Plan Overview

A Data Management Plan created using DMPTool

Title: Identificación de transacciones financieras fraudulentas mediante técnicas de machine learning

Creator: Julian Gomez

Affiliation: Universidad de Los Andes (uniandes.edu.co)

Principal Investigator: jhon adarve, giovanny andres cardona rojo

Data Manager: jhon adarve, giovanny andres cardona rojo

Project Administrator: jhon adarve, giovanny andres cardona rojo

Contributor: jhon adarve, giovanny andres cardona rojo

Funder: Digital Curation Centre (dcc.ac.uk)

Template: Digital Curation Centre

Project abstract:

El proyecto tiene como objetivo detectar transacciones financieras fraudulentas a través del uso de técnicas de Machine Learning, utilizando datos generados sintéticamente por BankSim, un simulador basado en agentes de pagos bancarios. Estos datos permiten experimentar con escenarios que incluyen tanto pagos normales como transacciones fraudulentas conocidas. El enfoque se centra en desarrollar modelos predictivos para detectar patrones de fraude en grandes volúmenes de datos transaccionales, lo que contribuye a la prevención de fraude en el sector financiero.

Objetivo principal:

Desarrollar e implementar un modelo de aprendizaje automático capaz de identificar con alta precisión transacciones financieras fraudulentas en un conjunto de datos bancarios

Objetivos secundarios:

1. Analizar y preprocesar el conjunto de datos de transacciones bancarias para identificar patrones y características relevantes en la detección de fraudes.
2. Evaluar y comparar diferentes algoritmos de aprendizaje automático para determinar el más efectivo en la detección de transacciones fraudulentas, considerando métricas como precisión, recall y F1-score.

Metodología

Tomando como referencia los 8 pasos del ciclo de vida de los datos (Harvard Business School, 2021), se estructura la metodología del proyecto.

- Generación: Obtención del conjunto de datos simulados de transacciones bancarias de Kaggle.
- Recolección: Descarga y almacenamiento seguro de los datos en un entorno controlado.
- Procesamiento: Limpieza, normalización y preprocesamiento de los datos para su análisis.
- Almacenamiento: Estructuración y organización de los datos procesados en una base de datos segura.
- Gestión: Implementación de protocolos de acceso y seguridad para los datos.
- Análisis: Aplicación de técnicas de aprendizaje automático para la detección de patrones fraudulentos.
- Visualización: Creación de gráficos y dashboards para representar los resultados del análisis.
- Interpretación: Evaluación de los resultados y extracción de insights para la toma de decisiones.

Start date: 10-10-2024

End date: 11-28-2024

Last modified: 10-20-2024

Identificación de transacciones financieras fraudulentas mediante técnicas de machine learning

Data Collection

What data will you collect or create?

Se utilizará un conjunto de datos simulados de transacciones bancarias obtenido de Kaggle, que incluye información como montos de transacciones, tipos de transacciones, fechas, ubicaciones y etiquetas de fraude/no fraude.

How will the data be collected or created?

Los datos serán descargados directamente de la plataforma Kaggle, asegurando la integridad y autenticidad de la fuente.

Documentation and Metadata

What documentation and metadata will accompany the data?

Se creará un documento detallado que describa la estructura del conjunto de datos, incluyendo definiciones de variables, tipos de datos y cualquier transformación aplicada durante el preprocesamiento.

Ethics and Legal Compliance

How will you manage any ethical issues?

Aunque los datos son simulados, se tratarán como si fueran reales para mantener prácticas éticas. Se anonimizará cualquier información que pudiera identificar a individuos.

How will you manage copyright and Intellectual Property Rights (IP/IPR) issues?

Se respetarán los términos de uso establecidos por Kaggle para el conjunto de datos. El modelo desarrollado y los resultados del análisis serán propiedad del equipo de investigación, con el debido reconocimiento a la fuente de los datos.

Storage and Backup

How will the data be stored and backed up during the research?

Tiene suficiente almacenamiento?. Se descargara un volumen de datos que pueda ser manejado por los equipos de los integrantes sin incurrir en gastos adicionales.

Como será la data almacenada? se almacenará en nuestros computadores.

Quién será el responsable para salvar la información y su recuperación: Julián Gómez se encargará de hacer copias de seguridad y recupererarla.

Como será la data recuperada en el evento de un incidente? La información se almacenará en dos discos duros de modo que se pueda recuperar fácilmente.

How will you manage access and security?

El riesgo es que los datos se pierdan o se borren por lo que se almacenarán en dos discos duros.

Todos los miembros del equipo podran acceder a la información sin cambiar los datos originales.

Selection and Preservation

Which data are of long-term value and should be retained, shared, and/or preserved?

Que data será preservado o destruido para propósitos legales contractuales o regulatorios? Una vez terminada la tarea la información será mantenida en los computadores de los miembros, pues la fuente es de carácter publica y no reviste ningún riesgo de seguridad el compartirla.

What is the long-term preservation plan for the dataset?

Los datos son de origen publico por lo que no tienen problema de preservación, pueden ser adquiridos en cualquier momento de la fuente por cualquier persona que necesite recopilar la data.

Data Sharing

How will you share the data?

La data puede ser compartida a través de medio magnético y estará disponible una vez se termine la tarea. Sin embargo la data original es de fuente pública y puede ser accesada por cualquier en cualquier momento.

Are any restrictions on data sharing required?

No hay ninguna restricción para compartir la data.

Responsibilities and Resources

Who will be responsible for data management?

Quién es responsable por implementar el DMP y asegurar que sea revisado y modificado? Giovanny Andrés Cardona Rojo

Quién será responsable por cada actividad de manejo de datos? Julian Alberto Gomez Gomez

Como se dividirán las responsabilidades a través de los diferentes proyectos de investigación? las responsabilidades se asignaran según el conocimiento, la experiencia y la disponibilidad de tiempo de cada uno de los miembros del equipo.

What resources will you require to deliver your plan?

Se necesitarán servidores en la nube (AWS), herramientas de desarrollo (Python) y bibliotecas especializadas en Machine Learning como Scikit-learn y TensorFlow.

Planned Research Outputs

Interactive resource - "Modelo Analitico de deteccion de fraudes financierso"

Modelo realizado en python que permite tomar la data he identificar el posible riesgo de fraude en una situación dada. La idea es presentar gráficos y resultados de los modelos de machine learning y que puedan ser accesados a través de internet.

Planned research output details

Title	Type	Anticipated release date	Initial access level	Intended repository(ies)	Anticipated file size	License	Metadata standard(s)	May contain sensitive data?	May contain PII?
Modelo Analitico de deteccion de fraudes financier ...	Interactive resource	Unspecified	Open	None specified		None specified	None specified	No	No