1. Diseño de caso de uso de ML en Ciberseguridad

Problema de Negocio:

En una institución financiera con una base de clientes de 1 millón de usuarios, se enfrentan a un problema de detección de fraude en transacciones con tarjeta de crédito y débito. Actualmente, están experimentando una tasa de fraude del 1.5%, lo que resulta en pérdidas económicas significativas debido a transacciones fraudulentas, así como en la pérdida de confianza de los clientes en la seguridad de sus servicios financieros.

Acciones Actuales:

La institución financiera está utilizando reglas de detección de fraude basadas en patrones conocidos y transacciones sospechosas. Sin embargo, estas reglas tienen una tasa de detección del 85% con un 15% de falsos positivos, lo que significa que algunas transacciones legítimas están siendo identificadas erróneamente como fraudulentas.

Acción Buscada:

La institución financiera busca implementar un sistema de detección de fraude basado en Machine Learning para mejorar la precisión y eficacia de la detección de transacciones fraudulentas. Este sistema utilizará algoritmos avanzados de ML para analizar el comportamiento de compra, patrones de gasto, ubicaciones de transacciones y otros datos relevantes para identificar transacciones sospechosas con mayor precisión.

KPIs - Indicadores de Negocio:

El KPI directo definido es la tasa de detección de fraudes verdaderos (TPR - True Positive Rate), que representa el porcentaje de transacciones fraudulentas detectadas correctamente por el sistema de detección. Un TPR objetivo del 95% se consideraría satisfactorio, lo que indica que el sistema es capaz de identificar la gran mayoría de las transacciones fraudulentas.

Mínimos Esperados:

Se espera un aumento del TPR del 10% en comparación con las métricas actuales, lo que significaría reducir la cantidad de fraudes no detectados y mejorar la capacidad del sistema para identificar transacciones fraudulentas con mayor precisión.

Validación:

Para determinar si la solución es aceptable, se establecerá un umbral máximo para la tasa de falsos positivos, asegurando que el sistema no esté generando una cantidad excesiva de alertas falsas que puedan afectar negativamente la experiencia del cliente y aumentar los costos operativos de la institución financiera.

Experimentación:

Se realizarán pruebas experimentales utilizando conjuntos de datos históricos para evaluar el rendimiento del sistema de detección de fraude basado en Machine Learning. La frecuencia de las

pruebas será mensual, y se esperará ver una mejora continua en la precisión y eficacia del sistema a lo largo del tiempo.

Productivización:

En caso de una experimentación satisfactoria, la solución se implementará en producción y se integrará en los sistemas de detección de fraude existentes de la institución financiera. Se proporcionará una interfaz de usuario para que los analistas de fraude revisen y gestionen las alertas generadas por el sistema de ML de manera recurrente.

Equipo de Trabajo:

El equipo de trabajo estará compuesto por científicos de datos, ingenieros de software y analistas de fraude, quienes trabajarán conjuntamente de manera continua y diaria para desarrollar, implementar y mantener la solución de detección de fraude basada en Machine Learning.

Detalle del Caso de Uso:

Se analizarán los patrones de comportamiento de compra, las ubicaciones de transacciones, la cantidad de transacciones y otros datos relevantes para identificar señales de actividad fraudulenta en las transacciones financieras de los clientes. Los datos transaccionales y los registros de actividad serán los principales orígenes de datos utilizados en el análisis y entrenamiento del modelo de ML.

Identificación de Orígenes de Datos:

Desde un punto de vista funcional, los principales orígenes de datos que se utilizarían para la detección de fraude en transacciones financieras incluirían:

- 1. Registros de Transacciones: Este conjunto de datos contendría información detallada sobre todas las transacciones realizadas por los clientes, incluyendo la fecha, hora, ubicación, monto de la transacción, tipo de tarjeta utilizada (crédito/débito), etc.
- 2. Historial de Actividad del Cliente: Este conjunto de datos incluiría información sobre el comportamiento histórico de compra de cada cliente, como la frecuencia de transacciones, los patrones de gasto, las categorías de compra preferidas, etc.
- 3. Datos de Ubicación: Los datos de ubicación podrían provenir de dispositivos móviles o sistemas de geolocalización asociados con las transacciones, lo que proporcionaría información sobre la ubicación física del cliente al realizar la transacción.
- 4. Datos Demográficos y de Perfil del Cliente: Estos datos incluirían información demográfica, como edad, género, ingresos, historial crediticio, etc., así como cualquier otra información relevante del perfil del cliente.

Desarrollo del Caso de Uso:

Para mejorar la detección de fraude en transacciones financieras, se realizarán los siguientes puntos intermedios o de seguimiento:

1. Análisis de Patrones de Comportamiento: Se investigará si ciertos patrones de comportamiento, como cambios inusuales en los hábitos de gasto o la aparición de nuevas categorías de compra,

podrían indicar actividad fraudulenta.

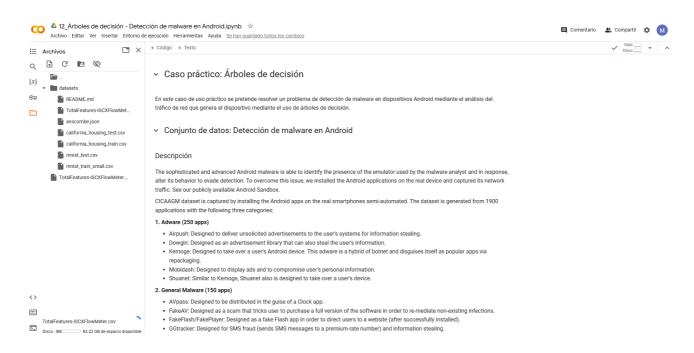
- 2. Validación de Alertas: Se realizarán pruebas para verificar la validez de las alertas generadas por el sistema de detección de fraude basado en Machine Learning, utilizando conjuntos de datos históricos y simulaciones de casos de fraude conocidos.
- 3. Optimización de Modelos de Machine Learning: Se explorarán diferentes algoritmos de ML y técnicas de preprocesamiento de datos para mejorar la precisión y eficacia del sistema de detección de fraude.
- 4. Evaluación de Rendimiento: Se evaluará el rendimiento del sistema en términos de tasa de detección de fraudes verdaderos, tasa de falsos positivos, etc., y se realizarán ajustes según sea necesario.

Aporte Esperado por Big Data:

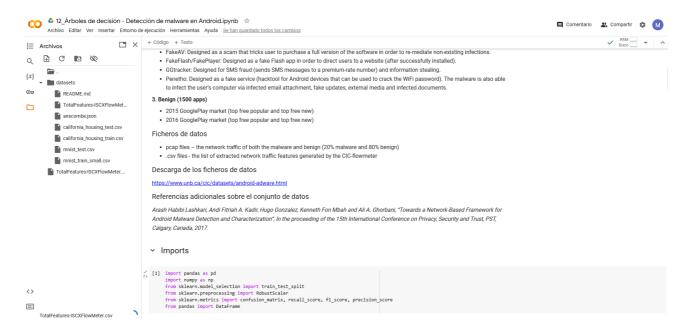
El equipo de Big Data espera mejorar significativamente la capacidad de la institución financiera para detectar y prevenir fraudes en transacciones financieras mediante el uso de técnicas avanzadas de análisis de datos y Machine Learning. Se espera que esto reduzca las pérdidas económicas asociadas con transacciones fraudulentas, mejore la confianza de los clientes y proteja la reputación de la institución financiera en el mercado. Además, se espera que el uso de Big Data permita una detección más temprana de fraudes y una respuesta más rápida ante posibles amenazas, lo que minimizará el impacto negativo en la institución financiera y sus clientes.

2. Resolver usando python el siguiente problema de ML:

Detección de malware en Android



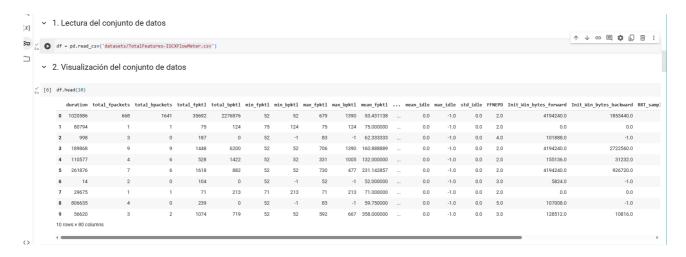
Cargamos las librerías



ejecuto la función train_val_test_split y divido la variable objetivo. También me dará la métrica con los datos preparados o sin preparar:



leemos el conjunto de datos del fichero datasets/TotalFeatures-ISCXFlowMeter.csv

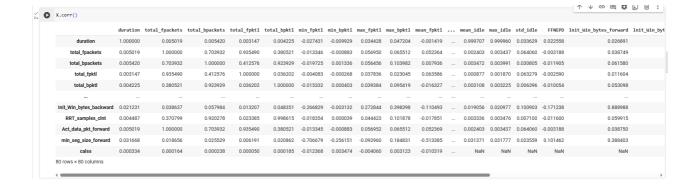


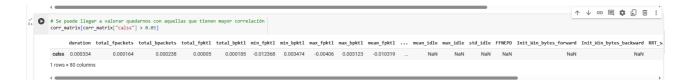
la información de cada variable (nombre, no nulo y tipo de dato):

```
| Cass | Fands | Color | Color
```

buscamos correlaciones



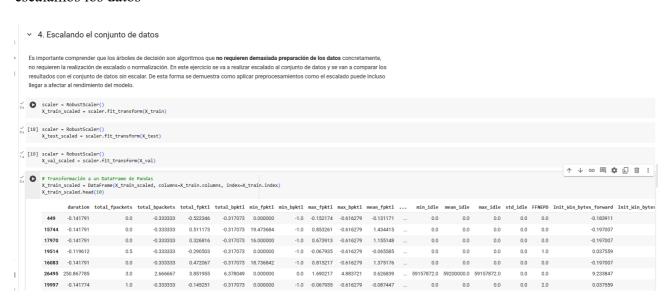




dividimos el conjunto de datos



escalamos los datos



	duration	total_fpackets	total_bpackets	total_fpktl	total_bpktl	min_fpktl	min_bpktl	max_fpkt1	max_bpktl	mean_fpktl	 min_idle	mean_idle	max_idle	std_idle	F
count	16877.000000	16877.000000	16877.000000	16877.000000	16877.000000	16877.000000	16877.000000	16877.000000	16877.000000	16877.000000	 1.687600e+04	1.687600e+04	1.687600e+04	1.687600e+04	16876.0
mean	62.408948	2.541625	2.019533	8.299416	39.583684	3.729538	0.019233	0.360583	1.690014	0.327258	 1.410876e+07	1.427730e+07	1.449960e+07	2.357698e+05	0.3
std	1256.573685	103.637866	64.865012	759.545244	1459.301115	7.533917	1.758742	0.709434	4.393656	0.633235	 2.972311e+08	2.974012e+08	2.973193e+08	5.124242e+06	0.9
min	-0.141791	-0.500000	-0.333333	-0.667598	-0.317073	-2.789474	-1.000000	-0.296196	-0.616279	-0.351199	 0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.0
25%	-0.141791	0.000000	-0.333333	-0.494413	-0.317073	0.000000	-1.000000	-0.152174	-0.616279	-0.131171	 0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.0
50%	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	 0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.0
75%	0.858209	1.000000	0.666667	0.505587	0.682927	1.000000	0.000000	0.847826	0.383721	0.868829	 0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.0
max	143957.841224	12048.500000	4748.000000	92588.206704	120702.756098	70.421053	25.245283	3.483696	15.558140	5.530324	 3.406373e+10	3.410000e+10	3.406373e+10	2.110000e+08	18.0