logística Con Penalización



fppt.com