Detección de anomalías de tráfico en servidores web

Mitsiu Alejandro Carreño Sarabia
E23S-18014

01

PROPUESTA CIENTÍFICA

ANTECEDENTES

Se propone desarrollar una solución integral de **monitoreo de tráfico en servidores web** así como la detección automatizada de tráfico anómalo mediante la **implementación de técnicas de análisis topológico** así como **aprendizaje automático** las cuales en conjunto permitan por una parte el constante monitoreo de los usos y la toma de decisiones preventivas y correctivas.

Aplicando estas metodologías, es posible evaluar nuevas peticiones basado en el tráfico histórico del servidor y obtener un índice de similitud respecto a solicitudes pasadas, con ello es posible detectar anomalías o contenido malicioso y tomar tanto acciones correctivas (protección ante uso anómalo) como preventivas (detectar picos o valles de tráfico y ajustar la infraestructura acorde).

OBJETIVOS

Enumerar las características y casos de uso de sistemas de monitoreo y alerta efectivos

Analizar las técnicas y procesos tanto tradicionales como de aprendizaje automático mediante los cuales se analiza tráfico web actualmente

Evaluar el sistema desarrollado implementandolo en un entorno controlado

Desarrollar un sistema detección de anomalías basado en aprendizaje automático

PREGUNTAS DE INVESTIGACIÓN

¿De qué manera se analiza el tráfico web actualmente?

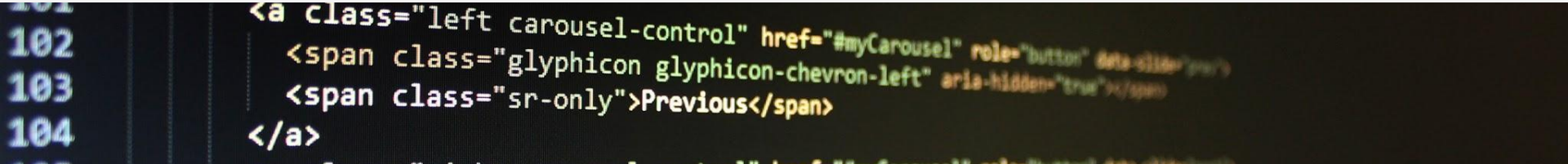
¿Actualmente cómo se ha implementado el aprendizaje automático en análisis de tráfico web?

¿Qué elementos debe tener un sistema de alertas para ser útil (falsos negativos/falsos positivos, canales de comunicación, protocolos extras)?

JUSTIFICACIÓN

El tráfico a un servidor web provee datos confiables sobre la información y el contexto bajo el que se usan sus recursos, pero la cantidad de información generada es tan grande que un análisis manual no es viable. Entender los usos típicos y diferenciarlos de los atípicos es una herramienta poderosa que aplicada en tiempo real permitirá mejorar la calidad, y resguardo de la información contenida.

Analizar los registros de tráfico web permite no solo entender la manera en que se consume la información que contiene un servidor, sino también detectar si el uso generalizado se transforma, o si existen anomalías e incluso calcular un parámetro de probabilidad de ser malintencionadas. Dado el volumen de información que se genera, y la creciente sensibilidad de los datos alojados, aplicar herramientas de aprendizaje automático permitirá agilizar y perfeccionar cualquier proceso manual.



VIABILIDAD

01

Arquitectura de red neuronal:

Ingesta de datos de entrenamiento y validación mediante integración a sistema de almacenamiento aws S3 o similar. Entorno de ejecución Python 3 o superior, empleando una instancia de procesamiento aws EC2 (c7a.medium) o rentar una instancia dedicada según se ajuste mejor al presupuesto.

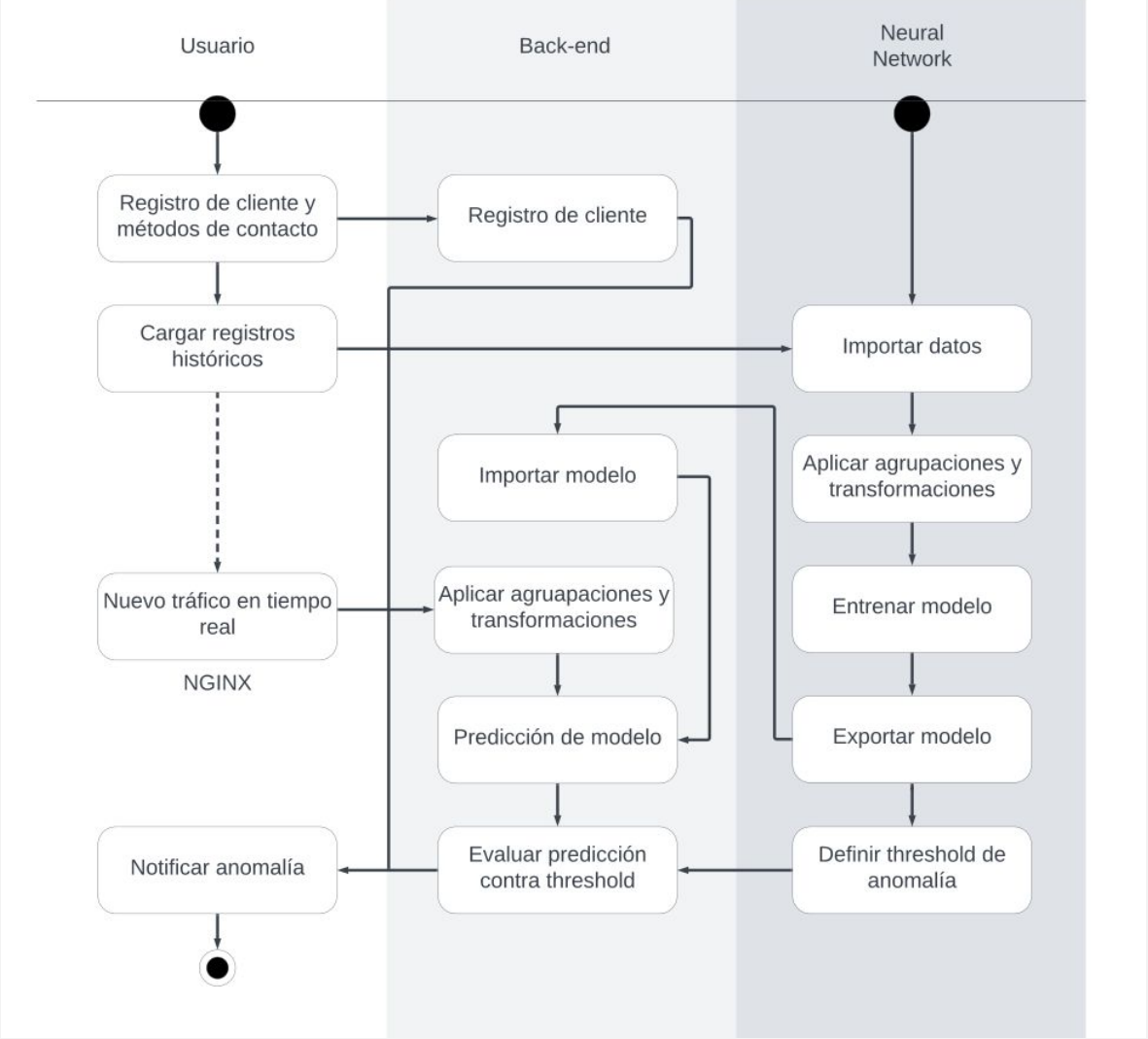
02

Backend:

Punto de conexión para los clientes y el público en general es responsable del manejo de flujos en tiempo real, la autenticación y administración de usuarios, y realiza predicciones sobre el modelo entrenado, también se encarga de procesar el envío de notificaciones una vez que se detecte actividad anómala.

VIABILIDAD

Diagrama general del sistema



02
PROPUESTA FINANCIERA

PROPUESTA FINANCIERA

01

Recursos humanos

Equipo de ocho personas interdisciplinario integrando las áreas de dirección, desarrollo, ciencia de datos, finanzas y manejo de clientes

02

Productos/Servicios personales

Coworking: Espacio de oficina compartida por distintas empresas, con diversos servicios como internet, muebles, bebidas, limpieza, etc.

03

Recursos computacionales

Amazon Web Services - S3 storage

Servicio de almacenamiento empleado para almacenar los datos de entrada (tráfico en tiempo real), así como assets o versiones del modelo entrenado.

Amazon Web Services - EC2 instance

Servicio de cómputo dedicado a ejecutar el proyecto en ambiente de producción

Google Colab Pro:

Servicio de computo enfocado en el desarrollo del código de desarrollo de la red neuronal

Dominio de internet

Cadena de caracteres único que identifica un ámbito de autonomía, autoridad y control, con la finalidad de identificar servicios en internet

PRESUPUESTO

A continuación se detalla el procedimiento mediante el cual se obtuvieron las cantidades expuestas.

Recursos humanos

Se requiere un equipo con 9 integrantes, con sueldos entre $15,000 y $25,000 pesos mensuales
Total = $389,000.00

Licencia google colab pro

Permitirá realizar el desarrollo y entrenamiento de la red neuronal usando hardware dedicado (unidades de procesamiento gráfico GPU e incluso unidades de procesamiento tensorial TPU) de esta manera se renta el hardware durante el periodo necesario en lugar de invertir en equipo que obsolesce, y requiere condiciones especiales, $3,500 al mes, 2 meses, 2 licencias
Total = $14,000

EC2 instance (c7a.medium)

EC2 es el servicio que ofrece AWS para rentar maquinas virtuales para realizar procesamiento, de esta manera se ahorran costos de adecuación de espacio para un servidor físico así como no se exponen redes domésticas a internet. Costo por hora bajo demanda = $0.852, 2352 horas (14 semanas)
Total = $2003.90

PRESUPUESTO

Total de infraestructura/servicios= $43,143.87

S3 storage (S3 Standard)

Es la tecnología de almacenamiento que nos permitirá almacenar los datos completos de entrenamiento de la red, sino también el flujo de datos de clientes reales, ofreciendo un servicio 24/7, tolerante a fallos, autoincrementable.
Costo por Gb = $0.34, Costo por Gb en transferencia = $0.34, 500 Gb,
Total = $340

Dominio de internet (loggart.com)

Permite a los clientes contactarnos de manera rápida en internet, ofreciendo una identidad del servicio, además usualmente incluyen un registro de email lo que facilita la comunicación dentro y fuera del equipo de trabajo.
Proveedor godaddy
Total = $1,199.97 + impuestos

Coworking (alda - private desk)

Corresponde a la renta de un espacio físico de trabajo, amueblado, con servicios de comida, y limpieza, ofreciendo un ambiente colaborativo entre los integrantes del equipo, así como promover la comunicación, socialización y camaradería, factores que se consideran esenciales para el éxito del proyecto.
Costo$3,200 mensual, 4 meses, 4 personas
Total = $25,600.00

FUENTES DE FINANCIAMIENTO

Se considera que existen diversas opciones capaces de generar el financiamiento necesario para cubrir los costos de desarrollo y operación del proyecto.

- Entrar a concursos es una gran oportunidad para dar visibilidad al proyecto, además de ser una puerta para realizar tareas de networking con gente inmersa en el desarrollo tecnológico o en el desarrollo de proyectos en general, además existe este potencial de ganar y obtener reconocimiento y apoyos económicos al ganar.
- Apoyos del gobierno a MIPyMES puede ser una opción viable, en los que además se obtienen asesorías sobre el manejo de negocio además de contactos con gente que conoce la regulación y administración de negocios que si bien no ofrecen una fuente de financiamiento directo, pueden tener un impacto significativo en la toma de decisiones del negocio y son gratuitas.
- Una opción que requiere una inversión inicial son las incubadoras, que ofrecen lo mejor de los concursos y de los apoyos del gobierno, ya que permiten realizar mucho networking, se reciben asesorías especializadas y se conoce gente inmersa en el desarrollo de tecnologías emergentes y de riesgo.
- Finalmente es posible considerar una inversión inicial propia o de conocidos cercanos que permita tener la liquidez inicial necesaria con una baja tasa de interés.

Independientemente de la fuente de financiamiento, cabe resaltar que el objetivo principal es generar un producto útil y deseable que capture el interés del mercado objetivo, que ofrezca un beneficio real y cuantificable y que sea superior a la competencia, para de esta manera poder generar ingresos.

03

PROPUESTA DE GESTIÓN DEL PROYECTO

PROPUESTA DE GESTIÓN DEL PROYECTO

Para la correcta realización del proyecto se deberá contar con un equipo **multidisciplinario** con principal énfasis en las áreas de ingeniería de software y ciencia de datos, además se consideran algunos roles de coordinación como son líder de desarrollo y responsable de financiamiento. Dichos roles ejecutarán la toma de decisiones en sus áreas correspondientes y ayudarán a los distintos equipos a solucionar problemas brindando guía y experiencia. Finalmente se cuenta con el responsable del proyecto que es quién tiene la visión general del proyecto.

Además se consideran **roles auxiliares** para tareas de soporte como ingeniero devops o un responsable de manejo de clientes que apoyen en tareas asociadas a sus cargos. Con un equipo de dichas características y habilidades se estima que se poseerán las capacidades técnicas, de operación de negocio y de finanzas necesarias para resolver los retos y problemas que se presenten derivado del desarrollo mismo del proyecto.

EQUIPO DE TRABAJO Y ESTRUCTURA ORGANIZATIVA

Para la correcta realización del proyecto se considera pertinente integrar un equipo de trabajo que cubra los siguientes roles y necesidades:

Responsable de proyecto:

Es la figura central que entiende todas las dimensiones del proyecto (técnica, financiera, de negocio, de gestión, etc) y es capaz de tomar decisiones y delegar responsabilidades así como definir y priorizar tareas.

Responsable de financiamiento:

Es la persona que entiende y está inmerso en áreas de manejo financiero, tiene entendimiento legal y de administración de recursos por lo que es capaz de solventar las cuestiones financieras del proyecto.

Lider de desarrollo:

Es el rol a cargo de gestionar el desarrollo general del proyecto por lo que debe tener una base fuerte en desarrollo de tecnología y ciencia de datos en sus diferentes áreas (tecnologías, integraciones, servicios, plataformas de desarrollo y despliegue etc). Es la figura responsable de que el proyecto desde el enfoque tecnológico se lleve a cabo de manera exitosa.

Científico de datos:

Es el cargo responsable de concretar las tareas y actividades concernientes a la ciencia de datos con acompañamiento y dirección del líder de desarrollo, se espera que tenga conocimiento fundamental sobre redes neuronales así como conocimiento de buenas prácticas de desarrollo y capacidad de trabajo en equipo.

EQUIPO DE TRABAJO Y ESTRUCTURA ORGANIZATIVA

Ingeniero de software:

Es el cargo responsable de generar el código para las distintas integraciones necesarias (alertas, procesamiento de información en tiempo real, manejar almacenamiento de información, etc), por lo que debe tener fuertes bases de conocimiento en sistemas y flujos de información, así como poder aplicar e integrar tecnologías existentes.

Ingeniero devops:

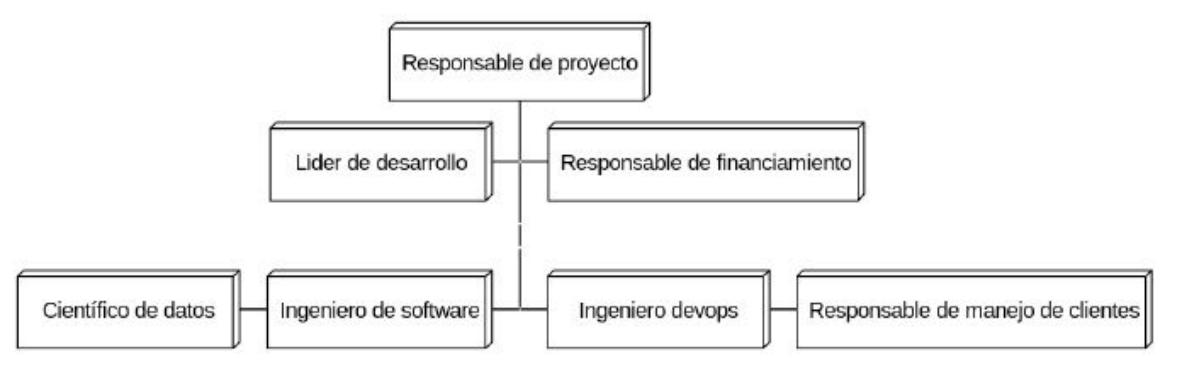
Rol con la capacidad de integrar los distintos sistemas generados (redes neuronales, despliegues automatizados, versionado automático, bitácoras de registros para los distintos sistemas, replicabilidad de ambientes) así como ser capaz de generar y replicar ambientes de desarrollo, pruebas y producción así como soluciones de tolerancia a fallos

Responsable de manejo de clientes:

Es el rol responsable de manejar y mantener el contacto con los clientes, permitiendo la transparentización de los procesos tanto internos como externos y dando seguimiento al avance del equipo de trabajo.

EQUIPO DE TRABAJO Y ESTRUCTURA ORGANIZATIVA

Además se considera esencial que todos los perfiles cuenten con aptitudes sociales, comunicativas y de trabajo en equipo para que las los problemas y dudas se puedan escalar y clarificar de manera adecuada, además se espera que los distintos responsables y líderes sean capaces de instruir a sus equipos para que logren desarrollar sus habilidades más allá de sus tareas rutinarias.
A continuación se muestra el organigrama propuesto para el equipo de trabajo:



PLAN DE TRABAJO

Para el desarrollo del proyecto se dividirán las actividades en dos etapas, desarrollo e implementación, donde durante el desarrollo se realizan pruebas de concepto de las tecnologías así como de las ideas, mientras que en la fase de implementación se considera emplear las tecnologías y soluciones creadas y evaluadas durante la fase de implementación aplicado a clientes con datos reales, a continuación se listan las actividades que se consideran necesarias realizar en cada una de estas etapas.

DESARROLLO

1. Definir requisitos y alcances del desarrollo (Responsable de proyecto [2 semanas])
2. Realizar cotizaciones para fase de entrenamiento (Responsable de financiamiento [3 semanas])
3. Fase de preprocesamiento (Líder desarrollo [3 semanas])
4. Análisis topológicos (Científico de datos [2 semanas])
5. Desarrollo de red neuronal (Científico de datos / Líder desarr. [5 semanas])

DESARROLLO

6

Validación de predicciones y gráficos (Científico de datos [3 semanas])

7

Validar entrada, tratamiento y predicciones (Líder desarrollo [1 semana])

8

Desarrollar conexión con stream en tiempo real (Ingeniero de software [4 semanas])

9

Evaluar tecnologías para sistema de alertas (Líder desarrollo [2 semanas])

10

Desarrollar integración de sistema de alertas (Ingeniero de software [4 semanas])

11

Despliegue de prueba de concepto (Ingeniero devops / Líder desarr. [2 semanas])

12

Validación general del sistema (Lider de desarrollo [1 semana])

IMPLEMENTACIÓN

Solicitar insumos a cliente (Responsable de cliente [4 semanas]):

Preparar sistema de almacenamiento (Ingeniero devops / Líder desarr. [3 semanas])

Realizar entrenamiento específico (Científico de datos [3 semanas])

Desplegar implementación de sistema de alerta (Ingeniero devops [3 semanas])

RIESGOS Y MITIGACIÓN

Todos los proyectos tienen ciertos riesgos y este no es la excepción, tras un análisis del concepto del sistema a realizar se detectaron los siguientes riesgos los cuales se categorizaron por su probabilidad, gravedad y tipo de riesgo.

	Gravedad				
	1 - Insignificante	2 - Menor	3 - Moderada	4 - Importante	5 - Catastrófica
	5	10	15	20	25
5 - Muy Probable		e- El cliente no sabe cómo reaccionar a alertas del sistema			
	4	8	12	16	20
4 - Probable		e- Las alertas del sistema son ignoradas por los clientes	b- El sistema tiempo real tenga latencia	b- El desarrollo del sistema rebasa las estimaciones de tiempo	
	3	6	9	12	15
3 - Posible			c- Los costos de operación rebasan el financiamiento d- El equipo no tiene los conocimientos/expertise	b- El almacenamiento es costoso	a- El cliente usa tecnologías no compatibles a- Exponer información de clientes de manera accidental
	2	4	6	8	10
2 - No es probable				d- El sistema es lento	a- El cliente no quiere compartir información/accesos
	1	2	3	4	5
1- Muy improbable					

PLAN DE COMUNICACIÓN

La comunicación se considera un factor principal para lograr los objetivos y para mejorar la capacidad de cada uno de los integrantes del equipo por lo que se consideran una serie de acciones, tecnologías y protocolos para mejorarla. En primer lugar se considera en el presupuesto un espacio de coworking que permita a los integrantes reunirse, conocerse y convivir como estímulo para la comunicación, si bien no se rechazan estratégias de trabajo remoto, el postulante debe ser sobresaliente en habilidades comunicativas, para considerarlo una opción viable.

Además se considera apropiado implementar Jira o Github issues para dar seguimiento a las tareas, pendientes, nuevas funcionalidades etc ya que ambas nos permiten:

- Visibilidad del progreso y pendientes
- Asignar responsables
- Dar seguimiento de la tarea a nivel código
- Dar feedback o solicitar cambio de manera persistente, abierta y directa
- Realizar evaluaciones de código de manera jerárquica en el área

Se considera oportuno generar reuniones re arranque de día muy rápidas (15-20 min) y segmentadas por área en las que se aborden los problemas persistentes que se topa el equipo, mediante esta junta, se puede escalar los problemas con los responsables de área y abre el espacio a un análisis detallado del problema y una solución concreta.

Se considera desarrollar una junta con todo el equipo cada vez que se cumpla un hito (completar el preprocesamiento, entrenar la red, completar el sistema de alertas, etc) para mantener una visión clara del progreso, así como realizar ajustes al cronograma.

Finalmente durante las fases de implementación se considera realizar una reunión semanal entre los integrantes del equipo involucrados y el cliente para ofrecer actualizaciones sobre el avance, coordinar esfuerzos y solicitar insumos.

ÉTICA Y CUMPLIMIENTO

Para garantizar el correcto uso de la información así como buenas prácticas de consentimiento y transparencia es necesario realizar varias acciones comenzando por los integrantes del equipo mismo, se debe elaborar un acuerdo de confidencialidad y no divulgación, acompañado de un taller de capacitación sobre el correcto uso de la información recibida. Además es necesario evaluar los términos de servicio con los proveedores de servicio (procesamiento, almacenamiento, etc) asegurando que los datos que manejan y alojen no sean divulgados ni empleados en otras tareas. Respecto al cliente, es necesario elaborar un acuerdo en el que el cliente exprese su consentimiento para compartir información del tráfico de sus servidores y establecer como su responsabilidad notificar a sus usuarios que la información que generen en sus plataformas será compartida para su evaluación. Además el acuerdo con el cliente debe expresar puntualmente qué información y cuánto tiempo se guardará en nuestro sistema, si se podrá usar o no para futuros re-entrenamientos o posibles usos de otra naturaleza.

Por otra parte se debe establecer si los usuarios que generan la información (que acceden a los servidores monitorizados) pueden o no tener derecho a rechazar este tratamiento y queda en responsabilidad del cliente únicamente compartir la información de clientes que hayan aceptado en caso de ser opcional.

Finalmente es necesario establecer políticas y protocolos de seguridad internos que aseguren la integridad, confidencialidad y disponibilidad de los datos al mismo tiempo que se protege ante accesos no autorizados o uso indebido de la información recopilada y almacenada.

SIGUIENTES PASOS

Una vez que las fases de desarrollo e implementación han sido completadas cabe resaltar que para seguir siendo un servicio atractivo, se deben mejorar las capacidades logradas, así como entender las necesidades de los clientes para realizar nuevas propuestas de valor, mejorar la precisión de la detección de anomalías, detectar mejor distintos tipos de ataques cibernéticos, ofrecer alternativas en el sistema de alertas, realizar entrenamientos que permitan evaluar rangos mayores de tiempo, incluso refinar estratégias de marketing, manejo de clientes solo por nombrar algunas, además se pueden explorar maneras de abaratar los costos de operación, o mejorar el performance logrado, en fin al tratarse de un servicio ofertado, siempre se deben buscar maneras de mantenerse en el mercado.

CONCLUSIONES

Los sistemas de aprendizaje por computadora han llegado a cambiar muchos aspectos y flujos de la vida moderna, hay tareas en que las computadoras en general son increíblemente eficientes, una de ellas es la capacidad de procesar información en grandes volúmenes de manera tan rápida que se puede realizar en flujos en tiempo real, esta capacidad es perfecta para integrarse en un sistema de monitoreo, a pesar de ello, poco se ha explorado en las áreas de aprendizaje por computadora y ciberseguridad. Es por ello que se considera que este proyecto es pionero en un campo tan fértil y proporciona una ventaja real sobre otros sistemas de monitoreo y alerta.