UNIVERSIDAD TECNOLÓGICA DE CHIHUAHUA

DESARROLLO DE SOFTWARE

EXTRACCIÓN DE CONOCIMIENTO EN BASES DE DATOS

II.1. REPORTE DE LIMPIEZA DE DATOS

PRESENTA:

ANGEL RICARDO CHAVEZ ZARAGOZA

DOCENTE:

ING. LUIS ENRIQUE MASCOTE CANO

GRUPO:

IDGS91N

Chihuahua, Chihuahua, 5 de octubre de 2025

Contenido

Introducción

SecureBank, una entidad financiera con presencia nacional, enfrenta el reto creciente del fraude en transacciones electrónicas debido al auge del comercio electrónico y la banca móvil. El incremento en fraudes por transferencias no autorizadas y pagos con tarjetas clonadas ha impulsado la necesidad de un sistema robusto de Data Warehouse (DW) y analítica en tiempo real. Este sistema buscará identificar patrones de comportamiento anómalos y reaccionar de manera inmediata, previniendo pérdidas económicas y garantizando la seguridad de sus clientes. Este reporte aborda la procedencia y tipos de datos involucrados, las técnicas de limpieza de datos utilizadas, y la fundamentación técnica y estructural para implementar una solución efectiva contra el fraude bancario.

Procedencia de los datos

Los datos con los que se alimenta el sistema antifraude de SecureBank provienen de distintas fuentes, principalmente transacciones financieras electrónicas generadas tanto por humanos como por sistemas automáticos. La mayoría de estas transacciones son el resultado de acciones de los clientes a través de plataformas digitales tales como aplicaciones móviles, sitios web de comercio electrónico y cajeros automáticos conectados. Además, se reciben datos máquina a máquina que registran interacciones del sistema bancario con otras redes y servicios financieros. También se integran datos biométricos (como reconocimiento facial o huellas digitales para autenticación), datos generados en redes sociales y perfiles digitales que ayudan a identificar comportamientos sospechosos. La combinación de estas fuentes ofrece una visión integral y detallada del comportamiento transaccional en tiempo real (IBM, 2025; SEON, 2025).

Tipos y fuentes de datos

Los datos involucrados en la detección de fraude tienen características heterogéneas que requieren clasificación para su análisis y procesamiento adecuado. En general, estos datos se pueden categorizar de la siguiente forma:

- **Datos cuantitativos:** Incluyen variables numéricas tales como el monto de la transacción, frecuencia de transacciones, duración de sesiones o número de intentos por inicio de sesión.

- **Datos cualitativos:** Se refiere a características cualitativas como el tipo de transacción (transferencia, pago, retiro), categoría del comercio, o el canal utilizado para la operación (móvil, web, cajero).
- **Datos nominales:** Identificadores únicos del cliente, números de cuenta, números de tarjeta, direcciones IP desde donde se accede, o identificadores de dispositivos.
- **Datos ordinales:** Clasificaciones de riesgo asignadas a las transacciones según su nivel de sospecha, categorizadas en rangos (bajo, medio, alto).
- **Datos estructurados:** Registros organizados dentro de bases de datos relacionales o de tipo Data Warehouse, con esquemas definidos que permiten consultas optimizadas.
- **Datos no estructurados:** Mensajes de texto en chats, registros de llamadas, publicaciones o actividad en redes sociales que requieren procesamiento con técnicas de minería de texto o análisis de sentimiento para extraer valor (InnoWise, 2025; SEON, 2025; IBM, 2025).

Técnicas de limpieza de datos

La calidad y confiabilidad de los datos es un factor crítico para la efectividad de los sistemas de detección de fraude. Los conjuntos de datos suelen presentar problemas típicos que deben ser corregidos mediante técnicas de limpieza específicas:

- **Manejo de valores nulos:** Es común encontrar registros incompletos donde ciertos campos están vacíos o no fueron capturados. Se aplican estrategias como la imputación mediante medias, medianas o modelos predictivos, o bien la eliminación de registros si la falta de información es crítica.
- **Detención y tratamiento de valores atípicos:** Transacciones con montos excesivamente altos o ubicaciones geográficas inusuales son analizadas para distinguir entre fraude o errores de captura. Se usan métodos estadísticos y de aprendizaje automático para identificar estos casos y tratarlos apropiadamente.
- **Corrección de errores de formato:** Se normalizan formatos de fechas, números telefónicos, códigos postales, y direcciones IP para garantizar consistencia en el procesamiento y garantizar correspondencia al compararlos.

- **Eliminación de duplicados:** Registros repetidos o múltiples reportes de la misma transacción pueden distorsionar análisis estadísticos y modelos predictivos, por lo que deben ser filtrados para mantener la integridad del dataset.

Estas técnicas de limpieza se complementan con procesos automáticos de validación y normalización que aseguran que los datos que ingresan al Data Warehouse estén en condiciones óptimas para ser analizados con algoritmos de detección de fraude en tiempo real (IBM, 2025; SEON, 2025; InnoWise, 2025).

Fundamentación y estructura

El desarrollo del sistema de Data Warehouse y analítica en tiempo real para la detección de fraude en SecureBank se fundamenta en las mejores prácticas y avances tecnológicos en inteligencia artificial y big data aplicada a la seguridad financiera. Estudios recientes resaltan la importancia de integrar múltiples fuentes de datos, tanto estructurados como no estructurados, para construir perfiles ricos y comportamientos de referencia que permitan identificar anomalías y patrones de fraude (IBM, 2025).

El uso de algoritmos de machine learning facilita la clasificación automática de transacciones y la identificación de posibles fraudes con alta precisión, mediante el entrenamiento con históricos y la actualización en tiempo real con datos nuevos. La arquitectura propuesta utiliza un Data Warehouse para consolidar información centralizada, mientras que un sistema de analítica en streaming analiza las transacciones en el momento de su ocurrencia, permitiendo acciones inmediatas para bloquear operaciones sospechosas y notificar a clientes y personal de seguridad (SEON, 2025; InnoWise, 2025).

La estructura documental del reporte sigue un orden lógico partiendo de la definición y origen de datos, pasando por la clasificación, identificación de problemas y técnicas aplicadas para la limpieza y preparación, finalizando con la fundamentación teórica y tecnológica que sustenta el diseño e implementación del sistema antifraude, garantizando claridad, coherencia y respaldo académico con citas confiables en formato APA.

Conclusión

La detección de fraude bancario en línea es fundamental frente al aumento de transacciones digitales y ataques sofisticados. La integración de datos provenientes de transacciones, dispositivos y perfiles digitales permite identificar patrones anómalos con mayor precisión. La limpieza adecuada de datos —eliminando nulos, duplicados y atípicos— es esencial para evitar errores en el análisis. Finalmente, el uso de inteligencia artificial y análisis en tiempo real habilita respuestas inmediatas que minimizan pérdidas y protegen a los clientes, fortaleciendo la seguridad bancaria en un entorno digital cada vez más complejo.

Referencias

IBM. (2025). Detección de fraude con IA en la banca. IBM Think. Recuperado de https://www.ibm.com/mx-es/think/topics/ai-fraud-detection-in-banking

Sukhadolski, S. (2025, April 15). *Dominar la detección y prevención del fraude en banca y FinTech*. Innowise. https://innowise.com/es/blog/financial-fraud-detection-software/

Kadar, T. (2025, March 6). *Fraude en transacciones: cómo detectarlo y reducirlo*. SEON ES. https://seon.io/es/recursos/reducir-el-fraude-en-transacciones-bancarias/