**UNIVERSIDAD PRIVADA ANTENOR ORREGO**

**Facultad de Ingeniería**

**Escuela de Ingeniería de Computación y Sistemas**

****

**Machine Learning para la detección de fraudes de transacciones con tarjetas de crédito**

**Proyecto**

***Curso:***

                Administración y Arquitectura de Mainframes

***Alumno:***

Azabache Medina, Jean Pierre

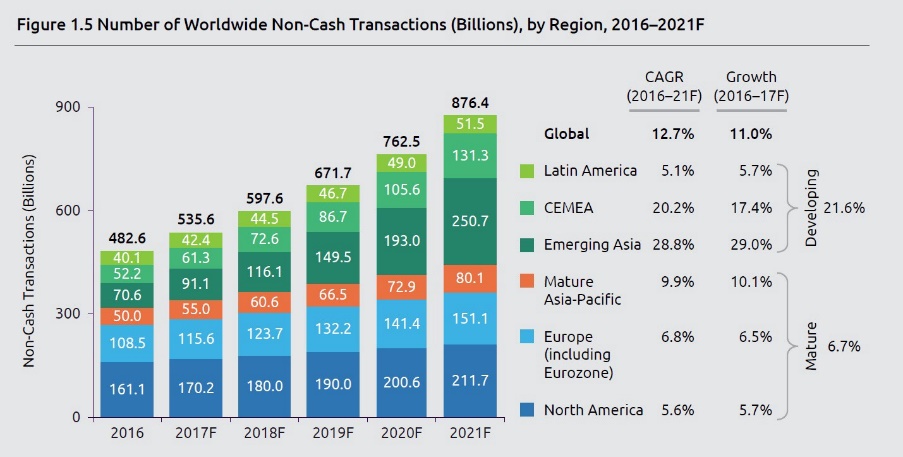
***Docente:***

                Cueva Chávez, Walter Manuel

**Trujillo – 2020**

Machine Learning para detección de fraude en transacciones con tarjeta de crédito

Según el [Informe Mundial de Pagos](https://worldpaymentsreport.com/wp-content/uploads/sites/5/2018/10/World-Payments-Report-2018.pdf) , en 2016 el total de transacciones no monetarias aumentó en un 10.1% desde 2015 para un total de 482.6 mil millones de transacciones. Además, se espera que en los próximos años haya un crecimiento constante de las transacciones no monetarias, como se muestra a continuación:



Modelos para emplear la compasión de los resultados con mayor precisión son : Regresión Logística, Bosque Aleatorio y Naive Bayes.

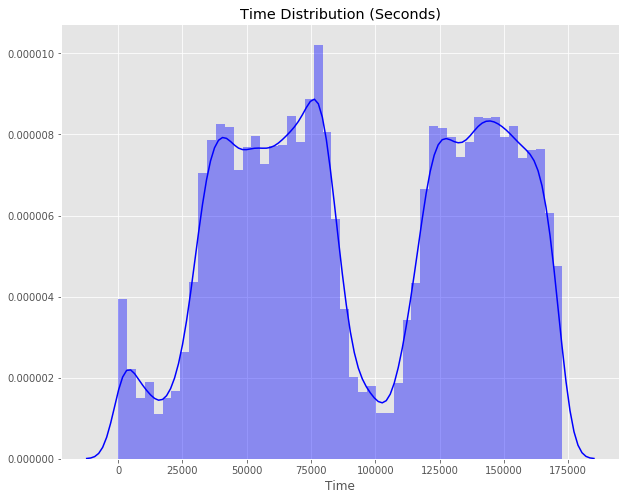
# ****Recopilando los datos****

Para comenzar, reuní mis datos de un [conjunto](https://www.kaggle.com/mlg-ulb/creditcardfraud) de datos de [Kaggle](https://www.kaggle.com/mlg-ulb/creditcardfraud) que contenía 285,000 filas de datos y 31 columnas. De todas las columnas, las únicas que tenían más sentido eran Tiempo, Cantidad y Clase (fraude o no fraude). Las otras 28 columnas se transformaron utilizando lo que parece ser una reducción de dimensionalidad de PCA para proteger las identidades de los usuarios.

# Análisis exploratorio de datos

Ahora que tenemos los datos, quería ejecutar algunas comparaciones iniciales entre las tres columnas que mencioné anteriormente (Tiempo, Cantidad y Clase).

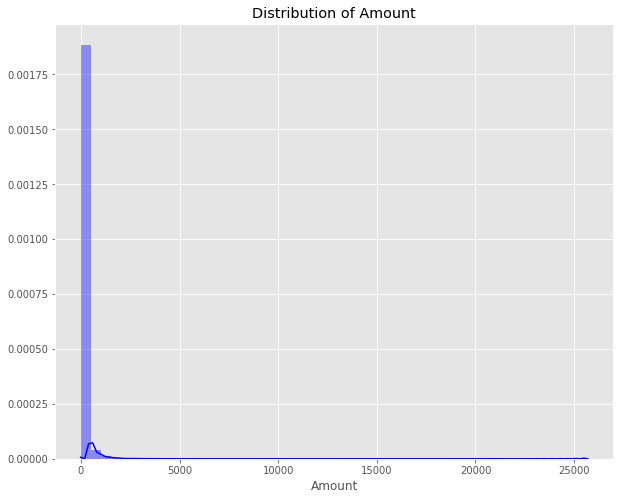
**Hora**



Cantidad Distribución de datos de la tarjeta de crédito

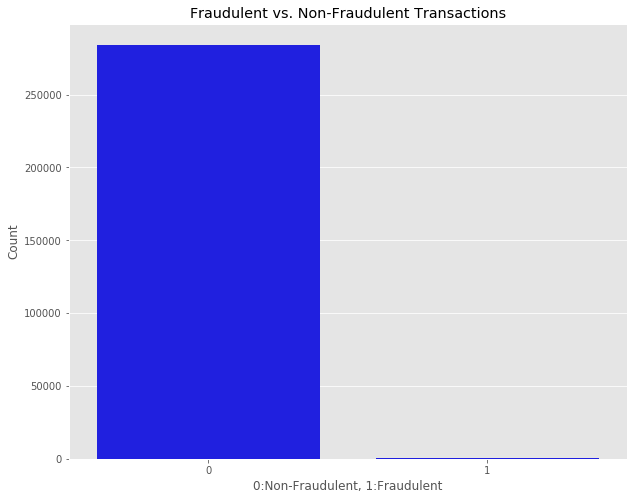
Dado que esta distribución es de dos días de datos, seguiría la tendencia que esperaría ver para los consumidores normales.

**Cantidad**



Cantidad Distribución de datos de la tarjeta de crédito

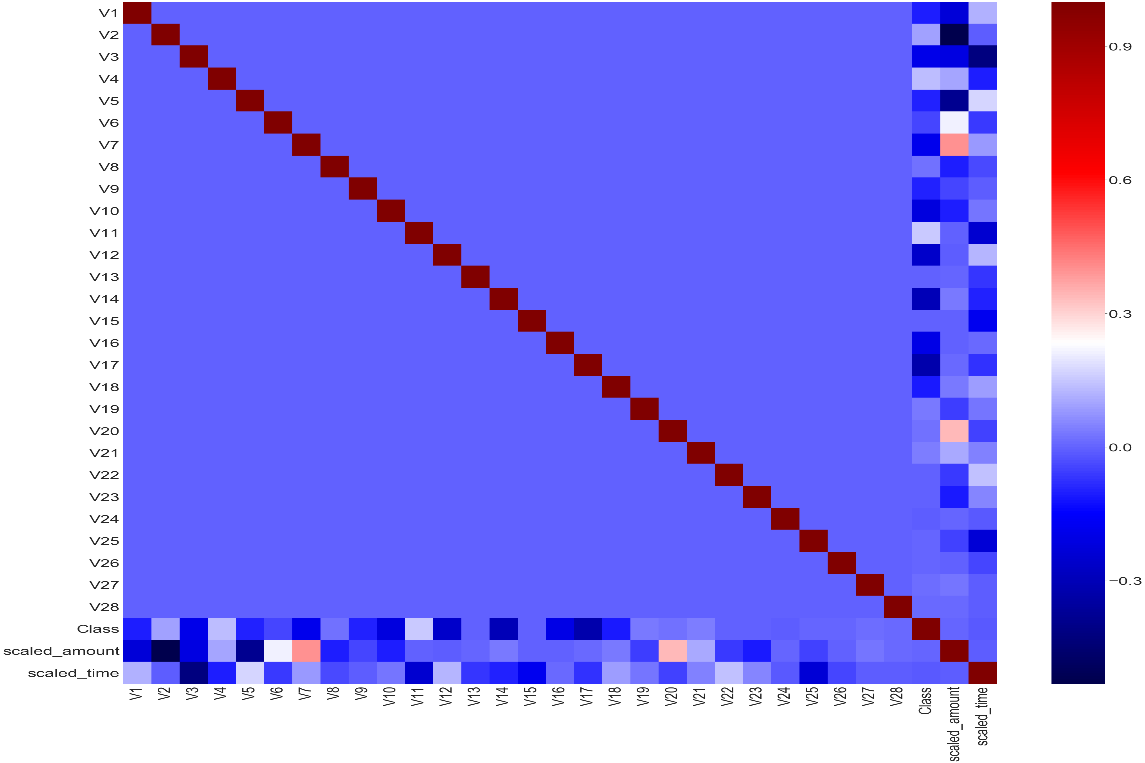
Si bien la gran mayoría de las transacciones son muy bajas, también se espera esta distribución. La mayoría de las transacciones diarias no son extremadamente caras (la mayoría son <$ 50), pero es probable que también se produzcan transacciones más fraudulentas.

**Clase (Fraude / No Fraude)**

En este conjunto de datos, solo hay 492 transacciones fraudulentas. Aun así, mi objetivo será capturar a tantos de los estafadores molestos como sea posible.

# ****Procesamiento de datos****

Comencé asegurándome de que no hubiera ninguna colinealidad fuerte en los datos usando un mapa de calor.



En este ejemplo, no agregué ningún número, ya que sería muy difícil de ver para los lectores. Estoy buscando algo que muestre una alta correlación, y algunos que vi fueron:

* Tiempo y V3 (-0.42)
* Cantidad y V2 (-0.53)
* Cantidad y V4 (0.4)

Si bien estas correlaciones son relativamente altas en comparación con las otras, no sentí que fueran lo suficientemente altas como para correr el riesgo de multicolinealidad.

**Standard Scaler en Tiempo y Cantidad**

Standard Scaler es una herramienta común cuando se trabaja con problemas de clasificación como este. Transforma los datos a donde hay una media de 0 y una desviación estándar de 1, estandarizando así los datos en una distribución normal. Especialmente al trabajar en un rango tan amplio de cantidades y tiempo, vi que escalar mis datos antes de ejecutar las pruebas proporcionaba mejores resultados.

Inicialicé el ajuste usando el conjunto de entrenamiento, luego escalé el conjunto de entrenamiento, validación y prueba antes de ejecutar todo en los modelos.

# Construyendo el modelo

Con tantos puntos de datos con los que trabajar, me decidí por una división de prueba del 60% / 20% de validación / 20% de prueba antes de crear cualquier modelo.

Otro problema para tratar es el conjunto de datos altamente desequilibrado. Con tantas transacciones no fraudulentas implementadas, implementé el muestreo aleatorio para disminuir el número de transacciones no fraudulentas y hacer que coincida con la cantidad de transacciones fraudulentas.

## **Regresión logística y bosques al azar**

**La regresión logística es un modelo estadístico que intenta minimizar el costo de cuán incorrecta es una predicción. Random Forests es un conjunto de árboles de decisión que predicen colectivamente si una transacción es fraudulenta o no.**

# Exactitud, recuperación, precisión y puntaje F1

Para esta sección, se usan estas abreviaturas en las fórmulas:

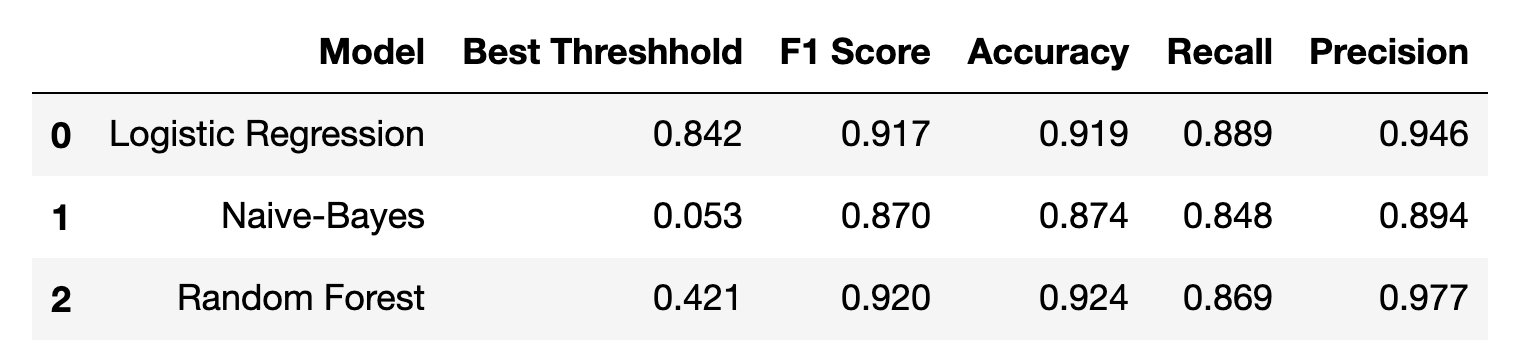
TP = Verdadero Positivo. Transacciones **fraudulentas** que el modelo predice como **fraudulentas** .

TN = Verdadero negativo. Transacciones **normales** que el modelo predice como **normales** .

FP = Falso Positivo. Transacciones **normales** que el modelo predice como **fraudulentas** .

FN = Falso negativo. Transacciones **fraudulentas** que el modelo predice como **normales** .

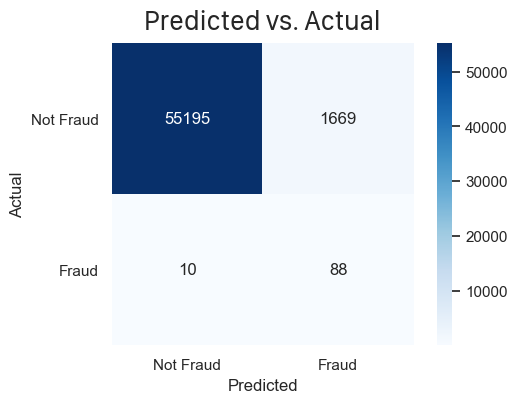
## **Los resultados de la comparación de los modelos**



Regresión logística y Bosque aleatorio son muy prometedoras para nuestro conjunto de datos. Cada modelo tiene una alta tasa de verdadero positivo y una baja tasa de falso positivo, que es exactamente lo que estamos buscando.

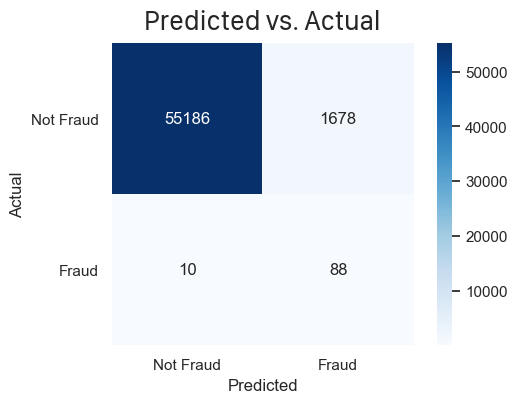
En estos umbrales, capturamos la cantidad óptima de transacciones fraudulentas mientras mantenemos los falsos positivos lo más bajo posible. Con estos umbrales, podemos visualizar los efectos de cada modelo utilizando una matriz de confusión:

Matriz de confusión - Regresión logística



Con un umbral de 0,842 en este conjunto de prueba fuera de muestra, capturamos 88 de las 98 transacciones fraudulentas y marcamos por error 1,669 transacciones como fraudulentas.

Matriz de confusión - Bosque aleatorio

Matriz aleatoria de confusión forestal

En un umbral de 0.421, el modelo de bosque aleatorio tiene resultados similares al modelo de regresión logística. El modelo Random Forest captura 88/98 transacciones fraudulentas, pero indica 9 transacciones normales más como fraudulentas en comparación con la Regresión logística. Ambos modelos siguen siendo prometedores y al final dan buenos resultados.

# Conclusión

Encontrar transacciones de tarjetas de crédito fraudulentas es en realidad importante, especialmente en la sociedad actual. Hay muchos métodos para capturar estas instancias, y es realmente sobresaliente ver cómo las empresas lidian con esto en el día a día.