**Project Goal:**

This project aims to develop an efficient and effective model to detect fraudulent cryptocurrencies transactions on the Ethereum using a combination of a tree-based model and graph neural network.

In semester 1, the focus will be on the dataset, which includes:

* Obtaining a suitable dataset
* Data cleaning
* Feature selection and dataset balancing
* Determining and start building the model pipeline

In semester 2, the focus will be on the model, including:

* Continue building the model pipeline
* Optimizing and evaluating the model

**Problem Statement:**

Ethereum is one of the most used blockchains in the world, which is transparent, trust-less and does not need a trusted third party for monitoring. It is the backbone of decentralized apps (dApps) through smart contracts, which is essentially automated computer programs deployed on a blockchain. Nowadays, finance, healthcare, and gaming organizations often adopt blockchains as services. As its usages grow, blockchains face fraudulent activities, such as bribery, money laundering and phishing, causing major threat to user assets – their cryptocurrencies. Unlike banks and other centralized organizations, there is no third party to block these actions. An efficient and accurate algorithm to detect fraudulent activities on Ethereum can reduce financial losses for wallet owners and reputation damage to the blockchain ecosystem, where Ethereum plays a significant role. Investor confidence and industrial reputation can be improved, encouraging a wider usage of blockchain systems in general.

**Project Significance:**

This project can benefit in terms of end-users, regulation, and reputation.

Phishing activities are prevalent on Ethereum, lowering user confidence on usage on Ethereum. This project benefits users by increasing trust and confidence on Ethereum-based services and platforms since suspicious activities can be flagged, providing a more secure environment for transactions and prevent losing Ether through fraud. It also assists developers protecting their user base, as flagging fraud can enhance user trust in their applications.

This project may help ease authorities’ difficulty to track down illicit behaviour, identifying potential illegal transactions, assisting them in anti-fraud activities. It also promotes ethical usage of blockchains, as improper forms of transactions are identified as fraud.

At last, the reputation of Ethereum and other blockchains can be raised. Investors can be assured their investments are not grabbed by fraudulent schemes. With the rise of decentralize finance and Initial Coin Offering applications which uses blockchain as their backbone, protecting their interest provides stability and credibility to the blockchain this Ethereum ecosystem.

**Proposed Solutions:**

For the data acquisition, the aim is to obtain a relatively recent, large, and detailed dataset. Many research papers use this Kaggle dataset 1 for their model. However, with sufficient feature engineering, a larger dataset obtained by [Al-E’mari](https://link.springer.com/chapter/10.1007/978-981-33-6835-4_5#auth-Salam-Al_E_mari) et al. (2021) can be used to train our model more effectively.

For the model, tree-based models are well-suited for tabular data with its inductive biases (Grinsztajn et al., 2022), and graph neural networks (GNN) can refine predictions with relational information. A combination of Gradient Boosting Decision Tree (GBDT) and graph neural network called Boosted GNN (BGNN) can be used to improve the accuracy of the model, since it inherits the flexibility of neural networks and inductive biases of tree-based models (Ivanov & Prokhorenkova, 2021). Since there is yet a model which inherits the usages of both trees and GNNs on Ethereum fraud detection, this project aims to use this model to further improve precision and accuracy on predictions.

For the toolset, Python will be the main programming language used. Pytorch is a Python tensor library used for deep learning, and other major libraries used for data processing includes Pandas, SciKitLearn, MatplotLib. The toolset will expand as the project goes on. A complete report of the project will be delivered after the model is successfully built, trained, and tested then deployed.

**Proposed Timelines (subject to changes later):**

Focus of semester 1: dataset collection & preparation:

* Data Acquisition (1-2 weeks)
* Data Description (1 week)
* Quality Check (1-2 weeks)
* Exploratory Data Analysis (2 weeks)
* Feature Processing and selection (2 weeks)

Focus of semester 2: modelling:

* Model design (4 weeks)
* Hyperparameter tuning (2 weeks)
* Model Evaluation (3 weeks)

**References:**

[Léo Grinsztajn](https://arxiv.org/search/cs?searchtype=author&query=Grinsztajn,+L), [Edouard Oyallon](https://arxiv.org/search/cs?searchtype=author&query=Oyallon,+E) and [Gaël Varoquaux](https://arxiv.org/search/cs?searchtype=author&query=Varoquaux,+G). Why do tree-based models still outperform deep learning on tabular data? [arXiv:2207.08815](https://arxiv.org/abs/2207.08815) [cs.LG]. July 2022

[Sergei Ivanov](https://arxiv.org/search/cs?searchtype=author&query=Ivanov,+S), [Liudmila Prokhorenkova](https://arxiv.org/search/cs?searchtype=author&query=Prokhorenkova,+L). Boost then Convolve: Gradient Boosting Meets Graph Neural Networks.  [arXiv:2101.08543](https://arxiv.org/abs/2101.08543) [cs.LG]. March 2021

[Salam Al-E’mari](https://link.springer.com/chapter/10.1007/978-981-33-6835-4_5#auth-Salam-Al_E_mari), [Mohammed Anbar](https://link.springer.com/chapter/10.1007/978-981-33-6835-4_5#auth-Mohammed-Anbar), [Yousef Sanjalawe](https://link.springer.com/chapter/10.1007/978-981-33-6835-4_5#auth-Yousef-Sanjalawe) and [Selvakumar Manickam](https://link.springer.com/chapter/10.1007/978-981-33-6835-4_5#auth-Selvakumar-Manickam). A Labeled Transactions-Based Dataset on the Ethereum Network. In Advances in Cyber Security: Second International Conference, ACeS 2020, Penang, Malaysia, December 8-9, 2020, Revised Selected Papers 2 (pp. 61-79). Springer Singapore.