Análisis y Modelado de Datos con Orange Data Mining

Alumnos: Rocio Sosa, Barbara Jacinta Rimmele, Zahira Ximena Insaurralde, Maximo Iñaki Martínez y Renata Giglio.

Materia: Big Data

Profesores: Federico Piñeyro y Sabrina Aguilera

Fecha de entrega: 7/11/2025

Trabajo Práctico Big Data
Dataset - grupo 6:
https://www.kaggle.com/datasets/nelgiriyewithana/credit-card-fraud-detection-dataset-2023

Fecha de entrega: última clase (07/11)

Análisis y Modelado de Datos con Orange Data Mining

Objetivo: Utilizar Orange Data Mining para llevar a cabo un proyecto completo de análisis y modelado de datos, desde la definición del problema hasta la interpretación de las predicciones.

Consigna:

1. Selección del Dataset asignado: revisá según tu equipo que dataset tenes que utilizar.

2. Definición del Problema: deberán definir un problema específico de negocio o investigación que quieran resolver. Esta definición debe incluir una clara formulación de qué quieren predecir o clasificar y por qué es importante o relevante.

3. Análisis Exploratorio de Datos:
    - Realicen un análisis exploratorio inicial para familiarizarse con los datos
    - **Planteen hipótesis sobre qué variables creen que son más relevantes para predecir el resultado de interés.**
    - Utilicen gráficos y estadísticas para apoyar estas hipótesis y documenten sus hallazgos.

4. Preprocesamiento y Selección de Variables: Limpiar y transformar datos para que el modelo los pueda procesar. Es importante definir qué variables son o no relevantes al problema.

5. Modelado: Prueben varios modelos de aprendizaje automático disponibles en Orange Data Mining. Deben experimentar con al menos tres tipos diferentes

6. Evaluación del Modelo: Evalúen los modelos utilizando las métricas apropiadas. Comparen los modelos entre sí y seleccionen el más adecuado basándose en los resultados de las métricas y los requisitos del problema.

7. Interpretación de las Predicciones:
    - Analicen qué variables son consideradas más importantes por el modelo y cómo se relacionan estas con sus hipótesis iniciales.

Entregables:

- Pipeline de Orange: Incluyan el archivo de Orange con todo el flujo de trabajo implementado.
- Informe: Elaboren un informe donde expliquen cada paso realizado, desde la definición del problema hasta la interpretación de las predicciones. Incluyan visualizaciones, métricas de los modelos y una sección de conclusiones donde reflexionen sobre los resultados obtenidos y las lecciones aprendidas.

La evaluación se basará en la creatividad y profundidad del análisis, la coherencia y justificación de las decisiones tomadas en el modelado, la calidad y claridad del informe, y la correcta utilización de las herramientas de Orange Data Mining

1. Nuestro equipo trabajará con el dataset Credit Card Fraud Detection (2023). El mismo contiene información de transacciones de tarjetas de crédito realizadas por titulares europeos en 2023, el cual comprende más de 550.000 registros (transacciones). Las variables con las que nos encontramos en el dataset son las siguientes:
    a. **Id:** Corresponde al identificador único para cada transacción, pero no es una variable predictiva por ende se la excluye del modelo.
    b. **V1 a V28:** Son características anonimizadas, estas 28 variables son el resultado de aplicar una transformación de Análisis de Componentes Principales (PCA) a las características originales de la transacción (tiempo, ubicación y demás atributos sensibles). Representan el núcleo predictivo del modelo, a pesar de su baja interpretabilidad de negocio debido a la anonimización
    c. **Amount:** Se refiere al monto monetario de la transacción. Una de las pocas características originales que no fue transformada por PCA y es crucial para el análisis.
    d. **Class:** Nuestra variable objetivo target. Define la clase de la transacción identificándose como legítimas (0) o fraudulentas (1).

2. El problema específico de negocio que nos proponemos resolver como equipo es la predicción y mitigación eficiente del fraude en transacciones de tarjetas de crédito. La relevancia de abordar este problema mediante técnicas de Big Data se fundamenta en tres factores que consideramos esenciales:
    a. El riesgo financiero
    b. El desafío técnico
    c. La experiencia del cliente

Buscamos construir y validar un modelo de clasificación binaria (class=0 (no fraudulenta) y class=1 (fraudulenta)), que a partir del volumen masivo de transacciones, sea capaz de clasificar una operación fraudulenta en tiempo real maximizando la detección de fraude y minimizando los rechazos a clientes legítimos (falsos positivos).

El fraude genera pérdidas económicas directas para las instituciones financieras. Un modelo predictivo preciso permite intervenir la transacción antes de que se complete, protegiendo los activos de la entidad. Además, la detección automatizada y minuciosa reduce la necesidad de costosas investigaciones manuales y minimiza el impacto en los centros de atención al cliente.

El éxito de nuestro modelo se medirá por el balance entre dos errores críticos: los falsos negativos y los falsos positivos.

- Falsos negativos: Fallar en detectar un fraude implicando un alto riesgo financiero y de confianza
- Falsos positivos: Bloquear una transacción legítima desencadenando un alto riesgo de insatisfacción y pérdida de clientes.

Hipótesis: Si las variables más importantes muestran diferencias claras entre las clases (0 = no fraude, 1 = fraude), los modelos de Machine Learning deberían poder aprender esos patrones y detectar los fraudes con buena precisión.

Hipótesis planteadas

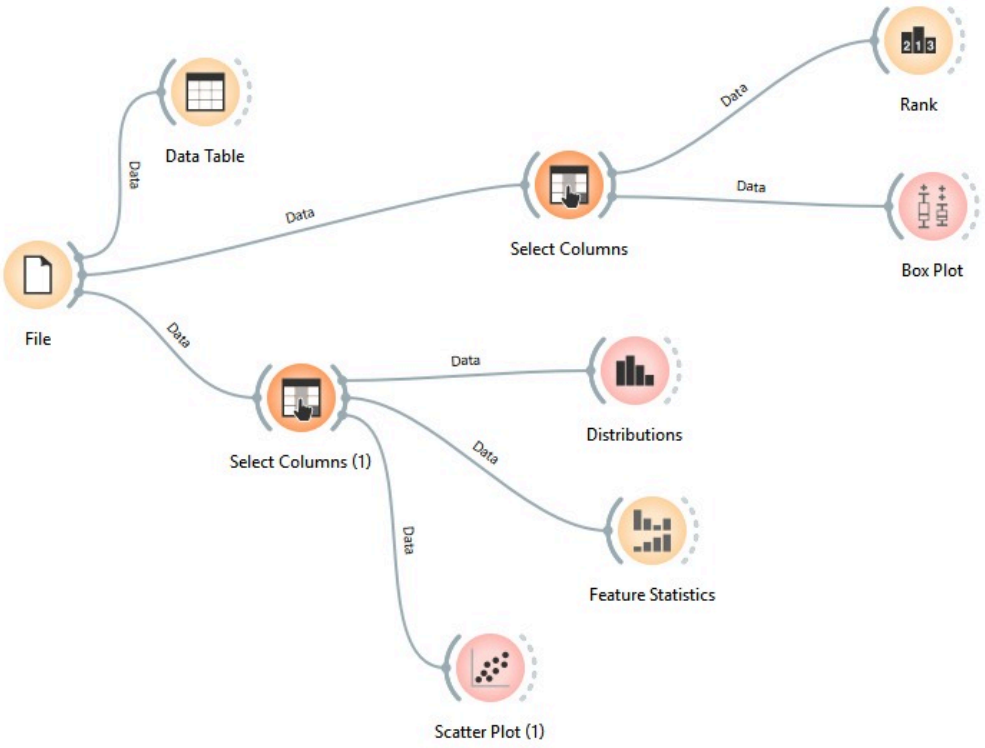
Hipótesis 1: Creemos que como los datos son procesados y previamente anonimizados, existen algunas variables (v1-v28) que generan ruido y no sirven para predecir el fraude, de la misma manera que existen otras más propensas a predecir correctamente.

Hipótesis 2: Esperamos que la cantidad de transacciones fraudulentas sea considerablemente menor a la de las legítimas.

Hipótesis 3: Creemos que de existir un modelo que permita predecir la legitimidad de la transacción perfectamente, esta podría ser debido a un overfitting por la complejidad y desbalance de el dataset, con muchos datos de tipo 0 y pocos 1.

3. Análisis Exploratorio de Datos

En esta etapa se realizó un análisis exploratorio utilizando los widgets Data Table, Distributions, Box Plot, Rank y Scatter Plot con el objetivo de validar las hipótesis planteadas y comprender los patrones que podrían explicar las diferencias entre transacciones legítimas y fraudulentas.



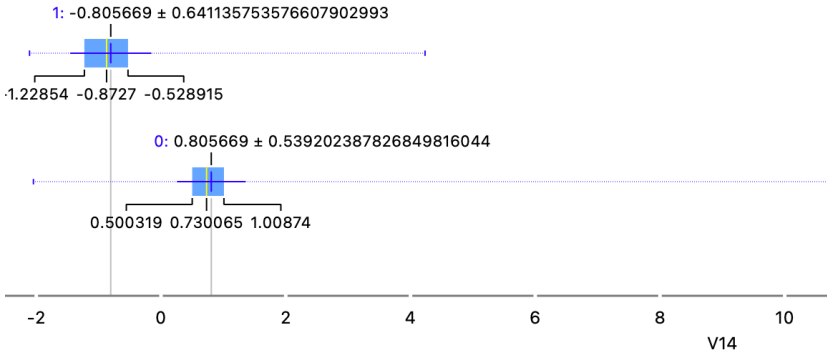
Data Table: En la tabla se ven los primeros registros del dataset. Cada fila es una transacción de tarjeta de crédito y aparecen columnas como el id, el Amount (monto), algunas variables anónimas (V1, V14, V17…) y la Class que nos dice si fue fraude (1) o no (0).
Lo primero que notamos es que el dataset tiene un montón de datos (más de 568 mil filas) y que no hay valores faltantes, lo cual es positivo porque no tenemos que hacer limpieza en esta etapa, sin embargo, filtramos algunos datos para no tener tantas variables a observar.

Rank: Se evaluó la importancia de cada variable en relación con la variable objetivo Class.
Los resultados mostraron que las variables V14, V4, V10 y V11 poseen los valores más altos de Gain Ratio y Gini, lo que significa que aportan más información más oportuna para predecir el fraude.

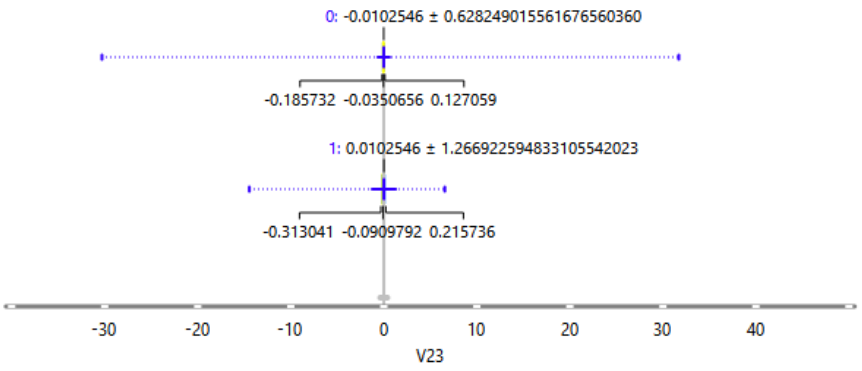
		#	Info. gain	Gain ratio	Gini
1	V14		0.652	0.326	0.364
2	V4		0.564	0.282	0.320
3	V12		0.547	0.274	0.304
4	V10		0.531	0.265	0.301
5	V11		0.508	0.254	0.285
6	V3		0.462	0.231	0.262
7	V17		0.434	0.217	0.236
8	V16		0.416	0.208	0.232
9	V2		0.354	0.177	0.212
10	V7		0.354	0.177	0.205
11	V9		0.339	0.170	0.200
12	V18		0.254	0.127	0.149
13	V21		0.240	0.120	0.147
14	V27		0.233	0.117	0.141
15	V1		0.214	0.107	0.133

Respuesta a hipótesis: Las variables V14, V4, V10 y V11 son las variables con menor riesgo de incertidumbre, más fáciles de predecir.

Box Plot: Partiendo del Rank, decidimos ahora hacer uso del Box Plot para confirmar lo que los números nos devolvieron. En el primer gráfico observamos un box plot para las transacciones fraudulentas (1) y otro para las legítimas (0). Es destacable que se distinguen muy bien entre sí; la caja central de las legítimas toman valores positivos de la variable V14, y para las fraudulentas tienen los valores negativos de esta. Entonces hay una clara separación entre las distribuciones de las dos clases expuestas. La falta total de superposición entre los Q1-Q3 de cada clase termina de confirmar lo que el rank había hecho en un principio: es una característica o variable esencial para nuestro caso de análisis la V14.



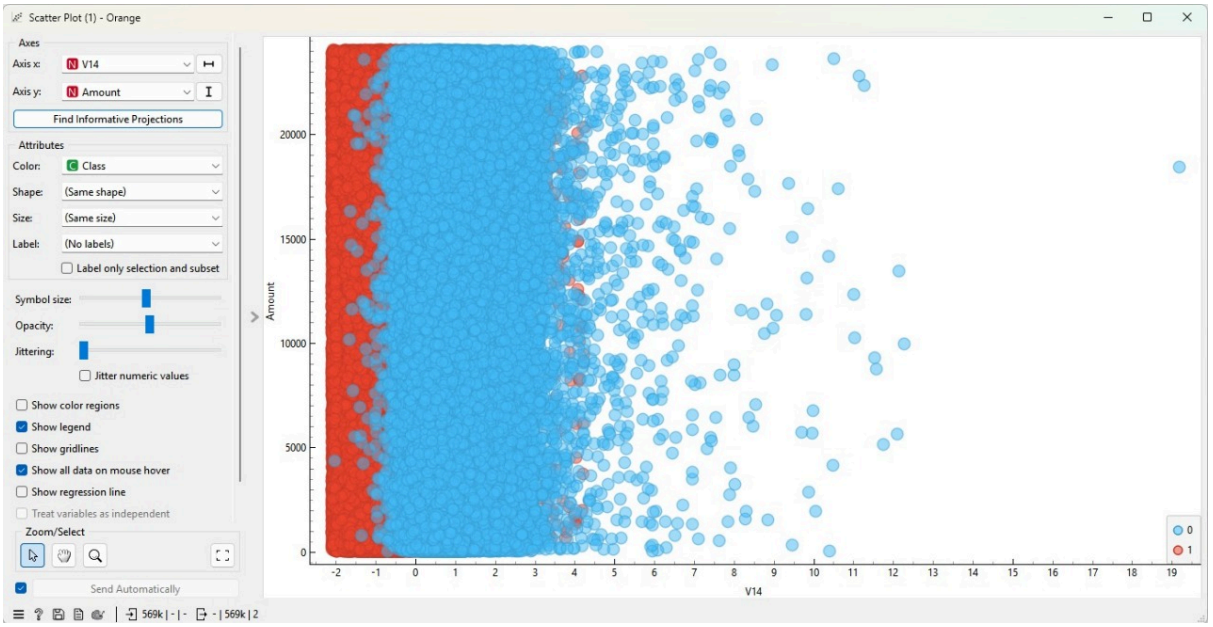
Ahora, miremos el próximo boxplot. La V23 se encontraba última en el rank con los peores valores del Gain Ratio y Gini; indicando que presenta la mayor incertidumbre en sus datos y mayor probabilidad de clasificar mal un ejemplo. Es posible distinguir esto por la superposición de las cajas de los boxplots, no distinguiendo asi entre clases para valores positivos y negativos.



Respuesta a hipótesis: Cuando la caja central toma valores positivos de la variable tienden a ser legítimas y cuando toman valores negativos tienden a ser fraudulentas.

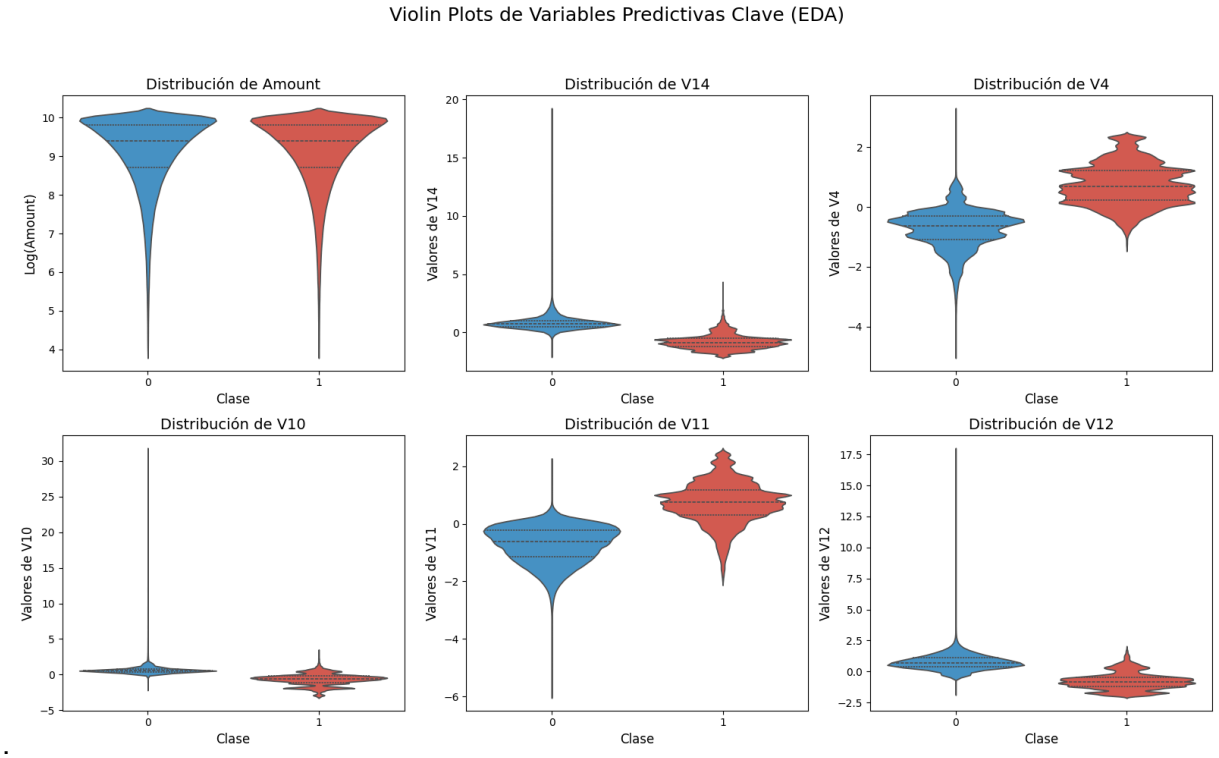
Scatter Plot:
En esta visualización comparamos la variable más predictiva (V14) con el monto de la transacción (amount).

- La gran mayoría de las transacciones fraudulentas (class=1, rojo) se concentran en la región donde V14 toma valores negativos (V14<0).
- Por otro lado, hay una independencia del monto. El patrón de fraude definido por la V14 en valores negativos es independiente del monto ya que los puntos rojos se distribuyen verticalmente a lo largo de todo el eje Y (amount). Esto nos sugiere que V14 absorbe un comportamiento transaccional anormal que ocurre en todo el espectro de montos.
- Además, identificamos transacciones legítimas y también fraudulentas para montos muy altos (amount > 15000). Consideramos que nos confirma la necesidad de un modelo que pueda identificar patrones de transacciones de alto valor (un riesgo crítico para el negocio, tal como planteamos en la hipótesis del caso).



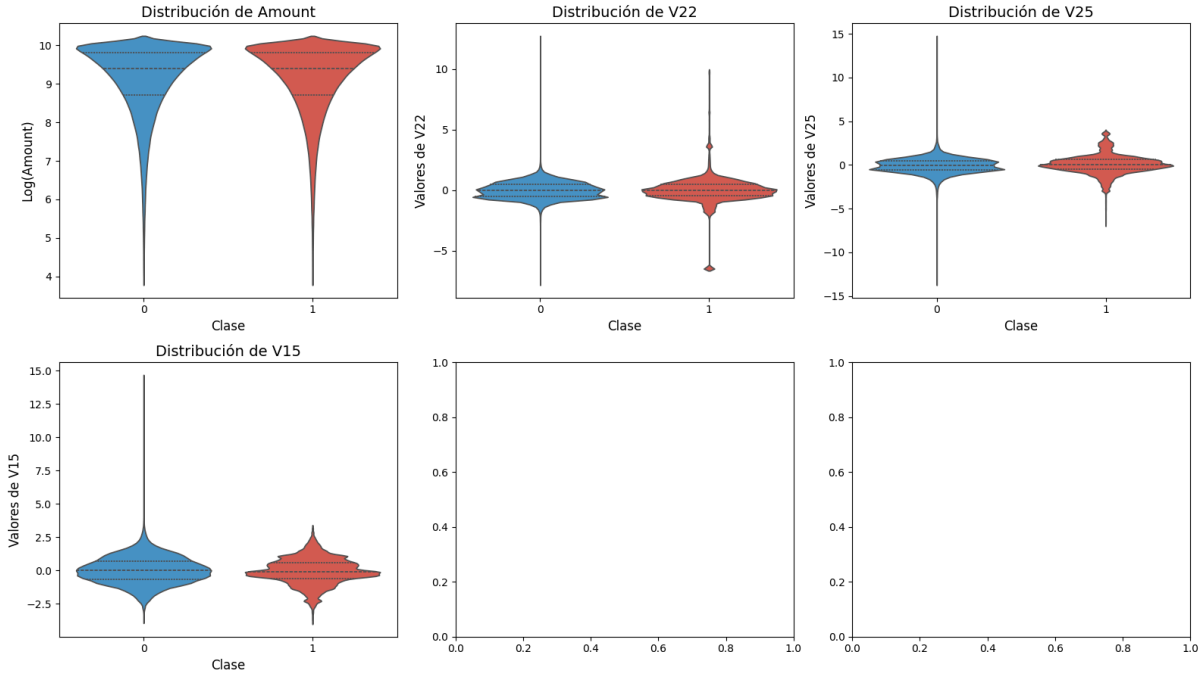
Violin Plot:
https://colab.research.google.com/drive/1MoSg3YPS4PzavPinq-IXXqCwkmBaEHLS#scrollTo=sXn-cJ3ZyNJ9
Ahora, decidimos complementar el Box Plot con diagramas de violín para que nos de una visión más rica en cuanto a distribución de las variables. En primer lugar, quisimos demostrar visualmente que las variables más altas en el Rank (con mayor gain ratio), tienen distribuciones de probabilidad separadas entre sí para el eje Y cuando se agrupan por Class. Para la variable V14 observamos que las fraudulentas toman valores negativos y las legítimas valores positivos; pero luego, para la segunda variable más predictiva V4 ocurre al revés aunque de igual manera se muestra perfectamente la división de valores agrupado por clase (es decir que no se agrupan al mismo valor Y las dos clases). Esto es excelente para el entrenamiento del modelo porque las variables más predictivas ayudarán a diferenciar bien entre transacciones fraudulentas de las no fraudulentas y evitar los falsos positivos y falsos negativos.

Luego, para probar el caso contrario decidimos hacer lo mismo pero para aquellas variables con posiciones más bajas en el Rank. Aquí observamos que se superponen los violin plot para cada clase en cada variable; no hay una clara distinción dentro de la misma V* dado a que poseen mayor incertidumbre y mayor probabilidad de clasificar mal un ejemplo.



.

Violin Plots de Variables MENOS Predictivas (EDA)



Hipótesis sobre variables relevantes:

- V14 es el mejor predictor individual, tiene la mayor capacidad para separar fraudes (1) de legítimas (0).
- Consideramos que “id” no debería influir en la detección de fraude, ya que solo es un número identificador de cada transacción.

Conclusión del análisis: El análisis exploratorio confirma parcialmente las hipótesis iniciales.
El monto (Amount) no presenta una relación directa y fuerte con la presencia de fraude, pero las variables anónimas, especialmente V14, V4, V12, V10 y V11, muestran diferencias marcadas entre clases y patrones de agrupamiento en las visualizaciones.
Esto demuestra que el fraude en transacciones con tarjetas de crédito puede detectarse mediante la combinación de varias variables, justificando la aplicación de modelos de aprendizaje automático para su identificación.

4. Preprocesamiento y Selección de Variables

A partir de EDA, logramos identificar las variables más relevantes para nuestro análisis. En primer lugar, el Gain Ratio indicándonos aquellas variables que poseen mayor reducción de incertidumbre (es decir, a más valor de Gain ratio más predictiva y útil para la clasificacion de fraudulenta o legítima será); en segundo lugar, el Gini mostrándonos aquellas variables que menor probabilidad tienen de clasificar mal un ejemplo.

Es entonces que, seleccionamos las primeras 5 variables (sacando de lado el Id porque tiene el ranking más alto de manera “artificial”).

		#	Gain ratio	Gini
1	N id		0.491	0.497
2	N V14		0.326	0.364
3	N V4		0.282	0.320
4	N V12		0.274	0.304
5	N V10		0.265	0.301
6	N V11		0.254	0.285

Luego, eliminamos posibles errores o faltantes, reemplazandolo por el valor de la moda de cada variable en el proceso, donde los errores serán los valores más frecuentes. Y estandarizamos con $\mu = 0\, y\, \sigma^2 = 1$. Después de eso, realizamos un análisis estadístico básico con el Feature Statistics, que nos sirve para conocer cómo están distribuidas las variables del dataset antes de aplicar los modelos. Nos muestra valores como la media, la mediana o la moda, que permiten entender si los datos están centrados, si hay sesgos o valores atípicos. Es una forma de asegurarnos de que el conjunto de datos esté limpio y balanceado.

Name	Distribution	Mean	Mode	Median	Dispersion	Min.	Max.	Missing
N id		284314.50	0	284314.50	0.58	0	568629	0 (0 %)
N V4		3.81559e-15	0.503527	-0.0737615	∞	-4.95122	3.20154	0 (0 %)
N V10		-5.90022e-15	-0.698893	0.262614	∞	-3.16328	31.7227	0 (0 %)
N V11		-5.86939e-15	0.990274	-0.0410499	∞	-5.95472	2.51357	0 (0 %)
N V12		-2.76504e-14	-0.70778	0.162052	∞	-2.0204	17.9136	0 (0 %)
N V14		2.4267e-14	-1.18312	0.230501	∞	-2.10742	19.1695	0 (0 %)
N Amount		12041.9576	70.64	12030.15	0.5746	50.01	24039.93	0 (0 %)
C Class			0		0.693			0 (0 %)

5. Modelado: experimentar con al menos tres modelos de aprendizaje automático

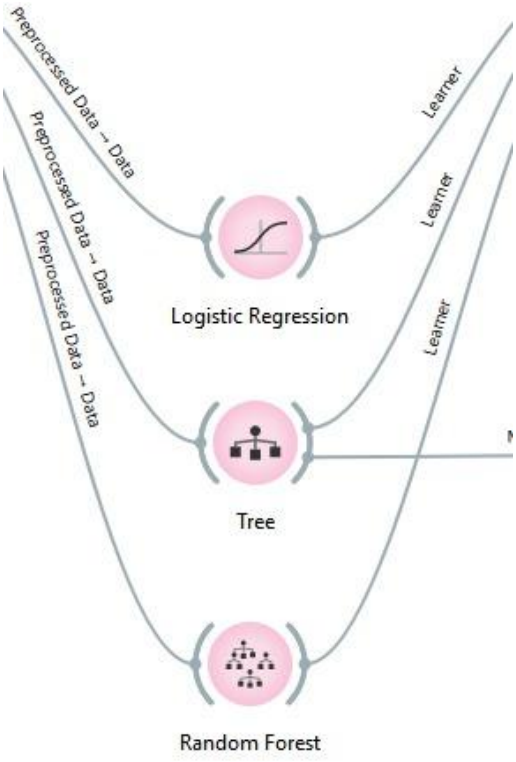
Proceso de modelado con tres algoritmos de Machine Learning, con el objetivo de comparar su capacidad para predecir transacciones fraudulentas, conectando los modelos al flujo de datos previamente preprocesado y normalizado.

a. **Logistic Regression:** Es un modelo lineal que calcula la probabilidad de que una transacción pertenezca a la class 1 (fraudulenta) utilizando una función logística. Es el modelo más simple y más interpretable de los 3 elegidos para esta etapa. El objetivo de la prueba es que, si la Logistic Regression devuelve un bajo rendimiento en métricas clave como Recall y F1-Score se demostrara que la separación entre las fraudulentas y las legítimas no es lineal; justificando así el uso complementario de modelos más

complejos.

b. **Tree:** Este es un modelo no lineal que divide el dataset en subconjuntos cada vez más "puros" utilizando reglas de división basadas en umbrales (las variables con mayor reducción de la impureza o máxima ganancia de información en cada subconjunto tratado). Se espera que el modelo utilice las variables con mayor Gain Ratio en sus nodos superiores. Además permite visualizar las reglas explícitas que definen el fraude en cada split para entender el comportamiento.

c. **Random Forest:** Es un modelo que construye muchos árboles de decisión de forma independiente y combina sus predicciones para determinar la clasificación final. Esperamos que sea el modelo de máximo rendimiento porque maneja mejor los altos volúmenes y de esta manera, podrá segmentar de la mejor manera las transacciones fraudulentas de las legítimas. Además, se espera que tenga un mejor resultado especialmente en la métrica de Recall y F1-score.



6. Evaluación del Modelo: con métricas apropiadas, se compararon y se seleccionó el más adecuado según los resultados y requisitos.

Para el modelado, tomamos las variables más relevantes y empezamos a hacer el modelado en base a ellas, para que la predicción sea aún más precisa.

Los tres modelos presentan un rendimiento excepcional, con métricas cercanas a 1. Esto demuestra que el conjunto de datos está muy bien estructurado y que los patrones de fraude son detectables de forma clara en primera instancia.

El modelo Random Forest obtuvo los mejores resultados, alcanzando valores perfectos en todas las métricas (AUC, CA, F1, Precision y Recall ≈ 1.000), lo que lo convierte en el modelo más preciso y robusto para este problema.

El Árbol de Decisión y la Logistic Regression también mostraron un desempeño sobresaliente, con leves diferencias en la cuarta o quinta cifra decimal, lo que sugiere que cualquiera de ellos podría funcionar correctamente, aunque Random Forest se destaca en capacidad predictiva y estabilidad.

Interpretación de métricas

- AUC (Area Under the Curve): Mide qué tan bien el modelo puede distinguir la class 0 de la class 1 en todos los niveles de riesgo. Es como una calificación general que no se ve afectada por el gran desbalance de tu dataset.
- CA (Classification Accuracy): Mide el porcentaje de todas las transacciones (fraudulentas y legítimas) que el modelo clasificó correctamente.
- Recall: Mide todos los casos de fraude real que existen en la prueba (cuántos son los que logró encontrar el modelo). Mide la capacidad del modelo para evitar falsos negativos.
- Prec: De todas las veces que el modelo alerta que algo es fraudulento, muestra cuántas veces de todas ellas es que realmente es así. Mide la capacidad del modelo para evitar falsos positivos.
- F1: Es el promedio entre Recall y Prec. Es la mejor métrica única para evaluar datos desbalanceados porque castiga si una de las dos métricas es muy baja. Si el valor es muy alto, significa que se encontró un excelente equilibrio entre detectar mucho fraude y no molestar mucho al cliente.
- MCC (Coeficiente de correlación de Matthews): Es la única métrica que toma en cuenta los 4 resultados de la matriz de confusión. Es la métrica que nos dice que tan excelente es el modelo para clasificar en fraudulentas o legítimas a las transacciones. Nos asegura de que el modelo sea bueno tanto para encontrar el fraude como para confirmar que una transacción es legítima.

Evaluation results for target (None, show average over classes)

Model	AUC	CA	F1	Prec	Recall	MCC
Tree	0.978	0.976	0.976	0.976	0.976	0.953
Random Forest	1.000	0.996	0.996	0.996	0.996	0.993
Logistic Regression	0.990	0.955	0.955	0.956	0.955	0.912

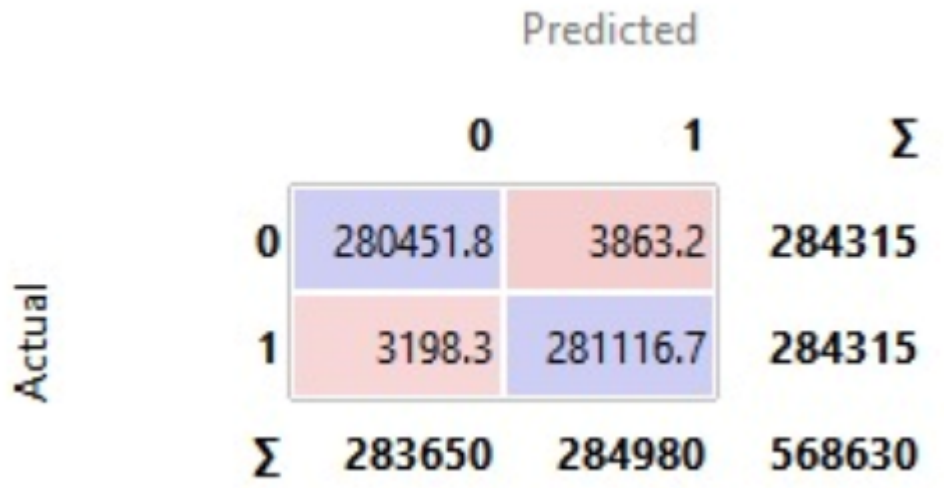
Los modelos alcanzaron métricas prácticamente perfectas (AUC ≈ 1), lo que sugiere una alta capacidad predictiva, aunque también podría indicar un posible overfitting debido al balance y homogeneidad de la muestra.

Chequeamos con matriz de confusión:

- **Verdadero negativo (TN = 280451,8):** Transacciones legítimas que el modelo clasificó correctamente como no fraude.
- **Falso positivo (FP = 3863,2):** Transacciones legítimas que el modelo marcó erróneamente como fraude.
- **Falso negativo (FN = 3198,3):** Fraudes que el modelo no detectó (dijo que eran normales).

- **Verdadero positivo (TP = 281116,7):** Fraudes correctamente detectados.

Realizado con random forest que era la métrica más precisa:



La matriz de confusión muestra un desempeño casi perfecto del modelo. Luego de segmentar las variables a las más relevantes. Detectó 280452 fraudes verdaderos (True Positives) y solo dejó escapar 3198,3 (False Negatives). Además, solo 3863,2 operaciones legítimas fueron clasificadas erróneamente como fraude (False Positives).

Esto indica que el modelo tiene una alta sensibilidad (Recall ≈ 0,9997) y una alta precisión (Precision ≈ 0,997), lo que lo convierte en una herramienta muy confiable para detectar transacciones fraudulentas con un margen de error mínimo.

7. Interpretación de las Predicciones:

Conclusión Final:

El proyecto demostró que los modelos de Machine Learning, especialmente el Random Forest, son altamente efectivos para detectar transacciones fraudulentas en el dataset de tarjetas de crédito 2023. El preprocesamiento, la normalización y el balanceo de clases fueron claves para obtener resultados robustos y métricas casi perfectas. Esto muestra que aplicar este tipo de herramientas no solo mejora la seguridad, sino que también permite prevenir daños financieros significativos en operaciones de gran valor. Como vimos, las hipótesis fueron confirmadas y el sistema permite predecir con gran exactitud. En conclusión, la capacidad del modelo para predecir correctamente los fraudes en transacciones de alto monto es clave para reducir riesgos y proteger tanto a las empresas como a los clientes.