Task Overview

The task involves performing an exploratory data analysis (EDA) on a dataset of smart contracts deployed on a blockchain. This dataset includes the contracts' bytecode, opcodes, and related information. The goal is to understand the dataset, identify patterns and inconsistencies, and propose potential data preprocessing and model training approaches.

A smart contract is a self-executing program that runs on the blockchain. While smart contracts can perform legitimate operations, they can also be used maliciously in fraudulent schemes, such as financial scams like Ponzi schemes and exploits.

Dataset Overview

The dataset contains labeled data of smart contracts deployed on various blockchain networks. It includes the following columns:

address - Address of the contract

chain_id - ID of the blockchain where the contract is deployed

bytecode - Compiled code that runs on the blockchain

opcode - Human-readable instructions derived from bytecode

malicious_category_l1 - First level of the malicious category

malicious_category_l2 - Second level of the malicious category

data_source - Generalized data source of the sample

man_validated - Boolean flag indicating if the contract label was manually reviewed

man_validated_by - Reviewer who performed the validation

comment - Comment provided by the reviewer

label - Label of the smart contract

Deliverables

Please provide the following:

1. A Jupyter Notebook (or similar format) containing:
    - EDA of the dataset with visualizations.
2. A brief report (can be in the notebook or separate) explaining:
    - A discussion of patterns or anomalies discovered in the data.
    - A proposed strategy for training data preprocessing steps.
    - The models or approaches you would use
    - Justification for the chosen methods