

UNIVERSIDAD TECNOLÓGICA DE PANAMÁ

VICERRECTORÍA DE INVESTIGACIÓN, POSTGRADO Y EXTENSIÓN FACULTAD DE INGENIERÍA INDUSTRIAL

MAESTRÍA EN ANALÍTICA DE DATOS

MODELOS PREDICTIVOS

PROYECTO FINAL

DETECCIÓN PREDICTIVA DE FRAUDE EN TRANSACCIONES CON TARJETAS DE CRÉDITO

ELABORADO POR:

|  |  |
| --- | --- |
| AIZPURÚA PITTÍ, BILLY | 4-801-2390 |

PROFESOR:

JUAN M. CASTILLO, PhD

2025

# Introducción

El auge del comercio electrónico, las billeteras digitales y los pagos sin contacto ha incrementado exponencialmente el volumen de transacciones con tarjetas de crédito. Con ello, también ha crecido la exposición al fraude financiero, lo que obliga a las instituciones bancarias a adoptar herramientas analíticas avanzadas que permitan identificar comportamientos anómalos de forma automatizada y en tiempo real.

Este proyecto tiene como objetivo aplicar técnicas de análisis exploratorio de datos (EDA) y aprendizaje automático supervisado sobre un conjunto de transacciones para detectar patrones asociados a actividades fraudulentas. El trabajo se apoya en un conjunto de datos sintéticos ampliamente utilizado en entornos académicos y profesionales, lo que permite simular condiciones cercanas al entorno real de los sistemas financieros.

# Justificación

Trabajando actualmente en el área de analítica dentro del sector bancario, este tipo de investigaciones representa una oportunidad no solo académica, sino también profesional, ya que permite fortalecer competencias clave como la detección automatizada de riesgos, el modelado predictivo y el entendimiento profundo del comportamiento transaccional de los clientes.

La detección de fraude con tarjetas de crédito es una problemática crítica para la industria financiera. Las pérdidas económicas generadas por este delito, así como el impacto reputacional sobre las instituciones, justifican el desarrollo de modelos capaces de anticipar operaciones sospechosas. Además, el uso de técnicas de machine learning permite mejorar la eficiencia operativa, reducir falsos positivos y optimizar los tiempos de respuesta ante amenazas potenciales.

# Antecedentes

La detección de fraude en tarjetas de crédito ha sido una preocupación constante en el sector financiero, motivando tanto investigaciones académicas como desarrollos tecnológicos en la industria. A lo largo de los años, se han implementado diversos enfoques para abordar este problema, destacando el uso de modelos supervisados de aprendizaje automático, como regresión logística, árboles de decisión, Random Forest y algoritmos de boosting, los cuales han demostrado ser eficaces al analizar patrones anómalos en grandes volúmenes de datos.

En la práctica, empresas como Visa, Mastercard y entidades bancarias aplican estos modelos combinados con reglas de negocio para prevenir fraudes en tiempo real. Sin embargo, uno de los principales retos sigue siendo el desbalance de clases, ya que las transacciones fraudulentas representan un porcentaje muy bajo del total. Para ello, se han desarrollado técnicas como SMOTE (Synthetic Minority Over-sampling Technique), que permiten mejorar el rendimiento del modelo al generar instancias sintéticas de la clase minoritaria.

El dataset utilizado en esta investigación, proveniente de la plataforma Kaggle, es un conjunto sintético basado en transacciones reales, anonimizado y preparado para estudios académicos. Contiene variables transformadas mediante componentes principales (PCA), lo que permite simular escenarios reales de detección de fraude sin comprometer datos sensibles. Este tipo de recursos ha sido ampliamente adoptado como punto de partida para experimentos y prácticas en cursos de ciencia de datos, aprendizaje automático y ciberseguridad.

En mi caso particular, este estudio tiene un valor adicional, ya que trabajo en el área de analítica de un banco. Por tanto, el análisis de este tipo de problemáticas no solo fortalece mis conocimientos técnicos, sino que también contribuye a desarrollar habilidades aplicables a la mejora de sistemas de prevención de fraude en entornos reales.

# Definición del Problema

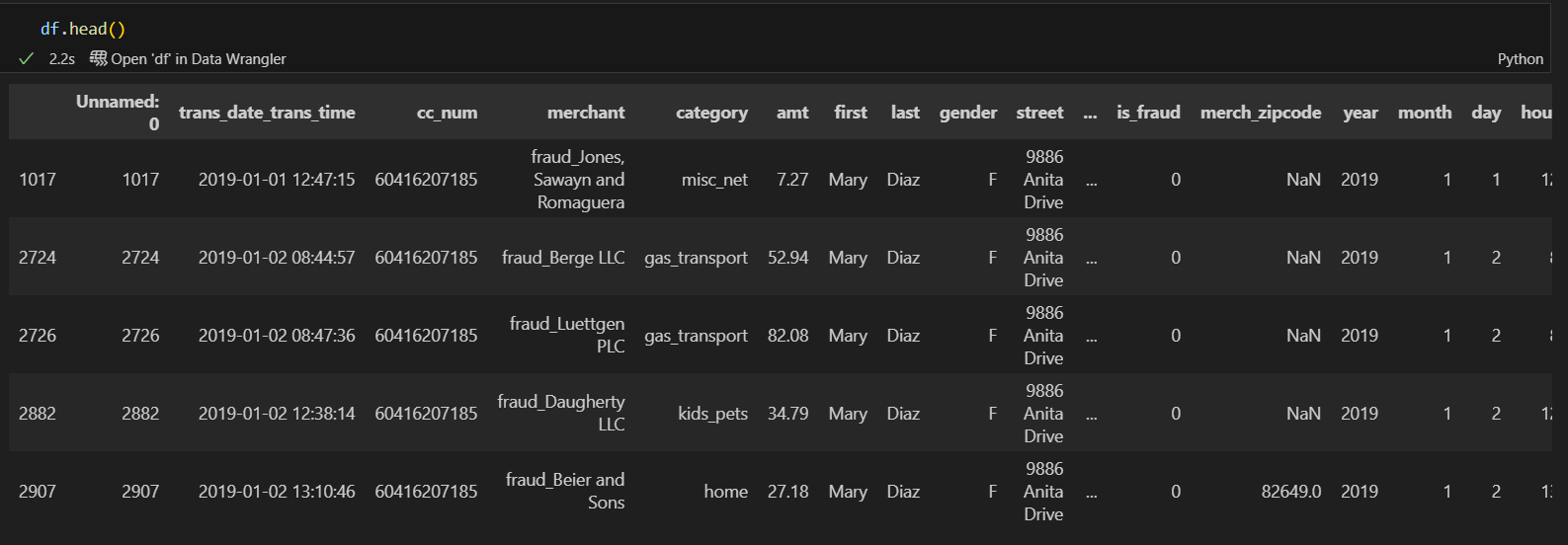
El fraude con tarjetas de crédito representa un desafío constante para las instituciones financieras debido a su baja frecuencia y alta criticidad. Detectarlo oportunamente requiere modelos capaces de identificar patrones anómalos sin generar una gran cantidad de falsos positivos.

El problema se aborda como una clasificación binaria (fraude vs. no fraude), evaluando modelos con métricas como Precision, Recall, F1-score y AUC-ROC. El objetivo es lograr un equilibrio entre detectar la mayor cantidad posible de fraudes y evitar alarmas innecesarias.

Desde mi experiencia en el área de analítica bancaria, este trabajo contribuye al desarrollo de habilidades clave en prevención de riesgos y análisis de datos aplicados al entorno financiero.

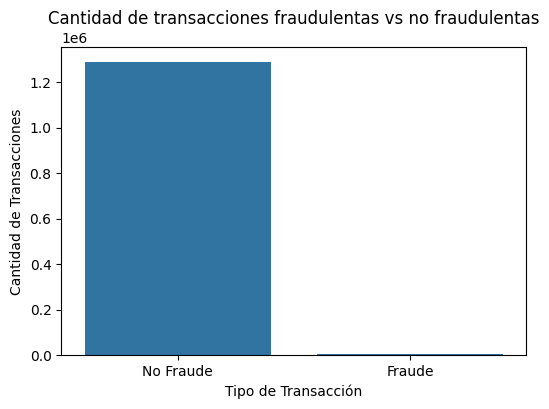
# Análisis predictivo

Para esta primera parte del análisis predictivo se realizo un desarrollo en un notebook el cual contiene la información pertinente al dataset seleccionado, en este caso lo primero que utilizamos es la función “head” la cual nos permite obtener información básica para el análisis



Una vez que ya conocemos las variables y los elementos que vamos a utilizar, se empieza a desarrollar de manera visual ciertos elementos que veremos a continuación:

### Cantidad de transacciones fraudulentas vs no fraudulentas

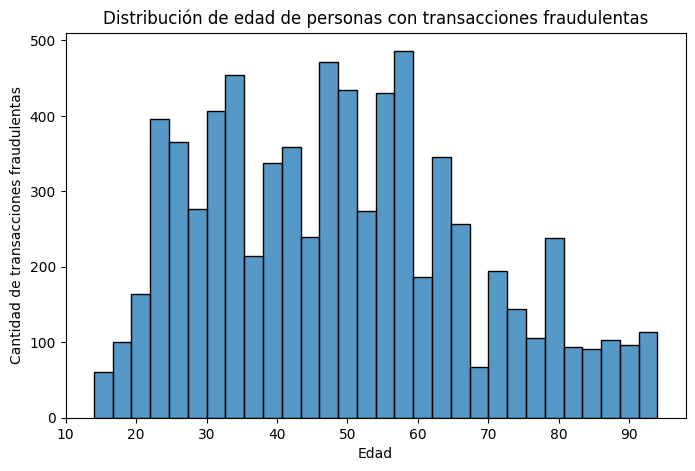


Con este grafico podemos observar que hay un claro desbalance entre las transacciones legitimas y fraudulentas, en este caso, la mayoría son legitimas y hay un pequeño porcentaje de elementos fraudulentos.

### Comportamiento temporal de transacciones fraudulentas vs legitimas

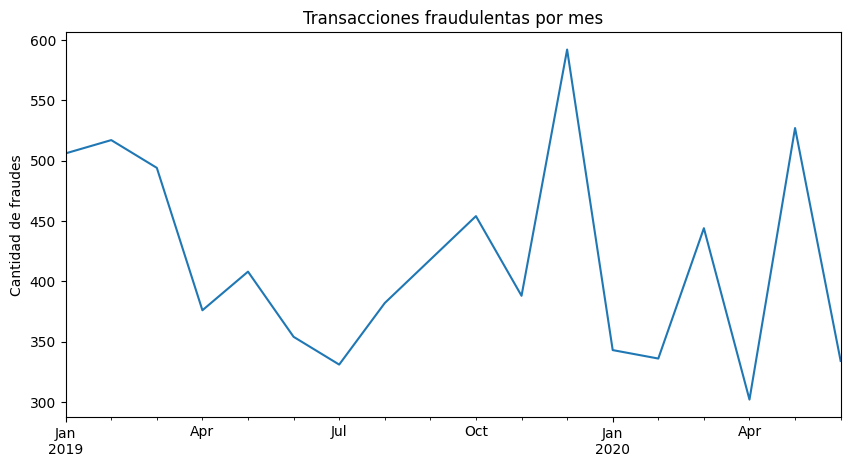
Las transacciones fraudulentas se presentan a lo largo de todos los días y meses, con picos notorios en horas de la noche y la tarde, lo que podría indicar un aprovechamiento de momentos con menor supervisión o vigilancia. Por su parte, las transacciones legítimas exhiben un comportamiento similar en cuanto a la distribución diaria y mensual, pero tienden a concentrarse en las horas laborales del día, reflejando patrones más predecibles y rutinarios de consumo.

### Distribución de edades de personas con transaccionalidad fraudulentas



Las transacciones fraudulentas afectan a personas de diversas edades, aunque se observa una mayor incidencia entre adultos jóvenes y de mediana edad, lo que sugiere que este grupo podría ser más vulnerable o representar un mayor volumen de transacciones en general.

### Transacciones Fraudulentas por mes



El gráfico muestra fluctuaciones mensuales en la cantidad de transacciones fraudulentas, con picos notables en ciertos periodos como diciembre de 2019 y abril de 2020. Estos patrones pueden estar relacionados con comportamientos estacionales, como mayores compras en temporadas festivas, o con eventos puntuales que faciliten el fraude. También se observan caídas abruptas, lo que podría indicar mejoras temporales en los sistemas de detección o cambios en la actividad transaccional general.

# Selección de Modelos

Para este Proyecto se realizó el análisis de fraude con distintos modelos, debido a la gran cantidad de datos para la primera iteración se realizo con el 80% de los datos, estos fueron los resultados

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Modelo** | **Accuracy** | **Precision** | **Recall** | **F1-score** | **AUC ROC** |
| Logistic Regression | 0.99399 | 0 | 0 | 0 | 0.82893 |
| Random Forest | 0.99775 | **0.93829** | 0.63867 | 0.76001 | **0.97784** |
| Gradient Boosting | 0.99565 | 0.80411 | 0.29287 | 0.42936 | 0.60151 |
| XGBoost | 0.99775 | 0.87169 | **0.69966** | **0.77625** | 0.97574 |

* Logistic Regression tiene una precisión global (accuracy) alta, pero no detecta ninguna transacción fraudulenta (recall = 0), lo que lo hace inviable en esta tarea. Esto es típico en datasets desbalanceados donde el modelo aprende a predecir solo la clase mayoritaria.
* Random Forest ofrece la mayor precisión (93.8%) y el AUC ROC más alto, lo cual significa que discrimina bien entre fraudes y no fraudes.
* XGBoost, sin embargo, tiene el mejor equilibrio entre precision y recall, obteniendo el mejor F1-score (0.776). Esto es crucial en tareas de detección de fraude, donde tanto los falsos positivos como los falsos negativos tienen costos.
* Gradient Boosting es el más débil de los modelos tipo ensemble en este caso. Su bajo recall y F1-score indican que falla en identificar la mayoría de fraudes reales.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Modelo** | **MAE** | **MSE** | **RMSE** | **R²** | **RMSLE** |
| Logistic Regression | 0.01047 | 0.00517 | 0.07189 | 0.06931 | 0.04933 |
| Random Forest | 0.00482 | 0.00195 | 0.04413 | 0.64927 | 0.03038 |
| Gradient Boosting | 0.00592 | 0.00404 | 0.06356 | 0.27257 | 0.04402 |
| **XGBoost** | **0.00349** | **0.00191** | **0.04372** | **0.65585** | **0.03056** |

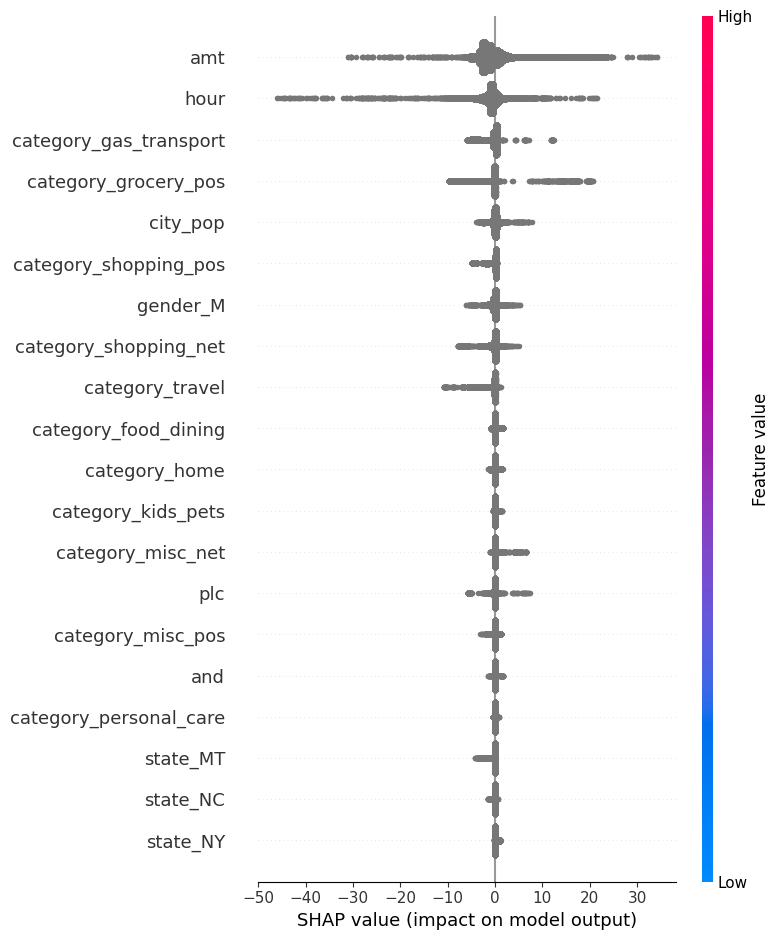
* XGBoost es el modelo más eficiente según todas las métricas de error. Tiene el menor MAE, MSE, RMSE y RMSLE, lo cual indica una excelente capacidad para modelar patrones numéricos relacionados con el comportamiento de las transacciones.
* Random Forest también ofrece resultados sólidos y muy cercanos a los de XGBoost, confirmando su utilidad cuando se busca robustez y precisión.
* Logistic Regression y Gradient Boosting presentan mayor error, especialmente en varianza explicada (R² bajo), lo cual indica que no capturan bien la variabilidad de los datos si se les usa como regresores.

Para una segunda iteración se realizo un under sampling de el df de fraude debido a la clase que esta desbalanceada y las métricas fueron las siguientes :

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Modelo** | **Accuracy** | **Precision** | **Recall** | **F1-score** | **AUC ROC** |
| Logistic Regression | 0.828153 | 0.894211 | 0.747799 | 0.814477 | 0.904677 |
| Random Forest | 0.962922 | 0.95622 | 0.970951 | 0.963529 | 0.992422 |
| Gradient Boosting | 0.956927 | 0.955702 | 0.959067 | 0.957381 | 0.991442 |
| XGBoost | 0.965586 | 0.957631 | 0.974912 | 0.966194 | 0.994789 |

* XGBoost fue el mejor modelo en todos los indicadores: precisión, sensibilidad (recall), F1-score y AUC ROC. Su desempeño sugiere que es el más balanceado y potente para este conjunto balanceado de datos.
* Random Forest también mostró un rendimiento excelente, apenas por debajo de XGBoost, lo que lo convierte en una alternativa robusta con excelente capacidad de generalización.
* Gradient Boosting tuvo métricas muy competitivas, pero ligeramente inferiores en comparación con los dos anteriores. Aun así, es un modelo viable.
* Logistic Regression mejoró significativamente con respecto al entrenamiento sin balanceo (especialmente en recall y AUC), pero sigue siendo el de menor rendimiento entre los cuatro.

### Análisis del Gráfico SHAP



Variables con mayor impacto positivo

* category\_gas\_transport: Tiene el mayor impacto positivo en la predicción del modelo.
* category\_grocery\_pos: También muestra un impacto positivo significativo.
* city\_pop: La población de la ciudad parece influir positivamente en el resultado.

Variables con impacto negativo

* state\_MT (Montana): Muestra el impacto negativo más fuerte.
* state\_NC (Carolina del Norte) y state\_NY (Nueva York): También tienen impactos negativos, aunque menores que MT.

Otras observaciones

* Variables como category\_shopping\_pos, category\_travel y category\_food\_dining tienen impactos positivos moderados.
* gender\_M (género masculino) parece tener un pequeño impacto positivo.
* plc ( "purchase location code") muestra un impacto negativo moderado.
* Las categorías category\_home, category\_kids\_pets y category\_misc\_pos tienen impactos relativamente neutros.

El modelo parece estar más influenciado por categorías de gastos específicas (especialmente transporte/gasolina y compras de comestibles) que por datos demográficos o de ubicación. Los estados particulares (MT, NC, NY) parecen ser predictores negativos en este modelo.

# Conclusión Final del Análisis de Detección de Fraude

### 1. Comparación de Modelos en Datos Desbalanceados vs. Balanceados

Primera Iteración (Datos Desbalanceados - 80% del Dataset)

* Logistic Regression falló completamente en detectar fraudes (\*Recall = 0\*), lo que lo hace inútil para esta tarea.
* Random Forest tuvo la mejor Precisión (93.8%) y AUC ROC (97.8%), pero un Recall moderado (63.9%), lo que significa que aún dejaba pasar muchos fraudes.
* XGBoost logró el mejor equilibrio entre Precisión (87.2%) y Recall (70.0%), con el mejor F1-score (77.6%), lo que lo convierte en el modelo más equilibrado.
* Gradient Boosting tuvo el peor desempeño entre los modelos ensemble, con un Recall muy bajo (29.3%), lo que indica que no detectaba suficientes fraudes.

🔹 Conclusión en datos desbalanceados:

-XGBoost fue el mejor modelo en términos de equilibrio entre detección de fraudes y minimización de falsas alarmas.

-Random Forest fue el mejor en precisión, pero con menor capacidad para capturar fraudes reales.

-Logistic Regression fue inútil en este escenario debido al desbalance.

### Segunda Iteración (Datos Balanceados con Undersampling)

* XGBoost siguió siendo el mejor modelo en todas las métricas:
* Recall (97.5%): Detectó casi todos los fraudes.
* F1-score (96.6%): Excelente equilibrio entre precisión y recall.
* AUC ROC (99.5%): Máxima capacidad de discriminación entre clases.
* Random Forest también tuvo un rendimiento excelente, muy cercano a XGBoost.
* Gradient Boosting mejoró significativamente, pero siguió siendo inferior a XGBoost y Random Forest.
* Logistic Regression mejoró notablemente (Recall pasó de 0% a 74.8%), pero aún fue el peor.

Conclusión en datos balanceados:

- XGBoost sigue siendo el mejor modelo, ahora con un Recall extremadamente alto (97.5%), lo que es crítico en detección de fraude.

- Random Forest es una excelente alternativa, casi igual de buena.

-Logistic Regression mejoró, pero sigue siendo el menos efectivo.

### Recomendaciones Futuras para Mejorar el Modelo de Detección de Fraude

Para futuros desarrollos, recomiendo profundizar en el análisis de características (feature engineering), incorporando variables temporales y de comportamiento del usuario (patrones de gasto, ubicaciones frecuentes, etc.). También sería valioso probar técnicas avanzadas de balanceo como SMOTE o ADASYN, ya que el undersampling tradicional puede perder información valiosa. Además, optimizar hiperparámetros con métodos como Bayesian Optimization podría mejorar aún más el rendimiento de XGBoost y Random Forest, los modelos más prometedores en este estudio.

Otra línea de trabajo importante es la implementación en producción y monitoreo continuo. Desplegar el modelo con una API (FastAPI/Flask) permitiría analizar transacciones en tiempo real, mientras que un sistema de feedback ayudaría a reentrenar el modelo con nuevos casos confirmados de fraude. Finalmente, explorar técnicas de explicabilidad (SHAP/LIME) y combinar el modelo con reglas de negocio simples mejoraría la transparencia y eficacia del sistema en la detección de fraudes complejos.

### Bibliografía

<https://www.sciencedirect.com/science/article/abs/pii/S0167923610001326?via%3Dihub>

<https://www.jair.org/index.php/jair/article/view/10302>

<https://www.sciencedirect.com/science/article/abs/pii/S0306457309000259?via%3Dihub>

<https://scikit-learn.org/stable/>

<https://xgboost.readthedocs.io/en/stable/>

<https://imbalanced-learn.org/stable/>