# ¿Impagos a la vista?

Propuesta de modelos de Machine Learning para identificar posibles impagos

Autor: Rebeca Herrejón
Fecha: 23 de julio de 2024

## Tabla de contenidos

# Propósito

Este proyecto es la presentación final de **Rebeca Herrejón** para el curso **De Cero a Ciencia de Datos**.

Presenta el proceso que se siguió para elegir entre dos modelos de *machine learning* que permitieran identificar si los clientes de una empresa financiera incumplirán en sus pagos, con base en las siguientes características:

- Historial de pagos
- Montos de facturación
- Límite de crédito
- Características demográficas

Las características están incluidas en un *dataset* llamado `data.csv`.

# Introducción

Este proyecto es la presentación final de **Rebeca Herrejón** para el curso **De Cero a Ciencia de Datos**.

Presenta el proceso que se siguió para elegir entre dos modelos de *machine learning* que permitieran identificar si los clientes de una empresa financiera incumplirán en sus pagos, con base en las siguientes características:

- Historial de pagos
- Montos de facturación
- Límite de crédito
- Características demográficas

Las características están incluidas en un *dataset* llamado `data.csv`.

# Metodología

Esta sección presenta la metodología que se siguió para elegir el modelo de machine learning con mejor precisión en predicción.

## Explicación de los datos

El dataset data.csv contiene información de 30 mil clientes de una empresa de servicios financieros, agrupados en 25 columnas que se describen a continuación:

- **ID**: Identificación del cliente.
- **LIMIT_BAL:** Monto del límite de crédito (asumí que son pesos).
- **SEX**: Género del cliente.
- **EDUCATION**: Nivel educativo del cliente.
- **MARRIAGE**: Estado civil del cliente.
- **AGE**: Edad del cliente.
- **PAY_0** a **PAY_6**: Historial de pagos de los últimos seis meses.
- **BILL_AMT1** a **BILL_AMT6**: Monto de facturación de los últimos seis meses.
- **PAY_AMT1 a PAY_AMT6**: Monto de pago de los últimos seis meses.
- **default.payment.next.month**: Indicador de si el cliente incumplió el pago el próximo mes (columna objetivo).

## Análisis exploratorio de los datos

De acuerdo con las estadísticas descriptivas obtenidas con Jupyter Notebook y el siguiente código:

- El promedio del límite de crédito es de $167,484 pesos, con un mínimo de $10,000 y un máximo de un millón de pesos.
- El 60% de los clientes son mujeres.
- La edad promedio es de 35.5 años.
- El 22% de los clientes incurrieron en incumplimiento de pago en el último mes.
- Los clientes con historial de retrasos en los pagos tienen mayor probabilidad de incumplimiento.

- Los clientes que incumplen con el pago probablemente tienen las facturas más altas.
- Los clientes que incumplen con el pago tienden a realizar pagos más bajos en comparación con aquellos que no incumplen.

## División del conjunto de datos para entrenamiento y prueba

Para este proyecto, elegí regresión logística (PCA) y árbol de decisión como los algoritmos para evaluar si un cliente podría incumplir en el pago o no.

Para ambos casos usé el 20 por ciento de los datos para prueba y el resto para entrenamiento.

### PCA

Eliminé los datos nulos o vacíos.

Eliminé del dataset las variables o columnas que no nos sirven, como el ID del cliente y la columna que predice si hay o no incumplimiento del pago.

Usé la técnica de Análisis de Componentes Principales (PCA), reduciendo la dimensionalidad a dos componentes (lo intenté con 10-por el número de columnas que se analizan- y obtuve una exactitud similar).

La matriz de confusión predijo los siguientes datos:

- True Negatives (TN): 4580. El modelo predijo 'No' y la realidad es 'No'.
- False Positives (FP): 107. El modelo predijo 'Sí' pero la realidad es 'No'.
- False Negatives (FN): 1108. El modelo predijo 'No' pero la realidad es 'Sí'.
- True Positives (TP): 205. El modelo predijo ‘Sí’ y la realidad es ‘Sí’.

El modelo tiene una exactitud o *accuracy* de 80%, lo que significa que clasifica correctamente el 80% de las instancias del conjunto de datos. Sin embargo, puede mejorarse con otras técnicas que aún estoy analizando.

### Árbol de Decisión

Eliminé los datos nulos o vacíos.

Eliminé del dataset las variables o columnas que no nos sirven, como el ID del cliente y la columna que predice si hay o no incumplimiento del pago.

La matriz de confusión para este modelo arrojó los siguientes resultados:

- **True Negatives (TN)**: 3824
- **False Positives (FP)**: 863
- **False Negatives (FN)**: 781
- **True Positives (TP)**: 532

Este modelo tiene una exactitud de 72.6%.

# Conclusiones

Elijo el modelo de PCA por tener la mayor exactitud 80% para predecir si los clientes de la empresa financiera incumplirán en su próximo pago, con base en las siguientes características:

- Historial de pagos
- Montos de facturación
- Límite de crédito
- Características demográficas

El modelo aún puede mejorarse, incrementando el conjunto de datos de entrenamiento y prueba.

Las recomendaciones incluyen seguir evaluando modelos. *Random Forest*, que es un conjunto de árboles de decisión, podría funcionar para este caso de uso.

# Ligas a Documentación

Las siguientes ligas dirigen a los documentos de trabajo que viven en el siguiente repositorio: `rebecaheva/DeCeroACien`

- Regresión Logística (PCA)
- Arbol de Decisión