UNIVERSIDAD DE BUENOS AIRES

PROPUESTA DE TRABAJO PROFESIONAL

DeltaML
Machine Learning descentralizado utilizando Blockchain para fomentar participación en la red

---

Author:
Fabrizio GRAFFE 93158
Agustín ROJAS 91462

Supervisor:
Dr. Mariano BEIRÓ

20 de enero de 2019

Índice general

Capítulo 1

Introducción

El siguiente documento presenta la propuesta de Trabajo Profesional de Ingeniería en Informática de los estudiantes Fabrizio Sebastian Graffe (padrón 93158) y Agutín Rojas (padrón 91462).

El objetivo es aplicar los conocimientos adquiridos durante la carrera, para lo cual el tema elegido es **"Machine Learning descentralizado utilizando Blockchain para fomentar participación en la red"**.

Machine learning (o "aprendizaje automatico") es la rama del area de la inteligencia artificial centrada en el estudio y construcción de sistemas capaces de aprender de los datos, identificar patrones y realizar decisiones con mínima intervención humana.

Blockchain o DLT (Descentralized Ledger Technology), por su lado, es una tecnología que consta de un registro contable distribuido en una red de nodos. Este registro almacena las transacciones que se realizan entre los nodos de la red y cada nodo tiene una copia completa. A su vez, para evitar fraude y alteraciones al registro por cualquier nodo malicioso la tecnología cuenta con algoritmos de concenso y de prueba de trabajo realizado (Proof of Work). Las ventajas de la tecnología Blockchain por sobre bases de datos tradicionales son:

- **Mayor transparencia:** Todas las transacciones son publicas, por lo que cualquier participante de la red puede verlas.
- **Criptograficamente segura:** Debido a los algoritmos antes nombrados, para poder cometer fradue se necesitaria un poder de computo mayor al 51 % de la red (algo que no es posible hoy en dia, al menos contra las redes de Bitcoin o Ethereum, ni por las mayores empresas de software en el mundo).
- **Irreversibilidad:** Una transacción registrada en la blockchain no puede ser alterada por ninguno de los participantes de la red (otra vez, se necesitaria un poder de computo mayor al 51 % de la red).

Capítulo 2

Descripción del problema

Tanto el area de Machine Learning como la tecnología Blockchain son elementos que estan disrumpiendo la industria del software en la actualidad y cada vez tienen mas importancia en la vida diaria de las personas. En algunos casos su aplicación ha resultado en avances con un impacto positivo en la humanidad, pero tambien existen casos en los que se su impacto ha sido negativo.

Un ejemplo muy importante de mal uso de la tecnologia es facebook, que ha tenido varios casos de violación de la privacidad de los datos de sus usuarios, venta e intercambio de éstos con otras companias, uso de técnicas de aprendizaje automatico para generar adicción a las novedades en su plataforma y generación de cámaras de eco (echo chambers) por medio de filtrado de contenido (para mostrar a los usuarios solo opiniones similares a la suya, provocando así un refuerzo de su propia visión aprovechandose de sesgos propios de los humanos como el Sesgo de Confirmación o Confirmation Bias), por nombrar algunos.

Así como facebook incurrió en estas malas prácticas que debilitan y manipulan a los usuarios de su plataforma, muchas otras empresas e incluso estados también lo hacen día a día.

La creciente necesidad de mantener la privacidad de nuestros datos, que dia a dia son mas valiosos, está llevando a diferentes gobiernos a plantear regulaciones mas estrictas en lo relativo al uso de los mismos. Esto a su vez está impulsando una gran cantidad de iniciativas en el mundo para cambiar el paradigma actual donde las empresas son dueñas de los datos de las personas, a uno donde las personas sean dueñas de sus propios datos y puedan venderlos para un uso determinado y ningun otro. En el ámbito local se tiene el ejemplo de Wibson como una empresa que va en ese camino.

Debido a todo lo anterior mencionado, se eligió la temática de este trabajo teniendo en mente que en los años por venir se necesitarán maneras de entrenar modelos de Machine Learning sin violar la privacidad de las personas, es decir, sin tener una copia de sus datos en bases de datos de las companias. Por lo cual el presente trabajo trata de contribuir en este paso hacia un futuro donde las personas sean dueñas de su información.

Capítulo 3

Estado del arte

Actualmente existen varios proyectos en desarrollo y técnicas en investigación que intentan atacar la misma problematica de diferentes maneras.

- **Federated Learning:** El aprendizaje federado es un método de aprendizaje automático en el que el objetivo es entrenar a un modelo centralizado de alta calidad con datos de entrenamiento distribuidos entre un gran número de clientes, cada uno con conexiones de red poco confiables y relativamente lentas. Los algoritmos de aprendizaje para esta configuración funcionan de la siguiente manera: en cada iteración, cada cliente calcula de forma independiente una actualización del modelo actual en función de sus datos locales y comunica esta actualización a un servidor central, donde las actualizaciones del lado del cliente se agregan para calcular una nueva versión global del modelo. Los clientes típicos en este entorno son los teléfonos móviles, y la eficiencia de la comunicación es de suma importancia. Actualmente, Google está investigando esté método de aprendizaje automático y publicó varios papers al respecto, analizando mejoras y aplicaciones para dispositivos móbiles (tales como predicción de palabras para GBoard).

- **OpenMined:** Marketplace descentralizado de modelos de machine learning. Opera sobre la red de Ethereum. Utiliza Federated Learning para el entrenamiento de los modelos predictivos sobre Tensorflow y Pytorch ya que permiten entrenamiento desde clientes web. Modelos encriptados con homomorphic encryption para una transmición de datos segura. Smart contracts regulan el marketplace y pagan a quienes gastaron poder de computo entrenando un proporcional de acuerdo a cuanto aportaron a la mejora del modelo global. Actualmente en desarrollo.

- **DML:** Muy similar a OpenMined.

- **NumerAI:** Fondo de inversion que genera predicciones utilizando modelos predictivos crowd-sourced. Transforma y regulariza datos financieros a Transforms and regularizes financial data into machine learning problems for global network of data scientists

  Los intercambios de Numerai están determinados por una IA, que es alimentada por una red de miles de científicos de datos anónimos. La innovación tecnológica que Numerai proporciona está en su uso del cifrado de preservación de la estructura que aplican en sus fuentes de datos. Su objetivo es evitar sesgos y sobreajuste, también hace posible que Numerai comparta sus fuentes de datos de forma gratuita con sus usuarios.

- **Augur:** Es un protocolo de predicción de mercados destinado a ser propiedad de las personas que lo usan. Augur es un oraculo descentralizado y un protocolo peer to peer para la predicción de mercados. Augur es proyecto opensource. Es un conjunto de smart contracts escritos en Solidity que operan sobre Ethereum.

- **Golem Network:** Primera supercomputadora descentralizada que crea un marktplace global de poder de computo. Golem conecta computadoras en una red peer-to-peer, permitiendo tanto a dueños de aplicaciones como a usuarios individuales (requestors") alquilar recursos de las maquinas de otros usuarios ("providers"). Los pagos entre requestors, providers y desarrolladores de apliaciones se realizan por medio de un sistema de transacciones basado en la red de Ethereum. El sistema está orientado a competir con los proveedores de infraestructura o servicios cloud para aplicaciones que requieran gran capacidad de computo reduciendo el precio de dicha capacidad. Como consecuencia, aplicaciones complejas como la representación CGI, el cálculo científico y el aprendizaje automático se volverían más accesibles.

- **OceanProtocol:** Ecosistema destinado a comprartir activos y servicios. Los activos son datos y algoritmos. Los servicios son la integración, procesamiento, computación y almacenamiento. Ocean Protocol es un protocolo de intercambio de datos descentralizado, que permite que personas compartan y monetizen sus datos a la vez que garantiza control, auditabilidad, transparencia y is a decentralized data exchange protocol that lets people share and monetize data while guaranteeing control, auditability, transparency y conformidad a todos los actores involucrados to all actors involved.

- **Wibson:** Mercado de datos descentralizado, basado en blockchain, que proporciona a las personas una forma segura y anónima de vender información privada validada en un entorno confiable. Opera sobre la red de Ethereum. Las personas pueden vender sus datos por medio de la aplicacion movil de Wibson.

Capítulo 4

Objetivos

El presente trabajo consta de los siguientes objetivos principales:

- Desarrollar un marketplace de modelos de ML que permita a usuarios dueños de estos modelos entrenarlos de manera descentralizada sin necesidad de violar la privacidad de los usuarios dueños de los datos que serán utilizados para dicho entrenamiento.
- Desarrollar una forma de recompensar a los usuarios que provean datos y poder de computor para el entrenamiento de los modelos.
- Desarrollar un framework de Federated Learning de código libre para generar un aporte a la comunidad.
- Desarrollar un caso de uso de un modelo que requiera entrenamiento para hacer prueba del sistema desarrollado durante este trabajo.

Capítulo 5

Carácteristicas del trabajo

5.1. Módulo de Machine Learning descentralizado

Este modulo tendrá como responsabilidad realizar entrenamientos de modelos de Machine Learning siguiendo el esquema de Federated Learning. Tendrá una interfaz lo suficientemente genérica como para poder usarse con un servidor central o en una red de nodos descentralizados (como se pretende hacer con la red de Ethereum).

5.2. Módulo de encriptacíon

Este modulo tendrá como responsabilidad la encriptación y desencriptación de los modelos que se transfieran en la red para su entrenamiento.
Para evitar el envío de datos por la red, y así preservar la privacidad, el esquema de Federated Learning establece que se envíen los modelos a los clientes que los entrenarán. Dichos modelos deben estar encriptados, ya que de otra forma cualquier participante malicioso de la red podría robarlo.
Para poder operar sobre un modelo encriptado existen técnicas tales como Hommomorphic Encryption.
El modulo deberá minimizar el riesgo de violación de la privacidad tanto del usuario que entrena el modelo con sus datos, como del usuario que desea obtener un modelo entrenado (sin que nadie se lo robe en el proceso), para esto se explorarán técnicas de Secure Multi party computation y Differential Privacy.

5.3. La red

La red constará de uno o mas smart contracts desplegados en la red de Ethereum que estarán integrados con IPFS como método de almacenamiento descentralizado. La red proveerá medios para recompensar a los participantes de la misma, tales como pueden ser los usuarios que entrenan los modelos, según el nivel de su aporte a la calidad de las predicciones del modelo.

5.4. Interfaz web del marketplace

El sistema debe tener una interfaz web que permita que usuarios puedan construir un modelo de machine learning que use determinado tipo de datos y enviarlo para su entrenamiento a los diferentes nodos de la red que cumplan con las carácteristicas pedidas por el usuario dueño del modelo.

5.5. Cliente web/mobile

Este cliente deberá permitir que los usuarios decidan cuales de sus datos quieren que sean usados para entrenar modelos de machine learning.

Capítulo 6

Tecnologías

6.1. Blockchain

6.1.1. Ethereum network

(**Ethereum**) Ethereum es una plataforma descentralizada que ejecuta contratos inteligentes (smart contracts): aplicaciones que se ejecutan exactamente como fueron programadas sin ninguna posibilidad de downtime, censura, fraude o interferencia de una tercera parte.

6.1.2. Solidity

Solidity es un lenguaje de alto nivel, orientado a objetos para implementar smart contracts. Los smart contracts son programas que gobiernan el comportamiento de cuentas dentro del ecosistema Ethereum.

6.1.3. Zeppelin OS

ZeppelinOS es una plataforma de desarrollo diseñada para proyectos que involucren smart contracts. Permite realizar actualizaciones de los mismos de manera sencilla y provee incentivos económicos para crear un ecosistema saludable de aplicaciones seguras.

6.1.4. Open Zeppelin

OpenZeppelin es un framework de smart contracts reutilizables para Ethereum y otras blockchains de EVM y eWASM. Reduce el riesgo de vulnerabilidades mediante el uso de un estándar probado y revisado por la comunidad.

6.1.5. Truffle

Framework de desarrollo de smart contracts para blockchains que utilizan la Máquina Virtual Ethereum (EVM).

6.1.6. Ganache

Ganache es una blockchain personal para el desarrollo en Ethereum (testnet local) que se puede utilizar para implementar contratos, desarrollar sus aplicaciones y ejecutar pruebas.

6.1.7. Metamask

MetaMask es una extensión que permite ejecutar aplicaciones descentralizadas (dApps) de Ethereum en web browsers sin la necesidad de tener un full node de Ethereum.

6.2. Machine Learning

6.2.1. Python

6.2.2. Sci-kit Learn

Es un conjunto de herramientas de codigo abierto, de facil uso, destinandas a la mineria de datos, al análisis de los mismos y al aprendizaje automatico, construido sobre NumPy, SciPy, y matplotlib.

6.2.3. PyTorch

Es una biblioteca de código abierto de aprendizaje automatico para Python, basada en Torch, usada para aplicaciones tales como el procesamiento del lenguaje natural. Es desarrollada principalmente por el equipo de investigación de inteligencia artificial de Facebook. PyTorch provee 2 funcionalidades de alto nivel: Computo usando Tensores (como NumPy) con aceleración usando GPUs y Redes Neuronales profundas.

6.2.4. Tensorflow

Es una biblioteca de software libre que se utiliza para realizar cálculos numéricos mediante diagramas de flujo de datos. Los nodos de los diagramas representan operaciones matemáticas y las aristas reflejan las matrices de datos multidimensionales (tensores) comunicadas entre ellas. Se utiliza en gran medida para tareas de aprendizaje automatico.

6.2.5. Keras

Es una API de alto nivel de redes neuronales, escrita en Python y capaz de ser ejecutada por encima de Tensorflow, CNTK o Theano. Fue desarrollada para permitir experimentación rapida con modelos de aprendizaje automatico basados en redes neuronales.

6.2.6. Tensorflow.js

Es una biblioteca de JavaScript que permite entrenar y desplagar modelos de Machine Learning en los web browsers y en aplicaciones Node.js. Posee comptibilidad con modelos de Keras y Tensorflow, por lo que permite importar y exportar modelos de, y a, estos frameworks.

6.3. Frontend

6.3.1. React

6.4. Storage

6.4.1. IPFS

IPFS (InterPlanetary File System) es un sistema distribuido para almacenar y acceder a archivos, sitios web, aplicaciones y datos cuyo funcionamiento sintetiza e integra ideas exitosas previas de los sistemas peer-to-peer, incluyendo DHTs (Distributed Hash Tables), BitTorrent, Git y SFS (Self-Certified FileSystems).

Capítulo 7

Alcance

El alcance del presente Trabajo Profesional comprende:

- Desarrollo de un framework de machine learning descentralizado.
- Desarrollo de un modulo de encriptación.
- Desarrollo de marketplace de modelos de ML por medio de smart contracts con las reglas de negocio para monetizar la contribución de los diferentes participantes.
- Desarrollo de integración entre smart contracts e IPFS para almacenar datos de manera descentralizada.
- Desarrollo de interfaz web para el marketplace.
- Pruebas del sistema en testnet de ethereum local y posibles pruebas en testnets globales (Kovan, Rinkeby, Ropsten).

Capítulo 8

Plan de Trabajo

8.1. Equipo de trabajo

Equipo	
Integrante	**Rol**
Mariano Beiró	Tutor
Fabrizio Graffe	Desarrollador
Agustín Rojas	Desarrollador

8.2. Metodología

Para la ejecución del proyecto se usará una metodología ágil basada en SCRUM. La misma consistirá en definir una serie de iteraciones con fechas pautadas de reuniones de avance y entregas.

Al inicio de cada iteración se conformará una priorización de los requerimientos pendientes de desarrollo. Posteriormente, se procederá a su implementación y, para finalizar cada iteración, se hará una presentación de los entregables pautados. Las reuniones, entregas, presentaciones y la priorización de los requerimientos para cada iteración se realizará en con el tutor del trabajo.

8.3. Estimación

Se muestra a continuación el listado de tareas necesarias para alcanzar los objetivos descritos anteriormente. Todos los esfuerzos se encuentran expresados en horas.

It.	Descripción	Esf.
1	Aprendizaje desarrollo de aplicaciones descentralizadas	40
	Aprendizaje de uso de IPFS como storage descentralizado	10
	Desarrollo de prueba de concepto (IPFS + Ethereum)	20
2	Investigación sobre métodos de encripción	40
	Desarrollo de modulo de encriptación de modelos de ML	20
3	Investigación sobre Federated Learning y alternativas	40
	Desarrollo de modulo de Federated Learning	20
	Desarrollo de caso de uso para modulo de Federated Learning	20
4	Desarrollo de aplicación web para marketplace descentralizado	20
	Desarrollo de smart contracts para marketlace descentralizado	20
	Integración de los modulos desarrollados con el marketplace descentralizado	60
5	Pruebas del sistema en funcionamiento	20
	Reuniones	15
	Presentación	15
	Reuniones	15

Además se debe tener en cuenta el tiempo dedicado a la administración del proyecto, estimando un 15 % del tiempo de desarrollo, lo cual resulta en:

Descripción	Esf.
Iteración 1	70
Iteración 2	60
Iteración 3	80
Iteración 4	100
Iteración 5	20
Otros	45
Administración	105
Esfuerzo total	805

8.4. Cronograma de entregables

A continuación se hace un cronograma tentativo de entregables al finalizar cada una de las iteraciones. Las mismas pueden ser modificadas en caso de verse necesario, por el tutor o por los desarrolladores del proyecto. La fecha de cada una de las entregas serán determinadas en conjunto con el tutor.

It.	**Entregables**	**Hitos**
1	Prueba de concepto (IPFS + Ethereum)	Curso de aplicaciones descentralizadas terminado Videos y lecturas sobre IPFS terminadas Prueba de concepto desarrollada
2	Modulo de encriptación de modelos	Investigación sobre hommomorphic encryption, multy-party encryption y alternativas terminada. Método de encriptación seleccionado Modulo de encriptación terminado
3	Modulo de Federated Learning	Investigación sobre Federated Learning y alternativas terminado. Pruebas de conepto usando Federated Learning terminadas. Desarrollo de framework de ML usando Federated Learning terminado.
4	Marketplace descentralizado usando smart contracts sobre Ethereum	Desarrollo de smart contracts terminado. Desarrollo de frontend terminado. Integración de frontend con smart contracts e IPFS terminado. Desarrollo de caso de prueba terminado
5	Casos de uso probados	Prueba exhaustiva de uno o dos casos de uso terminada