INFORMATICS INSTITUTE OF TECHNOLOGY

In Collaboration with

UNIVERSITY OF WESTMINSTER

PureChain

Decentralized Data Quality for Machine Learning

A Project Proposal by

Mr. D.D Gayantha Sandeep Dewasiri

Supervised by

Mr. Devon Wijesinghe

ABSTRACT

Problem:

Cooperative data-sharing systems are critical in areas that involves handling of sensitive information, especially areas such as health, finance, and supply chain. However, these sectors experience some difficulties in the field of data quality, especially, if data are provided by several contributors. Low-quality or incorrect data can result in noisy predictions in ML models and become a threat to ethical performance and data quality. Decentralized systems using blockchain technology are transparent and secure therefore can support ethically sensitive data. However, they do not have the built-in means to check for the quality of the data being stored. This research seeks to fill these gaps by adopting the use of blockchain smart contracts and machine learning methods to develop a sound framework that can provide quality data for training of ML models.

Methodology:

This research adopts an architecture called PureChain, which uses blockchain on the machine learning platform for better data quality. Data contributors provide raw data that are fed into a decentralized platform where smart contract standards facilitate the quality of the data being processed. The main steps of the approach are data preprocessing, Smart Contract rule implementation and training/validation of the ML model. In PureChain, smart contracts play a double role of data protection and preventing the submission of low-quality data. Validation of the performance of the ML models trained on high quality data is then carried out by machine learning operators. The prototype system was used to measure the feasibility, efficiency and effectiveness of the data at the decentralized system in its scaled-up state.

Subject Descriptors:

• Security and privacy ➞ Security services ➞ Pseudonymity, anonymity and untraceability

• Information systems ➞ Data management systems ➞ Database management system engines➞ Triggers and rules

• Computing methodologies ➞Machine learning➞ Machine learning approaches➞ Classification and regression trees