Propuesta de investigación

1. El nombre de la investigación

**"Detección de lavado de dinero mediante análisis predictivo y estructural en el conjunto de datos SAML-D"**

1. La descripción de la base de datos a utilizar.

El conjunto de datos SAML-D es una simulación avanzada de transacciones financieras diseñada específicamente para investigar patrones asociados al lavado de dinero (AML). Contiene 9,504,852 transacciones, de las cuales aproximadamente el 0.1039 % están etiquetadas como sospechosas, simulando de forma realista el fuerte desbalance que ocurre en entornos financieros reales.

Dado el tamaño del archivo original y las limitaciones prácticas de análisis en entornos de escritorio, se realizó un recorte del conjunto de datos, extrayendo una muestra representativa del 1 % mediante muestreo estratificado, manteniendo la proporción real de transacciones sospechosas. Esta versión reducida facilita el análisis exploratorio, predictivo y visual, sin comprometer la integridad estadística del problema.

Incluye 12 características como:

* Fecha y hora
* Identificadores de cuenta remitente y receptor
* Monto de la transacción
* Moneda del remitente y receptor
* País del remitente y del receptor
* Tipo de pago (tarjeta, efectivo, transferencias, etc.)
* Indicador binario "es sospechoso"
* Clasificación de 28 tipologías (11 normales y 17 sospechosas)

Además, incorpora 15 estructuras de red, representando flujos transaccionales entre entidades.

1. El motivo de la selección de los datos

La selección del conjunto de datos SAML-D se basa en una motivación principalmente laboral y profesional.

Este proyecto me permite simular el tipo de desafíos que enfrentan analistas de riesgos y data scientists en bancos.

Adicionalmente, la disponibilidad pública del conjunto SAML-D, su escala realista y su complejidad estructural (tipologías, etiquetas, grafos), ofrecen una oportunidad ideal para desarrollar habilidades técnicas valiosas en minería de datos, machine learning y análisis de redes, aplicadas a un contexto de gran relevancia laboral.

1. Una breve introducción del contexto

El lavado de dinero representa una amenaza global para la integridad de los sistemas financieros, facilitando actividades ilícitas como corrupción, narcotráfico y financiamiento al terrorismo. Los métodos tradicionales de monitoreo bancario se basan en reglas estáticas, con alta tasa de falsos positivos y detección tardía.

Este proyecto busca aplicar técnicas modernas de análisis de datos, machine learning y redes complejas para construir modelos que identifiquen transacciones potencialmente sospechosas, ayudando así a fortalecer los sistemas de cumplimiento normativo en instituciones financieras.

1. Una línea de tiempo de investigación

|  |  |  |
| --- | --- | --- |
| Fecha | Actividades | Entregables |
| 25/03 | -Lectura del enunciado completo del proyecto -Estudio del dataset SAML-D y su documentación -Inicio de redacción: Introducción, Justificación, Antecedentes | Borradores de las 3 primeras secciones del informe |
| 26/03 | - Análisis exploratorio inicial (EDA).  -Limpieza básica de datos y preprocesamiento | Script inicial de limpieza + primeras gráficas |
| 27/03 | -Análisis descriptivo: ubicación geográfica, monedas, anomalías  - Continuar redacción: Definición del problema | Gráficas de análisis descriptivo + redacción |
| 28/03 | - Aplicar técnicas para balanceo (como SMOTE)  - Selección de variables relevantes | Dataset procesado + resumen de variables clave |
| 29/03 | -Entrenamiento de modelos básicos: regresión logística, árbol de decisión  - Evaluación de desempeño | Primeros resultados predictivos |
| 30/03 | -Comparación con modelos avanzados (XGBoost, Random Forest)  - Curvas ROC, matriz de confusión | Resultados y comparación de modelos |
| 31/03 | - Descanso o día de recuperación si hay retrasos | - |
| 1/04 | - Análisis de grafos: creación de redes con estructuras del dataset  - Visualizaciones de redes | Grafo(s) y visualizaciones para incluir |
| 2/04 | - Redacción de secciones: Análisis Predictivo y Conclusiones | Informe casi completo |
| 3/04 | - Redacción de Recomendaciones y futuros estudios  - Creación de bibliografía (estilo APA o similar) | Secciones finales del informe |
| 4/04 | - Subida de scripts, notebooks y archivos a GitHub  - Organización de carpetas y documentación (README.md) | Repositorio completo |
| 5/04 | - Preparación de la presentación (diapositivas)  - Redacción de respuestas personales (7 preguntas) | Presentación casi lista |
| 6/04 | - Ensayo de la presentación  - Última revisión del informe y GitHub | Versión final |
| 7/04 | - Día libre o repaso final (respaldo, enlaces de GitHub, copias) | - |
| 9/04 | -Entrega y presentación final del proyecto | Informe + GitHub + Presentación |