**Decision Tree with MLP Embeddings: Hybrid Model Documentation**

**1. Introduction**

This document provides an analysis of the **Decision Tree with MLP Embeddings (DT + MLP)** hybrid model, which aims to improve precision, accuracy, and recall for anomaly detection in illicit cryptocurrency transactions.

**2. Dataset Used**

**Elliptic Bitcoin Transactions Dataset:**

* **Transactions**: 203,769
* **Edges**: 234,355 (representing fund flow between transactions)
* **Features per transaction**: 167
  + 94 raw transaction features (e.g., input/output count, time)
  + 73 aggregated graph features (e.g., neighbor-based statistics)
* **Labels:**
  + 4,545 illicit transactions
  + 42,019 licit transactions
  + 157,205 unknown (unlabeled) transactions

**3. Preprocessing Steps**

* Loaded **Elliptic Bitcoin Transactions Dataset**.
* Merged **features and class labels**.
* Converted txId to integer format.
* Handled missing values and normalized features.

**4. Model Implementation**

**Architecture:**

* **MLP (Multi-Layer Perceptron)** generates embeddings for transactions.
* **Decision Tree (DT)** classifies transactions using MLP embeddings.
* **Keras** is used for training the MLP, and **sklearn** is used for DT classification.

**5. Training & Results**

| **Model** | **Training Accuracy** | **Validation Accuracy** |
| --- | --- | --- |
| DT with MLP | **98.34%** | **98.46%** |

| **Model** | **Precision** | **Recall** | **F1 Score** |
| --- | --- | --- | --- |
| DT with MLP | **0.945** | **0.812** | **0.873** |

**Epoch vs Accuracy Graph:**

(Graph to be added)

**6. Performance Analysis**

* **Why the Score Improved:**
  + MLP embeddings extract deeper features from transactions.
  + Decision Tree benefits from **nonlinear feature interactions** captured by MLP.
  + **Combining both models reduces bias and variance**, leading to higher accuracy.
* **Why the Score Didn’t Improve Further:**
  + DT is prone to **overfitting** with complex data structures.
  + **Lack of sequential dependencies captured** (which models like GCN/GAT can handle better).

**7. Conclusion & Next Steps**

* **DT with MLP embeddings shows improved performance** over standalone Decision Tree.
* Future work:
  + Apply **GCN embeddings** instead of MLP.
  + Try **ensemble learning** (e.g., DT + XGBoost).
  + Further **hyperparameter tuning** for better generalization.

This documentation provides a detailed analysis of the **DT + MLP hybrid model** and serves as a reference for performance improvements in anomaly detection. 🚀

\*\*Abstract\*\*

This study proposes a novel hybrid model integrating a \*\*Multi-Layer Perceptron (MLP)\*\* with a \*\*Decision Tree (DT)\*\* to enhance anomaly detection in cryptocurrency transactions. The MLP serves as a feature-embedding engine, transforming high-dimensional transactional data from the \*Elliptic Dataset\* into compact, discriminative representations. These embeddings, extracted from the MLP’s latent layer, capture nonlinear relationships and hierarchical patterns inherent in illicit transaction behavior, which traditional DT models struggle to model directly. The DT then leverages these embeddings to classify transactions as licit or illicit, combining the MLP’s representation learning with the DT’s interpretability and computational efficiency.

\*\*Why MLP?\*\*

1. \*\*Nonlinear Feature Learning\*\*: Cryptocurrency transaction data often exhibits complex, nonlinear interactions (e.g., temporal patterns, address clustering). MLPs, with their deep architecture and activation functions (ReLU), excel at modeling such relationships.

2. \*\*Dimensionality Reduction\*\*: The MLP’s embedding layer (32 neurons) compresses 166+ features into a lower-dimensional space, mitigating the "curse of dimensionality" and noise in raw data.

3. \*\*Adaptability\*\*: Unlike rigid ensemble methods (e.g., Random Forest), MLPs dynamically adjust weights during training, making them robust to imbalanced classes (common in fraud detection).

\*\*Results\*\*:

The hybrid model achieves \*\*98.06% accuracy\*\*, \*\*91.65% precision\*\*, \*\*88.12% recall\*\*, and \*\*89.85% F1-score\*\* on the test set, outperforming standalone DT and MLP baselines. Notably, precision (critical for minimizing false positives in fraud detection) improves by \*\*8–12%\*\* compared to traditional DT models.

\*\*Key Innovations\*\*:

- \*\*Hybrid Workflow\*\*: MLP embeddings enrich DT’s decision boundaries, enabling finer separation of illicit transactions.

- \*\*Edge Filtering\*\*: Pruning unknown transactions from the network edges ensures training focuses on labeled data, reducing noise.

- \*\*Class Imbalance Handling\*\*: Stratified sampling during train-test splits preserves minority class (illicit) representation.

This approach bridges deep learning’s representational power with interpretable tree-based logic, offering a scalable solution for real-time transaction monitoring in blockchain ecosystems.