# PRIMERA ENTREGA DEL PROYECTO

**Por:**

Jonatan Stiven Restrepo Lora

CC:1018376574

Andres Felipe Graciano Monsalve

CC: 71375739

Ricardo Osorio Castro

CC: 1036785264

**Materia:**

Introducción a la Inteligencia Artificial

Métodos Estadísticos

**Profesor**

Raul Ramos Pollan


Facultad de Ingeniería

Universidad de Antioquia - Colombia

2023

1. **Planteamiento del problema**

   ¿Cómo saben los emisores de tarjetas que devolveremos lo que cobramos? La predicción del incumplimiento crediticio es fundamental para gestionar el riesgo en un negocio de préstamos al consumo. La predicción del incumplimiento crediticio permite a los prestamistas optimizar las decisiones crediticias, lo que conduce a una mejor experiencia del cliente y a una economía empresarial sólida.

   El objetivo del modelo es predecir la probabilidad de que un cliente no pague el saldo de su tarjeta de crédito en el futuro en función de su perfil de cliente mensual. La variable binaria objetivo se calcula observando una ventana de rendimiento de 18 meses después del último extracto de la tarjeta de crédito, y si el cliente no paga el monto adeudado en 120 días después de la última fecha del extracto, se considera un evento de incumplimiento.

2. **Dataset**

   Vamos a usar el dataset de kaggle de esta competición (https://www.kaggle.com/competitions/amex-default-prediction/data?select=train_data.csv),El objetivo de esta competencia es predecir la probabilidad de que un cliente no pague el saldo de su tarjeta de crédito en el futuro en función de su perfil de cliente mensual. La variable binaria objetivo se calcula observando una ventana de rendimiento de 18 meses después del último extracto de la tarjeta de crédito, y si el cliente no paga el monto adeudado en 120 días después de la última fecha del extracto, se considera un evento de incumplimiento.El conjunto de datos contiene características de perfil agregadas para cada cliente en cada fecha de estado de cuenta. Las funciones son anónimas y normalizadas y se clasifican en las siguientes categorías generales:

   D_* = Delinquency variables (Variables de morosidad)
   S_* = Spend variables (Variables de Gasto)
   P_* = Payment variables (Variables de Pago)
   B_* = Balance variables (Variables de Balance)
   R_* = Risk variables (Variables de Riesgo)

   La tarea es predecir para cada customerID la probabilidad de un incumplimiento de pago futuro (target=1). Tenga en cuenta que la clase negativa se ha submuestreo para este conjunto de datos al 5 % y, por lo tanto, recibe una ponderación de 20 veces en la métrica de puntuación.

   tiene los siguientes archivos:

   **train_data.csv** - training data with multiple statement dates per customer_ID (datos de entrenamiento con múltiples fechas de extracto por customer_ID)

   **train_labels.csv** - target label for each customer_ID (etiqueta target para cada unocustomer_ID)

**test_data.csv** - corresponding test data; your objective is to predict the target label for each customer_ID (datos de prueba correspondientes; su objetivo es predecir la etiqueta target para cada customer_ID)

**sample_submission.csv** - a sample submission file in the correct format (un archivo de envío de muestra en el formato correcto)

tiene 924621 número de muestras y 396 columnas de todo el dataset (poner la lista de columnas, o si es muy grande, poner las que se consideren más representativas para dar una idea de cómo es el dataset).

3. **Métricas**

La métrica de evaluación será la media de dos medidas de ordenación de rangos: Coeficiente de Gini normalizado, **G**, y tasa de incumplimiento capturada en 4%, **D**:

$$M = 0.5 \cdot (G + D)$$

La tasa de incumplimiento capturada al 4 % es el porcentaje de etiquetas positivas (valores predeterminados) capturadas dentro del 4 % de las predicciones mejor clasificadas y representa una estadística : ”Sensitivity/Recall”

Para ambas sub métricas **G** y **D**, a las etiquetas negativas se les asigna un peso de 20 para ajustarlas a la reducción de resolución.

Esta métrica tiene un valor máximo de 1,0.

4. **Desempeño**

El desempeño deseable en producción para el modelo de predicción de incumplimiento en el pago de tarjetas de crédito será tener un Recall Alto, dado que el objetivo principal es identificar a los clientes que no pagarán a tiempo (incumplimientos), un recall alto es fundamental. Esto significa que el modelo está capturando la mayoría de los incumplimientos, lo que ayuda a reducir el riesgo de pérdidas financieras para la institución financiera.

5. **Cibergrafía**

Addison Howard, AritraAmex, Di Xu, Hossein Vashani, inversion, Negin, Sohier Dane. (2022). American Express - Default Prediction. Kaggle. https://kaggle.com/competitions/amex-default-prediction