# PREDICCIÓN DEL **SCORE CREDITICIO** MEDIANTE TÉCNICAS DE **APRENDIZAJE SUPERVISADO**


**Integrantes:**
Paula Cárdenas
Sebastián Giraldo
Alejandro Rendón
Alejandro Velásquez


Técnicas de Estadística Multivariada Avanzada
2022 - 2

# Agenda

1. Comprensión del negocio
2. Comprensión de los datos
3. Preparación de los datos
4. Modelado
5. Evaluación
6. Conclusiones

# 1. Comprensión del negocio

### Compañía financiera

Información básica de las personas y créditos en un banco de Estados Unidos (Kaggle)

### Planteamiento del problema

Clasificación del score crediticio con técnicas de aprendizaje supervisado
¿Cuál es el mejor clasificador?

### Objetivo

Disminuir el riesgo crediticio
Mejor asignación de recursos
Reducir esfuerzos manuales

# 2. Comprensión de los datos

**Registros**

Transacciones de 12.500 clientes en una ventana temporal de 8 meses.

100.000 registros

**Variable objetivo**

Poor

Standard

Good

**Variables**

**Categóricas**

Profesión

Tipo de préstamo

Comportamiento de pago

**Numéricas**

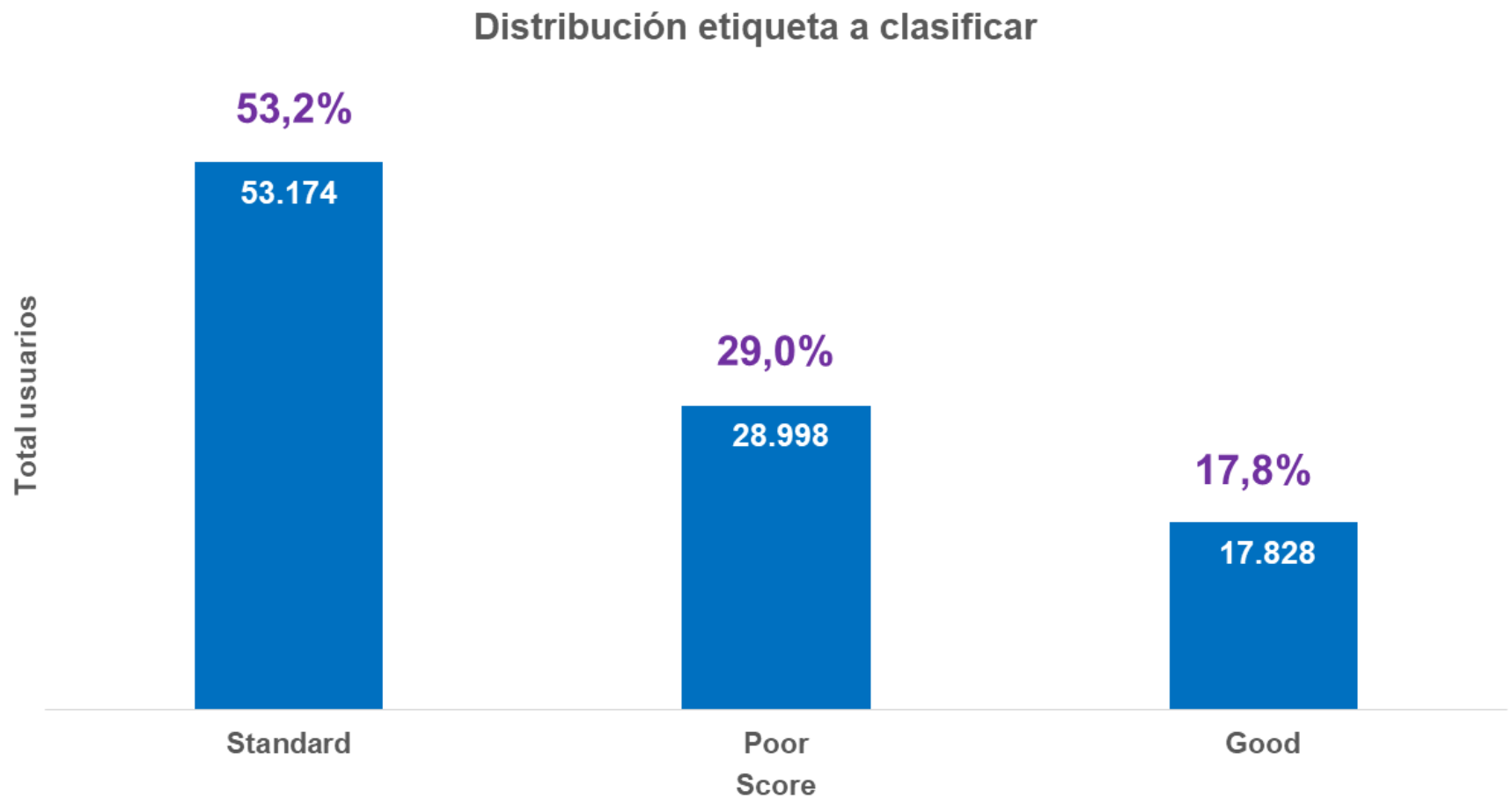
Edad

Ingresos mensuales

Tasa interés

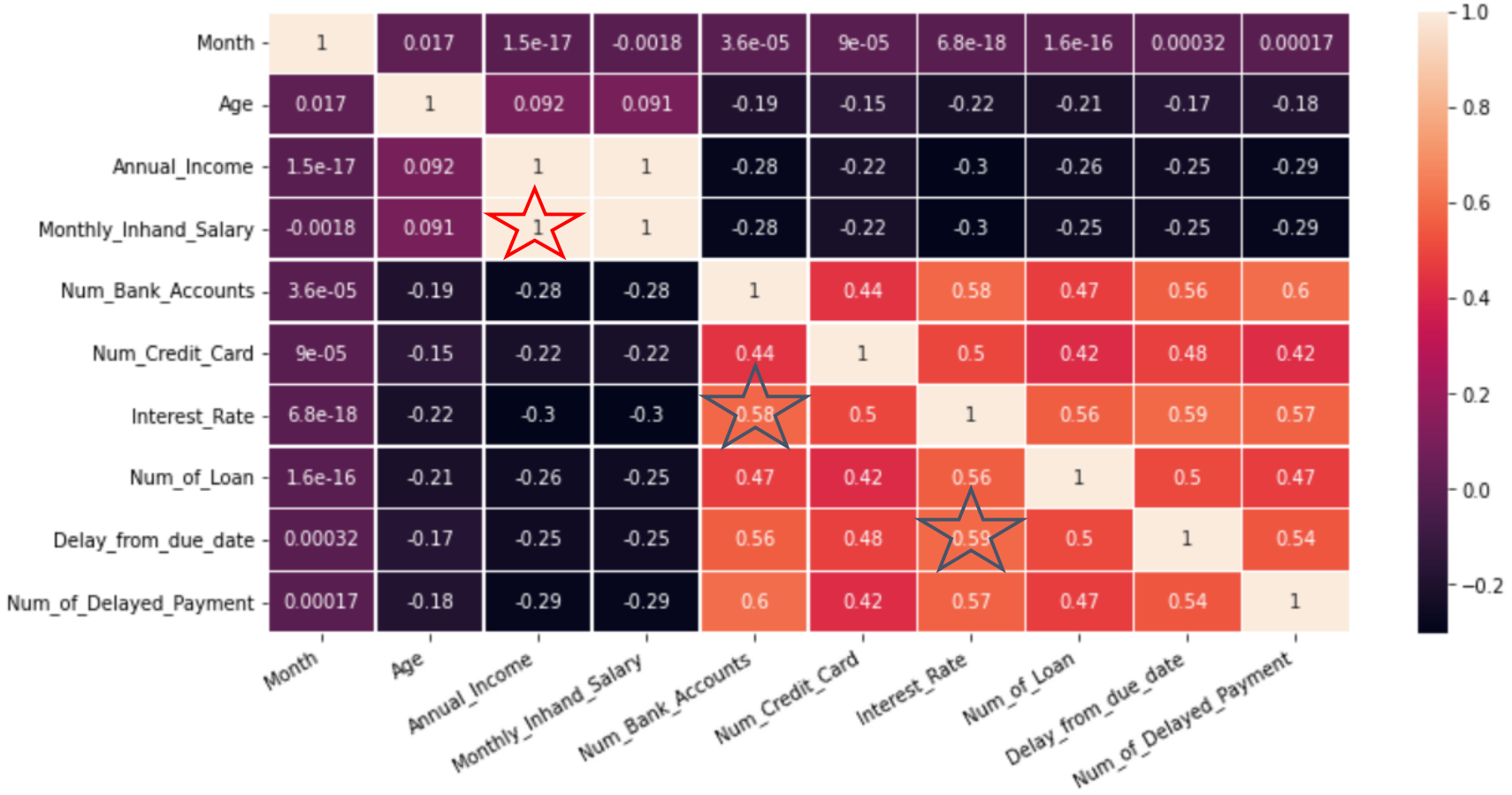
Número de pagos retrasados

# Exploración de datos



# Exploración de datos

## Análisis multivariante



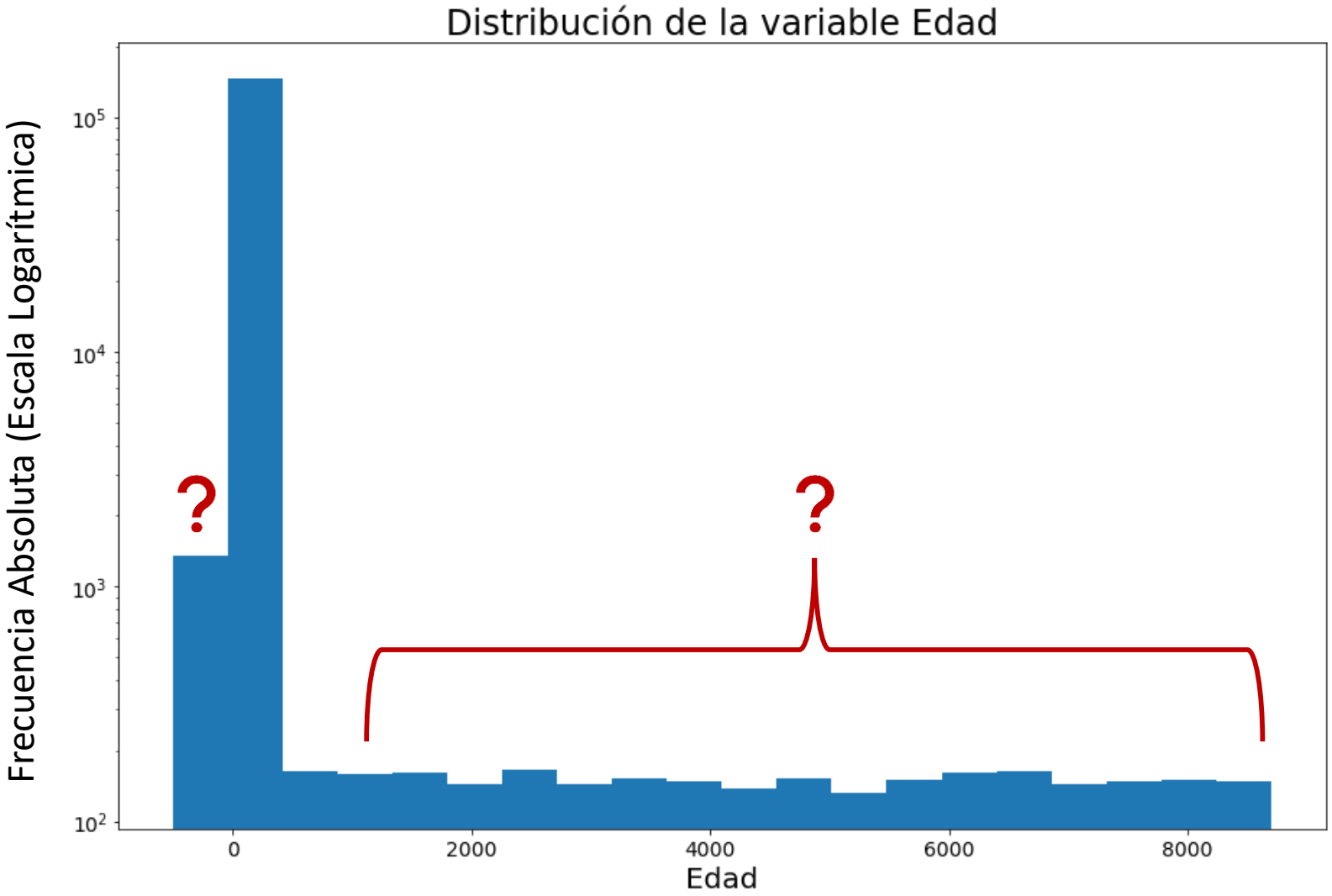

Cantidad de cuentas bancarias
vs
Tasa de interés

Retraso en pagos
vs
Tasa interés

# 3. Preparación de los datos

| ID Cliente | Mes | Nombre | Edad | Ocupación | Salario mensual (USD) | Historial crediticio | Comportamiento de pago | Balance mensual |
|---|---|---|---|---|---|---|---|---|
| CUS_0xd40 | January | Aaron Maashoh | 23 | Scientist | 1.824 | 22 Years and 1 Months | High spent - Small value | 31.249 |
| CUS_0xd40 | February | Aaron Maashoh | 23 | Scientist | | NA | Low spent - Large value | 28.462 |
| CUS_0xd40 | March | Aaron Maashoh | -500 | Scientist | | 22 Years and 3 Months | Low spent - Medium value | 33.120 |
| CUS_0xd40 | April | Aaron Maashoh | 23 | Scientist | | 22 Years and 4 Months | Low spent - Small value | 22.345 |
| CUS_0xd40 | May | Aaron Maashoh | 23 | Scientist | 1.824 | 22 Years and 5 Months | High spent - Medium value | 34.248 |
| CUS_0xd40 | June | Aaron Maashoh | 23 | _____ | | 22 Years and 6 Months | !@9#%8 | 34.047 |
| CUS_0xd40 | July | Aaron Maashoh | 23 | Scientist | 1.824 | 22 Years and 7 Months | Low spent - Small value | 24.456 |
| CUS_0xd40 | August | | 23 | Scientist | 1.824 | NA | High spent - Medium value | 35.812 |

# Datos atípicos



# Ingeniería de características

## Cantidad de Variables

Variables iniciales: 25

One-hot Encoding

↓

Variables finales: 46

## Principal Component Analysis (PCA)

Factor Analysis of Mixed Data

Variance ratio: 80%

PCA: 27

↓

¿Vale la pena hacer PCA?

# 4. Modelado

## Entrenamiento

Train 70%

Test 30%

**División estratificada**

GroupShuffleSplit

## Grid Search

Hiperparámetros óptimos para predicciones más precisas.

✓ Sesgo

✓ Varianza

## Modelos

**Regresión Logística**

```
C=0.25,penalty='l1'
```

**Random Forest**

```
max_depth=20
```

**LightGBM**

```
max_depth=15
```

**XGBoost**

```
max_depth=5
```

# 5. EVALUACIÓN

## Matriz de Confusión – SIN PCA

**LGBM**

```
[[6692  975 1110]
 [3637 8793 3617]
 [ 126  796 4254]]
              precision    recall  f1-score   support

           0    0.76245   0.64008   0.69592     10455
           1    0.54795   0.83236   0.66085     10564
           2    0.82187   0.47367   0.60097      8981

    accuracy                        0.65797     30000
   macro avg    0.71076   0.64870   0.65258     30000
weighted avg    0.70471   0.65797   0.65515     30000
```

## Precisión en test

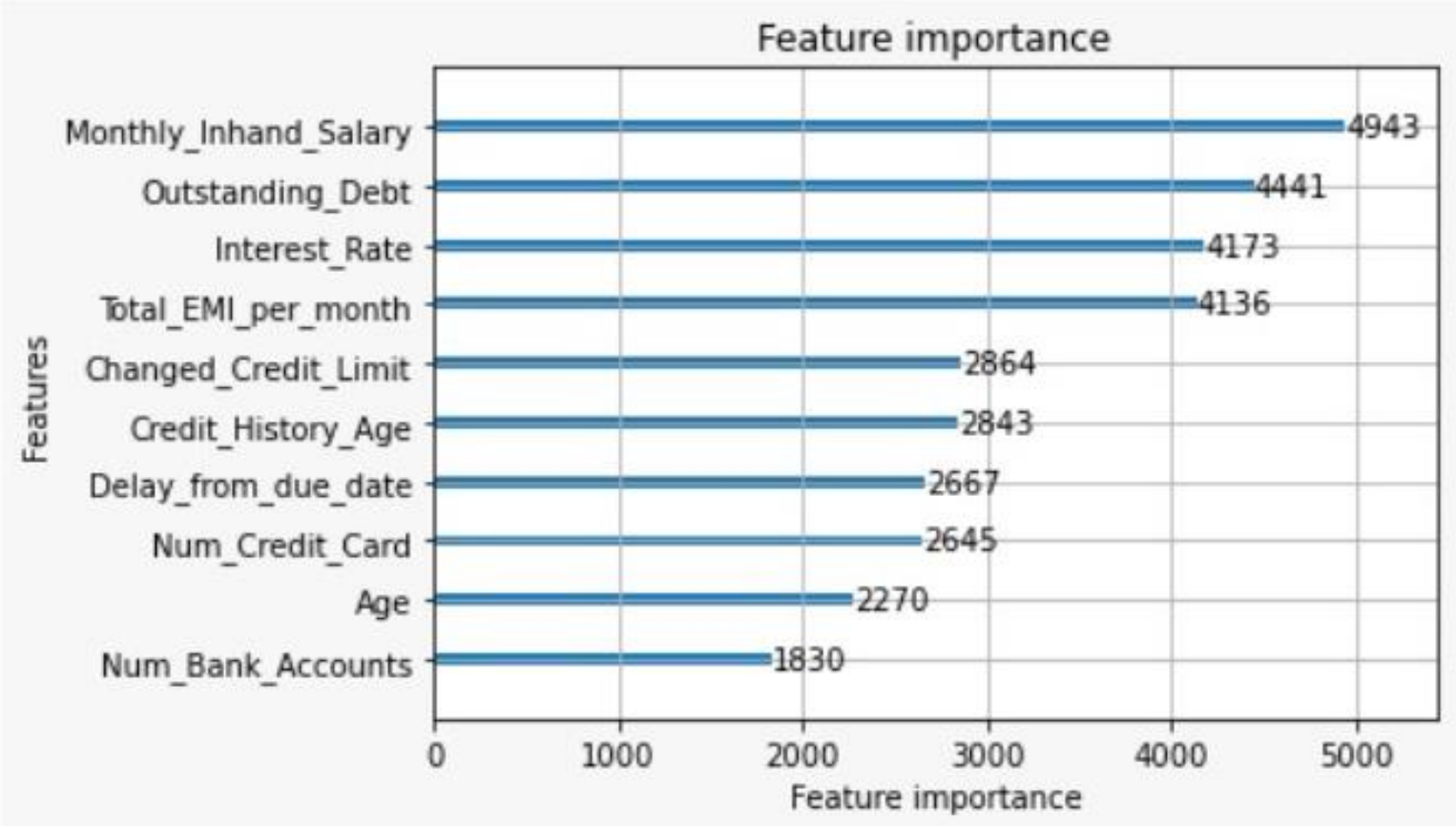
Para la clase – **Poor (0)**

| SIN PCA | |
|---|---|
| **Modelo** | **Precisión** |
| Regresión Logística | 0,66 |
| Random Forest | 0,70 |
| LGBM | 0,76 |
| XGBoost | 0,63 |

| CON PCA | |
|---|---|
| **Modelo** | **Precisión** |
| Regresión Logística | 0,70 |
| Random Forest | 0,68 |
| LGBM | 0,70 |
| XGBoost | 0,60 |

# Feature importance - LGBM

**SIN PCA**



# CONCLUSIONES

- Las técnicas usadas para el tratamiento de datos disminuyen la probabilidad de perdida de información, por tanto, los métodos abordados son una forma de reconstrucción de información.
- Aplicar técnicas de reducción de dimensionalidad en matrices con pocas características es innecesario, ya que genera perdida de interpretabilidad.
- Con la generación de todos los modelos se comprueba que el Grid Search sí mejora la precisión.
- Las técnicas de Gradient Boosting fueron más precisas que las técnicas lineales y el random forest tradicional.
- LGBM es mejor que XGBoost por ser más rápido y preciso, además de tener menos costo computacional.
- Los modelos generados son para clientes que ya tienen historial crediticio en la entidad financiera.
- Aplicar redes neuronales con el volumen de datos actual es costoso computacionalmente y la precisión no mejora sustancialmente.

# GRACIAS