INSTRUMENTOS PARA LA CONSTATACIÓN DEL ESTADO DEL PROBLEMA DE INVESTIGACIÓN

Implementación de un Modelo de Machine Learning
para la Detección de Transacciones Fraudulentas y Anómalas
en Pagos Digitales de la Empresa TechSport

Gestión 2025

Ing. Ada Condori Callisaya

Noviembre 2025

1. Título de la Tesis

"Implementación de un Modelo de Machine Learning para la Detección de Transacciones Fraudulentas y Anómalas en Pagos Digitales de la Empresa TechSport, Gestión 2025"

2. Pregunta de Investigación

¿Cómo mejorar la detección de transacciones fraudulentas y anómalas en pagos digitales de la empresa TechSport durante la gestión 2025?

3. Objetivo General

Implementar un modelo de Machine Learning supervisado basado en Random Forest para la detección de transacciones fraudulentas y anómalas en pagos digitales, logrando un F1-Score $\geq 85\,\%$, Recall $\geq 90\,\%$, Precision $\geq 80\,\%$ y AUC-ROC ≥ 0.92, mediante validación temporal estricta sobre datos históricos de la empresa TechSport, gestión 2025.

4. Tabla de Variables, Dimensiones, Indicadores, Técnicas e Instrumentos

Marco Conceptual: Técnicas Cuantitativas de Investigación

Según Hernández Sampieri (2014), las técnicas cuantitativas incluyen:

- **Análisis de datos secundarios:** Uso de datasets existentes con fines de investigación
- **Análisis estadístico descriptivo:** Medidas de tendencia central, dispersión, distribuciones
- **Análisis estadístico inferencial:** Pruebas de hipótesis, intervalos de confianza, bootstrap
- **Machine Learning supervisado:** Algoritmos de clasificación con métricas de evaluación
- **Análisis exploratorio de datos (EDA):** Visualizaciones, correlaciones, detección de outliers

VARIABLES	DIMENSIONES	INDICADORES	TÉCNICAS	INSTRUMENTI	ACTIVIDADES CONCRETAS
VAR. INDEP.: Modelo de ML	1.1. Arquitectura y configuración del modelo	1.1.1. Feature Importance por variable	**Análisis de datos secundarios + ML supervisado**	Scripts Python (pandas, scikit-learn), Random Forest	Extracción de 15.4M+ transacciones, feature engineering de 15+ variables, cálculo de importancias con `feature_importances_`
		1.1.2. Distribución de features transformadas		Scripts de preprocesamiento, visualizaciones (matplotlib)	Histogramas, boxplots, análisis de outliers, normalización Min-Max
		1.1.3. Métricas de entrenamiento: F1, Precision, Recall		Funciones scikit-learn: `classification_` `confusion_matri`	Cálculo de métricas en train set (70 %), validation set (15 %)
		1.1.4. Ratio de balanceo de clases (%)		SMOTE (imblearn) o class_weight (scikit-learn)	Análisis de distribución fraude/no fraude, aplicación de SMOTE si desbalance > 10:1
		1.1.5. Tiempo de inferencia (ms)		Medición con `time.time()` en Python	Promedio de tiempo de predicción en 10,000 transacciones del test set
VAR. INDEP.: Modelo de ML	1.2. Selección y optimización de algoritmo	1.2.1. Comparación RF vs. XGBoost vs. SVM	**Análisis comparativo cuantitativo**	Grid Search CV (scikit-learn), métricas F1 y AUC-ROC	Entrenamiento de 3 algoritmos, comparación de F1-Score y AUC-ROC, selección del mejor
		1.2.2. Hiperparámetros optimizados: `max_depth`, `n_estimators`		GridSearchCV con validación cruzada k-fold (k=5)	Búsqueda exhaustiva en espacio de hiperparámetros, selección por mejor F1 en validation

VARIABLES	DIMENSIONES	INDICADORES	TÉCNICAS	INSTRUMENTO	ACTIVIDADES CONCRETAS
		1.2.3. Error train vs. validation (gap de overfitting)		Logs de entrenamiento, gráficos learning curve	Cálculo de Train Accuracy - Validation Accuracy, objetivo: gap $< 5\,\%$
		1.2.4. Tamaño del modelo serializado (MB)		Serialización con joblib/pickle, comando `ls -lh`	Guardar modelo entrenado como .pkl, verificar tamaño < 500 MB
VAR. DEP.: Transacciones fraudulentas y anómalas	2.1. Desempeño de detección del modelo	2.1.1. F1-Score en test set ($\geq 85\,\%$)	**Análisis estadístico inferencial**	Test set temporal (2025: 15.5M trans.), matriz de confusión, métricas scikit-learn	Aplicación del modelo final en test set, cálculo de VP, VN, FP, FN, F1-Score $= 2 \times \frac{P \times R}{P+R}$
		2.1.2. Recall / Sensibilidad ($\geq 90\,\%$)			Recall $= \frac{VP}{VP+FN} \times 100$, prioridad: detectar fraudes
		2.1.3. Precision ($\geq 80\,\%$)			Precision $= \frac{VP}{VP+FP} \times 100$, minimizar falsos positivos
		2.1.4. AUC-ROC (≥ 0.92)		Curva ROC con `roc_curve()` y `roc_auc_score()`	Área bajo curva ROC, evaluación de discriminación global del modelo
		2.1.5. Intervalos de confianza 95 % (bootstrap)		Bootstrap con 1000 muestras de test set	IC del 95 % para F1, Precision, Recall $\rightarrow$ validar robustez estadística

VARIABLES	DIMENSIONES	INDICADORES	TÉCNICAS	INSTRUMENTO	ACTIVIDADES CONCRETAS
VAR. DEP.: Transacciones fraudulentas y anómalas	2.2. Caracterización del fraude en el dataset	2.2.1. Tasa de fraude en dataset (%)	**Análisis estadístico descriptivo**	Análisis exploratorio de datos (EDA) con Python (pandas, numpy, matplotlib)	Cálculo de $\frac{\text{Fraudes}}{\text{Total trans.}} \times 100$, distribución por canal, gateway, hora del día
		2.2.2. Pérdidas económicas por fraude (USD)		Suma agregada de columna `amount` filtrada por `is_fraud==1`	Cálculo de $\sum \text{monto}_{\text{fraude}_i}$, análisis de percentiles (P50, P90, P95)
		2.2.3. Distribución temporal de fraudes		Gráficos de series de tiempo, heatmaps por hora/día	Identificación de patrones horarios, días de mayor incidencia
		2.2.4. Top 3 patrones de fraude identificados		Análisis de clusters, reglas de asociación (si aplica)	Documentación de: (1) tarjetas robadas, (2) duplicados sospechosos, (3) comportamientos anómalos
VAR. DEP.: Transacciones fraudulentas y anómalas	2.3. Proceso de etiquetado (Ground Truth)	2.3.1. Tiempo de etiquetado (0-5 meses)	**Análisis documental cuantitativo**	Revisión de documentación interna de TechSport sobre proceso de etiquetado	Documentación del proceso: chargebacks (0-5 meses), disputas, reportes. NOTA: Esto NO es entrevista cualitativa formal

VARIABLES	DIMENSIONES	INDICADORES	TÉCNICAS	INSTRUMENTI	ACTIVIDADES CONCRETAS
		2.3.2. Criterios de etiquetado (frecuencia por tipo)			Chargebacks confirmados (60 %), disputas (25 %), reportes manuales (15 %) - valores ilustrativos del proceso
		2.3.3. Cobertura de etiquetado (%)			Porcentaje de transacciones etiquetadas: 100 % (dataset completo tiene etiquetas)
VAR. DEP.: Transacciones fraudulentas y anómalas	2.4. Comparación con benchmarks de literatura	2.4.1. F1-Score vs. benchmarks (Hafez et al., 2025)	**Análisis comparativo cuantitativo**	Revisión bibliográfica sistemática, tabla comparativa de métricas	Comparación con F1-Scores reportados: Hafez (85-94 %), Feng (90-94 %), Hernández Aros (85-92 %)
		2.4.2. Significancia práctica de mejora		Análisis de efecto práctico (no solo estadístico)	Evaluar si F1 $\geq$ 85 % es suficientemente alto para aplicación real
		2.4.3. Tiempo de inferencia vs. literatura			Comparar <200ms con tiempos reportados en estudios similares

5. Análisis Exploratorio de Datos (EDA)

5.1. Objetivo del EDA

El Análisis Exploratorio de Datos es una técnica cuantitativa fundamental que permite:

- Comprender la estructura y distribución del dataset histórico de TechSport
- Identificar patrones, tendencias y anomalías en las transacciones
- Validar la calidad de los datos (valores faltantes, outliers, duplicados)
- Fundamentar decisiones de preprocesamiento y feature engineering

5.2. Actividades Cuantitativas del EDA

N°	ANÁLISIS	INSTRUMENTO	OBJETIVO
1.	Estadísticas descriptivas del dataset	`pandas.describe()`, medidas de tendencia central y dispersión	Caracterizar distribución de montos, frecuencias, tiempos de transacción
2.	Análisis de distribución de clases (fraude/no fraude)	Tabla de frecuencias, gráfico de barras	Determinar desbalance de clases, decidir estrategia de balanceo (SMOTE o class_weight)
3.	Análisis de correlación entre features	Matriz de correlación de Pearson, heatmap (seaborn)	Identificar multicolinealidad, seleccionar features independientes
4.	Detección de outliers	Boxplots, IQR (Rango Intercuartílico), Z-score	Identificar transacciones con montos atípicos, decidir si son fraude o errores de datos
5.	Análisis temporal de transacciones	Series de tiempo, gráficos de línea por fecha	Identificar estacionalidad, tendencias, picos de actividad fraudulenta
6.	Distribución por canal de pago	Tabla de frecuencias, gráfico de pastel	Comparar tasa de fraude por canal (Web, App, POS)
7.	Distribución por gateway	Tabla de frecuencias, gráfico de barras horizontales	Identificar gateways con mayor incidencia de fraude
8.	Análisis de valores faltantes	`pandas.isnull().sum()`, heatmap de missingness	Cuantificar % de datos faltantes por variable, decidir estrategia de imputación

N°	ANÁLISIS	INSTRUMENTO	OBJETIVO
9.	Análisis de transacciones duplicadas	`pandas.duplicated()`, conteo de duplicados	Detectar posibles errores de registro o intentos de fraude por duplicación
10.	Feature importance preliminar	Correlación con variable target, análisis univariado	Seleccionar features candidatas para el modelo (top 15-20)

5.3. Entregables del EDA

- **Reporte estadístico:** Documento PDF con estadísticas descriptivas, distribuciones, gráficos
- **Dataset limpio:** Archivo CSV procesado sin valores faltantes, outliers tratados
- **Notebook Jupyter:** Código Python documentado con todo el análisis exploratorio
- **Visualizaciones:** Conjunto de gráficos (PNG/PDF) para incluir en Capítulo 2 de la tesis

6. Segmentación de Comportamiento Transaccional

6.1. Segmentación Basada en Riesgo de Fraude

El análisis cuantitativo del dataset histórico permite identificar diferentes segmentos de riesgo mediante técnicas de clustering (K-Means, DBSCAN) y scoring basado en features del modelo:

Segmento	Nivel de Riesgo	% Estimado	Características Cuantitativas
1	Riesgo muy bajo	5-10 %	Score ML: 0.0-0.2, monto en rango P25-P75, frecuencia usuario < 5 trans/día, gateway confiable (Stripe), sin chargebacks previos
2	Riesgo bajo	10-15 %	Score ML: 0.2-0.4, monto típico, usuario con historial > 6 meses, velocidad transaccional normal
3	Riesgo moderado	30-40 %	Score ML: 0.4-0.6, 1-2 señales de alerta: monto $>$ P90, nuevo usuario (< 30 días), horario inusual (2-6 AM)
4	Riesgo alto	35-45 %	Score ML: 0.6-0.8, 3+ señales: múltiples intentos fallidos, cambio geolocalización IP, monto $>$ P95, velocidad anormal (> 10 trans/hora)
5	Fraude confirmado	5-8 %	Score ML: 0.8-1.0, etiqueta Ground Truth $= 1$ (fraude), chargebacks confirmados, disputas resueltas
Total: 15,492,846 transacciones (100 %)			

6.2. Nota Metodológica sobre Segmentación

Técnica cuantitativa utilizada:

- **Clustering no supervisado:** K-Means con k=5 segmentos, basado en features normalizadas (monto, frecuencia, velocidad transaccional)
- **Scoring supervisado:** Probabilidades generadas por Random Forest (`predict_proba()`) como score de riesgo [0,1]
- **Validación:** Análisis de silueta (silhouette score) para evaluar calidad de clusters, comparación con Ground Truth

Esta segmentación es **cuantitativa** porque se basa en análisis estadístico de features numéricas y categóricas, NO en percepciones o interpretaciones cualitativas.

7. Cronograma de Actividades de Constatación

Semana	Actividad	Instrumentos Cuantitativos	Entregables
Semana 1	Extracción de dataset histórico y análisis exploratorio inicial	Scripts SQL/Python para extracción, `pandas.read_csv()`, estadísticas descriptivas	Dataset limpio de 15.4M+ transacciones con 15+ variables etiquetadas
Semana 2	Análisis exploratorio de datos (EDA) completo	Visualizaciones (matplotlib, seaborn), correlaciones, boxplots, heatmaps	Notebook Jupyter documentado, reporte estadístico con gráficos, identificación de patrones de fraude
Semana 3	Análisis documental del proceso de etiquetado	Revisión de documentación interna de TechSport (PDFs, guías internas)	Documento resumen: criterios de etiquetado (chargebacks, disputas), tiempos (0-5 meses), proceso del equipo de contabilidad. **NOTA:** Esto NO es entrevista cualitativa, es revisión documental
Semana 4	Feature engineering y transformación de variables	Scripts de preprocesamiento, normalización Min-Max, SMOTE, cálculo de ratios y agregaciones	Dataset con 15+ features comportamentales, análisis de feature importance preliminar
Semana 5	Segmentación de comportamiento transaccional	Clustering K-Means (scikit-learn), análisis de silueta, scoring de riesgo	Segmentación en 5 niveles de riesgo con porcentajes y características cuantitativas
Semana 6	Validación del dataset y división temporal	División Train (2024: 9.7M) / Test (2025: 15.5M), verificación de estratificación	Datasets finales listos para entrenamiento, documento de validación de calidad de datos

Justificación de Actividades 100 % Cuantitativas

Todas las actividades del cronograma utilizan **técnicas cuantitativas**:

- **Semana 1-2:** Análisis de datos secundarios (dataset histórico)
- **Semana 3:** Análisis documental cuantitativo (NO entrevistas)
- **Semana 4:** Feature engineering (transformaciones numéricas)

- **Semana 5:** Clustering y scoring (técnicas de ML no supervisado)
- **Semana 6:** Validación estadística del dataset

NO se realizan:

× Entrevistas formales estructuradas o semiestructuradas

× Grupos focales

× Observación participante o shadowing

× Análisis de contenido cualitativo (codificación, categorías emergentes)

× Triangulación cualitativa-cuantitativa

8. Referencias Metodológicas

- Hernández Sampieri, R., Fernández Collado, C., & Baptista Lucio, P. (2014). *Metodología de la investigación* (6ª ed.). McGraw-Hill.
- Martínez González, R. (2020). *Método AQP/CCA para la elaboración de tesis*. Universidad Mayor de San Andrés.
- Hafez, A. I., et al. (2025). Random Forest for Credit Card Fraud Detection. *Journal of Financial Crime*, F1-Score: 85-94 %.
- Carcillo, F., Dal Pozzolo, A., Le Borgne, Y. A., Caelen, O., Mazzer, Y., & Bontempi, G. (2018). SCARFF: A scalable framework for streaming credit card fraud detection with Spark. *Information Fusion*, 41, 182-194.
- Dal Pozzolo, A., Caelen, O., Johnson, R. A., & Bontempi, G. (2015). Calibrating probability with undersampling for unbalanced classification. In *2015 IEEE Symposium Series on Computational Intelligence* (pp. 159-166). IEEE.