



**UNIVERSIDAD DE GUAYAQUIL  
FACULTAD DE INGENIERÍA INDUSTRIAL  
DEPARTAMENTO ACADÉMICO DE GRADUACIÓN**

**TRABAJO DE TITULACIÓN  
PREVIO A LA OBTENCIÓN DEL TÍTULO DE  
INGENIERA EN TELEINFORMÁTICA**

**ÁREA  
TECNOLOGÍAS APLICADAS**

**TEMA**

**DETECCIÓN DE ANOMALÍAS EN LA RED DE LA  
EMPRESA NEWOFFICE UTILIZANDO ALGORITMOS  
DE APRENDIZAJE AUTOMÁTICOS.**

**AUTOR  
VERA CÓRDOVA ANA LUISA**

**DIRECTOR DEL TRABAJO  
ING. CASTILLO LEÓN ROSA ELIZABETH, MG.**

**GUAYAQUIL, JULIO 2020**



**ANEXO XI.- FICHA DE REGISTRO DE TRABAJO  
DE TITULACIÓN**

**FACULTAD DE INGENIERÍA INDUSTRIAL  
CARRERA INGENIERÍA EN TELEINFORMÁTICA**



REPOSITORIONACIONAL EN CIENCIA Y TECNOLOGÍA			
FICHA DE REGISTRO DE TRABAJO DE TITULACIÓN			
TÍTULO Y SUBTÍTULO:			
Detección de anomalías en la red de la empresa NewOffice utilizando algoritmos de aprendizaje automáticos			
AUTOR(ES) (apellidos/nombres):		Vera Córdova Ana Luisa	
REVISOR(ES)/TUTOR(ES) (apellidos/nombres):		Ing. Iván Leonel Acosta Guzmán, MSIG / Ing. Castillo León Rosa Elizabeth, Mg.	
INSTITUCIÓN:		Universidad de Guayaquil	
UNIDAD/FACULTAD:		Facultad Ingeniería Industrial	
MAESTRÍA/ESPECIALIDAD:			
GRADO OBTENIDO:		Ingeniera en Teleinformática	
FECHA DE PUBLICACIÓN:		22 de Octubre 2020	No. DE PÁGINAS: 100
ÁREAS TEMÁTICAS:		Tecnologías Aplicadas	
PALABRAS CLAVES/ KEYWORDS:		Machine Learning, IDS, Algoritmos de aprendizaje, Recolección de datos, sistema de detección.	

En el presente trabajo de titulación se detectan anomalías y ataques informáticos en una red LAN y se analizan si los mismos son de carácter maliciosos y dañinos para los datos que se manejan en la empresa. Para llevar a cabo la propuesta presentada en la empresa NewOffice se procedió a recolectar información mediante el uso de un sistema de detección de intrusiones (IDS) comercial, el cual permitió establecer la base del análisis. Como complementos se detallan en el marco teórico los diferentes tipos de ataques que se pueden encontrar en la red, sistemas operativos, y lenguajes de programación que serán las herramienta principales para el desarrollo de la propuesta, definición y tipos de algoritmos que se manejan en Machine Learning, se describe el paso a paso de la implementación de la data en el algoritmo no supervisado para obtener así el resultado de las anomalías recolectada por los IDS.

In this degree work, computer anomalies and attacks are detected in a LAN network and it is analyzed whether they are malicious and harmful to the data handled in the company. To carry out the proposal presented in the NewOffice company, information was collected using a commercial intrusion detection system (IDS), which allowed to establish the basis of the analysis. As complements, the different types of attacks that can be found on the network, operating systems, and programming languages that will be the main tools for the development of the proposal, definition and types of algorithms that are handled in Machine are detailed in the theoretical framework. Learning, it describes the step-by-step implementation of the data in the unsupervised algorithm to obtain the result of the

anomalies collected by the IDS.		
ADJUNTO PDF:	SI	X NO
CONTACTO CON AUTOR/ES:	Teléfono: 0989966116	E-mail: ana.verac@ug.edu.ec
CONTACTO CON LA INSTITUCIÓN:	Nombre: Ing. Ramón Maquilón Nicola, Mg.	
	Teléfono: 593-2658128	
	E-mail: direccionTi@ug.edu.ec	



**ANEXO XII.- DECLARACIÓN DE AUTORÍA Y DE  
AUTORIZACIÓN DE LICENCIA GRATUITA  
INTRANSFERIBLE Y NO EXCLUSIVA PARA EL USO NO  
COMERCIAL DE LA OBRA CON FINES NO ACADÉMICOS**



**FACULTAD DE INGENIERÍA INDUSTRIAL  
CARRERA INGENIERÍA EN TELEINFORMÁTICA**

---

**LICENCIA GRATUITA INTRANSFERIBLE Y NO COMERCIAL DE LA OBRA CON  
FINES NO ACADÉMICOS**

Yo, **VERA CORDOVA ANA LUISA**, con C.C. No. **0953405867**, certifico que los contenidos desarrollados en este trabajo de titulación, cuyo título es “**DETECCIÓN DE ANOMALÍAS EN LA RED DE LA EMPRESA NEWOFFICE UTILIZANDO ALGORITMOS DE APRENDIZAJE AUTOMÁTICOS**” son de mi absoluta propiedad y responsabilidad, en conformidad al Artículo 114 del CÓDIGO ORGÁNICO DE LA ECONOMÍA SOCIAL DE LOS CONOCIMIENTOS, CREATIVIDAD E INNOVACIÓN, autorizo la utilización de una licencia gratuita intransferible, para el uso no comercial de la presente obra a favor de la Universidad de Guayaquil.

---

**VERA CORDOVA ANA LUISA**

**C.C.No. 0953405867**



**ANEXO VII.- CERTIFICADO PORCENTAJE DE SIMILITUD  
FACULTAD DE INGENIERÍA INDUSTRIAL  
CARRERA INGENIERÍA EN TELEINFORMÁTICA**



Habiendo sido nombrada ING. ROSA ELIZABETH CASTILLO LEÓN MG, tutora del trabajo de titulación certifico que el presente trabajo de titulación ha sido elaborado por VERA CÓRDOVA ANA LUISA, C.C.: 0953405867, con mi respectiva supervisión como requerimiento parcial para la obtención del título de INGENIERA EN TELEINFORMÁTICA

Se informa que el trabajo de titulación: DETECCIÓN DE ANOMALÍAS EN LA RED DE LA EMPRESA NEWOFFICE UTILIZANDO ALGORITMOS DE APRENDIZAJE AUTOMÁTICOS, ha sido orientado durante todo el periodo de ejecución en el programa Antiplagio URKUND quedando el 4% de coincidencia.

**URKUND**

Document: [VERA CORDOVA ANA LUISA.docx](#) (D79860542)

Submitted: 2020-09-24 19:20 (-05:00)

Submitted by: ana.verac@ug.edu.ec

Receiver: rosa.castillo@ug@analysis.arkund.com

Message: VERA CORDOVA ANA LUISA [Show full message](#)

4% of this approx. 32 pages long document consists of text present in 2 sources.

Sources		Highlights
	Rank	Path/Filename
<input type="checkbox"/>		<a href="#">Tesis-Aprendizaje.docx</a>
<input checked="" type="checkbox"/>		REYES MARTINEZ MIRIAM MONOGRAFIA.docx
<b>Alternative sources</b>		
<input type="checkbox"/>		Francis 15042020.docx
<b>Sources not used</b>		

0 Warnings   Reset   Export   Share

algoritmos de aprendizaje no supervisados permiten a los usuarios realizar tareas de procesamiento más complejas en comparación con el aprendizaje supervisado. Sin embargo, el aprendizaje no supervisado puede ser más impredecible en comparación con otros métodos de aprendizaje natural. Los algoritmos de aprendizaje no supervisados incluyen agrupamiento, detección de anomalías, redes neuronales, etc.

2.2.7.1. Clustering El clustering o agrupación puede considerarse el algoritmo de aprendizaje no supervisado más importante; entonces, como cualquier otro algoritmo de este tipo, se trata de encontrar una estructura en una colección de datos sin etiquetar. Asimismo, una definición poco precisa de agrupamiento podría

<https://secure.arkund.com/view/76427143-226782-405671>

**ING. ROSA ELIZABETH CASTILLO LEÓN**  
**DOCENTE TUTOR**  
**C.C. 0922372610**

**FECHA: 01 DE OCTUBRE DE 2020**



**ANEXO VI. - CERTIFICADO DEL DOCENTE-TUTOR DEL  
TRABAJO DE TITULACIÓN  
FACULTAD DE INGENIERÍA INDUSTRIAL  
CARRERA INGENIERÍA EN TELEINFORMÁTICA**



Guayaquil, 01 de octubre de 2020

Sra.

**Ing. Annabelle Lizarzaburu Mora, MG.**

Directora de Carrera Ingeniería en Teleinformática / Telemática

**FACULTAD DE INGENIERÍA INDUSTRIAL DE LA UNIVERSIDAD DE  
GUAYAQUIL**

Ciudad. -

De mis consideraciones:

Envío a Ud. el Informe correspondiente a la tutoría realizada al Trabajo de Titulación **DETECCIÓN DE ANOMALÍAS EN LA RED DE LA EMPRESA NEWOFFICE UTILIZANDO ALGORITMOS DE APRENDIZAJE AUTOMÁTICOS** de la estudiante **VERA CORDOVA ANA LUISA**, indicando que ha cumplido con todos los parámetros establecidos en la normativa vigente:

- El trabajo es el resultado de una investigación.
- El estudiante demuestra conocimiento profesional integral.
- El trabajo presenta una propuesta en el área de conocimiento.
- El nivel de argumentación es coherente con el campo de conocimiento.

Adicionalmente, se adjunta el certificado de porcentaje de similitud y la valoración del trabajo de titulación con la respectiva calificación.

Dando por concluida esta tutoría de trabajo de titulación, **CERTIFICO**, para los fines pertinentes, que la estudiante está apta para continuar con el proceso de revisión final.

Atentamente,

**ING. ROSA ELIZABETH CASTILLO LEÓN, MG.**

**C.C. 0922372610**

**FECHA: 01 DE OCTUBRE DE 2020**



**ANEXO VIII.- INFORME DEL DOCENTE REVISOR**  
**FACULTAD DE INGENIERÍA INDUSTRIAL**  
**CARRERA INGENIERÍA EN TELEINFORMÁTICA**



Guayaquil, 14 de Octubre del 2020.

Sr (a).

**Ing. Annabelle Lizaraburu Mora, MG.**

Director (a) de Carrera Ingeniería en Teleinformática / Telemática

**FACULTAD DE INGENIERÍA INDUSTRIAL DE LA UNIVERSIDAD DE GUAYAQUIL**  
 Ciudad. -

De mis consideraciones:

Envío a Ud. el informe correspondiente a la REVISIÓN FINAL del Trabajo de Titulación **“DETECCIÓN DE ANOMALÍAS EN LA RED DE LA EMPRESA NEWOFFICE UTILIZANDO ALGORITMOS DE APRENDIZAJE AUTOMÁTICOS”** de la estudiante, **VERA CÓRDOVA ANA LUISA**. Las gestiones realizadas me permiten indicar que el trabajo fue revisado considerando todos los parámetros establecidos en las normativas vigentes, en el cumplimiento de los siguientes aspectos:

Cumplimiento de requisitos de forma:

El título tiene un máximo de 22 palabras.

La memoria escrita se ajusta a la estructura establecida.

El documento se ajusta a las normas de escritura científica seleccionadas por la Facultad.

La investigación es pertinente con la línea y sublíneas de investigación de la carrera.

Los soportes teóricos son de máximo 5 años.

La propuesta presentada es pertinente.

Cumplimiento con el Reglamento de Régimen Académico:

El trabajo es el resultado de una investigación.

El estudiante demuestra conocimiento profesional integral.

El trabajo presenta una propuesta en el área de conocimiento.

El nivel de argumentación es coherente con el campo de conocimiento.

Adicionalmente, se indica que fue revisado, el certificado de porcentaje de similitud, la valoración del tutor, así como de las páginas preliminares solicitadas, lo cual indica el que el trabajo de investigación cumple con los requisitos exigidos.

Una vez concluida esta revisión, considero que el estudiante está apto para continuar el proceso de titulación. Particular que comunicamos a usted para los fines pertinentes.

Atentamente,



Firmado electrónicamente por:

**IVAN LEONEL ACOSTA GUZMAN**

---

**REVISOR**

**ING. IVAN ACOSTA GUZMAN, MSIG**

**FECHA: 14 OCTUBRE DEL 2020**

## **Dedicatoria**

El presente trabajo está a mis padres, el Sr. Justo Vera y la Sra. Esperanza Córdova que fueron, son y serán el pilar fundamental en cada uno de los pasos que he dado hasta este momento, la paciencia y el amor con el que me educaron e inculcaron a ser una persona responsable en mis actividades, son la motivación y la razón de cada uno de mis logros.

A mi esposo el Ing. Willis Mero quien ha sabido complementar las diferentes destrezas a lo largo de la carrera universitaria de los cuales se han obtenido grandes y fructíferos resultados.



## **Agradecimiento**

A mis padres por guiarme por el mejor camino desde pequeña, por ser perseverantes con sus enseñanzas, por no dejarme sola y apoyarme siempre.

A mis maestros de este camino recorrido en mi ámbito estudiantil, cada uno formo cada parte de mis conocimientos en cada etapa que surgía.

A mis amigos, los que aparecieron en cada etapa, hasta los que se fueron quedando en el camino, perseverantes y ambiciosos de conocimiento que extendieron su mano para ayudarme cuando lo necesitaba.

## Índice General

No	Descripción	Pág.
	Introducción	1

### Capítulo I El Problema

No	Descripción	Pág.
1.1	Planteamiento del Problema	3
1.2	Formulación del Problema	4
1.3	Sistematización del Problema	4
1.4	Objetivos	4
1.4.1	Objetivo General	4
1.4.2	Objetivos Específicos	4
1.5	Justificación e Importancia	4
1.6	Delimitación del Problema	5
1.7	Alcance del Problema	5
1.8	Premisa de la Investigación	6
1.9	Operacionalización	6

### Capítulo II Marco Teórico

No	Descripción	Pág.
2.1	Antecedente	8
2.2	Referencias Teóricas	9
2.2.1	Importancia del Dominio de Detección de Intrusiones en la Seguridad de la Red	9
2.2.2	Selección de Funciones y Proceso de Clasificación para Predecir el Comportamiento del Atacante	10
2.2.3	Detección de Intrusiones en la Red	11
2.2.4	Selección de Atributos	11
2.2.5	Proceso de Clasificación	13
2.2.5.1	Exactitud	15
2.2.5.2	Descripción de Datos	16

<b>No</b>	<b>Descripción</b>	<b>Pág.</b>
2.2.6	Algoritmo de Aprendizaje	17
2.2.6.1	Algoritmo de Aprendizaje Supervisado	17
2.2.6.2	Máquinas de Vector Soporte	17
2.2.6.3	Algoritmos de Árbol de Decisión	17
2.2.6.4	Bosques Aleatorios (Random Forests)	18
2.2.5.5	K Nearest Neighbors	18
2.2.7	Algoritmo de Aprendizaje no Supervisado	18
2.2.7.1	Clustering	19
2.2.7.2	Algoritmo DBSCAN	19
2.2.7.3	K-means	20
2.2.7.4	Algoritmos de Estimación - Algoritmo Isolation Forest	21
2.2.8	Modelos de Mezcla Gaussiana (MMG	21
2.2.9	Comparación entre Algoritmo de Aprendizaje Supervisado y no Supervisados	22
2.2.10	Algoritmo de Meanshift	23
2.2.11	Agrupación de Algoritmos de Mean Shift	23
2.2.12	Estimación de Gradiente de Densidad	24
2.2.13	Método MSB-CGH	25
2.2.14	Técnica de Q-learning fuera de la Política para la Respuesta a la Intrusión.	27
2.2.14.1	Aprendizaje Reforzado	27
2.2.14.2	Algoritmos Q-Learning	27
2.2.15	Política y Selección de Políticas	28
2.2.15.1	Exploración vs. Explotación	28
2.2.16	Lenguaje de Programación	29
2.2.16.1	Lenguaje de Programación Python	29
2.2.16.2	Lenguaje de Programación C#	29
2.2.16.3	Diferencias entre Python & C#	29
2.2.16.4	Error de Programación	30
2.2.17	Software	30
2.2.17.1	Sistemas Operativos	31

<b>No</b>	<b>Descripción</b>	<b>Pág.</b>
2.2.17.2	Software de la Aplicación	31
2.2.17.3	Compiladores e Intérpretes	31
2.2.17.4	Sistema de Monitorización	31
2.2.18	Redes LAN	32
2.2.19	Tipos de Anomalías de Red	32
2.2.19.1	Ataques DoS yDDoS	32
2.2.19.2	Flashcrowd	32
2.2.19.3	Escaneo.	33
2.2.19.4	Gusano	33
2.2.19.5	Punto a Multipunto	33
2.2.20	Firewall	33
2.2.20.1	PfSense	34
2.2.20.2	Snort IDS	34
2.2.20.3	Características Básicas de los IDS basados en firmas	34
2.2.20.4	Suricata IDS	35
2.2.21	Marco Legal	35

### **Capítulo III**

#### **Propuesta**

<b>No</b>	<b>Descripción</b>	<b>Pág.</b>
3.1	Modalidad de la Investigación	38
3.2	Tipos de Investigación	38
3.3	Metodología de la Investigación	38
3.4	Área de Estudio	39
3.4.1	Infraestructura de la Red	39
3.4.2	Infraestructura de Equipos	40
3.4.3	Esquema de Seguridad de la Red	42
3.4.3.1	Seguridad Virtual	42
3.4.3.2	Seguridad Física	42
3.5	Levantamiento del Tipo de Información a Proteger	42
3.5.1	Análisis de Posibles Amenazas	43

<b>No</b>	<b>Descripción</b>	<b>Pág.</b>
3.5.2	Resultado de Escaneo de la Red	46
3.6	Propuesta Tecnológica	48
3.7	Etapas de la Metodología del Proyecto	47
3.7.1	Diagrama de la Estructura Física de la Red	47
3.7.2	Requerimiento del Software	48
3.7.3	Activación de los IDS	49
3.8	Ambiente de Desarrollo Python	51
3.9	Ataque Realizado en la Red	53
3.10	Extracción de las Alertas Snort y Suricata.	54
3.10.1	Limpieza de Datos	55
3.10.2	Inserción de Data en Algoritmo	56
3.11	Creación de Algoritmo de Carga de Datos	58
3.11.1	Algoritmos de Estimación	60
3.12	Observación de Resultados	64

**Índice de Tablas**

<b>No</b>	<b>Descripción</b>	<b>Pág.</b>
1	Delimitación de Problema	5
2	Variables Dependientes e Independientes	6
3	Resultados de detención de ataque con respecto a VGA, WVGA y FV-GA	16
4	Comparación entre Algoritmo de aprendizaje supervisado y no supervisados	21
5	Comparación de características Python y C#	29
6	Infraestructura de Equipos Disponibles en la empresa NewOffice	41
7	Información que Proteger de la empresa.	42
8	Rangos de la IP de la Red	45
9	Requerimientos de Hardware	47
10	Requerimientos de Software	47
11	Requerimientos del Algoritmo	57

## Índice de Figuras

<b>No</b>	<b>Descripción</b>	<b>Pág.</b>
1	Esquema General de un Sistema de Detección de Intrusos	3
2	Procedimientos Nasca	12
3	Procedimiento de Clasificación de un Ataque	14
4	Podado de Árbol para más Exactitud	16
5	Matriz de Confusión	22
6	Representación de Datos en Algoritmo Meanshift	24
7	Parcela de Superficie KDE	25
8	Algoritmo Q-learnig	28
9	Selección de Acción Codiciosa	28
10	Diagrama de la Arquitectura en la Red actual NewOffice	40
11	Esquema de Equipos en la Red Actual	43
12	Escaneado del Servidor 192.168.0.11	44
13	Escaneado del Servidor 192.168.0.9	45
14	Vulnerabilidades de los Servidores	45
15	Vulnerabilidades de los Servidores NMAP	47
16	Pantalla Inicial de Control Pfsense	49
17	Pack de Instalación en Interfaz del Firewall Pfsense	50
18	Pfsense antes de Funcionamiento	50
19	IDS Activados y en Funcionamiento	50
20	Página Oficial PIP	52
21	Comando de Instalación en Windows	52
22	Zerophgemalware Contaminación de red	54
23	Sistema de Alertas. Ambiente de trabajo de Pfsense	54
24	Descargas de Archivos de Alertas en IDS.	55
25	Transformación de Archivos Txt a Csv	55
26	Formato Suricata csv	56
27	Formato Snort csv	56
28	Inserción de Dataset de Snort	58
29	Extracción al Dataset de Suricata	59
30	Revisión de Valores Insertados en Algoritmo	59
31	Paso de Variables Numéricas a Enteros	60

<b>No</b>	<b>Descripción</b>	<b>Pág.</b>
32	Comando de Estimación	60
33	Comando de Arreglos y Entrenamiento de Algoritmo	61
34	Comandos de Predicción en Dataset de Suricata	62
35	Comando de Predicción en Dataset de Snort	62
36	Matriz de Confusión de Algoritmo MeanShief	63
37	Comandos Finales de Algoritmo	63
38	Grafica de Conjunto de Datos Tomados de Snort	64
39	Evaluación de Falsos Positivos tomados de Dataset de Snort.	65
40	Grafica de Conjunto de Datos tomados de Suricata	66
41	Estimación de Falsos Positivos de datos obtenidos de Suricata	66



**Índice de Anexos**

<b>No</b>	<b>Descripción</b>	<b>Pág.</b>
1	Instalación de Pfsense	70
2	Instalación de Snort	72
3	Reglas de Snort	72
4	Interfax a Snort	73
5	Instalación de Suricata	75
6	Modelo de Algoritmo Meanshift	76



**FACULTAD DE INGENIERÍA INDUSTRIAL**  
**CARRERA DE INGENIERÍA EN TELEINFOMÁTICA**  
**UNIDAD DE TITULACIÓN**  
**“DETECCIÓN DE ANOMALÍAS EN LA RED DE LA EMPRESA**  
**NEWOFFICE UTILIZANDO ALGORITMOS DE APRENDIZAJE**  
**AUTOMÁTICOS”**

**Autor:** Vera Córdova Ana Luisa

**Tutor:** Ing. Castillo León Rosa Elizabeth, Mg.

En el presente trabajo de titulación se realizó la detección de anomalías en la red LAN de la empresa NewOffice la cual no cuenta con un sistema de seguridad y se encuentra expuesta a posibles ataques de ciberseguridad, se analizan si los mismos son de carácter maliciosos para los datos que se manejan en la empresa. Para llevar a cabo esta propuesta, se procedió con la recolección de información mediante el uso de un sistema de detección de intrusiones (IDS) comercial, se aplicó la modalidad bibliográfica para el desarrollo del trabajo y la investigación descriptiva para establecer la base del análisis de los datos obtenidos. A nivel teórico se describe los diferentes tipos de ataques en una red, sistemas operativos, y las herramientas principales para el funcionamiento del algoritmo no supervisado, como resultado se presentan diferentes categorías en forma de clúster, que aplicando nuevos comandos de ejecución se pueden evitar.

**Palabras Claves:** Machine Learning, IDS, Algoritmos de aprendizaje, Recolección de datos, sistema de detección.



**FACULTY OF INDUSTRIAL  
TELEINFORMATICS ENGINEERING CAREER**  

---

**“DETECTION OF ANOMALIES IN THE NEWOFFICE COMPANY  
NEWOFFICE USING AUTOMATIC LEARNING ALGORITHMS”**

**Author** Vera Córdova Ana Luisa

**Advisor:** Ing. Castillo León Rosa Elizabeth, Mg.

In this degree work, computer anomalies and attacks are detected in a LAN network and it is analyzed whether they are malicious and harmful to the data handled in the company. To carry out the proposal presented in the NewOffice company, information was collected using a commercial intrusion detection system (IDS), which allowed to establish the basis of the analysis. As complements, the different types of attacks that can be found on the network, operating systems, and programming languages that will be the main tools for the development of the proposal, definition and types of algorithms that are handled in Machine Learning, it describes the step-by-step implementation of the data in the unsupervised algorithm to obtain the result of the anomalies collected by the IDS.

**Keywords:** Machine Learning, IDS, Learning algorithms, Data collection, detection system.

## **Introducción**

Dado el avance continuo de las aplicaciones de red y nuestra creciente dependencia de los sistemas basados en software, existe una necesidad apremiante de desarrollar técnicas de seguridad mejoradas para defender los sistemas modernos de tecnología de la información (TI) de los ciber ataques maliciosos. De hecho, cualquiera puede verse afectado por tales actividades, incluidos individuos, corporaciones y gobiernos (Seguridad, 2016).

La expansión sostenida de la base de usuarios de la red y su conjunto de aplicaciones asociadas también está introduciendo vulnerabilidades adicionales que pueden conducir a infracciones penales y pérdida de datos críticos. Como resultado, el área más amplia del problema de ciberseguridad se ha convertido en una preocupación importante, con muchas estrategias de solución propuestas para la detección y prevención de intrusos. Ahora, en general, el dilema de ciberseguridad se puede tratar como una configuración de resolución de conflictos que implica un sistema de seguridad y un mínimo de dos agentes de decisión con objetivos en competencia (por ejemplo, el atacante y el defensor). Es decir, por un lado, el defensor se enfoca en garantizar que el sistema opere a un nivel adecuado (o más alto). Por el contrario, el atacante se centra en tratar de interrumpir o corromper la operación del sistema (Mogollon, 2017).

A la luz de lo anterior, esta disertación presenta metodologías novedosas para construir estrategias apropiadas para los administradores del sistema (defensores). En particular, se desarrollan modelos matemáticos detallados de sistemas de seguridad para analizar el rendimiento general y predecir el comportamiento probable de los tomadores de decisiones clave que influyen en la estructura de protección. El objetivo inicial aquí es crear un mecanismo confiable de detección de intrusos para ayudar a identificar ataques maliciosos en una etapa muy temprana, es decir, para minimizar las consecuencias potencialmente críticas y el daño a la privacidad y estabilidad del sistema. Además, otro objetivo clave es también desarrollar mecanismos efectivos de prevención de intrusiones (respuesta). En este sentido, se desarrolla un marco de solución basado en aprendizaje automático que consta de dos módulos. Específicamente, el primer módulo prepara el sistema para el análisis y detecta si hay un ciberataque o no. mientras tanto, el segundo módulo analiza el tipo de incumplimiento y formula una respuesta adecuada. Es decir, se utiliza un agente de decisión en el último módulo para investigar el entorno y tomar decisiones apropiadas en

caso de incertidumbre. Este agente comienza realizando su análisis en un entorno completamente desconocido, pero continuamente aprende a ajustar su toma de decisiones en función de los comentarios proporcionados. El sistema general está diseñado para operar de manera automatizada sin ninguna intervención de los administradores u otro personal de seguridad cibernética. La entrada humana solo se requiere esencialmente para modificar algunos parámetros y configuraciones clave del modelo (sistema). En general, el marco desarrollado en esta disertación proporciona una base sólida desde la cual desarrollar mecanismos mejorados de detección y protección de amenazas para configuraciones estáticas, con una mayor extensibilidad para manejar la transmisión de datos.

# Capítulo I

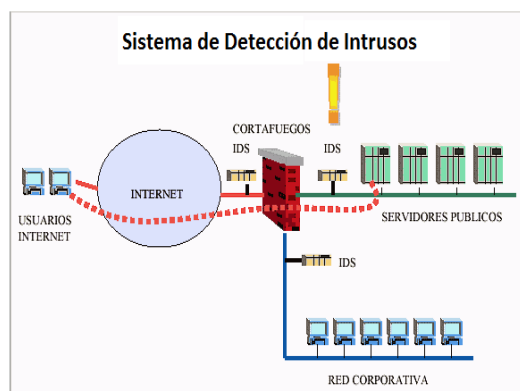
## El Problema

### 1.1 Planteamiento del Problema

El desarrollo significativo de los sistemas informáticos ha transformado por completo nuestra vida diaria e hizo que nuestra existencia dependiera de ellos. Según (Cisco, 2019), se esperan 3.5 dispositivos informáticos per cápita en todo el mundo en 2021 y casi 106 terabytes por segundo de tráfico global de Internet. Con el rápido progreso de Internet, nuestras estructuras informáticas están expuestas a un mayor número de amenazas. Aunque la investigación y las innovaciones tecnológicas están progresando rápidamente, una ciberseguridad absoluta sigue siendo un desafío.

El IDS observa el tráfico de la red, lo analiza e identifica posibles anomalías o acceso no autorizado al comportamiento de la red. Algunos de los IDS también responden a la intrusión, que es una medida necesaria para proteger nuestra red informática. Existen varias limitaciones y problemas de los métodos existentes que se indicara en este capítulo con el modelo propuesto de respuesta a la intrusión de Q-learning fuera de la política. Por un lado, la explotación y el mal uso de los recursos ocurren porque el IDS está diseñado para observar la red todo el tiempo; en consecuencia, los recursos se utilizan incluso si no se produce ningún ataque (Morales, 2018).

Por otro lado, aunque el tráfico que fluye se examina continuamente, una vez que se detecta un ataque, hay un tiempo considerable necesario para proporcionar una respuesta. El tráfico de red a menudo viaja una cierta distancia en forma de paquetes, además, el intruso puede alternarlo o incluso terminarlo antes de llegar al IDS. Otro desafío es proporcionar una forma confiable de proteger nuestro sistema o hasta qué punto se pueda confiar en los IDS.



**Figura 1.** Esquema general de un Sistema de detección de intrusos. Tomado de (Morales, 2018)  
Elaborado por Vera Córdova Ana Luisa

## **1.2 Formulación del Problema.**

¿Cómo mejorar la seguridad de la red LAN de la empresa NewOffice a través de la tecnología IDS?

## **1.3 Sistematización del Problema.**

Con la implementación de un sistema de detección y prevención de intrusos mejorará el resguardo de la red corporativa de la empresa NewOffice al localizar cualquier intento de intrusión, transferencia de código malicioso o amenazas a través de la red, sin repercusión alguna sobre su rendimiento. Esta utilidad funcionará de carácter transparente para el beneficiario, por lo que no requiere ninguna reconfiguración de la red existente a la que se conecta. Utilizando técnicas de identificación de protocolos, identificación y análisis de tráfico, es capaz de detectar e identificar al administrador de la red frente a amenazas informáticas.

## **1.4 Objetivos**

### **1.4.1 Objetivo General**

Crear un modelo basado en algoritmo de aprendizaje automático sobre un sistema de detección de intruso implementado en la empresa NewOffice para detectar futuros ataques en la red.

### **1.4.2 Objetivos Específicos**

- Realizar el levantamiento de información de la red establecida en la empresa.
- Analizar las herramientas técnicas idóneos disponibles en el mercado que permitirán el impulso de un sistema de detección de intrusos para una red.
- Capturar datos en la red mediante la utilización del sistema de detección de intrusos.
- Aplicar el algoritmo de aprendizaje automático sobre los datos obtenidos de la empresa NewOffice.

## **1.5 Justificación e Importancia.**

La investigación y desarrollo de esta permitirá poseer unas herramientas de seguridad que permita obtener información sobre posibles ataques o intentos de intrusión, en la red local, ayudará a tener informes de análisis de vulnerabilidades que permitan identificar los puntos vulnerables de los servidores y de la misma manera permita corregir y controlar la información de los usuarios.

La importancia del presente trabajo es implementar un sistema que permita observar el tráfico de la red, analizar e identificar posibles anomalías o acceso no autorizado al comportamiento de la red. Algunos de los IDS también responden a la intrusión, que es una medida necesaria para proteger nuestra red informática. Gracias a la investigación se podrá tener una documentación bien estructurada que permitirá a otros investigadores abordar temas de machine learning, métodos de aprendizaje, y estudios de redes y paquetes cifrados de red, la cual es la columna vertebral de una buena práctica de seguridad e integridad de datos.

Por otro lado, aunque el tráfico que fluye se examina continuamente, una vez que se detecta un ataque, hay un tiempo considerable necesario para proporcionar una respuesta. El tráfico de red a menudo viaja una cierta distancia en forma de paquetes; además, el intruso puede alternarlo o incluso terminarlo antes de llegar al IDS. Los administradores deben actualizar periódicamente sus mecanismos de protección; de lo contrario, una vez que el intruso reconozca debilidades y limitaciones específicas, enviará aún más ataques, desafiando así el sistema de detección.

### 1.6 Delimitación del Problema

**Tabla 1.** *Delimitación del problema*

Delimitación del Problema	
Campo	Tecnologías Aplicadas
Área	Inteligencia Artificial
Aspecto	Seguridad de la Información
Tema	Crear un modelo basado en algoritmo de aprendizaje automático sobre un sistema de detección de intruso implementado en la Empresa NewOffice para detectar futuros ataques en la red.

*Información adaptada en relación con la importancia de proteger la información. Elaborado por el Vera Córdova Ana Luisa.*

### 1.7 Alcance del Problema.

El alcance del proyecto comprende:

- Identificar la principal información sobre los sistemas de detección de Intrusos (IDS), en una red LAN.
- Profundizar los conocimientos sobre algoritmos de aprendizaje aplicados a la detección de intrusos en la red.



- Implementar un sistema de detección de intrusiones que permita detectar algún tipo anomalía en la red.
- Seleccionar un algoritmo idóneo para realizar el análisis de los datos resultantes de los IDS
- Analizar resultados obtenidos de implementación de algoritmos en datos recolectados del sistema de detección aplicados.

### 1.8 Premisa de la Investigación.

La propuesta permitirá tener un sistema de detección de intrusos no autorizados a la red de la Empresa NewOffice.

### 1.9 Operacionalización.

**Tabla 2** Variables independientes y dependientes

Variables	Dimensiones	Indicadores	Técnica & instrumento
<b>Independiente:</b> Sistema de detección y prevención de Intrusos	<ul style="list-style-type: none"> <li>• Hardware</li> <li>• Software</li> <li>• Acceso de usuarios no autorizados</li> </ul>	<ul style="list-style-type: none"> <li>• Router</li> <li>• Servidores</li> <li>• Estaciones de Trabajo</li> <li>• Anti virus</li> <li>• Firewall</li> <li>• Usuarios Externos</li> </ul>	<ul style="list-style-type: none"> <li>• Encuestas</li> <li>• Bloc de Notas</li> <li>• Encuestas</li> <li>• Bloc de Notas</li> <li>• Encuestas</li> <li>• Bloc de Notas</li> </ul>
<b>Dependiente:</b> Vulnerabilidad en los servidores	<ul style="list-style-type: none"> <li>• Servidores.</li> <li>• Configuraciones</li> <li>• Atacantes</li> <li>• Información o Servicios</li> </ul>	<ul style="list-style-type: none"> <li>• Capacidad</li> <li>• Modelo y marca</li> <li>• Cantidad</li> <li>• Sistema operativo</li> <li>• Servidores</li> <li>• Equipos</li> <li>• Routers</li> <li>• Tipos de ataque del atacante.</li> <li>• Herramientas del atacante.</li> <li>• Servidor Web</li> </ul>	<ul style="list-style-type: none"> <li>• Encuestas</li> <li>• Bloc de Notas</li> <li>• Encuestas</li> <li>• Bloc de Notas</li> <li>• Encuestas</li> <li>• Bloc de Notas</li> <li>• Documentación Bibliográfica.</li> </ul>

<b>Variables</b>	<b>Dimensiones</b>	<b>Indicadores</b>	<b>Técnica &amp; instrumento</b>
		<ul style="list-style-type: none"> <li>• Servicio de Motor de bases de datos</li> <li>• Servidor de correo</li> <li>• Aplicaciones web</li> <li>• Aplicaciones de Escritorio</li> </ul>	<ul style="list-style-type: none"> <li>• Encuestas</li> <li>• Bloc de Notas</li> </ul>

*Información adaptada en relación con la importancia de proteger la información. Elaborado por el Vera Córdova Ana Luisa.*

## **Capítulo II**

### **Marco Teórico**

#### **2.1 Antecedentes**

En la Investigación de Zheni (2018) en la cual esta titulada “Métodos de aprendizaje automático para la detección de intrusiones en red y los sistemas de prevención de intrusiones”. Además, la expansión sostenida de la base de usuarios de la red y su conjunto de aplicaciones asociadas también está introduciendo vulnerabilidades adicionales que pueden conducir a infracciones penales y pérdida de datos críticos. Como resultado, el área más amplia del problema de ciberseguridad se ha convertido en una preocupación importante, con muchas estrategias de solución propuestas para la detección y prevención de intrusos.

Ahora, en general, el dilema de ciberseguridad puede tratarse como una configuración de resolución de conflictos que implica un sistema de seguridad y un mínimo de dos agentes de decisión con objetivos en competencia (por ejemplo, el atacante y el defensor). Es decir, por un lado, el defensor se enfoca en garantizar que el sistema opere a un nivel adecuado (o más alto). Por el contrario, el atacante se centra en tratar de interrumpir o corromper la operación del sistema.

Según Bhattacharjee & Md Fujail, (2017), describe que, con el rápido aumento en el uso de computadoras en red en las últimas décadas, ha habido un aumento en muchos tipos diferentes de ataques a la red por parte de intrusos. Para detectar diferentes ataques a la red, en este documento se emplea el Sistema de detección de intrusiones (IDS) basado en Algoritmo genético (GA). El objetivo es encontrar una función de Vectorized Fitness adecuada para las evaluaciones de cromosomas para obtener una solución para IDS. Para lograr este objetivo, se proponen y evalúan IDS basados en Algoritmo genético con función ponderada de Vectorized Fitness sobre el conjunto de datos NSL-KDD. En el presente trabajo, la función de membresía Fuzzy se utiliza con la función Vectorized Fitness en GA para la detección de intrusiones eficientes. Los resultados experimentales muestran que la GA vectorizada difusa propuesta funciona mejor que la GA vectorizada y la GA vectorizada ponderada en la detección de ataques de red para el conjunto de datos NSL-KDD considerado.

## 2.2 Referencias Teóricas

### 2.2.1 Importancia del Dominio de Detección de Intrusiones en la Seguridad de la Red

Estos enfoques de aprendizaje automático estudian principalmente un conjunto de características de red (donde no hay orden o secuencia en los detalles de la red), y su objetivo es típicamente calcular un puntaje de categoría. Además, la pregunta que surge es qué tan bien se desempeñarán en la clasificación de la red, especialmente cuando hay más ruido y la estructura de información no es homogénea. Se necesita probar qué tan precisos son estos métodos para predecir categorías para redes grandes, donde la información se concentra en unas pocas características. Además, la atención se centrará en ayudar de manera accesible a los lectores potenciales de este trabajo a descubrir un mecanismo de defensa eficiente que proteja su sistema de red. Como ejemplo del papel crucial de la ciberseguridad, el uso progresivo de la tecnología inalámbrica está haciendo que las redes sean aún más vulnerables a los ataques. Sin embargo, los desafíos avanzados y sofisticados hacen que las medidas de seguridad clásicas, como un firewall, sean inadecuadas. Si un hacker quiere robar datos, no intentará penetrar en el firewall, pero buscará el acceso menos seguro para tomar el control del sistema. Los ciberdelincuentes continuamente inventan nuevas técnicas. Por lo tanto, se necesita encontrar soluciones efectivas que puedan defender de manera dinámica y adaptativa nuestros sistemas.

En la Investigación actual, se presenta un algoritmo de aprendizaje automático para sistemas de detección y prevención de intrusiones en red (IDPS) que servirá como una herramienta exitosa para defender nuestra red. El IDPS propuesto incorpora una capa: la capa brindará la oportunidad de monitorear la red y detectar una falla en el sistema, en base a técnicas de reducción y clasificación de dimensiones de aprendizaje automático. Esta red IDPS está totalmente adaptada a la transmisión de datos y se puede implementar en cualquier servidor, donde se requiere un análisis de datos en línea. La mayoría de los sistemas IDS funcionan por sí solos junto con un firewall. Sin embargo, están restringidos a sus funciones de detección y monitoreo. Por lo tanto, la introducción de un mecanismo individual de detección de intrusión que funcionará junto con el IDS permitirá una protección integral, que no solo asegurará el sistema, sino que también creará un agente único que tomará una decisión automatizada en un entorno desconocido.

El esquema propuesto es una técnica avanzada de sistemas de aprendizaje automático para la reducción y clasificación de características en la configuración de seguridad de red. La representación sugerida de la red informática como una estructura de probabilidad es un enfoque innovador en el que el agente de decisión operará en un entorno desconocido, obtendrá comentarios y, en función de la recompensa recibida, aprenderá a seleccionar una política óptima, por lo que red para estar continuamente protegida.

### **2.2.2 Selección de Funciones y Proceso de Clasificación para Predecir el Comportamiento del Atacante.**

El procedimiento de selección, clasificación y precisión de atributos de red (NASCA) es un procedimiento de cuatro pasos para sistemas de detección de intrusos. El modelo comienza primero extrayendo la información del Protocolo de Control de Transmisión (TCP) y luego clasificando la información relevante que caracteriza a la red. Durante la segunda etapa, que clasifica la red como estar bajo ataque o no. Por otra parte, si el modelo detecta que el sistema está bajo ataque, se ejecuta otro nivel de un enfoque de clasificación para identificar el tipo de ataque que se está produciendo. La tercera etapa proporciona resultados sobre la precisión de la estimación analítica realizada.

El método de selección de atributos comienza con la elección de un subconjunto de la información adecuada mediante la eliminación de redundancia, sin relación, y datos de ruido del conjunto de datos original. El método de clasificación comienza inicialmente con un conjunto existente de redes etiquetados; en consecuencia, se entera de la dependencia entre el contenido de la red y su etiqueta correspondiente y, a continuación, predice la etiqueta de un conjunto de redes no marcadas con la mayor precisión posible el uso de un tipo de árbol de decisión de análisis. Los objetivos del procedimiento NASCA se detallan de la siguiente manera;

- Proponer un procedimiento NASCA para sistemas de detección de intrusos;
- Para probarlo en un conjunto de datos reales;
- Informar los resultados obtenidos para proporcionar evidencia de por qué este procedimiento puede superar a los de clasificación y clasificación técnicas existentes.

### 2.2.3 Detección de Intrusiones en la Red: esquema del Procedimiento.

El agente de clasificación es capaz de tomar decisiones en un entorno en constante cambio y, por lo tanto, probar el modelo mientras evalúa la red. Además, una ventaja esencial es el hecho de que solo se utilizará información relevante de la red antes de que se tome la decisión de clasificación particular. En esas circunstancias, el agente puede aprender a clasificar la red de manera rápida, precisa y eficiente.

**Recopilación de Datos de red:** Recoge los datos de la red del TCP / IP, utilizado para ello cualquier instrumento de comunicación tipo conector directo para la lectura de los datos;

**Selección de Atributos:** Aplica un enfoque de evaluación de atributos de ganancia de información, que es una herramienta para la reducción de características y la clasificación de atributos.

**Clasificación:** Proporcione un procedimiento de clasificación, que aprenderá a etiquetar la red de manera precisa y rápida. En este trabajo, se aplicará un procedimiento de dos etapas:

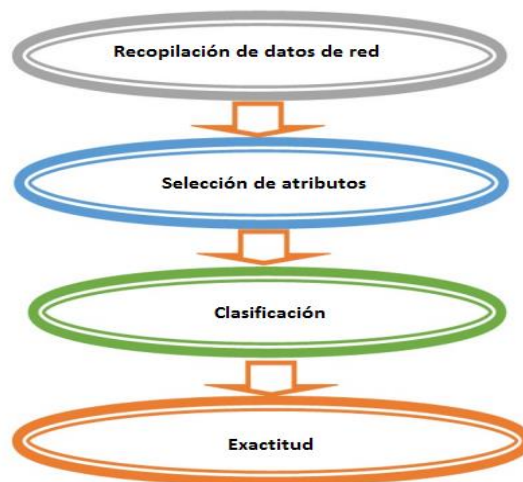
- a) Clasifique la red como atacada o no. En el primer paso de la clasificación, se utilizará un bosque aleatorio (RF) para etiquetar la red como atacada o no; si el paso (a) etiqueta la red como si estuviera bajo ataque, entonces use el paso (b).
- b) Clasifique aún más la red si se produce un ataque. Se aplicará un método de árbol de decisión parcial (PART) para clasificar la posible estimación adicional para el tipo de ataque que está sufriendo.

Precisión: se demostrará la precisión de la estimación para el conjunto de datos específico utilizado. Una vez que se entrena la información y se construye un modelo, se informan los factores de precisión.

### 2.2.4 Selección de Tributos, utilizando el Enfoque de Evaluación de Atributos de Ganancia de Información

La selección de atributos es el método para reconocer y eliminar la información irrelevante y redundante de un sistema evaluado. Si se eliminan algunas de las características no relacionadas, se podrá moderar la complejidad, eliminando dimensiones irrelevantes y mejorando el rendimiento del procedimiento de clasificación prospectivo. El agente de decisión es capaz de operar no solo más rápido y con menos información, sino también para mejorar el proceso de precisión de clasificación (Garcia, 2015), Existe una variedad de

diferentes métodos de selección de atributos propuestos. Por ejemplo, (Escalante, 2015). Analizó el número de estas técnicas de selección de atributos y describió las que lograron resultados notables, al saber seleccionar las características basadas en correlación de ganancia de información, Evaluación de subconjunto de envoltura, eliminación recursiva de características y Subconjunto basado en consistencia. La idea principal del procedimiento de selección de atributos es clasificar las variables relevantes y, en adelante, utilizar solo la información adecuada para realizar la clasificación de la red. La reducción de atributos es el proceso de mapear los datos de alta dimensión existentes en un espacio de menor dimensión.



**Figura 2.** Procedimientos Nasca. Tomada de (García, 2017). Elaborado por Vera Córdova Ana Luisa.

Se analiza el número de estas técnicas de selección de atributos y describen las que lograron resultados notables, La idea principal del procedimiento de selección de atributos es clasificar las variables relevantes y, en adelante, utilizar solo la información adecuada para realizar la clasificación de la red. La reducción de atributos es el proceso de mapear los datos de alta dimensión existentes en un espacio de menor dimensión. Por ejemplo, para un conjunto de datos dado de  $n$  variables  $\{x_1, x_2, x_3, \dots, x_n\}$ , se necesita calcular su representación dimensional  $x_i \in R^d \rightarrow y_i$  ( $p \ll d$ ). El criterio para la reducción de funciones puede ser diferente en función de diversas configuraciones de problemas. En esta investigación, indicaran algunos algoritmos de clasificación. En consecuencia, se proporcionarán resultados, de modo que se demuestre el rendimiento superior del filtro Clasificación de ganancia de información en comparación con otros algoritmos para la selección de atributos.

La ganancia de información (IG) cuantifica el volumen de información en bits sobre la estimación de clase. Mide la disminución esperada en la entropía de la variable de clase después de observar el valor de la característica (incertidumbre asociada con una

característica aleatoria). Es un filtro basado en entropía que clasifica la ganancia de los atributos. Por ejemplo, una Entropía para las clases  $i$  se puede definir como:

$$H(X) = - \sum (x_i) \log_2 P(x_i)$$

**Ecuación 1.** Ecuación de Ganancia de Atributos.

La entropía significa el nivel de inseguridad en el sistema. En la ecuación (1),  $P(x_i)$  es la función de densidad de probabilidad marginal para la variable aleatoria  $X$ , que se obtiene integrando la función de densidad de probabilidad conjunta. Primero se observan los valores de  $X$  en el conjunto de datos de entrenamiento  $S$  y los que se separan se mide de acuerdo con los valores de una segunda característica  $Y$ . En consecuencia, se mide la entropía de  $X$  con respecto a las particiones inducidas por  $Y$ . En el caso de que la medida de la entropía sea menor que la entropía de  $X$  antes de la partición, indica que existe una relación entre las características  $X$  e  $(x_i | y_i)$  es la probabilidad condicional de  $X$  dado  $Y$ . Teniendo en cuenta el hecho de que la entropía es una condición de impureza en el conjunto de entrenamiento  $S$ , se describe una medida que refleja información adicional sobre  $X$  proporcionada por  $Y$ , que representa la cantidad en la que disminuye la entropía de  $X$ . Esta medida se conoce como ganancia de información y viene dada por:  $IG\left(\frac{X}{Y}\right) = H\left(\frac{X}{Y}\right)$

Cuanto mayor es el valor de la ganancia informativa  $IG$ , más contribuye el atributo al conjunto de datos. Sin embargo, una desventaja del criterio de  $IG$  es que favorecerá los atributos con más valores porque está sesgado hacia la elección de atributos con un número más significativo de valores que producen un  $IG$  más alto. Sin embargo, dadas las características de nuestro conjunto de datos en particular,  $IG$  es un instrumento preferido para la selección de atributos. (Schiaffino, 2018)

### 2.2.5 Proceso de Clasificación

El método de clasificación es un paso esencial del procedimiento propuesto. Es un proceso de dos etapas con objetivos: precisión y clasificación más rápida. Por lo tanto, para acelerar el procedimiento, se completarán análisis adicionales, solo cuando la red se clasifique como atacada. El procedimiento propuesto combina dos técnicas principales de clasificación, a saber, si es aleatorio y, en consecuencia, árbol de decisión parcial. el procedimiento aleatorio, es un método de aprendizaje cooperativo que produce varios clasificadores y resume los resultados.



Además, se puede ejecutar si es necesario con dos procedimientos principales para realizar el análisis de clasificación o predicción, a saber, refuerzo y embolsado. Por un lado, al impulsar, los árboles sucesivos asignan un peso adicional a las instancias que fueron clasificadas incorrectamente por ensayos anteriores, y al final, se calcula una puntuación ponderada para fines de clasificación. Por otro lado, en el ensacado, los árboles siguientes son independientes de los árboles anteriores.



**Figura 3.** Procedimientos de clasificación de un ataque. Tomado de (Schiaffino, 2018). Elaborado Vera Córdova Ana Luisa.

La clasificación se realiza en base a la denominada división de puntaje mayoritario. El bosque aleatorio crece múltiples árboles, y cada uno de ellos produce una clasificación con un puntaje asignado para la clase específica. Como resultado, el bosque indica la clasificación con el puntaje más alto. El término se originó a partir de bosques de decisión aleatoria que fue propuesto por primera vez por Tin Kam Ho por Bell Labs en 1995. El bosque aleatorio (RF) es una combinación de predictores de árboles de modo que cada árbol depende de los valores de un vector aleatorio muestreado de forma independiente y con la misma distribución de todos los árboles en el bosque. El error de un bosque de clasificadores de árboles depende de la puntuación de los distintos árboles en el bosque y la correlación entre ellos. El proceso de Random Forest en nuestro análisis comienza con la creación de muchos árboles.

Introduce la aleatoriedad en los árboles de modo que cada árbol tenga una correlación mínima con los otros árboles. Cada árbol de la colección se forma seleccionando primero al azar, en cada nodo, un pequeño grupo de las características de entrada para dividir y, en segundo lugar, calculando la mejor división basada en estas características en el conjunto de entrenamiento. El método que se aplicara en el proceso de división es una técnica de aleatorización de dos pasos.

Inicialmente, el árbol se cultiva utilizando una muestra de entrenamiento, y luego se introducirá otra etapa de aleatorización, utilizando el enfoque de selección de características

aleatorias. En resumen, en lugar de dividir el nodo del árbol utilizando todas las características  $k$ , la selección será aleatoriamente en cada nodo de cada árbol un subconjunto de matrices, donde  $m \in [1, k]$  para dividir el nodo. Algunas de las sugerencias de división y el desarrollo del árbol. El método sugiere una estructura de aleatorización sub espacial que se combina con embolsado. La idea es volver a muestrear, con reemplazo, el conjunto de datos de capacitación cada vez que se construye un nuevo árbol individual (bookdown, 2020).

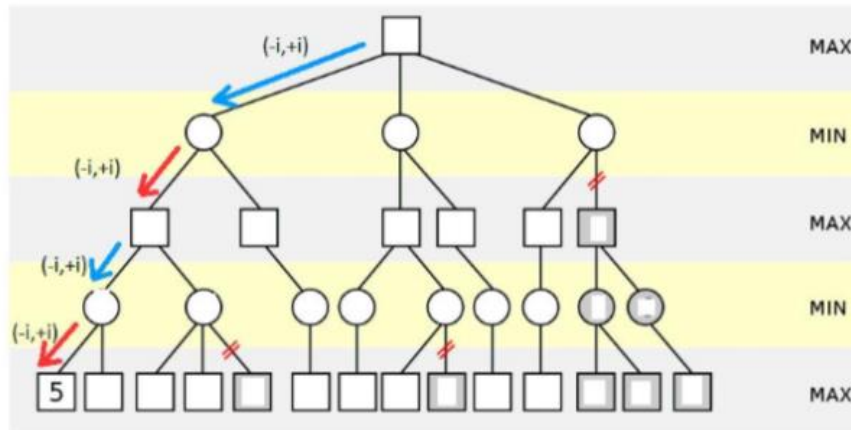
#### ***2.2.5.1 Exactitud***

El paso de precisión se relaciona con la prueba del rendimiento del modelo en función de dos criterios principales:

Precisión y complejidad: Un modelo demasiado complejo tomará más tiempo para clasificar la red, además, existe una compensación entre complejidad y precisión. Se realizan diferentes modelos de clasificación y selección de atributos, y se realiza una comparación de clasificación basada en esos dos criterios. Los resultados se informan con respecto al análisis de costos de la complejidad, la estadística de Kappa y la tasa de error para la precisión. El conjunto de datos se prueba sobre la base de una división del 80-20%, y los resultados se comparan e informan, según la metodología de clasificación aplicada.

Según la Web (bookdown, 2019) anuncia en su investigación:

Existen dos formas de poda muy comunes utilizadas en los diferentes algoritmos: la poda por coste-complejidad y la poda pesimista. En la poda por coste-complejidad se trata de equilibrar la precisión y el tamaño del árbol. La complejidad está determinada por el número de hojas que posee el árbol (nodos terminales). La poda pesimista utiliza los casos clasificados incorrectamente y obtiene un error de sustitución, eliminando los subárboles que no mejoran significativamente la precisión del clasificador



**Figura 4.** Podado de Árbol para más Exactitud, tomada de (Bookdown, 2019). Elaborado por Vera Córdova Ana Luisa.

### 2.2.5.2 Descripción de Datos

Una red está estructurada por paquetes de Protocolo de Control de Transmisión (TCP) que comienzan y terminan en un momento bien definido entre el cual los datos fluyen hacia y desde una dirección IP de origen a otra dirección IP de destino bajo un protocolo determinado. Cada red está etiquetada como normal o como un ataque con exactamente un tipo de ataque específico. En la investigación de (Bhattacharjee & Md Fujail, 2017), Los datos que se utilizan en este artículo científico son ISCX NSL-KDD Data Set, que es una versión mejorada de KDD CUP 99, DARPA, simularon un entorno para obtener nueve semanas de datos de volcado TCP sin procesar para una red de área local (LAN), simulando una red típica. Además, operaron la LAN como si fuera una verdadera atmósfera y simularon numerosos ataques los cuales les permitió obtener datos simulados y hacer pruebas a los algoritmos que ellos proponían.

**Tabla 3** Resultados de detección de ataque con respecto a VGA, WV-GA y FV-GA para el conjunto completo de datos NSLKDD usando 42 características

ALGORITMO GENETICO	F.P	NORMAL	DOS	PRUEBA	R21	U2R
VGA	0,9652	928	143	329	8	0
VGA	0,934	639	98	380	8	0
WV-GA	0,9547	849	115	372	13	0
WV-GA	0,9186	493	91	512	7	0
FV-GA	0,9902	1733	577	111	0	0
FV-GA	0,9918	1733	577	111	0	0

**Fuente:** Sistema de detección de intrusos para el conjunto de datos NSL-KDD utilizando la función de vectorización de la aptitud. Elaborado por Vera Córdova Ana Luisa

## **2.2.6 Algoritmos de Aprendizaje**

### ***2.2.6.1 Algoritmo de Aprendizaje Supervisado***

Un algoritmo de aprendizaje supervisado toma un conjunto conocido de datos de entrada (el conjunto de aprendizaje) y respuestas conocidas a los datos (la salida), y forma un modelo para generar predicciones razonables para la respuesta a los nuevos datos de entrada. Utiliza el aprendizaje supervisado si tiene datos existentes para la salida que está intentando predecir. (Mendoza S, 2020)

Los algoritmos de aprendizaje supervisados intentan modelar las relaciones y dependencias entre el resultado de predicción objetivo y las características de entrada de modo que se pueda predecir los valores de salida para los nuevos datos en función de las relaciones que aprendió de los conjuntos de datos anteriores.

### ***2.2.6.2 Máquinas de Vector Soporte***

En la década de 1990, se desarrolló un nuevo tipo de algoritmo de aprendizaje, basado en los resultados de la teoría del aprendizaje estadístico: Support Vector Machine (SVM). Esto dio lugar a una nueva clase de máquinas de aprendizaje teóricamente elegantes que utilizan un concepto central de núcleos SVM para una serie de tareas de aprendizaje.

Una máquina de vectores de soporte (SVM) es un modelo de aprendizaje automático supervisado que utiliza algoritmos de clasificación para problemas de clasificación de dos grupos. El objetivo de aplicar SVM es encontrar la mejor línea en dos dimensiones o el mejor hiperplano en más de dos dimensiones para ayudar a separar nuestro espacio en clases. El hiperplano (línea) se encuentra a través del margen máximo, es decir, la distancia máxima entre puntos de datos de ambas clases. (Bach, Schölkopf, & Smola, 2018)

### ***2.2.6.3 Algoritmos de Árbol de Decisión***

Los algoritmos de árbol de decisión son técnicas importantes y bien establecidas de aprendizaje automático que se han utilizado para una amplia gama de aplicaciones, especialmente para problemas de clasificación. Los árboles de decisión proporcionan un método no paramétrico para particionar conjuntos de datos. (Gangadhar & Shanta, 2018)

El árbol de decisión se utiliza para construir modelos de clasificación y regresión. Se utiliza para crear modelos de datos que predecirán etiquetas o valores de clase para el proceso de toma de decisiones. Los modelos se crean a partir del conjunto de datos de capacitación que se alimenta al sistema (aprendizaje supervisado). Usando un árbol de

decisión, se puede visualizar las decisiones que hacen que sea fácil de entender y, por lo tanto, es una técnica popular de minería de datos.

#### **2.2.6.4 Bosques Aleatorios (*Random Forests*)**

El bosque aleatorio es un algoritmo de aprendizaje supervisado que construye un conjunto de árboles de decisión, cada uno con los mismos nodos, pero utilizando diferentes datos que conducen a diferentes hojas. Fusiona las decisiones de múltiples árboles de decisión para encontrar una respuesta, que representa el promedio de todos estos árboles de decisión. (Chen & Li, 2018)

El uso de Algoritmo de bosque aleatorio ofrece muchos beneficios, pero una de las principales ventajas es que reduce el riesgo de sobreajuste y el tiempo de entrenamiento requerido; además, ofrece un alto nivel de precisión. El algoritmo Random Forest se ejecuta eficientemente en grandes bases de datos y produce predicciones altamente efectivas al estimar los datos faltantes.

#### **2.2.6.5 K Nearest Neighbors**

El principio detrás de los métodos del K Nearest Neighbors o vecino más cercano es encontrar un número predefinido de muestras de entrenamiento más cercanas en distancia al nuevo punto y predecir la etiqueta a partir de ellas. El número de muestras puede ser una constante definida por el usuario (k-aprendizaje vecino más cercano), o puede variar en función de la densidad local de puntos (aprendizaje vecino basado en el radio). La distancia puede ser, en general, cualquier medida métrica: la distancia euclidiana estándar es la opción más común. Los métodos basados en los vecinos se conocen como métodos de aprendizaje automático no generalizados, ya que simplemente "recuerdan" todos sus datos de entrenamiento (posiblemente transformados en una estructura de indexación rápida, como un árbol de bolas o un árbol KD). (Singh, Halgamuge, & Lakshmiganthan, 2017)

#### **2.2.7 Algoritmo de Aprendizaje no Supervisado**

El aprendizaje no supervisado es una técnica de aprendizaje automático en la que los usuarios no necesitan supervisar el modelo. En cambio, permite que el modelo funcione por sí solo para descubrir patrones e información que antes no se habían detectado. Se ocupa principalmente de los datos no etiquetados. (Mendoza S, 2020)

Los algoritmos de aprendizaje no supervisados permiten a los usuarios realizar tareas de procesamiento más complejas en comparación con el aprendizaje supervisado. Sin embargo, el aprendizaje no supervisado puede ser más impredecible en comparación con otros métodos de aprendizaje natural. Los algoritmos de aprendizaje no supervisados incluyen agrupamiento, detección de anomalías, redes neuronales, etc.

#### ***2.2.7.1 Clustering***

El clustering o agrupación puede considerarse el algoritmo de aprendizaje no supervisado más importante; entonces, como cualquier otro algoritmo de este tipo, se trata de encontrar una estructura en una colección de datos sin etiquetar. Asimismo, una definición poco precisa de agrupamiento podría ser el proceso de organizar objetos en grupos cuyos miembros son similares de alguna manera. Por otra parte, un clúster, es una colección de objetos que son similares entre ellos y son diferente a los objetos que pertenecen a otros grupos.

Por lo tanto, el objetivo de la agrupación es determinar la agrupación intrínseca en un conjunto de datos sin etiquetar. Pero ¿cómo decidir qué constituye un buen agrupamiento? Se puede demostrar que no existe un criterio mejor absoluto que sea independiente del objetivo final de la agrupación. En consecuencia, es el usuario quien debe proporcionar este criterio, de tal manera que el resultado de la agrupación se ajuste a sus necesidades. (Di Natale & Martinelli, 2019)

#### ***2.2.7.2 Algoritmo DBSCAN (Algoritmos de agrupamiento basados en densidad)***

La agrupación basada en la densidad se refiere a métodos de aprendizaje no supervisados que identifican grupos/agrupaciones distintivas en los datos, en base a la idea de que una agrupación en el espacio de datos es una región contigua de alta densidad de puntos, separada de otras agrupaciones similares por regiones contiguas de baja densidad de puntos.

La agrupación espacial basada en densidad de aplicaciones con ruido (DBSCAN) es un algoritmo base para la agrupación basada en densidad. Puede descubrir grupos de diferentes formas y tamaños a partir de una gran cantidad de datos, que contiene ruido y valores atípicos.

El algoritmo DBSCAN utiliza dos parámetros:

1. minPts: el número mínimo de puntos (un umbral) agrupados para que una región se considere densa.

2.  $\epsilon$  (eps): una medida de distancia que se usará para ubicar los puntos en la vecindad de cualquier punto.

Estos parámetros se pueden entender si se revisan dos conceptos llamados Densidad de accesibilidad y densidad de conectividad.

1. La accesibilidad en términos de densidad establece un punto al que se puede llegar desde otro si se encuentra dentro de una distancia particular ( $\epsilon$ ) de él.
2. La conectividad, por otro lado, implica un enfoque de encadenamiento basado en transitividad para determinar si los puntos están ubicados en un grupo particular. Por ejemplo, los puntos p y q podrían conectarse si  $p \rightarrow r \rightarrow s \rightarrow t \rightarrow q$ , donde  $a \rightarrow b$  significa que b está cerca de a.

#### **2.2.7.3 K-means**

Es un método popular de análisis de conglomerados. Es un proceso de organización de los objetos especificados en clases uniformes llamadas grupos basados en similitudes entre objetos basados en ciertos criterios. Resuelve el conocido problema de agrupamiento al considerar ciertos atributos y realizar un proceso de ajuste alternativo iterativo. Este algoritmo divide un conjunto de datos en grupos disjuntos de manera que cada observación pertenece al grupo con la media más cercana.

La agrupación de K-means se utiliza en una gran cantidad de aplicaciones que incluyen aprendizaje automático, detección de fallas, reconocimiento de patrones, procesamiento de imágenes, estadísticas e inteligencia artificial. El algoritmo k-means se considera uno de los algoritmos de agrupación más rápidos con una serie de variantes que son sensibles a la selección de los puntos iniciales y están destinados a resolver muchos problemas de k-means como la evaluación del número de clusters, el método de inicialización de los clústeres centroides, y la velocidad del algoritmo.

#### **2.2.7.4 Algoritmos de Estimación - Algoritmo Isolation Forest (Machine Learning)**

Devuelve el puntaje de anomalía de cada muestra usando el algoritmo IsolationForest 'aísla' las observaciones seleccionando aleatoriamente una característica y luego seleccionando aleatoriamente un valor dividido entre los valores máximo y mínimo de la característica seleccionada. Dado que la partición recursiva se puede representar mediante una estructura de árbol, el número de divisiones necesarias para aislar una muestra es equivalente a la longitud de la ruta desde el nodo raíz hasta el nodo de terminación.

Esta longitud del camino, promediada sobre un bosque de tales árboles aleatorios, es una medida de la normalidad y nuestra función de decisión. La partición aleatoria produce rutas notablemente más cortas para las anomalías. Por lo tanto, cuando un bosque de árboles aleatorios produce colectivamente longitudes de camino más cortas para muestras particulares, es muy probable que sean anomalías.

### 2.2.8 Modelos de Mezcla Gaussiana (MMG)

Un modelo de mezcla gaussiana es un modelo probabilístico que supone que todos los puntos de datos se generan a partir de una mezcla de un número finito de distribuciones gaussianas con parámetros desconocidos. Se puede pensar en los modelos de mezcla como una agrupación general de k-medias para incorporar información sobre la estructura de covarianza de los datos, así como los centros de los gaussianos latentes.

Los modelos de mezcla gaussiana se pueden usar para agrupar datos sin etiquetar de la misma manera que k-means.

Sin embargo, hay un par de ventajas al usar modelos de mezcla gaussiana sobre k-means tales como:

- k-means no tiene en cuenta la varianza. Por variación, se toma de referencia al ancho de la curva de forma de campana.
- En dos dimensiones, la varianza (la covarianza para ser exactos) determina la forma de la distribución.
- La diferencia entre los modelos de mezcla k-medias y gaussianos es que el primero realiza una clasificación dura mientras que el segundo realiza una clasificación suave.

### 2.2.9 Comparación entre Algoritmo de Aprendizaje Supervisado y no Supervisados

**Tabla 4.** *Comparación entre Algoritmo de aprendizaje supervisado y no supervisados*

Algoritmo de aprendizaje supervisado	Algoritmo de aprendizaje no supervisado
Utilizan datos etiquetados para ayudar a su lógica a tomar decisiones.	Son algoritmos que pretenden hallar similitudes sin entradas externas distintas de los datos sin procesar.



Es el tipo de algoritmo más popular.	Está vinculado a las actividades desarrolladas en la inteligencia artificial.
Incluye diferentes algoritmos tales como de regresión lineal y logísticos.	Se basa en agrupamientos o clustering, k-means y reglas de asociación.
Actúa como un patrón de guía para enseñar al algoritmo los resultados a los que debe llegar.	Los datos y resultados son desconocidos, se guían a través de operaciones lógicas.
Los algoritmos aprenden de los datos introducidos por una persona.	Los algoritmos aprenden de datos con elementos no etiquetados buscando patrones o relaciones entre ellos.
Trabajan con dos tipos de datos: por clasificación y regresión.	Trabajan con dos tipos de datos: por clustering (agrupación) y asociación.

***Fuente:** Algoritmo de aprendizaje supervisado y no supervisados. Tomado de (Comaniciu & Meer). Elaborado por Vera Córdova Ana Luisa*

### 2.2.9.1 Matriz de Confusión

Es una herramienta que permite realizar la visualización del desempeño de un algoritmo que emplea un determinado aprendizaje, sea este supervisado o no supervisado. Las columnas de una matriz representan el número de predicciones de cada clase, mientras que cada fila representa a las instancias en clase real.

Según Barrero G. (2018) en su investigación señala que:

“Si en los datos de entrada el número de muestras de clases diferentes cambia mucho la tasa de error del clasificador no es representativa de lo bien que realiza la tarea el clasificador. Si por ejemplo hay 990 muestras de la clase 1 y sólo 10 de la clase 2, el clasificador puede tener fácilmente un sesgo hacia la clase 1.”

También menciona que si se realiza la clasificación y la muestra es el número 1 su precisión será del 99%.

		Predicción	
		Positivos	Negativos
Observación	Positivos	Verdaderos Positivos (VP)	Falsos Negativos (FN)
	Negativos	Falsos Positivos (FP)	Verdaderos Negativos (VN)

***Figura 5.** Matriz de confusión, tomada de Investigación realizada por (Barrero G. 2018). Elaborado por Vera Córdova Ana Luisa*

### 2.2.10 Algoritmo de Meanshift

El algoritmo Meanshift es uno de los algoritmos de agrupamiento que se asocia con los puntos de mayor densidad o el valor de modo como el parámetro principal para desarrollar el aprendizaje automático, estando este inmerso en el grupo de algoritmo de aprendizaje automático no supervisado.

El algoritmo funciona sobre el concepto de Estimación de densidad de kernel conocido como KDE. También se conoce como algoritmo de búsqueda de modo. El Kernel está asociado con el cálculo matemático relacionado con el peso de los puntos de datos. Hay principalmente dos funciones populares del núcleo asociadas con el algoritmo de desplazamiento medio, como el núcleo plano y el núcleo gaussiano. Este algoritmo se usa principalmente para la visión por computadora y la segmentación de imágenes.

Según una investigación realizada por (Comaniciu & Meer), indican que:

“El algoritmo de agrupación Mean Shift es un algoritmo de agrupación no supervisado que agrupa los datos directamente sin estar capacitado en datos etiquetados. La naturaleza del algoritmo de agrupamiento Mean Shift es de naturaleza jerárquica, lo que significa que se basa en una jerarquía de clústeres, paso a paso.”

A diferencia del popular algoritmo de clúster K-Means, el cambio medio no requiere especificar el número de clústeres por adelantado. El número de grupos está determinado por el algoritmo con respecto a los datos.

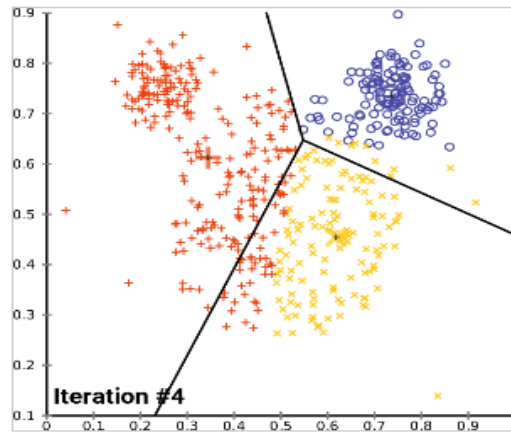
### 2.2.11 Agrupación de Algoritmos de MeanShift

Una técnica de aprendizaje no supervisada descubierta por Fukunaga y Hostetler para encontrar grupos:

- El cambio medio también se conoce como el algoritmo de búsqueda de modo que asigna los puntos de datos a los clústeres de una manera desplazando los puntos de datos hacia la región de alta densidad. La mayor densidad de puntos de datos se denomina modo en la región. El algoritmo Mean Shift tiene aplicaciones ampliamente utilizadas en el campo de la visión por computadora y la segmentación de imágenes.
- KDE es un método para estimar la distribución de los puntos de datos. Funciona colocando un núcleo en cada punto de datos. El núcleo en términos matemáticos es una función de ponderación que aplicará ponderaciones para puntos de datos individuales. Agregar todo el núcleo individual genera la probabilidad.

### 2.2.12 Estimación de Gradiente de Densidad

El primer paso al aplicar algoritmos de agrupación de turnos medios es representar sus datos de manera matemática, esto significa representar sus datos como puntos como el conjunto a continuación.

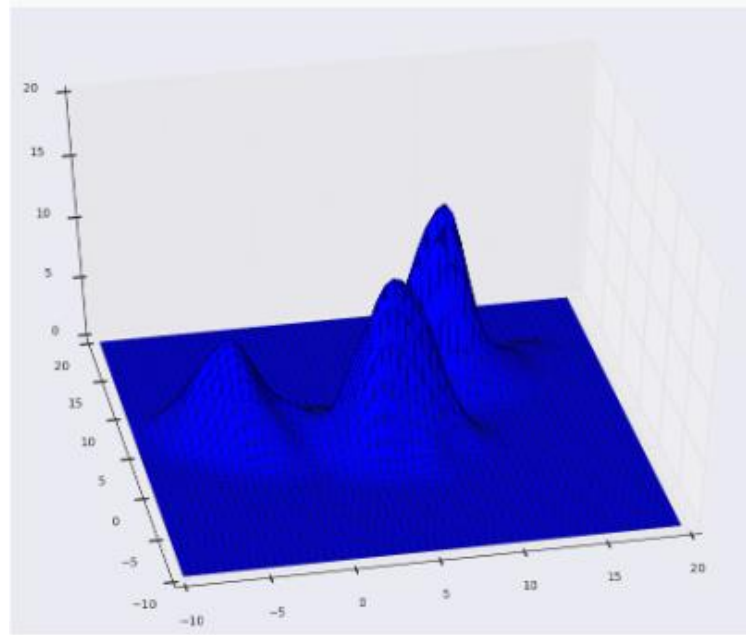


**Figura 6.** Representación de datos en Algoritmo Meanshift. Tomado de *Elaborado por Vera Córdova Ana Luisa*

La construcción de cambios medios sobre el concepto de estimación de densidad de kernel es ordenar KDE. Si se considera que los datos anteriores se muestrearon a partir de una distribución de probabilidad. KDE es un método para estimar la distribución subyacente, también llamada función de densidad de probabilidad para un conjunto de datos.

Funciona colocando un núcleo en cada punto del conjunto de datos. Un núcleo es una palabra matemática elegante para una función de ponderación generalmente utilizada en convolución. Hay muchos tipos diferentes de núcleos, pero el más popular es el núcleo gaussiano. Sumar todos los núcleos individuales genera una función de densidad de ejemplo de superficie de probabilidad. Dependiendo del parámetro de ancho de banda del núcleo utilizado, la función de densidad resultante variará.

A continuación, se muestra la superficie de KDE para sustentar puntos anteriores utilizando un núcleo gaussiano con un ancho de banda de núcleo de 2.



**Figura 7.** Parcela de superficie KDE. Algoritmo de aprendizaje. Tomado de (Comaniciu D. , 2002)  
Elaborado por el Vera Cordova Ana Luisa

### 2.2.13 Método MSB-CGH

Los algoritmos de suavizado tradicionales reemplazan los puntos en el centro de una ventana por el promedio ponderado de la ventana. Por lo tanto, difuminan indiscriminadamente las señales al eliminar no solo el ruido sino también la información más destacada. El suavizado de MSB, por el contrario, se basa en el uso de información local. Se ha demostrado que es un método de suavizado que preserva la discontinuidad, que reduce de manera adaptativa la cantidad de suavizado cerca de cambios abruptos (p. Ej., Bordes) en las estructuras locales.

Si  $x_i$  y  $z_i$  son puntos de datos en la entrada y la salida filtrada, respectivamente. Para cada punto  $x_i$ , asumir que un archivo externo que contiene una imagen, ilustración, etc. El nombre del objeto es 106inf3.jpges la posición en el dominio espacial en los perfiles de CGH-array, mientras que Un archivo externo que contiene una imagen, ilustración, etc. El nombre del objeto es 106inf4.jpges el dominio gama (relación log de la medición de la intensidad en los experimentos CGH).

Inicialice  $j = 1$  e  $y_1 = x_i$

1. Calcule  $y_{j+1} = (\sum_{i=1}^n g(\|y_j - x_i / h\|^2) / \sum_{i=1}^n g(\|y_j - x_i / h\|^2))$  hasta la convergencia, se obtiene  $e$ .

2. Asigne  $z_i = (x_i^s, y_c^s)$ , que son los datos filtrados. Esto significa que los datos filtrados en la ubicación espacial  $\sum_{i=1}^n$  tendrán el componente de rango (o dominio de intensidad para el array-CGH) del punto de convergencia  $y_c^s$ .

El núcleo en el procedimiento de desplazamiento medio se mueve en la dirección del aumento máximo en el gradiente de densidad de la junta. La característica clave del procedimiento de cambio medio es el uso de información local, que lo diferencia de los métodos de suavizado tradicionales. Cada punto está asociado con un modo significativo ubicado en su vecindario. La ventaja más importante del procedimiento de desplazamiento medio es que los puntos son atraídos por los modos (máximos locales) de la función de densidad subyacente. Por lo tanto, preserva efectivamente las discontinuidades y promueve la detección del punto de interrupción. Se puede extender directamente el procedimiento de desplazamiento medio y definir que los puntos vecinos en el cromosoma atraídos por el mismo modo en el dominio de intensidad pertenecen al mismo segmento de los perfiles de CGH.

#### **2.2.14 Técnica de Q-learning fuera de la Política para la respuesta a la Intrusión.**

Existen mecanismos de prevención de intrusiones que se basa en el aprendizaje automático y, más precisamente, en las técnicas de aprendizaje de refuerzo (RLT). El RLT la cual ayuda a crear un agente de decisión, que controlará el proceso de interacción con el entorno indeterminado. Un objetivo es obtener una política óptima, que representará la respuesta de intrusión al ataque y, por lo tanto, evitarlo. Resuelve el problema de aprendizaje de refuerzo, utilizando un enfoque de Q-learning. Este enfoque de Q-learning establecerá el equilibrio entre la exploración y la explotación y proporcionará un mecanismo de respuesta de inteligencia artificial único, de autoaprendizaje y estratégico para la detección de intrusos y los sistemas de prevención de intrusos.

##### **2.2.14.1 Aprendizaje Reforzado**

Hay un tomador de decisiones en la IDS y que interactúa regularmente con su entorno. En función de las acciones que realiza, puede modificar sus estados y, posteriormente, su rendimiento se evalúa mediante retroalimentación (recompensa). El objetivo es seleccionar un conjunto de acciones que optimicen su recompensa a largo plazo.

### 2.2.14.2 Algoritmos Q-Learning

A diferencia de un algoritmo de planificación, un algoritmo de aprendizaje como Q-learning implica determinar el comportamiento cuando el agente no sabe cómo funciona el mundo y puede aprender cómo comportarse por experiencia directa con el mundo. La siguiente figura ilustra un ejemplo típico de cómo el agente interactúa con el entorno. Como su nombre lo indica, Q-learning estima los valores Q óptimos de un MDP, lo que significa que el comportamiento se puede aprender tomando acciones con avidez con respecto a los valores Q aprendidos. En el algoritmo Q-learning, la forma más común de elegir una acción en los estados actuales es usar una política codiciosa. Es una fracción entre 0 y 1. Según la política, el agente selecciona aleatoriamente entre todas las acciones una fracción de tiempo, mientras que la acción con respecto a la diferencia del algoritmo de planificación, el agente de aprendizaje no tiene un conocimiento predefinido del entorno, lo que significa que la función de recompensa y la función de transición son desconocidas. En su lugar, el agente aprende cómo comportarse mediante la interacción con ambiente.



**Figura 8.** Algoritmo Q-learning. Tomado del autor. Elaborado por Vera Córdova Ana Luisa

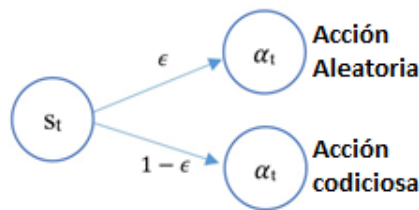
### 2.2.15 Política y Selección de Políticas

Casi todos los problemas de aprendizaje de refuerzo pueden formalizarse como MDP. El agente asigna el conjunto de estados en el espacio de probabilidad de tomar cada acción posible. Este proceso se hace referencia a mapeo como una política para el agente, que es una distribución de probabilidad formada a partir de posibles acciones, dados los estados actuales.

### 2.2.15.1 Exploración vs. Explotación

Por un lado, el agente inevitablemente debería explorar más oportunidades y, por lo tanto, desviarse del comportamiento habitual. Esta divergencia se llama exploración o acción no política. Por otro lado, debe seguir los procedimientos para estimar las funciones de valor. Cada vez que decide obedecer o seguir la política, se llama al proceso explotación o tomar medidas políticas. Existe una compensación entre ambos términos, y es desafiante y necesario encontrar un equilibrio adecuado, por lo que el agente debe poder decidir adecuadamente.

La selección de acciones  $\epsilon$ -codiciosas proporciona un enfoque heurístico simple para justificar entre explotación y exploración. El concepto es que el agente puede tomar una acción arbitraria a desde una distribución uniforme con probabilidad  $\epsilon$ ,  $0 \leq \epsilon \leq 1$ , y luego seleccionar con probabilidad  $1 - \epsilon$  la mejor acción (codiciosa). Es una práctica estándar disminuir el valor de  $\epsilon$  con el tiempo tan pronto como el agente de decisión tenga confianza y necesite menos exploración. La tasa baja implica un fuerte sesgo hacia la explotación sobre la exploración. La idea es asegurar continuidad. (Morillo, Moreno, & Diaz, 2015)



**Figura 9.** Selección de acción codiciosa. Tomado de (Fabrega, Fabrega, y Blair, 2016) Elaborado por Vera Córdova Ana Luisa

### 2.2.16 Lenguaje de Programación

Un lenguaje de programación es un idioma artificial diseñado para articular procesos que pueden ser llevadas a cabo por máquinas como las computadoras. Pueden usarse para inventar programas que controlen la actuación física y lógico de una máquina, para representar algoritmos con precisión, o como modo de comunicación humana. Está formado por un conjunto de símbolos y reglas sintácticas y semánticas que definen su estructura y el significado de sus elementos y expresiones. Al proceso por el cual se escribe, se prueba, se depura, se compila y se mantiene el código fuente de un programa informático se le llama programación. (Fabrega, Fabrega, y Blair, 2016)

### ***2.2.16.1 Lenguaje de Programación Python***

Python es un lenguaje multiusos, interactivo, orientado a objetos y de alto nivel. Fue creado por Guido van Rossum a lo largo de 1985-1990. Al igual que Perl, el archivo de texto ASCII de Python también se encuentra debajo de la Licencia pública general (GPL) de antílope. Python es un lenguaje de script de alto nivel, tomado, interactivo y orientado a objetos. Python pretende ser extremadamente claro. Utiliza palabras clave en inglés a menudo cuando los diferentes idiomas usan signos de puntuación, y tiene menos construcciones sintácticas que otros idiomas. (educba, 2018)

El lenguaje Python utiliza un intérprete, que es un programa que traduce y ejecuta las instrucciones en un programa de lenguaje de alto nivel. A medida que el intérprete lee cada instrucción individual en el programa, la convierte en instrucciones de lenguaje de máquina y luego las ejecuta de inmediato. Este proceso se repite para cada instrucción en el programa. Debido a que los intérpretes combinan traducción y ejecución, generalmente no crean programas de lenguaje de máquina separados.

### ***2.2.16.2 Lenguaje de Programación C#***

C # es desarrollado por Microsoft y aprobado por la Organización Internacional de Normalización (ISO). Es un lenguaje moderno y directo. C # fue desarrollado por Anders Hejlsberg y su equipo durante todo el evento de .Net Framework. C # está destinado a Common Language Infrastructure (CLI), que consiste en el código viable y la configuración de tiempo de ejecución que permite el uso de una variedad de lenguajes de alto nivel en plataformas y arquitecturas de PC totalmente diferentes.

### ***2.2.16.3 Diferencias entre Python & C#***

**Tabla 5** Comparación de Características Python & C#.

<b>Lenguajes</b>	<b>Paradigma</b>	<b>Características</b>	<b>Ventajas</b>	<b>Desventajas</b>
C#	Está orientado a objetos. Está estandarizado por Microsoft como parte de su plataforma net.	Sencillez de uso, compatible, moderno, Recolección de basura.	Se desempeña de forma plena en los sistemas operativos Windows. Sintaxis más en comparación	Requiere un mínimo de 4 GB para su instalación.



Lenguajes	Paradigma	Características	Ventajas	Desventajas
			con C y C++, Posibilidad de realizar aplicaciones web, de escritorio y móviles	
Python	Orientado a objetos	Permite la creación de todo tipo de programas incluso sitios web, no requiere de compilación es un código interpretado.	Libre y código fuente abierto, lenguaje de propósito general, portable.	Los lenguajes interpretados suelen ser relativamente lentos.

**Fuente:** Características de los lenguajes de programación. Tomado de (Ramirez, 2019). Elaborado por Vera Córdova Ana Luisa

#### **2.2.16.4 Error de Programación**

Las declaraciones que un programador escribe en un lenguaje de alto nivel se llaman código fuente, o simplemente código. Por lo general, el programador escribe el código de un programa en un editor de texto y luego guarda el código en un archivo en el disco de la computadora. Luego, el programador usa un compilador para traducir el código a un programa de lenguaje de máquina, o un intérprete para traducir y ejecutar el código. Sin embargo, si el código contiene un error de sintaxis, no se puede traducir. Un error de sintaxis es un error como una palabra clave mal escrita, un carácter de puntuación que falta o el uso incorrecto de un operador. Cuando esto sucede, el compilador o el intérprete muestra un mensaje de error que indica que el programa contiene un error de sintaxis. El programador corrige el error y luego intenta nuevamente traducir el programa.

#### **2.2.17 Software**

Si una computadora va a funcionar, el software no es opcional. Todo lo que hace una computadora, desde el momento en que enciende el interruptor de encendido hasta que apaga el sistema, está bajo el control del software. Hay dos categorías generales de software:

software de sistema y software de aplicación. La mayoría de los programas de computadora encajan claramente en una de estas dos categorías.

Los programas que controlan y administran las operaciones básicas de una computadora generalmente se denominan software de sistema. El software del sistema generalmente incluye los siguientes tipos de programas:

#### ***2.2.17.1 Sistemas Operativos***

Un sistema operativo es el conjunto más fundamental de programas en una computadora. El sistema operativo controla las operaciones internas del hardware de la computadora, administra todos los dispositivos conectados a la computadora, permite guardar y recuperar datos de los dispositivos de almacenamiento, y permite que otros programas se ejecuten en la computadora. Los sistemas operativos más populares son, Windows, Mac OS y Linux.

#### ***2.2.17.2 Software de la Aplicación***

Los programas que hacen que una computadora sea útil para las tareas cotidianas se conocen como software de aplicación. Estos son los programas que las personas normalmente pasan la mayor parte de su tiempo corriendo en sus computadoras. Ejemplos de Aplicaciones Microsoft Word, un programa de procesamiento de textos, y Adobe Photoshop, un programa de edición de imágenes. Algunos otros ejemplos de software de aplicación son programas de hojas de cálculo, programas de correo electrónico, navegadores web y programas de juegos.

#### ***2.2.17.3 Compiladores e Intérpretes***

Debido a que la CPU solo comprende instrucciones en lenguaje máquina, los programas que están escritos en un lenguaje de alto nivel deben traducirse al lenguaje máquina. Dependiendo del idioma en que se haya escrito un programa, el programador usará un compilador o un intérprete para realizar la traducción.

Un compilador es un programa que traduce un programa de lenguaje de alto nivel en un programa de lenguaje de máquina separado. El programa de lenguaje de máquina puede ejecutarse siempre que sea necesario.

#### ***2.2.17.4 Sistema de Monitorización***

Son conocidos también como sistema de vigilancia la cual se caracterizan por grandes flujos de datos que se necesitan estar analizados en tiempo real. Tiene como objetivo

distinguir situaciones no deseadas en el flujo de datos, esto a través de reconocimiento de acciones y sistemas de alerta de posibles casos que se presenten. (Mena, 2015)

### **2.2.18 Redes LAN**

Según Acosta, (2017). Una red de área local (LAN) es una red de datos de alta velocidad que cubre un área geográfica relativamente pequeña e interconecta estaciones de trabajo, computadoras personales, impresoras, servidores y otros dispositivos. Una LAN ofrece a los usuarios muchos beneficios, incluido el acceso compartido a dispositivos y aplicaciones, el intercambio de archivos entre usuarios conectados y la comunicación entre usuarios por correo electrónico y otras aplicaciones. Por lo general, se basa en estructuras cableadas tradicionales como su principal medio de transmisión.

### **2.2.19 Tipos de Anomalías de Red**

#### ***2.2.19.1 Ataques DoS y DDoS***

Un ataque de denegación de servicio tiene como objetivo inhabilitar el uso de un sistema, una aplicación o una máquina, con el fin de bloquear el servicio para el que está destinado. Este ataque puede afectar, tanto a la fuente que ofrece la información como puede ser una aplicación o el canal de transmisión, como a la red informática.

#### ***2.2.19.2 Flashcrowd***

Un flash-crowd attack (FCA) es un ataque DDoS que inunda una aplicación en la víctima con numerosas solicitudes de servicio. Dichos ataques son extremadamente difíciles de detectar y filtrar, porque las solicitudes legítimas y de ataque no se pueden distinguir entre sí. Los atacantes utilizan varios bots para enviar solicitudes a la víctima a tasas bajas. Los ataques de flash-crowd atraen a los atacantes porque pueden ser efectivos a un volumen bajo. Dado que muchas defensas DDoS operan a nivel de red y buscan grandes picos de tráfico en los agregados de la red, los ataques flash-crowd a menudo pasan desapercibidos. Un atacante puede utilizar solicitudes ligeras y regulares, como las de una página estática en un servidor web, o utilizar solicitudes costosas, que requieren más recursos del servidor, tales como solicitudes dinámicas, que implican búsquedas y actualizaciones de bases de datos.

#### ***2.2.19.3 Escaneo.***

Se utiliza para detectar qué servicios comunes está ofreciendo la máquina y posibles vulnerabilidades de seguridad según los puertos abiertos. Es usado por administradores de

sistemas para analizar posibles problemas de seguridad, pero también es utilizado por usuarios malintencionados que intentan comprometer la seguridad de la máquina o la red.

#### ***2.2.19.4 Gusano.***

Los gusanos son en realidad una subclase de virus, por lo que comparten características. Son programas que realizan copias de sí mismos, alojándolas en diferentes ubicaciones del ordenador.

El objetivo de este malware suele ser colapsar los ordenadores y las redes informáticas, impidiendo así el trabajo a los usuarios. A diferencia de los virus, los gusanos no infectan archivos.

El principal objetivo de los gusanos es propagarse y afectar al mayor número de dispositivos posible. Los gusanos suelen utilizar técnicas de ingeniería social para conseguir mayor efectividad. Para ello, los creadores de malware seleccionan un tema o un nombre atractivo con el que camuflar el archivo malicioso. Los temas más recurrentes son los relacionados con el sexo, famosos, temas de actualidad o software pirata.

Se replican los virus informáticos enviando archivos adjuntos infectados a través de correos electrónicos, mensajes instantáneos, a otros usuarios.

#### ***2.2.19.5 Punto a Multipunto.***

Es un término que se utiliza en el ámbito de las telecomunicaciones, que se refiere a la comunicación que se logra a través de un específico y distinto tipo de conexión multipunto, ofreciendo varias rutas desde una única ubicación a varios lugares. Una conferencia puede ser considerada una comunicación punto a multipunto ya que existe solo un orador (transmisor) y múltiples asistentes (receptor). Punto a multipunto es a menudo abreviado como P2MP, PTMP, o PMP.

### **2.2.20 Firewall**

El firewall se puede definir como una "colección de sistemas" instalados en el punto de conexión de la red de área protegida a otras redes, lo que requiere una política de seguridad predefinida. La instalación de un firewall en una organización tiene como objetivo la optimización del nivel existente de protección de datos y recursos informáticos de la organización contra los atacantes.

### ***2.2.20.1 PfSense***

Es una distribución de software de computadora de firewall / enrutador de código abierto basada en FreeBSD. Se instala en una computadora física o una máquina para hacer un firewall / enrutador dedicado para una red y se destaca por su confiabilidad y ofrece características que a menudo solo se encuentran en firewalls comerciales caros.

### ***2.2.20.2 Snort IDS***

Snort es un sistema de prevención de intrusiones de red (NIPS) gratuito y de código creado inicialmente por Martin Roesch en 1998 pero ahora desarrollado por sourcefire, del cual Roesch es el fundador y CTO. En 2009, Snort ingresó al Salón de la Fama del Código Abierto de InfoWorld como una de las "mejores piezas de software de código abierto de todos los tiempos". Snort tiene la capacidad de realizar análisis de tráfico en tiempo real e inicio de sesión de paquetes en redes de Protocolo de Internet (IP). Además, tiene la capacidad de realizar análisis de protocolo, búsqueda de contenido y coincidencia de contenido. TI implementa muchos servicios que se pueden usar para detectar sondas o ataques y tiene muchas modificaciones que se pueden configurar para que se ejecuten. Finalmente, snort se puede instalar en una variedad de sistemas operativos (Rodriguez C. , 2016).

Principalmente, un IDS se ocupa de la detección de acciones hostiles. Hay dos técnicas principales que están siendo utilizadas por un IDS. El primero, la detección de anomalías explora problemas en la detección de intrusos asociados con desviaciones del sistema normal o el comportamiento del usuario.

El segunda emplea la detección de firmas, esta es la técnica que se utiliza para discriminar entre patrones de anomalías o ataques (firmas) y firmas de detección de intrusos conocidas. Ambos métodos tienen sus distintas ventajas y desventajas, así como áreas de aplicación adecuadas de detección de intrusos.

### ***2.2.20.3 Características Básicas de los IDS basados en Firmas***

Un IDS basado en firmas examina el tráfico, la actividad, las transacciones o el comportamiento en curso para detectar coincidencias con patrones conocidos de eventos específicos de ataques conocidos. Un IDS basado en firmas requiere acceso a una base de datos actual de firmas de ataque para comparar y comparar activamente el comportamiento actual con una gran colección de firmas. Excepto cuando ocurren ataques completamente nuevos, sin catalogar, esta técnica funciona extremadamente bien. Otras ventajas de los IDS

basados en firmas: tasa de falsas alarmas muy baja, algoritmos simples, fácil creación de bases de datos de firmas de ataques, implementación sencilla y uso de recursos del sistema típicamente mínimo.

Cuando se considera que el área de aplicación es la fuente de datos utilizada para la detección de intrusiones, es posible distinguir los IDS en función de los tipos de actividades, tráfico, transacciones o sistemas que supervisan. En este caso, los IDS pueden dividirse en tipos de IDS basados en red (NIDS), basados en host (HIDS) y basados en aplicaciones. Los IDS que explotan la información obtenida de un segmento completo de una red local y buscan firmas de ataque se denominan IDS basadas en la red, mientras que las que operan en hosts defienden y monitorean los sistemas operativos y de archivos en busca de signos de intrusión y se denominan IDS basadas en host. Algunos IDS supervisan solo aplicaciones específicas y se denominan IDS basados en aplicaciones.

#### ***2.2.20.4 Suricata IDS***

Suricata es un motor de alto rendimiento de Network IDS, IPS y Network Security Monitoring. Es de código abierto y pertenece a una fundación sin fines de lucro dirigida por la comunidad Open Information Security Foundation (OISF). Suricata es desarrollado por la OISF. (Suricata, 2020)

Suricata implementa un lenguaje de firma completo para que coincida con amenazas conocidas, violaciones de políticas y comportamiento malicioso. Suricata también detectará muchas anomalías en el tráfico que inspecciona. Es capaz de usar el conjunto de reglas especializado Suricata de amenazas emergentes y el conjunto de reglas VRT. (Suricata, 2020)

Suricata es nativamente multiproceso y, por lo tanto, una sola instancia de Suricata es capaz de inspeccionar el tráfico de varios gigabits

#### **2.2.21 Marco Legal**

##### **Constitución de la República del Ecuador**

##### **Comunicación e Información**

Art. 16.-Todas las personas, en forma individual o colectiva, tienen derecho a:

1. Una comunicación libre, intercultural, incluyente, diversa y participativa, en todos los ámbitos de la interacción social, por cualquier medio y forma, en su propia lengua y con sus propios símbolos.
2. El acceso universal a las tecnologías de información y comunicación.
3. La creación de medios de comunicación social, y al acceso en igualdad de condiciones al uso de las frecuencias del espectro radioeléctrico para la gestión de estaciones de radio y televisión públicas, privadas y comunitarias, y a bandas libres para la explotación de redes inalámbricas.
4. El acceso y uso de todas las formas de comunicación visual, auditiva, sensorial y a otras que permitan la inclusión de personas con discapacidad.
5. Integrar los espacios de participación previstos en la Constitución en el campo de la comunicación.

### **Derechos de Libertad**

Art. 66 Numeral 19 indica lo siguiente: “El derecho a la protección de datos de carácter personal, que incluye el acceso y la decisión sobre información y datos de este carácter, así como su correspondiente protección. La recolección, archivo, procesamiento, distribución o difusión de estos datos o información requerirán la autorización del titular o el mandato de la ley”.

### **Código Orgánico Integral Penal**

- **Art. 230.- Interceptación ilegal de datos.** - Será sancionada con pena privativa de libertad de tres a cinco años:

- 1) La persona que, sin orden judicial previa, en provecho propio o de un tercero, intercepte, escuche, desvíe, grabe u observe, en cualquier forma un dato informático en su origen, destino o en el interior de un sistema informático, una señal o una transmisión de datos o señales con la finalidad de obtener información registrada o disponible.

- 2) La persona que diseñe desarrolle, venda, ejecute, programe o envíe mensajes, certificados de seguridad o páginas electrónicas, enlaces o ventanas emergentes o modifique el sistema de resolución de nombres de dominio de un servicio financiero o pago electrónico u otro sitio personal o de confianza, de tal manera que induzca a una persona a ingresar a una dirección o sitio de internet diferente a la que quiere acceder.

3) La persona que a través de cualquier medio copie, clone o comercialice información contenida en las bandas magnéticas, chips u otro dispositivo electrónico que esté soportada en las tarjetas de crédito, débito, pago o similares.

4) La persona que produzca fabrique, distribuya, posea o facilite materiales, dispositivos electrónicos o sistemas informáticos destinados a la comisión del delito descrito en el inciso anterior.

• **Art. 232.- Ataque a la integridad de sistemas informáticos.** - La persona que destruya, dañe, borre, deteriore, altere, suspenda, trabe, cause mal funcionamiento, comportamiento no deseado o suprima datos informáticos, mensajes de correo electrónico, de sistemas de tratamiento de información, telemático o de telecomunicaciones a todo o partes de sus componentes lógicos que lo rigen, será sancionada con pena privativa de libertad de tres a cinco años. Con igual pena será sancionada la persona que:

Diseñe, desarrolle, programe, adquiera, envíe, introduzca, ejecute, venda o distribuya de cualquier manera, dispositivos o programas informáticos maliciosos o programas destinados a causar los efectos señalados en el primer inciso de este artículo.

Destruya o altere sin la autorización de su titular, la infraestructura tecnológica necesaria para la transmisión, recepción o procesamiento de información en general.

Si la infracción se comete sobre bienes informáticos destinados a la prestación de un servicio público o vinculado con la seguridad ciudadana, la pena será de cinco a siete años de privación de libertad



## Capítulo III

### Propuesta

#### 3.1. Modalidad de la Investigación

A continuación, se indica las modalidades de investigación que se aplicaran para el desarrollo de esta propuesta:

**Bibliográfica:** La investigación bibliográfica se caracteriza por la utilización de los datos secundarios como fuente de información. Pretende encontrar soluciones a problemas planteados por una doble vía.

Esta modalidad se toma en consideración en base a las necesidades y requerimientos de seguridad expuestos en la empresa NewOffice, adicional de la información requerida para el desarrollo de la propuesta.

**Experimental:** Esta modalidad de estudio busca únicamente describir situaciones o acontecimientos; básicamente no está interesado en comprobar explicaciones, ni en probar determinadas hipótesis, ni en hacer predicciones.

En este proyecto se utiliza esta modalidad para documentar información obtenida en el proceso de implementación de herramientas y del sistema de detección de intrusos, ataque realizado con su respectivo análisis de cada ejecución.

#### 3.2 Tipos de Investigación

**Investigación Descriptiva:** es el procedimiento usado en ciencia para describir las características del fenómeno, sujeto o población a estudiar. Al contrario que el método analítico, no describe por qué ocurre un fenómeno, sino que se limita a observar lo que ocurre sin buscar una explicación. (Martinez C, 2017).

Este tipo de investigación permite identificar el problema de seguridad informática que presenta la red de la empresa NewOffice, la cual ha sido la problemática principal para el desarrollo de este proyecto, se realizó un análisis crítico de las vulnerabilidades presentadas en los sistemas y equipos instalados.

#### 3.3 Metodología de Desarrollo

Este capítulo se centra explícitamente en la metodología de investigación empleada para lograr los objetivos de la tesis. Es importante tener en cuenta que se emplearon varios

softwares de gestión de red de código abierto para administrar el acceso de los usuarios y monitorear los comportamientos de tráfico de los usuarios en la red

**Método Científico:** Su aplicación se basa en que es un conjunto de pasos ordenados empleados para adquirir nuevos conocimientos. Para poder ser calificado como científico debe basarse en el empirismo, en la medición y, además, debe estar sujeto a la razón. Mediante este método se procedió a realizar la respectiva recopilación de información sobre los ataques a la red, implementación de sistemas de detección y de algoritmos, y de esta manera obtener las herramientas necesarias para la debida ejecución de la propuesta establecida.

**Método Deductivo:** Es una estrategia de razonamiento que permite deducir conclusiones lógicas y con fundamentos.

Este método se aplicó para realizar el análisis y toma de decisión para proceder con la aplicación del sistema de detección de intrusos en la red y obtención de datos en la misma, ya que han sido bases para la utilización y ejecución de los IDS Suricata y Snort, que son los que utilizaron para capturar datos de tráfico y anomalías en la red

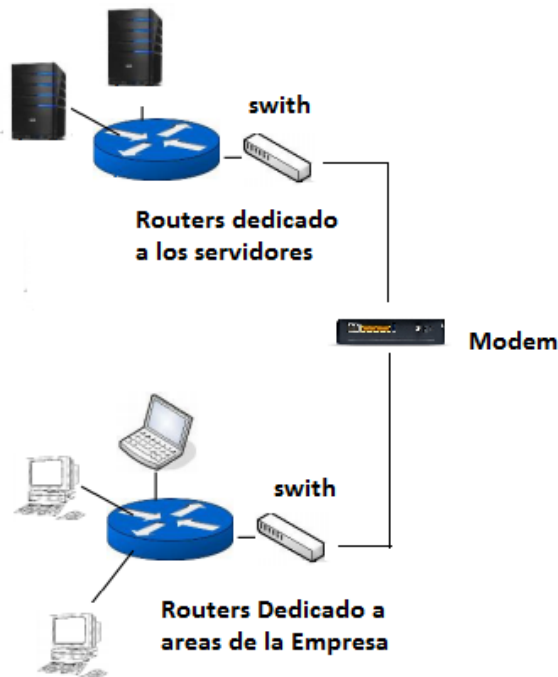
### 3.4 Área de Estudio

Es una empresa de capital cien por ciento nacional ubicada en Guayaquil, calle Los Ríos 200 entre Alejo Lascano y Luis Vernaza, cuentan con habilidades y destrezas en el sector de la confección, fabricación y montaje de muebles de oficina y decoraciones. Es una empresa joven dedicada a la elaboración y comercialización de todo tipo de accesorios para la decoración de ventanas, oficinas y casas. Basada en la excelencia en el servicio y en el desarrollo de las últimas tendencias, decoraciones engloba dentro de su oferta, las líneas principales de productos del sector. Poseen colecciones que abarcan un sin fin de soluciones para la decoración que van desde colecciones funcionales y profesionales.

#### 3.4.1 Infraestructura de la red

La infraestructura de la red de área local (LAN) implementada en la empresa es proporcionar instalaciones para apoyar y mejorar las comunicaciones internas y también interactuar con redes externas (es decir, Internet) para respaldar la información de las cuentas contables y financieras de la empresa; el intercambio de información; registro de compras y ventas; comprobación de resultados; colaboraciones del grupo de trabajo del personal para el flujo libre de la información interna.

Actualmente, la infraestructura LAN abarca una distancia máxima de aproximadamente 500 Metros al cuadrado e interconecta el bloque de administración, las oficinas de ventas, las unidades de mantenimiento, con enlaces ethernet de velocidad que funcionan sobre una red de Cable de red RJ45 Cat 6 dedicada como se muestra en la Figura 10. A continuación, se muestra una descripción general de la LAN:



**Figura 10.** Diagrama de la arquitectura en la red actual NewOffice Tomada del autor. Elaborado por Vera Cordova Ana Luisa

### 3.4.2 Infraestructura de Equipos

Actualmente la empresa NewOffice solo tiene una sucursal con todo el despliegue comercial y administrativo, la cual está ubicado en Guayaquil y posee los equipos descritos en la siguiente tabla:

**Tabla 6.** Infraestructura de Equipos disponibles en la empresa NewOffice

Hardware	Cantidad	Descripción
Servidor de Bases de Datos	1	<ul style="list-style-type: none"> <li>✓ Procesador Corel I3</li> <li>✓ 8gb de RAM DDR3</li> <li>✓ Unidades. SATA II, 8TB de espacio.</li> <li>✓ Sistema Operativo Windows.</li> </ul>
Servidor de Aplicaciones	1	<ul style="list-style-type: none"> <li>✓ Procesador Corel I5,</li> <li>✓ 16gb de RAM DDR4</li> <li>✓ Unidades. SATA II, 32TB de espacio</li> <li>✓ Sistema Operativo Windows</li> </ul>

Hardware	Cantidad	Descripción
Router inalámbrico TL-WR841ND	2	<ul style="list-style-type: none"> <li>✓ Velocidad inalámbrica ideal de 300 Mbps para las aplicaciones sensibles como la interrupción de difusión de vídeo HD</li> <li>✓ Dos antenas aumentan en gran medida la solidez y la estabilidad inalámbrica</li> <li>✓ Fácil Encriptado de la seguridad inalámbrica al presionar el botón QSS</li> <li>✓ Control de ancho de banda basado en IP permite a los administradores determinar la cantidad de ancho de banda asignado a cada PC</li> </ul>
Switch Administrable TP-LINK SG5412F	2	<ul style="list-style-type: none"> <li>✓ 12 puertos SFP/SFP+</li> <li>✓ Tecnología de cableado ethernet de cobre</li> <li>✓ Frecuencia de entrada AC 50/60 Hz</li> <li>✓ Voltaje de entrada AC 100-240 V</li> </ul>
Computador Corel I3	7	<ul style="list-style-type: none"> <li>✓ Placa Base ASUS H81M-A LGA1150</li> <li>✓ Procesador INTEL CORE i3-4170, 3.70GHz</li> <li>✓ Memoria RAM Kingston 4GB DDR3 1600MHZ</li> <li>✓ Disco Duro SEAGATE 500GB 7200RPM SATA</li> <li>✓ Lector/Grabador SAMSUNG Multigrabador CD/DVD</li> <li>✓ Tarjeta de Video integrado</li> <li>✓ Monitor 18.5" LG LED 1366X768</li> <li>✓ Teclado/Mouse KIT GENIUS Multimedia USB</li> <li>✓ Case/Fuente ATX 600W</li> <li>✓ Estabilizador FORZA 4 SALIDAS 1000VA</li> <li>✓ Sistema Operativo Windows 10</li> </ul>

**Fuente:** Elementos y equipos de red. Tomada de empresa NewOffice, Elaborado por Vera Córdova Ana Luisa.

Se cuenta con dos servidores que presentan de forma individual uno con las bases de datos necesarias para el control del negocio y otra las aplicaciones correspondientes a servidor web y las aplicaciones tanto de escritorio, que permiten a su vez la interacción de ambas para facilitar la usabilidad en cada una de las estaciones de trabajo.

Las estaciones de trabajo son 7 las cuales interactúan a diario con los sistemas alojados en el servidor, y al mismo tiempo interactúan con los clientes a través de correos electrónicos, búsquedas de materiales y nuevos proveedores en la red lo que influye poca seguridad al momento de explorar y navegar en la web, también se realiza mucho el uso de las redes sociales para el marketing de los productos y servicios que la empresa brinda.

### 3.4.3 Esquema de Seguridad de la Red

#### 3.4.3.1 Seguridad Virtual.

Actualmente no existe niveles de seguridad de la red, solo con los mismos que provee los sistemas operativos y la configuración por defecto de los routers, se mantienen los anti-virus actualizados, el firewall activados y los demás componentes necesarios para el uso de protocolos de red.

#### 3.4.3.2 Seguridad Física.

Los servidores se encuentran en un espacio abierto y ventilado en la oficina de gerencia, los cuales permanecen la mayor parte del tiempo cerrado lo que implica que no son de fácil acceso.

El perímetro de uso de los equipos de las estaciones de trabajo está en oficinas por estructuras adecuadas, pero sin niveles de seguridad, lo cual facilita el acceso a cualquiera de los computadores, por parte de los trabajadores.

### 3.5 Levantamiento del Tipo de Información a Proteger

Según la información recopilada se describen cómo interactúan los servidores con las aplicaciones y los puestos de trabajo, teniendo interacción también con carpetas y recursos compartidos en la red.

**Tabla 7** *Informacion a proteger de la empresa.*

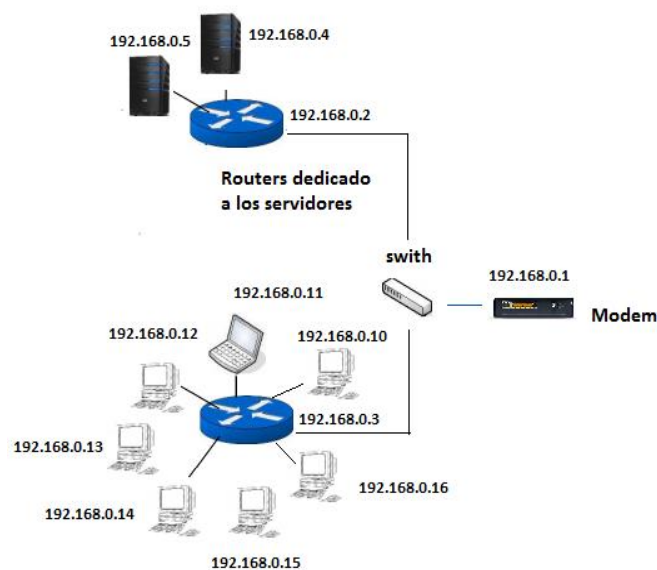
Área	Descripción
Gerencia	<ul style="list-style-type: none"> <li>✓ Carpetas compartidas en la red.</li> <li>✓ Sistema de Toma de Decisiones</li> <li>✓ Sistema Contable</li> <li>✓ Análisis de Marketing</li> <li>✓ Informe de Ventas</li> <li>✓ Informes de Compras</li> <li>✓ Informe de Inventarios</li> <li>✓ Informes contables</li> <li>✓ Informes Financieros</li> <li>✓ Estados de pérdidas y Ganancias</li> <li>✓ Auditorias</li> <li>✓ Listados de Clientes</li> </ul>
Contabilidad	<ul style="list-style-type: none"> <li>✓ Listado de Cuentas Contables</li> <li>✓ Estado de Cuentas contables</li> </ul>

Área	Descripción
	<ul style="list-style-type: none"> <li>✓ Pagos SRI</li> <li>✓ Balance General</li> <li>✓ Cuentas por pagar</li> <li>✓ Cuentas por cobrar</li> </ul>
Ventas	<ul style="list-style-type: none"> <li>✓ Listado de Ventas</li> <li>✓ Listado de Stock</li> <li>✓ Informe de Facturación</li> <li>✓ Estado de ventas por vendedor</li> </ul>
Almacén	<ul style="list-style-type: none"> <li>✓ Inventarios</li> <li>✓ Activos</li> <li>✓ Consumibles</li> </ul>
Compras	<ul style="list-style-type: none"> <li>✓ Listado de Proveedores</li> <li>✓ Listado de Entrada Stock</li> <li>✓ Informe de Recepción de consumibles</li> <li>✓ Requisiciones</li> </ul>

**Fuente:** Departamentos de empresa, tomada de empresa NewOffice, Elaborado por Vera Córdova Ana Luisa

### 3.5.1 Análisis de Posibles Amenazas - Internas y Externas

A través de diferentes herramientas se realizan una serie de pruebas para verificar el nivel de acceso a los sistemas desde la misma red local y a través de la WAN. Detallando lo visualizado en las herramientas que a continuación se presentan el diagrama de la red actual



**Figura 11.** Esquema de equipos en la red actual. Tomada de la empresa NewOffice, Elaborado por Vera Córdova Ana Luisa

Para el análisis de las vulnerabilidades de la red interna se utilizó dos herramientas: NMAP y LANGUARD NSS, las cuales son versiones gratuitas.

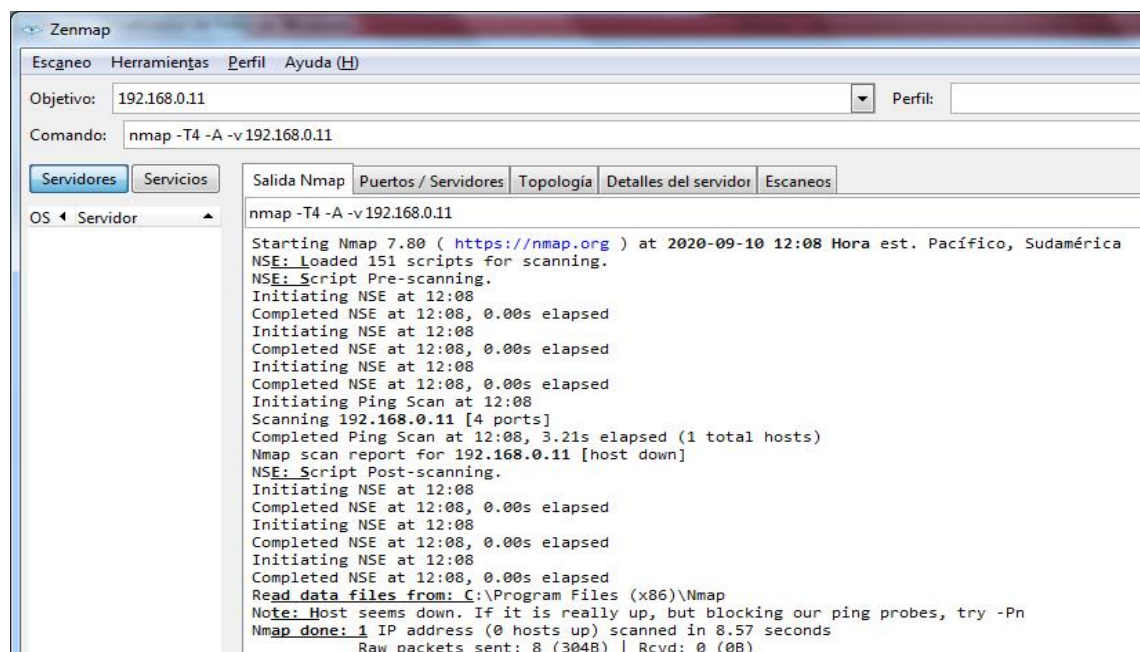
A continuación, se detallan la clasificación de las IP de la red:

**Tabla 8** Rangos de Ip de la Red

	Rangos de IP
Servidores y Dispositivos de Red	192.168.0.1 -192.168.0.9
Pc y Dispositivos Móviles	192.168.0.10 - 192.168.0.250

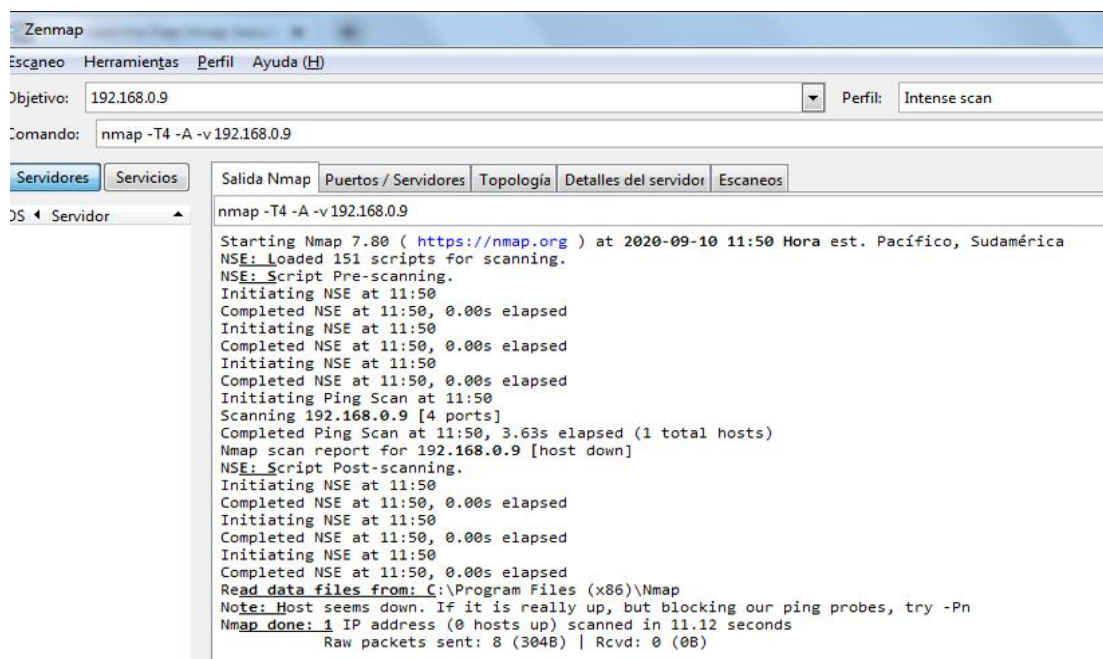
**Fuente:** Propia, tomada de empresa NewOffice, Elaborado por Vera Córdova Ana Luisa

Estas herramientas permiten conocer los puertos abiertos y protocolos disponibles de una ip determinada en la red.



**Figura 12.** Escaneado del Servidor 192.168.0.11. Tomado del autor. Elaborado por Vera Córdova Ana Luisa

En la figura 12 se observa el servidor de bases de datos la cual tiene abierto el puerto de base de datos, el puerto 3306 y el puerto 135, puertos muy usados por los hackers como agujero para entrar en los ordenadores, esto hace que los sistemas y la integridad de los datos estén vulnerables frente a ataques.



**Figura 13.** Escaneado del Servidor 192.168.0.9. Tomado del autor. Elaborado por Vera Córdova Ana

En la figura número 13 se comprueba el escaneo de la IP del servidor de aplicaciones, 192.168.0.9, lo cual es aceptable varios puertos que están activos como: 8, 9, 443 y 80 que son del apache o servidor web, pero se vuelve a comprobar la novedad del puerto 135 que está abierto lo que implica la vulnerabilidad a través de este puerto.

El scanner de LAN Guard 2012 permitió escanear los 9 equipos de la red incluyendo en ella los dos servidores los cuales se presentan a continuación en la siguiente imagen

Group by: Severity	Bulletin ID	Severity	QNumber	Date posted	Title
<input checked="" type="checkbox"/> All Patches	<input checked="" type="checkbox"/> APSB13-16	Critical	APSB13-16	2020-01-11	Adobe F
<input checked="" type="checkbox"/> Critical	<input checked="" type="checkbox"/> APSB13-16	Critical	APSB13-16	2020-01-11	Adobe F
<input checked="" type="checkbox"/> Important	<input checked="" type="checkbox"/> APSB13-16	Critical	APSB13-16	2020-01-11	Adobe F
<input type="checkbox"/> Moderate	<input checked="" type="checkbox"/> APSB13-16	Critical	APSB13-16	2020-01-11	Adobe F
<input type="checkbox"/> Low	<input checked="" type="checkbox"/> APSB13-16	Critical	APSB13-16	2020-01-11	Adobe F
<input type="checkbox"/> Undefined	<input checked="" type="checkbox"/> APSB13-16	Critical	APSB13-16	2020-01-11	Adobe F
	<input checked="" type="checkbox"/> APSB13-16	Critical	APSB13-16	2020-01-11	Adobe F
	<input checked="" type="checkbox"/> APSB13-16	Critical	APSB13-16	2020-01-11	Adobe F
	<input checked="" type="checkbox"/> APSB13-16	Critical	APSB13-16	2020-01-11	Adobe F
	<input checked="" type="checkbox"/> MS13-047	Critical	2838727	2020-01-11	Cumulat
	<input checked="" type="checkbox"/> MS13-047	Critical	2838727	2020-01-11	Cumulat
	<input checked="" type="checkbox"/> MS13-047	Critical	2838727	2020-01-11	Cumulat
	<input checked="" type="checkbox"/> MS13-047	Critical	2838727	2020-01-11	Cumulat
	<input checked="" type="checkbox"/> MS13-047	Critical	2838727	2020-01-11	Cumulat
	<input checked="" type="checkbox"/> MS13-047	Moderate	2838727	2020-01-11	Cumulat
	<input checked="" type="checkbox"/> MS13-047	Moderate	2838727	2020-01-11	Cumulat
	<input checked="" type="checkbox"/> MS13-047	Moderate	2838727	2020-01-11	Cumulat
	<input checked="" type="checkbox"/> MS13-047	Moderate	2838727	2020-01-11	Cumulat

**Figura 14.** Vulnerabilidades de los Servidores. Tomado del autor. Elaborado por el autor.



En la figura 14, se observa el mapeo del servidor 192.168.0.5 correspondiente al servidor de Aplicaciones web en donde se ve los diferentes nombres de NetBIOS, usuario validado, la MAC correspondiente y el tiempo de demora del ping, paquetes de aplicaciones vulnerables no actualizados, cabe destacar que son datos que en el servidor son un punto vulnerable dentro de la red ya que se sabe qué dirección apunta el servidor, su dominio y la dirección MAC, datos suficientes para tratar de ingresar a dicho servidor.

### **3.5.2 Resultado de Escaneo en la Red**

En los resultados del escaneo realizado se encontró vulnerabilidades en los servidores, se aprecia que no solo tiene puertos abiertos en base al manejo y tráfico de información, también existen espacios libres por el cual puede ser atacada la red.

No existen un firewall "Corta fuegos" que analice los procedimientos ni políticas de seguridad que guíen al usuario en el análisis y manejo de aplicaciones, manejo de la información, esto ha ocasionado la pérdida de información y vulnerabilidades dentro de la red por virus o ataques de usuarios mal intencionados, así como producir que la velocidad de la conexión disminuya por el mal funcionamiento.

Adicional a ello no se cuenta con un plan de detección de ataques a la red, o de instrucciones no permitidas en los accesos tanto de la red local, como de la WAN, se evidencia que no existe mantenimiento para equipos, en donde se tenga presente la actualización de aplicaciones y sistemas operativos, y la revisión del hardware para prevenir posibles daños en el mismo de forma periódica.

### **3.6 Propuesta Tecnológica**

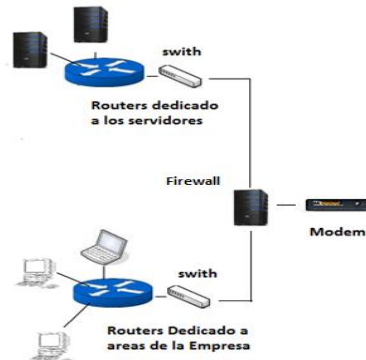
El presente proyecto pretende mejorar el escenario en el que están involucrados los usuarios mediante la aplicación y uso de herramientas comerciales que permiten mejorar el manejo de información, aplicando métodos de firewall, escaneo de las redes a través de Snort y Suricata aplicando reglas de IDS, esta a su vez apoyado a la aplicación de análisis de los paquetes de red a través de un algoritmo de Machine Learning.

### **3.7 Etapas de la Metodología del Proyecto**

#### **3.7.1 Diagrama de la Estructura Física de la Red Propuesta**

Para la implementación de la propuesta es necesario la instalación de un servidor adicional que fungirá como firewall, la cual es necesario que se anteponga a cualquier conexión con los servidores, esta se instalara entre la conexión del módem y a la vez con las

conexiones entre la red local de los router dedicados tanto a los servidores y las áreas comunes en la empresa.



**Figura 15.** Vulnerabilidades de los Servidores. Tomado de esquema de red de empresa NewOffice.  
Elaborado por Vera Cordova Ana Luisa

Las especificaciones técnicas del nuevo equipo servidor son:

**Tabla 9** Requerimientos del Hardware

	Cantidad	Descripción
Disco duro interno	1	500GB 7200 RP SATA
Memoria RAM	1	DDR3 ADATA 8 GB 1333Mhz UDIMM
CPU	1	Min. Intel Celeron g 460 1.8 ghz 5gt/s 1.5 mb
Tarjeta de Red	2	10/100 MBPS FAST ETHERNET PCI

**Fuente:** Requerimientos a nivel físico. (netgate, 2020). Elaborado por Ana Vera Córdova

### 3.7.2 Requerimientos del Software

A continuación, se describe tanto el sistema operativo y los sistemas necesarios para la implementación de la propuesta.

**Tabla 10** Requerimientos del Software

	Descripción
Sistema Operativo freebsd	La base del sistema operativo la cual es necesario, para poder instalar el firewall

	Descripción
Pfsense	Es una distribución personalizada de FreeBSD adaptado para su uso como Firewall y Router. Se caracteriza por ser de código abierto, puede ser instalado en una gran variedad de ordenadores, y además cuenta con una interfaz web sencilla para su configuración
IDS Snort	Es un sistema de detección de intrusos en red, libre y gratuito. Ofrece la capacidad de almacenamiento de bitácoras en archivos de texto y en bases de datos abiertas
IDS Suricata	Es un sistema de detección de intrusos basado en código abierto y un sistema de prevención de intrusos.
Python 2.7	Es un lenguaje de programación interpretado cuya filosofía hace hincapié en la legibilidad de su código. Se trata de un lenguaje de programación multiparadigma, ya que soporta orientación a objetos, programación imperativa y, en menor medida, programación funcional
Anacoda (Jupyter)	Es la interfaz de usuario de próxima generación. Ofrece todos los bloques de construcción familiares del clásico Jupyter Notebook (notebook, terminal, editor de texto, navegador de archivos, salidas enriquecidas) en una interfaz de usuario flexible y potente. Este es necesario para la perfecta sincronización de Machine Learning, limpieza de los paquetes de red, clasificación de los dataset y análisis de los paquetes tomados de la propuesta

**Fuentes:** (Snort, 2020), (suricata, 2020), (anaconda, 2020). Elaborado por Vera Cordova Ana Luisa

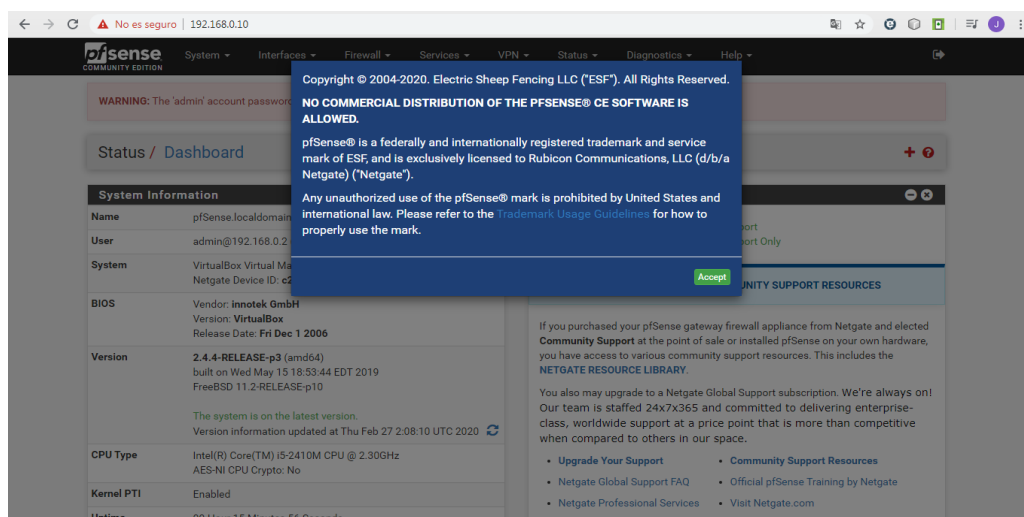
### 3.7.3 Activación de los IDS

Para la implementación de los IDS de la red de la empresa NewOffice se tomó como herramienta el UTM PfSense que trabaja en este caso como un Firewall.

Snort y Suricata son IDS de código abierto que se puede instalar fácilmente en un firewall pfSense para proteger una red de intrusos. Snort también se puede configurar para funcionar como un sistema de prevención de intrusiones (IPS), lo que lo hace muy flexible.

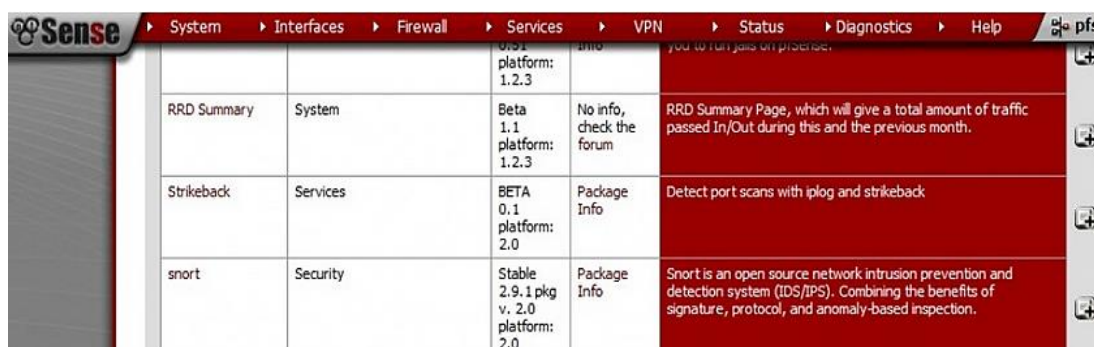
Como primer paso se debe ingresar al sitio oficial <https://www.pfsense.org/download/> proceder a realizar la descarga de la imagen ISO del sistema operativo, al seleccionar el tipo de instalación ya se ingresa a través de la BIOS y cambia el tipo de boteo si es pendrive, luego se realiza la configuración predeterminada del PfSense y complementos de manera automática de acuerdo a las casillas que permiten seleccionar. (Ver Anexo 1)

Puesta en funcionamiento la herramienta se tiene acceso a través de la ip que genera una conexión apache en la ip de V4/DHCP4 para este caso es la ip 192.168.0.10/24



**Figura 16.** Pantalla inicial de control Pfsense. Instalación de herramientas. Elaborado por Vera Cordova Ana Luisa

Para comenzar con Snort, deberá instalar el paquete utilizando el administrador de paquetes pfSense, el cual se encuentra en el menú del sistema de la GUI web de pfSense. (Ver Anexo 2)

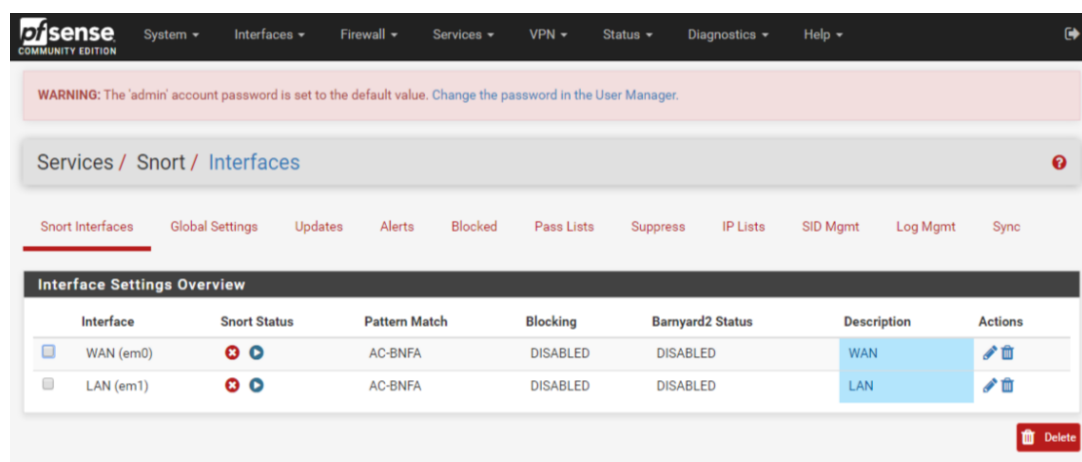


**Figura 17.** Pack de Instalación en interfaz. Tomado de Firewall Pfsense. Elaborado por Vera Córdova Ana Luisa

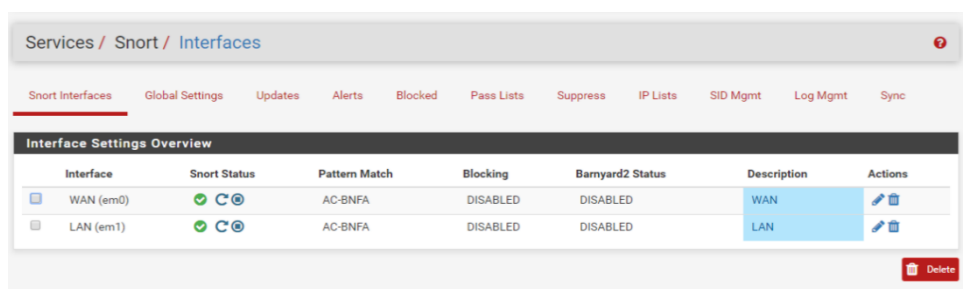
Una vez completada la instalación, Snort aparecerá en el menú de servicios.

Para que Snort reconozca los nuevos ataques es importante actualizar las firmas de la herramienta, y para ello se puede utilizar oinkmaster, el cual es un programa que actualiza estas reglas. (Ver Anexo 3)

El paquete descargará los últimos conjuntos de reglas de Snort.org y también Amenazas emergentes si es seleccionada las opciones de detección. Una vez finalizadas las actualizaciones, las reglas se extraerán y estarán listas para su uso.



**Figura 18.** Antes de puesta en funcionamiento. Tomado de Pfsense. Elaborado por Vera Córdova Ana Luisa



**Figura 19.** IDS ya funcionando y activado, Tomado del autor. Elaborado por Vera Cordova Ana Luisa

Antes de que Snort pueda comenzar a funcionar como un sistema de detección de intrusos, debe asignar interfaces para que pueda monitorear. La configuración típica es que Snort monitoree cualquier interfaz WAN. La otra configuración más común es que Snort monitoree la interfaz WAN y LAN. (Ver Anexo 4)

Al finalizar la utilización de Snort los detalles de la red se puede visualizar en alertas, la cual después de que Snort se haya configurado e iniciado correctamente, debe comenzar a ver alertas una vez que se detecte el tráfico que coincida con las reglas. Si no ve ninguna alerta, hay que esperar un poco y luego verifique nuevamente. Puede pasar un tiempo antes de que se vean las alertas, según la cantidad de tráfico y las reglas habilitadas.

Al visualizar los datos necesarios se procede a la descarga la cual enviara a la ruta que uno desee la Dataset en formato Txt archivo plano para su utilización en el algoritmo de Machine Learning.

### **Suricata**

La instalación de Suricata es igual que la de Snort, se dirige hacia la pestaña de pack manager e instala el paquete, Lo primero que se nota después de habilitar Suricata fue mucho ruido de las alertas. Las primeras alertas que inundaron el feed fueron cosas como "suma de comprobación no válida UDP4" y algunas otras que, aunque no necesariamente eran buenas para ver, a diferencia de Snort que presenta sus alertas en base a la configuración de alertas. Estos pueden ser importantes para ver en un entorno diferente, sin embargo, hay que agregar rápidamente algunas reglas de "supresión" que ocultaban estos numerosos y poco interesantes tipos de alertas. (Ver Anexo 5)

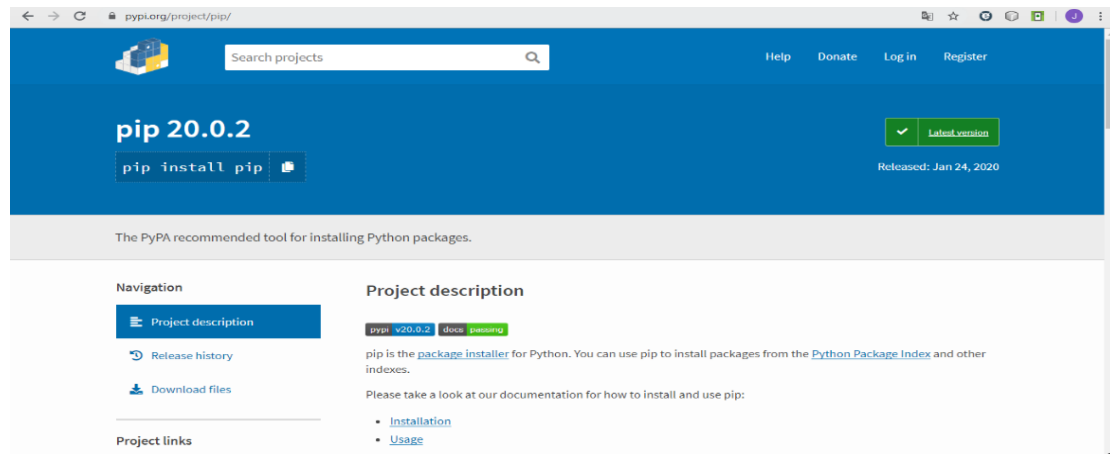
### **3.8 Ambiente de Desarrollo Python**

Existen un conjunto muy amplio de herramientas que coexisten para instalar y administrar paquetes de Python. Es posible instalar paquetes en su intérprete de Python base, pero tarde o temprano obtendrá paquetes de Python en conflicto ya que los paquetes tienen dependencias de variables y librerías distintas. Entonces, el lugar de este proyecto de investigación se crea en un entorno en su máquina local de la cloud de google, donde los paquetes de Python se pueden instalar y organizar. Para cada proyecto se recomienda crear un nuevo entorno. Otra forma de manejar paquetes Python en conflicto es con el administrador de paquetes de distribución de Ubuntu. Se ha probado que los paquetes Python disponibles del administrador de distribución funcionan juntos, pero algunos inconvenientes

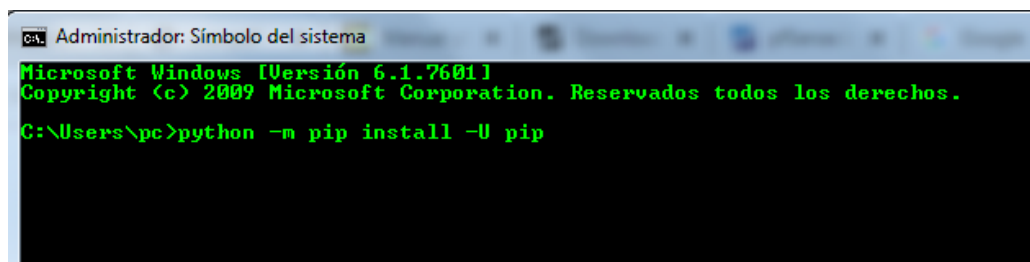
son que no tiene todos los paquetes y, a veces, tiene paquetes obsoletos. En el siguiente ítem explica como instalar los paquetes necesarios.

### ✓ **Paquete Pip**

Paquete que permite Instalar de Forma sencilla las librerías de Python. Para instalarle en Windows se debe abrir la consola y tipear `Python -m pip install -U pip`



**Figura 20.** Página Oficial PIP. Tomado de (Pip, 2020) Elaborado por Vera Cordova Ana Luisa



**Figura 21.** Comando de Instalación en Windows. Tomado del autor. Elaborado por Vera Cordova Ana Luisa

Luego de la instalación de pip ya se puede solo tipear pip y la librería y automáticamente se instalarán las librerías.

### ✓ **Pandas**

Comando de instalación: `pip install pandas`

Según (Pip, 2020), lo describe como un paquete de Python que proporciona estructuras de datos rápidas, flexibles y expresivas diseñadas para hacer que trabajar con datos estructurados (tabulares, multidimensionales, potencialmente heterogéneos) y series temporales sea fácil e intuitivo.

Esta librería necesaria para agilizar los procesos de Machine Learning en el proyecto ya que permite manipular los Dataset obtenidos del análisis de la red

#### ✓ **Numpy**

Comando de Instalación: `pip install Numpy`

Según (Pip, 2020), describe que “se puede utilizar como un contenedor eficiente multidimensional de datos genéricos. Se pueden definir tipos de datos arbitrarios. Esto permite que NumPy se integre sin problemas y rápidamente con una amplia variedad de bases de datos.”

#### ✓ **Matplotlib**

Comando de Instalación: `pip install matplotlib`

Según (Pip, 2020), matplotlib lo describe: se esfuerza por producir gráficos 2D y 3D de calidad de publicación para gráficos interactivos, publicación científica, desarrollo de interfaz de usuario y servidores de aplicaciones web dirigidos a múltiples interfaces de usuario y formatos de salida impresos. Hay un modo 'pylab' que emula gráficos matlab.

Gracias a esta librería se podrán graficar los elementos de clasificación y estimaciones de los algoritmos que se utilicen en el proyecto.

#### • **Scikit-learn**

Comando de Instalación: `pip install scikit-learn`

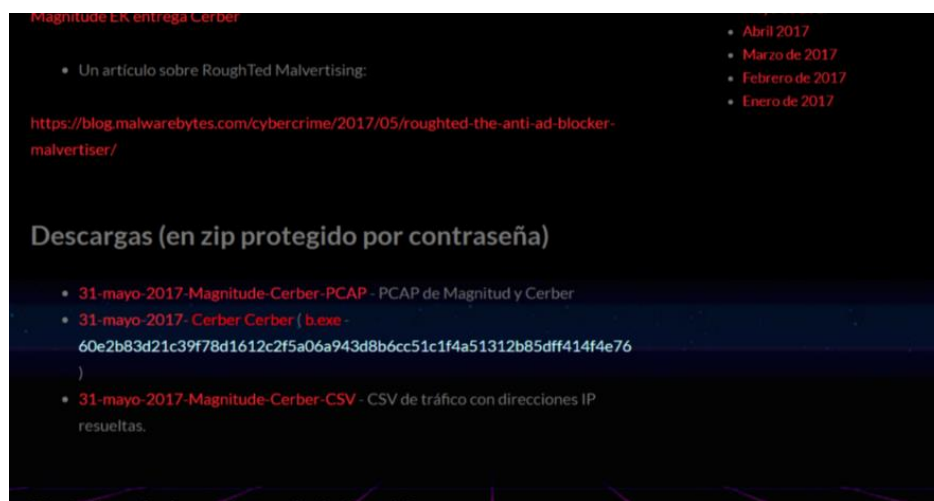
Este es una de las librerías más importante del proyecto ya que contiene todas las funciones, clases y objetos necesarios para desarrollar algoritmos de machine learning.

### **3.9 Ataque Realizado en la Red**

Para poner en marcha la propuesta se necesita que los IDS realicen la detención de algún ataque, registre una alerta en la red, para esto se llevó a cabo la simulación de los mismo.

Se utilizo el ransomware Cerber que es un tipo de malware (software malicioso) que encripta sus archivos y luego los retiene como rehenes, exigiendo un pago de rescate a cambio de devolvérselos, se pudo localizar toda la información en referencia al virus y en un archivo comprimido el virus para realizar las pruebas del IDS, se ejecutó en una máquina virtual de prueba, y se eliminó inmediatamente la máquina virtual para no contaminar la red,





**Figura 22.** Página Oficial. Fuente: (zerophagemalware, 2020) Elaborado por Vera Cordova Ana Luisa

### 3.10 Extracción de las Alertas Snort y Suricata.

Luego de realizar las configuraciones de funcionamiento en Pfsense de los IDS, este Firewall al ser un ambiente amigable posee en su interfaz de instalación tanto de Snort y Suricata una opción que se llama Alertas, en esta casilla se puede verificar el tipo de tráfico que tiene la red configurada y las anomalías que se presentan, también permite realizar la descarga del archivo de todos los datos obtenidos durante la ejecución del firewall.

En este proceso se toman 1000 alertas de Suricata y 1000 de Snort, las cuales mantienen el mismo formato en dos archivos distintos

Services / Snort / Alerts

Short InterfacesGlobal SettingsUpdatesAlertsBlockedPass ListsSuppressIP ListsSID MgmtLog MgmtSync

Clear all interface log files

Alert Log View Settings

Interface to InspectWANChoose interface..Auto-refresh view1000SaveAlert lines to display.

Alert Log ActionsDownloadClear

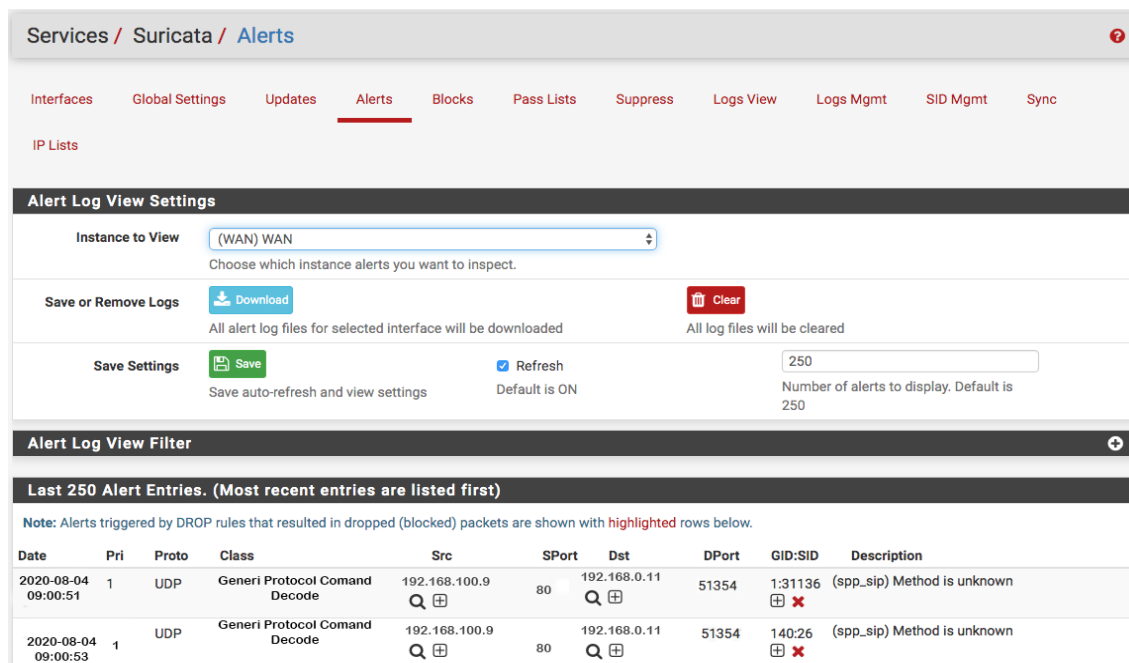
Alert Log View Filter

Last 1000 Alert Log Entries

Date	Pri	Proto	Class	Source IP	SPort	Destination IP	DPort	SID	Description
2020-08-04 09:00:51	1	UDP	Generi Protocol Comand Decode	192.168.100.9	80	192.168.0.11	51354	1:31136	(spp_sip) Method is unknown
2020-08-04 09:00:53	1	UDP	Generi Protocol Comand Decode	192.168.100.9	80	192.168.0.11	51354	140:26	(spp_sip) Method is unknown
2020-08-04 09:00:55	1	UDP	Generi Protocol Comand Decode	192.168.100.9	80	192.168.0.11	51354	140:26	(spp_sip) Method is unknown
2020-08-04 09:00:58	1	UDP	Generi Protocol Comand Decode	192.168.100.9	80	192.168.0.11	51354	140:26	(spp_sip) Method is unknown
2020-08-04 09:01:12	1	UDP	Generi Protocol Comand Decode	192.168.100.9	80	192.168.0.11	51354	140:26	(spp_sip) Method is unknown
2020-08-04 09:01:15	1	UDP	Generi Protocol Comand Decode	192.168.100.9	80	192.168.0.11	51354	140:26	(spp_sip) Method is unknown

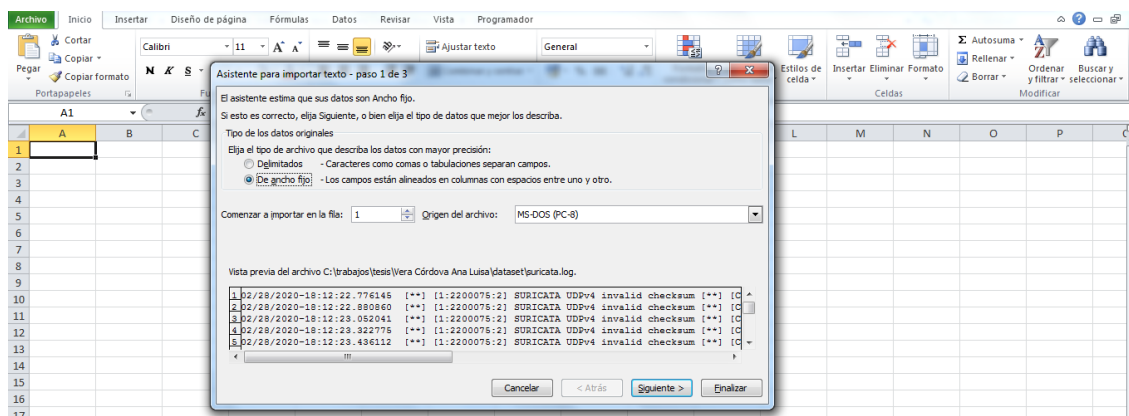
**Figura 23.** Sistema de Alertas. Ambiente de trabajo de Pfsense Elaborado por Vera Cordova Ana Luisa

Los archivos descargados presentan una compilación de tipo tar.gz las cuales son creados con el algoritmo de compilación GNU.



**Figura 24.** Descargas de Archivos. Ambiente de trabajo Pfsense. Elaborado por Vera Cordova Ana Luisa

Los archivos al descomprimirlos son en formato plano txt y se transforman a csv por medio de Excel para poder trabajarlos como dataset.



**Figura 25.** Transformación de archivo Txt a Csv. Elaborado por Vera Cordova Ana Luisa

### 3.10.1 Limpieza de Datos (Dataset).

Los archivos se visualizan y eliminan errores de transformación, campos null.

	A	B	C	D	E	F	G	H
1	fecha	code	protocolo	clase	ip_ini	port_ini	ip_fin	port_fin
2	02/06/2020	3	TCP	Generic Protocol Command Decode	162.208.119.40	4036	192.168.0.9	443
3	02/04/2020	3	TCP	Generic Protocol Command Decode	192.168.0.9	80	192.168.0.11	51354
4	02/04/2020	3	TCP	Generic Protocol Command Decode	192.168.0.9	80	192.168.0.11	51348
5	02/04/2020	3	TCP	Generic Protocol Command Decode	192.168.0.9	80	192.168.0.11	51343
6	02/04/2020	3	TCP	Generic Protocol Command Decode	192.168.0.9	80	192.168.0.11	51342
7	02/04/2020	3	TCP	Generic Protocol Command Decode	192.168.0.9	80	192.168.0.11	51339
8	02/04/2020	3	TCP	Generic Protocol Command Decode	192.168.0.9	80	192.168.0.11	51334
9	02/04/2020	3	TCP	Generic Protocol Command Decode	192.168.0.9	80	192.168.0.11	51321
10	02/04/2020	3	TCP	Generic Protocol Command Decode	192.168.0.9	80	192.168.0.11	51089
11	02/04/2020	3	TCP	Generic Protocol Command Decode	192.168.0.9	80	192.168.0.11	51075
12	02/04/2020	3	TCP	Generic Protocol Command Decode	192.168.0.9	80	192.168.0.11	51066
13	02/04/2020	3	TCP	Generic Protocol Command Decode	192.168.0.9	80	192.168.0.11	51065
14	02/04/2020	3	TCP	Generic Protocol Command Decode	192.168.0.9	80	192.168.0.11	51064
15	02/04/2020	3	TCP	Generic Protocol Command Decode	192.168.0.9	80	192.168.0.11	51062
16	02/04/2020	3	TCP	Generic Protocol Command Decode	192.168.0.9	80	192.168.0.11	51061
17	02/04/2020	3	TCP	Generic Protocol Command Decode	192.168.0.9	80	192.168.0.11	51059
18	02/04/2020	3	TCP	Generic Protocol Command Decode	192.168.0.9	80	192.168.0.11	51058

**Figura 26.** Formato Suricata csv. Tomado del autor. Elaborado por Vera Cordova Ana Luisa

Se deben eliminar caracteres que no representen información como (\*, [], {}) o campos que no tengan información necesaria

	A	B	C	D	E	F	G	H
1	fecha	protocolo	ip_ini	port_ini	ip_fin	port_fin	clase	prioridad
2	28/08/20-00:06:57.578292	TCP	192.168.0.11	64733	162.208.119.40	80	Misc activity	3
3	28/08/20-00:06:57.696876	TCP	192.168.0.11	64733	162.208.119.40	80	Misc activity	3
4	28/08/20-00:06:57.697806	TCP	192.168.0.11	64733	162.208.119.40	80	Misc activity	3
5	28/08/20-00:06:57.808467	TCP	192.168.0.9	64733	162.208.119.40	80	Misc activity	3
6	28/08/20-00:06:57.833944	TCP	192.168.0.11	33022	162.208.119.40	80	Misc activity	3
7	28/08/20-00:06:57.950956	TCP	192.168.0.11	33022	162.208.119.40	80	Misc activity	3
8	28/08/20-00:06:57.951676	TCP	192.168.0.11	33022	162.208.119.40	80	Misc activity	3
9	28/08/20-00:06:57.952566	TCP	192.168.0.9	33022	162.208.119.40	80	Misc activity	3
10	28/08/20-00:06:57.952610	TCP	192.168.0.11	33022	162.208.119.40	80	Misc activity	3
11	28/08/20-00:06:57.952648	TCP	192.168.0.9	33022	162.208.119.40	80	Misc activity	3
12	28/08/20-00:06:57.952684	TCP	192.168.0.11	33022	162.208.119.40	80	Misc activity	3
13	28/08/20-00:06:57.953223	TCP	192.168.0.11	33022	162.208.119.40	80	Misc activity	3
14	28/08/20-00:06:57.953263	TCP	192.168.0.11	33022	162.208.119.40	80	Misc activity	3
15	28/08/20-00:06:57.953299	TCP	192.168.0.11	33022	162.208.119.40	80	Misc activity	3
16	28/08/20-00:06:57.954059	TCP	192.168.0.11	33022	162.208.119.40	80	Misc activity	3
17	28/08/20-00:06:58.067199	TCP	192.168.0.9	33022	162.208.119.40	80	Misc activity	3
18	28/08/20-00:06:58.068220	TCP	192.168.0.11	33022	162.208.119.40	80	Misc activity	3
19	28/08/20-00:06:58.068309	TCP	192.168.0.11	33022	162.208.119.40	80	Misc activity	3
20	28/08/20-00:06:58.068750	TCP	192.168.0.11	33022	162.208.119.40	80	Misc activity	3

**Figura 27.** Formato Snort csv. Tomado del autor. Elaborado por Vera Cordova Ana Luisa

### 3.10.2 Inserción de Data en Algoritmo

Existen muchos algoritmos supervisados y no supervisados que permiten obtener reconocimientos de patrones o anomalías en el tráfico de red. Para poder estudiar esta serie de patrones, se ha tomado 4933 datos de la red que son parte de las alertas que se visualizaron de cada uno los IDS, de las cuales se tomaron 1000 líneas son de alertas de Suricata y 3933 alertas tomadas de Snort descargados previamente en la sección de alertas.

A continuación, se describe los diferentes algoritmos que evalúen de manera óptima las anomalías.

En la investigación de (blogthinkbig, 2020), para entender la clasificación de las anomalías hay que centrarla en características especiales y la perspectiva de los dataset como, por ejemplo:

Básica: que identifica los campos netflow, ip y puertos de origen y destino

Derivadas: longitud de los flujos (fecha de inicio a fecha de fin), tamaño medio de los flujos (bytes/número de paquetes), tasa media de paquetes (número de paquetes/longitud), agregación por IP y carga de bytes, porcentaje de carga de tráfico en un nodo, etc.

Aplicación heurística específica: como patrones diarios o semanales, duración de la sesión de un cliente, etc.

Avanzada: similitud en el intervalo del flujo, entropía, información mutua, etc.

**Tabla 11.** Requerimientos del algoritmo

Algoritmo	Clases
SVM	svm.LinearSVC (Clasificación de vectores de soporte lineal). svm.LinearSVR(Regresión de vectores de soporte lineal.) svm.NuSVC (Clasificación de vectores de Nu-Support) svm.NuSVR(Nu Support Regresión de vectores). svm.OneClassSVM(Detección de valores atípicos sin supervisión). svm.SVC( C-Clasificación de vectores de soporte). Devuelve el límite más bajo para C, de modo que para C en (11_min_C, infinito) se garantiza que el modelo no estará vacío.
Discepción	algoritmo de refuerzo SAMME discreto algoritmo de refuerzo SAMME Real
Redes neuronales	Multi-layer Perceptron Classification Regularization Regression
PSO (Optimización por enjambre de partículas)	Binaria Discreta Combinatoria
Cluster	Mean Shift K-Means Affinity propagation Spectral clustering Ward hierarchical clustering Agglomerative clustering DBSCAN

Algoritmo	Clases
	OPTICS Gaussian mixtures Birch

**Fuente.** Características del requerimiento del algoritmo a implementar. Tomado del autor.  
Elaborado por Vera Cordova Ana Luisa.

### 3.11 Creación de Algoritmo de Carga de Datos (Dataset)

Después de realizar la clasificación y limpieza de la data, se inserta en el algoritmo, en este caso como es del IDS Snort, se nombró al archivo snortS.csv como se observa en la figura 30, para que puedan insertarse los datos a analizar.

De la data recolectada se toma una parte random para poder realizar el escaneo de información y en base a esta selección mostrar su predicción.

Las variables que se toman de referencia de snort son: fecha, protocolo, ip inicial, ip final, clase y la prioridad.

```
In [260]: import pandas as pd
import re
import sys
from operator import add
import pandas as pd
from sklearn.ensemble import IsolationForest
df = pd.read_csv('/home/tesisanaug/snort5.csv', sep=";", encoding='latin1')
df
```

Out[260]:

	fecha	protocolo	ip_ini	port_ini	ip_fin	port_fin	clase	prioridad
0	28/08/20-00:06:57.578292	TCP	192.168.0.11	64733	162.208.119.40	80	Misc activity	3
1	28/08/20-00:06:57.696876	TCP	192.168.0.11	64733	162.208.119.40	80	Misc activity	3
2	28/08/20-00:06:57.697806	TCP	192.168.0.11	64733	162.208.119.40	80	Misc activity	3
3	28/08/20-00:06:57.808467	TCP	192.168.0.9	64733	162.208.119.40	80	Misc activity	3
4	28/08/20-00:06:57.833944	TCP	192.168.0.11	33022	162.208.119.40	80	Misc activity	3
5	28/08/20-00:06:57.950956	TCP	192.168.0.11	33022	162.208.119.40	80	Misc activity	3
6	28/08/20-00:06:57.951676	TCP	192.168.0.11	33022	162.208.119.40	80	Misc activity	3
7	28/08/20-00:06:57.952566	TCP	192.168.0.9	33022	162.208.119.40	80	Misc activity	3
8	28/08/20-00:06:57.952610	TCP	192.168.0.11	33022	162.208.119.40	80	Misc activity	3
9	28/08/20-00:06:57.952648	TCP	192.168.0.9	33022	162.208.119.40	80	Misc activity	3
10	28/08/20-00:06:57.952684	TCP	192.168.0.11	33022	162.208.119.40	80	Misc activity	3
11	28/08/20-00:06:57.953223	TCP	192.168.0.11	33022	162.208.119.40	80	Misc activity	3
12	28/08/20-00:06:57.953263	TCP	192.168.0.11	33022	162.208.119.40	80	Misc activity	3
13	28/08/20-00:06:57.953299	TCP	192.168.0.11	33022	162.208.119.40	80	Misc activity	3

**Figura 28.** Inserción de Dataset de Snort. Tomado del autor. Elaborado por Vera Córdoba Ana Luisa

De la misma forma como se insertó la data de snort en el algoritmo, se realiza la misma acción para los datos obtenidos de suricata, como muestra la figura 29 a diferencia de snort este presenta una nueva columna que se llama código y las demás variables son de base para

el algoritmo el cual trabajara con una referencia en base al entrenamiento que a este se le dé, para ello se considera el aumento de 1 variable.

```
In [81]: import pandas as pd
import re
import sys
from operator import add
import pandas as pd
from sklearn.ensemble import IsolationForest
df = pd.read_csv('/home/tesisanaug/suricata4.csv', sep=";", encoding='latin1')
df
```

Out[81]:

	fecha	code	protocolo	clase	ip_ini	port_ini	ip_fin	port_fin
0	02/06/2020	3	TCP	Generic Protocol Command Decode	162.208.119.40	4036	192.168.0.9	443
1	02/04/2020	3	TCP	Generic Protocol Command Decode	192.168.0.9	80	192.168.0.11	51354
2	02/04/2020	3	TCP	Generic Protocol Command Decode	192.168.0.9	80	192.168.0.11	51348
3	02/04/2020	3	TCP	Generic Protocol Command Decode	192.168.0.9	80	192.168.0.11	51343
4	02/04/2020	3	TCP	Generic Protocol Command Decode	192.168.0.9	80	192.168.0.11	51342
5	02/04/2020	3	TCP	Generic Protocol Command Decode	192.168.0.9	80	192.168.0.11	51339
6	02/04/2020	3	TCP	Generic Protocol Command Decode	192.168.0.9	80	192.168.0.11	51334
7	02/04/2020	3	TCP	Generic Protocol Command Decode	192.168.0.9	80	192.168.0.11	51321
8	02/04/2020	3	TCP	Generic Protocol Command Decode	192.168.0.9	80	192.168.0.11	51089
9	02/04/2020	3	TCP	Generic Protocol Command Decode	192.168.0.9	80	192.168.0.11	51075
10	02/04/2020	3	TCP	Generic Protocol Command Decode	192.168.0.9	80	192.168.0.11	51066
11	02/04/2020	3	TCP	Generic Protocol Command Decode	192.168.0.9	80	192.168.0.11	51065
12	02/04/2020	3	TCP	Generic Protocol Command Decode	192.168.0.9	80	192.168.0.11	51064

**Figura 29** Extracción al Dataset de Suricata. Tomado del autor. Elaborado por Vera Cordova Ana Luisa

Se revisa que no falte ningún dato en la inserción de los archivos y que todo este correcto para la ejecución

```
18:14:07.141940 [1:65000000:2] SURICATA ICMPv4 Generic Protocol Command Decode] 3] 192.168.0.9 192.168.0.11 8 192.168.0.11 51354
```

```
2020-02-29 03:45:11.840410 [1:2200076:2] SURICATA ICMPv4 Generic Protocol Command Decode] [Priority: 3] (ICMP) 192.168.0.9 8 192.168.0.11 -99999
```

220 rows x 10 columns

```
In [122]: df.describe()
```

Out[122]:

	port_ini	port_fin
count	220.000000	220.000000
mean	31332.872727	-1141.700000
std	20399.978280	17451.259952
min	8.000000	-99999.000000
25%	13269.250000	53.000000
50%	33544.000000	53.000000
75%	45955.750000	53.000000
max	65386.000000	59725.000000

**Figura 30.** Revisión de valores insertados en algoritmo. Tomado del autor. Elaborado por Vera Córdoba Ana Luisa

Las variables de los Dataset deben ser enteros para poder hacer la comparación con el algoritmo seleccionado

```
In [266]: pd.unique(df['ip_ini'])
Out[266]: array([u'192.168.0.11', u'192.168.0.9'], dtype=object)

In [267]: #array(['192.168.0.11', '192.168.0.9'], dtype=object) ip_ini
d = {'192.168.0.11' : '1', '192.168.0.9' : '2'}
df['ip_ini'] = df['ip_ini'].map(d)
df.head()

Out[267]:
```

	protocolo	ip_ini	port_ini	ip_fin	port_fin	clase	prioridad
0	1	1	64733	162.208.119.40	80	Misc activity	3
1	1	1	64733	162.208.119.40	80	Misc activity	3
2	1	1	64733	162.208.119.40	80	Misc activity	3
3	1	2	64733	162.208.119.40	80	Misc activity	3
4	1	1	33022	162.208.119.40	80	Misc activity	3

```
In [276]: df[['protocolo','ip_ini','port_ini','ip_fin','port_fin','clase','prioridad']] = df[['protocolo','ip_ini','port_ini','ip_fin','port_fin','clase','prioridad']].astype(int)
```

**Figura 31.** Paso de variables numéricas a enteros. Tomado del autor. Elaborado por Vera Cordova Ana Luisa

### 3.11.1 Algoritmos de Estimación

Conocido en español como Aislamiento del Algoritmo Forestal.

```
estimador =IsolationForest(n_estimators=100, #utiliza arboles... n_estimators es el numero de arboles
                           contamination=0.01,# umbral donde se detectan las anomalias
                           max_samples=256)#
estimador

IsolationForest(bootstrap=False, contamination=0.01, max_features=1.0,
                max_samples=256, n_estimators=100, n_jobs=1, random_state=None,
                verbose=0)
```

**Figura 32.** Comando de estimación. Tomado del autor. Elaborado por Vera Córdoba Ana Luisa

Se hace la transformada que es una manipulación sobre los datos, ordenando el índice de forma ascendente, tiempo de inicio y tiempo final

Si no se devuelve ningún tipo de estimador no se aplica ningún filtro y se devuelven todos los estimadores. Los valores posibles son 'clasificador', 'regresor', 'agrupamiento' y 'transformador' para obtener estimadores solo de estos tipos específicos, o una lista de estos para obtener los estimadores que se ajustan al menos a uno de los tipos.

Para calcular la predicción y demás campos todos los campos deben estar categorizados y en números enteros

Gracias al código se puede realizar la categorización transformando los datos obtenidos en campos en enteros y clasificados se pasa a calcular la predicción de la transformada. La predicción presentara las anomalías con un número “-1” y las normales un “1”,



A continuación, se presenta el entrenamiento de la data tal como se muestra en la figura

33

```
array(['192.58.128.30', ' 202.12.27.33', ' 193.0.14.129',
      ' 198.97.190.53', ' 198.41.0.4', '162.159.200.123', ' 192.5.5.241',
      ' 199.7.83.42', '192.203.230.10', '193.0.14.129', ' 199.9.14.201',
      '143.255.251.226', ' 192.58.128.30', '162.159.200.1',
      ' 192.112.36.4', ' 192.36.148.17', ' 192.33.4.12', ' 199.19.56.1',
      ' 66.220.13.229', ' 199.249.120.1', ' 192.43.172.30',
      ' 66.220.13.230', ' 69.55.226.55', ' 199.19.57.1', ' 192.35.51.30',
      ' 108.59.165.1', ' 192.12.94.30', ' 192.52.178.30', ' 192.5.6.30',
      ' 192.54.112.30', ' 199.249.112.1', ' 199.19.54.1', ' 199.19.53.1',
      ' 204.42.254.5', ' 204.61.216.4', ' 192.31.80.30',
      ' 192.41.162.30', ' 192.55.83.30', ' 192.48.79.30',
      '72.249.171.254', ' 136.144.52.123', ' 136.144.52.122',
      ' 165.227.133.2', ' 107.170.182.1', ' 136.144.52.12',
      ' 188.166.56.96', ' 64.57.176.2', ' 178.128.191.1',
      ' 192.42.93.30', ' 199.249.223.5', ' 204.13.251.136',
      ' 162.88.61.21', ' 107.170.182.17', ' 72.249.171.25',
      ' 64.57.183.94', ' 46.227.203.69', ' 216.239.34.10',
      ' 85.214.195.29', ' 162.88.60.21', ' 204.13.251.13',
      ' 216.239.32.10', ' 216.239.34.109', ' 216.239.36.10',
      ' 45.127.113.23', ' 185.134.197.7', ' 31.3.105.98',
      ' 45.79.130.187', ' 212.25.19.23', '192.168.0.12640',
      '192.168.0.12', '192.168.0.1295', '192.168.0.1264',
      '192.168.0.1289', '192.168.0.1220', '192.168.0.11'], dtype=object)
```

**Figura 33.** Comando de arreglos y entrenamiento de algoritmo. Tomado del autor. Elaborado por Vera Córdova Ana Luisa

Ya esta parte del algoritmo se encarga de dar informes de predicción, estimación, número de categorías clasificadas, donde identifica cuantos grupos están conformado la Dataset, en base a la distancia euclidiana, los falsos-positivos la cual es la estimación de la predicción en base a el número 1 que sería el 100% de la estimación y las Normales que serían los Dataset que no están contaminados.

En la figura 36 muestra que las predicciones de los datos tomados de suricata en las columnas destinada para predicción aparece el número 1, que como se indica en líneas anteriores sería el 100% de estimación.



```
--
: falsopositivos = dftransformada[(dftransformada['prediccion']==-1)]
normales = dftransformada[(dftransformada['prediccion']==1)]
print (falsopositivos)
```

	protocolo	ip_ini	port_ini	ip_fin	port_fin	clase	prediccion
195	1	2	80	3	64673	1	-1
199	1	1	64632	1	139	1	-1
200	1	1	64632	1	139	1	-1
201	1	1	64605	1	139	1	-1
202	1	1	64605	1	139	1	-1
203	1	2	80	3	64290	1	-1
204	1	2	80	3	64280	1	-1
205	1	2	80	3	64279	1	-1
206	1	2	80	3	64154	1	-1
207	1	2	80	3	64150	1	-1

**Figura 34.** Comandos de predicción en dataset de Suricata. Tomado del autor. Elaborado por Vera Córdova Ana Luisa.

Con la inserción de la data de Snort, como indica en la figura 34 las anomalías se presentan con un -1 que a diferencia de Snort en esta selección de datos si se presenta en un puerto, esto permite tomar en cuenta un próximo evento.

```
En [290]: estimador . encajar ( X )
Prediccion = estimador . predict ( X )
dftransformada [ 'prediccion' ] = prediccion
dftransformada # en las predicciones las anomalías se presentan con un numero -1
```

Fuera [290]:

	protocolo	ip_ini	port_ini	ip_fin	port_fin	clase	prioridad	prediccion
0	1	1	64733	1	80	1	3	1
1	1	1	64733	1	80	1	3	1
2	1	1	64733	1	80	1	3	1
3	1	2	64733	1	80	1	3	-1
4	1	1	33022	1	80	1	3	1
5	1	1	33022	1	80	1	3	1
6	1	1	33022	1	80	1	3	1
7	1	2	33022	1	80	1	3	-1
8	1	1	33022	1	80	1	3	1
9	1	2	33022	1	80	1	3	-1
10	1	1	33022	1	80	1	3	1

**Figura 35.** Comando de predicción en dataset de Snort. Tomado del autor. Elaborado por Vera Córdova Ana Luisa.

Se puede visualizar el funcionamiento del algoritmo mediante la clasificación de los datos además ver los aciertos y fallos como se muestra en la figura 35, dando como resultado la ponderación de la precisión 100%, recall es el 100% permite encontrar mediante clasificación todas las muestras que sean positivas, f1-score da 100% es un valor promedio.

```

n_neighbors = 7

knn = KNeighborsClassifier(n_neighbors)
knn.fit(X_train, y_train)
print('Accuracy of K-NN classifier on training set: {:.2f}'
      .format(knn.score(X_train, y_train)))
print('Accuracy of K-NN classifier on test set: {:.2f}'
      .format(knn.score(X_test, y_test)))

pred = knn.predict(X_test)
print(confusion_matrix(y_test, pred))
print(classification_report(y_test, pred))

Accuracy of K-NN classifier on training set: 1.00
Accuracy of K-NN classifier on test set: 1.00
[[977  0]
 [ 0  7]]

```

	precision	recall	f1-score	support
1	1.00	1.00	1.00	977
2	1.00	1.00	1.00	7
avg / total	1.00	1.00	1.00	984

**Figura 36.** Matriz de confusión de algoritmo MeanShief. Tomado del autor. Elaborado por Vera Córdova Ana Luisa

En esta parte del algoritmo que se muestra en la figura se encarga de dar informes de predicción, estimación, número de categorías clasificadas, donde identifica cuantos grupos están conformado la Dataset, en base a la distancia euclidiana, los falsos-positivos las cual es la estimación de la predicción en base a el número 1 que sería el 100% de la estimación y las Normales que serían los Dataset que no están contaminados.

```

print(__doc__)

import numpy as np
from sklearn.cluster import MeanShift, estimate_bandwidth
from sklearn.datasets import make_blobs

centers = df

# #####
# Calcula la agrupación con MeanShift

# El siguiente ancho de banda se puede detectar automáticamente usando
bandwidth = estimate_bandwidth(X, quantile=0.2, n_samples=500)

ms = MeanShift(bandwidth=bandwidth, bin_seeding=True)
ms.fit(X)
labels = ms.labels_
cluster_centers = ms.cluster_centers_

labels_unique = np.unique(labels)
n_clusters_ = len(labels_unique)

print("Cantidad de Grupos estimados:")
print(n_clusters_)
print("Falsos Positivos:")
print ((falsospositivos.size/dftransformada.shape[1]/dftransformada.shape[0]))#calculo e
stimado para falsos positivos
print("Normal Estimada:")
print ((normales.size/dftransformada.shape[1]/dftransformada.shape[0]))#calculo estimad
o para falsos positivos
# #####
# Plot result
import matplotlib.pyplot as plt
from itertools import cycle

```

**Figura 37.** Comandos finales de algoritmo. Tomado del autor. Elaborado por Vera Córdova Ana Luisa

### 3.12 Observación de Resultados

Para la aplicación del algoritmo se utilizaron herramientas comerciales, para este caso el Pfsense fue base para el desarrollo de cada paso.

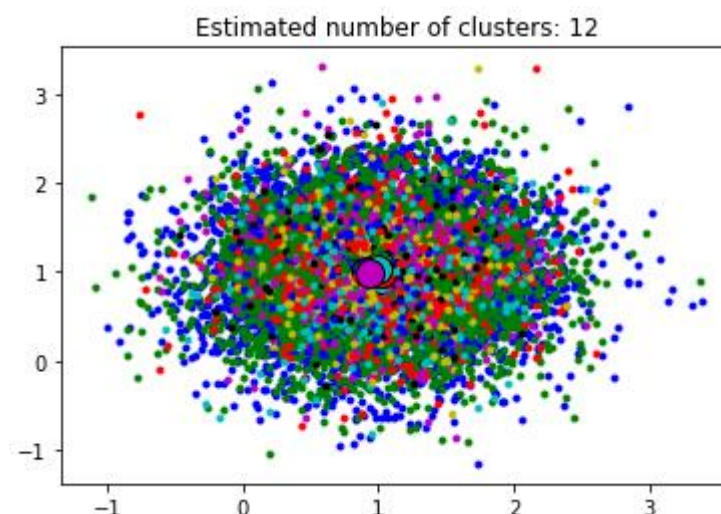
Snort ofrece herramientas de las cuales se destaca el uso de patrones de regla y muestra cuando y como se produce cada ataque que percibe, además que utiliza un lenguaje de descripción sencillo pero flexible y potente en su ejecución, en la extracción de la data se presentó en ocasiones lento y represento tiempo en la recolección de datos.

A diferencia de Snort, Suricata proporciona rapidez el momento de ejecutarse, esto gracias al ancho de banda que brinda, destaca sobre Snort por ser más controlable al momento de ejecutarse en el escaneo de la red

La ejecución de cada IDS fue realizada en una máquina virtual

#### ✓ Aplicación de Algoritmo

La aplicación del algoritmo MeanShift da como resultado la figura 38, la cual se presenta la estimación de clúster de la data recolectado en el IDS Snort, los colores representan la clasificación del número de categorías detectadas, en base al Dataset de Snort se estiman 12 categorías, este IDS obtiene muchas alertas las cuales no permiten de una forma gráfica distanciar las categorización se presentan en varios puntos del plano que representan un tiempo en referencias a las alertas recibidas en snort, se determinan rangos desde el -1 tomando la estimación de predicción y precisión realizada.

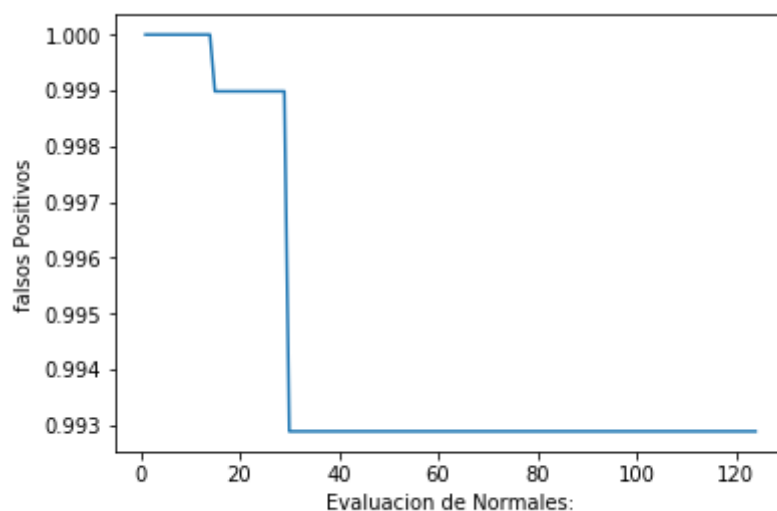


**Figura 38.** Grafica de conjunto de datos tomados de Snort. Tomado del autor. Elaborado por Vera Córdova Ana Luisa.

Tomando como referencia el número de clúster y su silueta (colores que se define cada grupo) se realiza la estimación de la cantidad de datos en las cuales se presentan como falsos positivos en base al entrenamiento del algoritmo sobre los datos.

La información está representada en los clústeres que se forman en el algoritmo, snort en este caso, toma la mayoría posible y no los ve tanto como vulnerabilidades, sino como cantidad de datos que tiene algo en común.

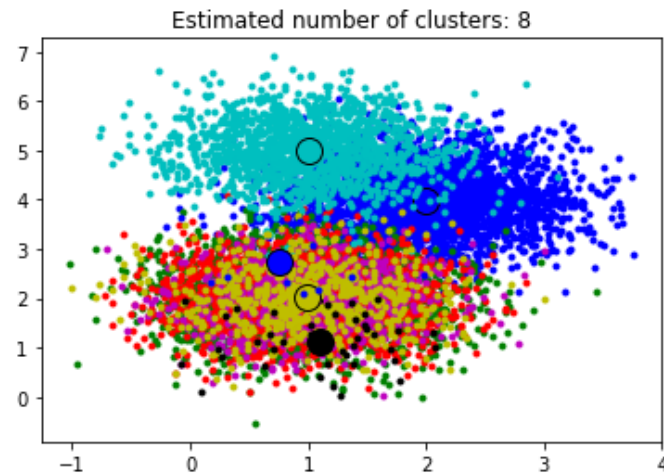
Lo común en el clúster es la anomalía presentada en ese paquete de información, de acuerdo con esa característica toman un color aleatorio que los representa gráficamente.



**Figura 39.** Evaluación de falsos positivos tomados de dataset de snort. Elaborado por Vera Cordova Ana Luisa

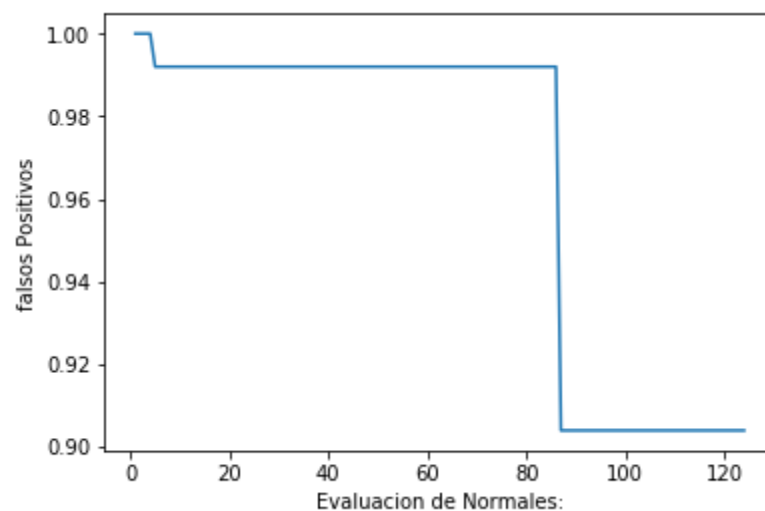
En base a los resultados obtenidos Suricata identifica con más exactitud los falsos positivos y la clasificación de los distintos tipos de alerta, las cuales se representan en la figura 40, indica que tiene solo 8 clúster, al ser más preciso en la lectura de datos se reduce el número de clúster, se evidencia tres zonas de colores puntuales las cuales se puede analizar que Suricata clasifica de una forma más efectiva los alertas.

Los colores no son de forma estándar, ni cada uno representa algo de forma permanente, estos se van dando y proyectando en base al número de alertas e intrusiones recibidas con los datos insertados, se evidencia la fuerza en dos colores la cual presenta reincidencia de una amenaza en específico, esto optimiza el nivel de predicción a futuros ataques.



**Figura 40.** Grafica de conjunto de datos tomados de Suricata. Tomado del autor. Elaborado por Vera Cordova Ana Luisa.

A continuación, se presenta la estimación de los falsos positivos de los datos de Suricata, la figura 41 indica un espacio más prolongado en el segmento seleccionado por el algoritmo, los datos con valores de -1 presentan una alerta.



**Figura 41.** Estimación de falsos positivos de datos obtenidos de Suricata. Tomada del autor. Elaborado por Vera Cordova Ana Luisa

Dada esta ejecución, sobre los datos de Suricata se maneja un pequeño porcentaje que indica dos posibles razones de su filtración, la primera es que por el rango en el que maneja realiza una selección de datos pero que no son vulnerables, el algoritmo toma los datos como si fuera un modelo de prueba y en base a su aprendizaje no refleja mayor detalle para las anomalías que presenta la red.

Se tomo en mismo ancho con respecto a la selección de datos en el algoritmo, la ventaja de Suricata es que presenta mayor filtro e indica una posible alerta lo que permitió mejor predicción y reducción de memoria en los clústeres que se forman.

Esto nos indica que los datos en la empresa NewOffice se encuentra vulnerables a nivel de red LAN, de los datos a proteger como se aprecia en la figura 40, los datos no se encuentra a nivel de ataques en su totalidad, lo cual se puede prevenir para así evitar futuros ataques y robo de información, en el análisis podemos ver que se destruyen ciertos datos y representa perdida para la empresa, ya que de forma interna cada información almacenada en los departamentos correspondientes tiene su valor.

## **Conclusión**

Los resultados obtenidos en el presente trabajo indican el cumplimiento de los objetivos planteados en el capítulo 1, lo cual se destaca que:

El algoritmo de aprendizaje implementado indicó dos tipos de predicciones con respecto a los datos obtenidos, estos resultan en base al IDS que se utilizó para la recolección de datos, Suricata logro tener mayor precisión con sus datos ya que presento menor número de clúster, una ventaja que tiene es su capacidad para comprender el nivel 7 del modelo OSI, lo que mejora su capacidad para detectar malware.

Una vez realizado el levantamiento de información de la red empresa reflejo que no mantiene una estructura robusta a nivel de red y seguridad, los datos se encuentran expuestos a manipulación de cualquier usuario, que si bien puede ser tomadas de manera física (dispositivos electrónicos) o mediante el uso de compartidos ya que no restringe de manera formal el acceso a los mismos, los datos que se manejan en la empresa son clasificados en base al nivel o rango de importancia que el usuario considere, pero no se encuentran libres de vulnerabilidades.

Entre las herramientas más reconocidas en el ámbito de IDS destacan Snort y Suricata, ambas se basan en conjuntos de reglas. La mayoría de las pruebas han demostrado que las reglas de Snort. son complementarias y ambas son necesarias para optimizar la detección de todos los tipos de ataques. Además, Snort como Suricata han demostrado su capacidad para detectar ataques basándose en firmas de reglas.

En el proceso de captura de datos se realizó la configuración de las reglas definidas para el proyecto en los IDS y se llevó a cabo la extracción de alertas de anomalías gracias al UTM Firewall Pfsense que es donde están instalados estos aplicativos lo cual facilita la descarga de los datos recolectados.

Se realizo la aplicación del algoritmo aprendizaje MeanShift porque no es necesario indicar el número de clúster como limitante, ya que una de sus ventajas es que durante su ejecución los descubre automáticamente y convergen hacia los puntos de mayor densidad que guardan mayor relación entre los datos obtenidos de los IDS y que guardan relación entre sí.

## **Recomendación**

En un trabajo de titulación, siempre se plantea que haya una mejora continua, para ellos se recomienda que en las futuras implementaciones y uso de la herramientas mencionadas en este trabajo se considere lo siguiente:

Ya que en la actualidad está tomando fuerza el uso de sistemas de detección con nuevas características y diseños las cuales ayudan a mantener la privacidad de los sistemas, se debe realizar actualizaciones constantes con relación al uso de algoritmos de aprendizaje, para de esta manera logra complementar nuevas líneas de comando dentro de un algoritmo, estas podrían predecir y evitar el ingreso de agentes anómalos en la red, que sería lo óptimo en el ámbito de seguridad informática.

Toda máquina, dispositivo o equipo informático que mantengan el servicio de internet debe contar con sistemas de seguridad que sirva como una barrera ante visitas inusuales en la red, los virus informáticos y anomalías se presentan como publicidad engañosa y los usuarios pueden verse perjudicados, aunque parezcan inofensivos. A nivel de seguridad en las empresas se debe implementar sistemas de uso autorizados en los que solo se permita el ingreso de personas calificadas.

Para el uso de herramientas Firewall con características que permitan brindar alertas o información del movimiento de datos, definir las reglas idóneas que la empresa o usuario maneja como política, esto permitirá que el personal autorizado en el sistema se mantenga al tanto de la seguridad de la red, y para procesos de análisis de datos se optimicen los resultados.

Recolectar la mayor cantidad posible de datos ya que así permitirá obtener mejores resultados de detección de anomalías, entre más tiempo este ubicado el firewall en una red LAN, mayor será la cantidad de información recolectada para futuros análisis en un algoritmo de aprendizaje.



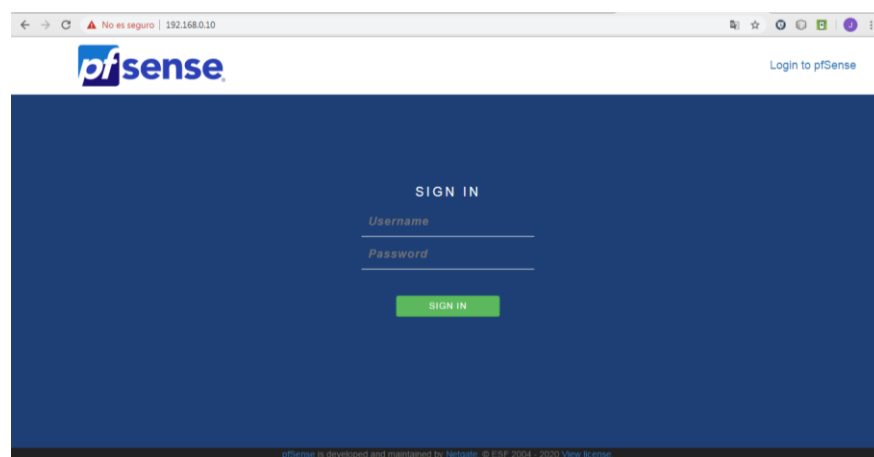
## Anexos

### Anexo 1 Instalación de PfSense

Snort y suricata son IDS de código abierto que se puede instalar fácilmente en un firewall pfSense para proteger una red doméstica o corporativa de intrusos. Snort también se puede configurar para funcionar como un sistema de prevención de intrusiones (IPS), lo que lo hace muy flexible.

Se debe bajar la imagen Iso de la página oficial (pfsense, 2020), la cual se descarga y se instala en un cd, pendrive o si se instalara en una máquina virtual el disco donde se valla a tomar la imagen iso.

- Selecciona la arquitectura.
- Luego selecciona el tipo de instalación de la imagen iso
  - Al seleccionar el tipo de instalación ya se ingresa a través de la bios y cambia el tipo de booteo si es pendrive o Cd/DVD.
  - Se aceptan las políticas de la empresa y continúa la instalación seleccionando luego Install.
  - Luego de ello se selecciona el tipo de idioma para el teclado. Anexo B-1
  - Seleccione Auto (UFS) para la instalación automática. Anexo B-2
  - Comenzará la instalación de forma automática, anexando todos los archivos necesarios. Anexo B-3.
  - Ya finalizada la instalación apaga el computador y cambia el proceso de arranque al disco duro para que inicie con el sistema operativo.
  - Ya instalada la herramienta puede tener acceso a través de la ip que te genera una conexión apache en la ip de V4/DHCP4 en nuestro caso es la ip 192.168.0.10/24

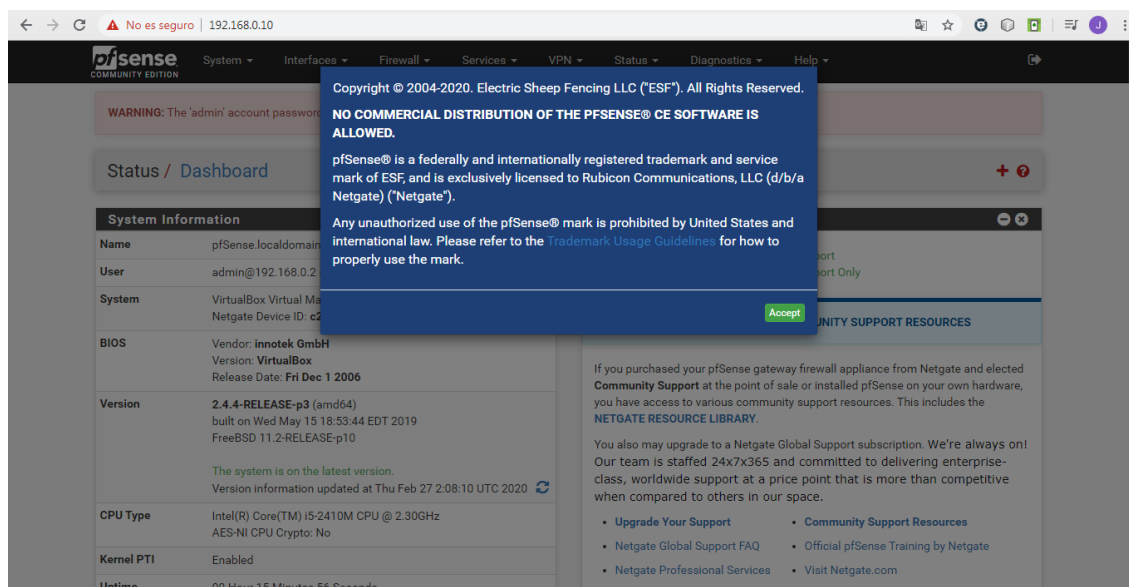


*Pantalla inicial de control Pfsense. Elaborada por Vera Córdova Ana Luisa*

La clave de acceso al servidor web es por defecto:

Usuario: admin

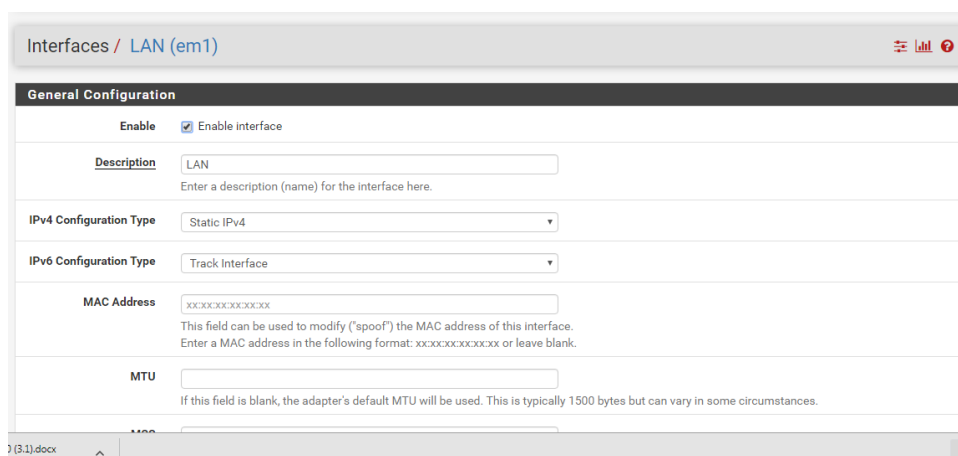
Contraseña: pfsense



*Pantalla inicial de control Pfsense. Elaborada por Vera Córdova Ana Luisa*

Mayormente se presenta el problema de que el firewall se instala con reglas de bloqueo, al momento de tener acceso por ello es aconsejable aplicar en la consola de la Shell y deshabilitar el firewall con todas su reglas.

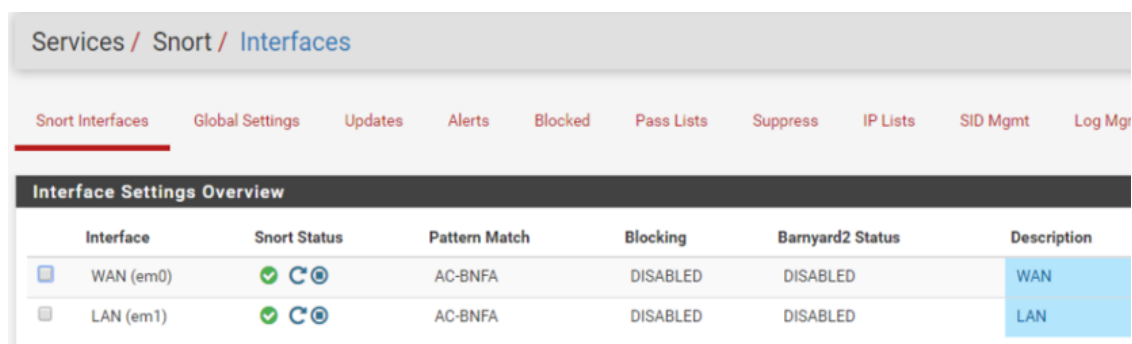
Otra novedad es que para tener acceso a los paquetes instalables debe de deshabilitar el IPv6 y direccionar toda la conexión de internet a la interfaz de IPv4







*Cambio de direccionamiento de Internet. Elaborada por Vera Córdova Ana Luisa*

## Anexo 2 Instalación de Snort

Para iniciar con Snort, deberá instalar el paquete utilizando el administrador de paquetes pfSense. El administrador de paquetes se encuentra en el menú del sistema de la GUI web de pfSense. Localice Snort de la lista de paquetes y luego haga clic en el símbolo más en el lado derecho para comenzar la instalación. Es normal que snort tarde un par de minutos en instalarse, tiene varias dependencias que pfSense primero debe descargar e instalar.



The screenshot shows the pfSense web interface for configuring Snort. The breadcrumb trail is 'Services / Snort / Interfaces'. Below the navigation tabs, there is a section titled 'Interface Settings Overview' containing a table with the following data:

Interface	Snort Status	Pattern Match	Blocking	Barnyard2 Status	Description
WAN (em0)	 	AC-BNFA	DISABLED	DISABLED	WAN
LAN (em1)	 	AC-BNFA	DISABLED	DISABLED	LAN

*Pack de Instalación. Elaborada por Vera Córdova Ana Luisa*

Una vez completada la instalación, Snort aparecerá en el menú de servicios.

Obteniendo un Código Oinkmaster Para que Snort sea útil, debe actualizarse con el último conjunto de reglas. El paquete Snort puede actualizar automáticamente estas reglas por usted, pero primero debe obtener un código Oinkmaster.

## Anexo 3 Reglas de Snort

El conjunto de versiones de suscriptores es el conjunto de reglas más actualizado disponible. El acceso en tiempo real a estas reglas requiere una suscripción anual pagada.

La otra versión de las reglas es el lanzamiento del usuario registrado, que es completamente gratuito para cualquier persona que se registre en el sitio Snort.org.

La principal diferencia entre los dos conjuntos de reglas es que las reglas en la versión de usuario registrado están 30 días detrás de las reglas de suscripción.

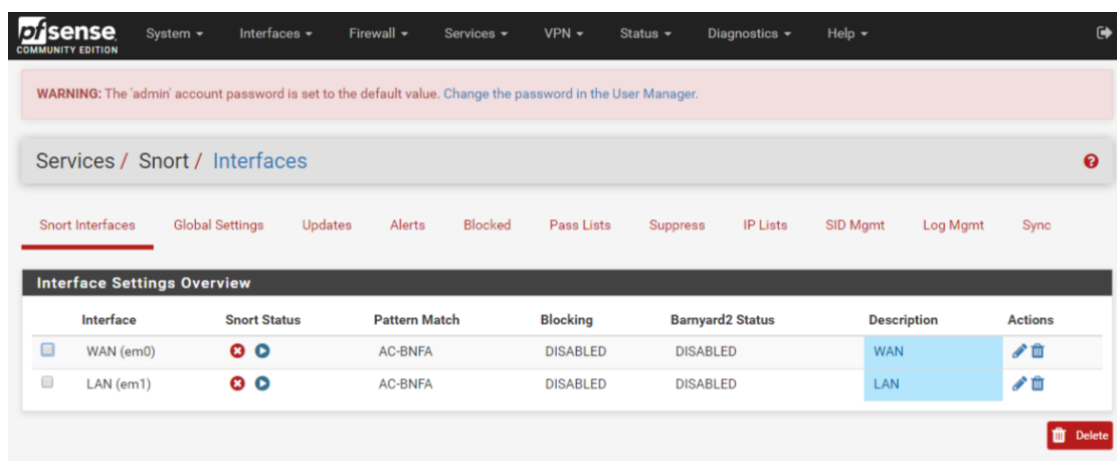
Si desea la protección más actualizada, debe obtener una suscripción. Para tener funcional las reglas de detección hay que realizar el registro de usuario en la página oficial de Snort y luego de confirmar el registro ingresar a la página oficial y obtener el código “oinkcodes” la cual permitirá bajar las reglas que permitirán crear las alerta y los filtros. Se ingresa en services/Snort y se habilita Snort VRT para colocar el código Oinkcodes

Después de obtener el Oinkcode, debe ingresarse en la configuración del paquete Snort. La página de configuración de Snort aparecerá en el menú de servicios de la interfaz web. Si no está visible, asegúrese de que el paquete esté instalado y vuelva a instalarlo si es necesario.

El Oinkcode debe ingresarse en la página de configuración global de la configuración de Snort. También me gusta marcar la casilla para habilitar las reglas de amenazas emergentes, también. Las reglas de ET son mantenidas por una comunidad de código abierto y pueden proporcionar algunas reglas adicionales que pueden no encontrarse en el conjunto Snort.

Por defecto, el paquete Snort no actualizará las reglas automáticamente. El intervalo de actualización recomendado es una vez cada 12 horas, pero puede cambiarlo para adaptarlo al entorno. Snort no viene con ninguna regla, por lo que se deberá actualizarlas manualmente la primera vez. Para ejecutar la actualización manual, haga clic en la pestaña de actualizaciones y luego haga clic en el botón actualizar reglas.

El paquete descargará los últimos conjuntos de reglas de Snort.org y también Amenazas emergentes si se seleccionada las opciones de detección. Una vez finalizadas las actualizaciones, las reglas se extraerán y estarán listas para su uso.



*Interfaz antes de funcionamiento. Elaborada por Vera Córdova Ana Luisa*

#### **Anexo 4 Interfax a Snort**

Antes de que Snort pueda comenzar a funcionar como un sistema de detección de intrusos, debe asignar interfaces para que pueda monitorear. La configuración típica es que Snort monitoree cualquier interfaz WAN.

El monitoreo de la interfaz LAN puede proporcionar cierta visibilidad a los ataques que ocurren desde su red. No es raro que muchas redes mayormente se infecten con malwares

en la red LAN para las pruebas de este IDS se utilizara un malware para luego clasificarlo ya que es el más común en la red estos mayormente comience a lanzar ataques en sistemas dentro y fuera de la red.

Existen reglas de categorización dentro de Snort Al dividir las reglas en categorías, puede habilitar solo las categorías particulares que le interesen. Se recomienda habilitar algunas de las categorías más generales. Si está ejecutando servicios específicos en su red, como un servidor web o de base de datos, también debe habilitar las categorías correspondientes. Es importante recordar que Snort requerirá más recursos del sistema cada vez que se active una categoría adicional. Esto también puede aumentar el número de falsos positivos tanto en las alertas como en la clasificación de las pruebas del algoritmo de Machine Learning, también. En general, es mejor activar solo los grupos que se necesitan.

Para probar luego el algoritmo de machine Learning con un ataque de malware es recomendable activar `snort_virus.rules` la cual da alerta de este tipo de virus.

Al momento de iniciar la captura de archivos de red se presentaron algunos errores, el de configuración del preprocesador, las cuales varias de las reglas requieren que la opción de inspección HTTP esté habilitado en la configuración del preprocesador, así que hay que asegurarse de tener esta función activada ya que presento fallas al momento de iniciarlo.

Ya para finalizar la utilización de Snort los detalles de la red lo podemos visualizar en alertas, la cual Después de que Snort se haya configurado e iniciado correctamente, debe comenzar a ver alertas una vez que se detecte el tráfico que coincida con las reglas. Si no ve ninguna alerta, hay que esperar un poco y luego verifique nuevamente. Puede pasar un tiempo antes de que se vean las alertas, según la cantidad de tráfico y las reglas habilitadas.

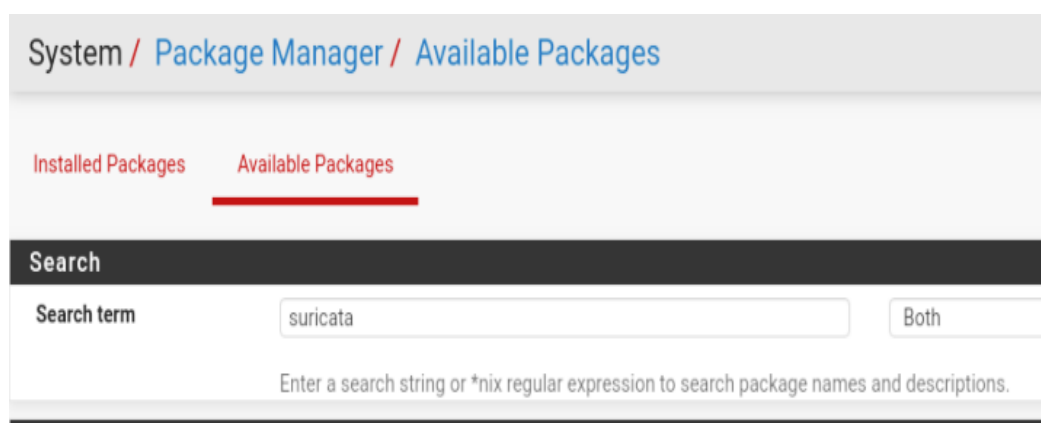
Al visualizar los datos necesario podemos descargarle en el botón descarga la cual enviara a la ruta que uno desee la Dataset en formato Txt archivo plano para su utilización en el algoritmo de machine Learning.



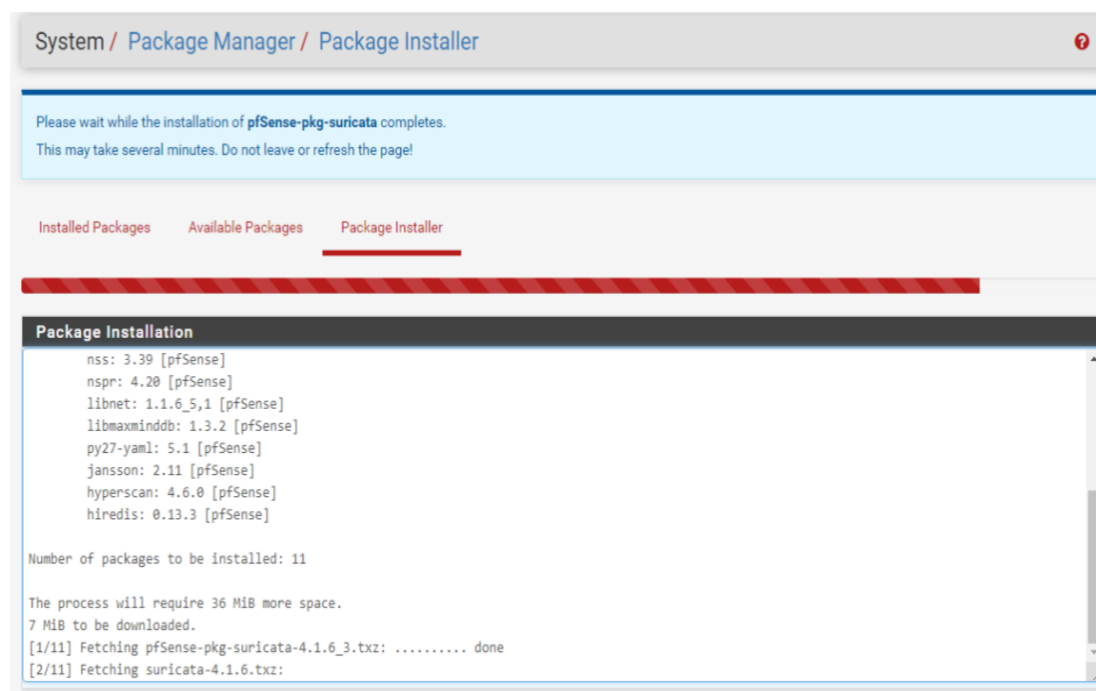
*Pantallas de snort. Elaborado por Vera Córdova Ana Luisa*

## Anexo 5 Instalación de Suricata

Se dirige hacia la pestaña de pack manager e instala el paquete, Lo primero que se nota después de habilitar Suricata fue mucho ruido de las alertas. Las primeras alertas que inundaron el feed fueron cosas como "suma de comprobación no válida UDP4" y algunas otras que, aunque no necesariamente eran buenas para ver, a diferencia de Snort que presenta sus alertas en base a la configuración de alertas. Estos pueden ser importantes para ver en un entorno diferente, sin embargo, hay que agregar rápidamente algunas reglas de "supresión" que ocultaban estos numerosos y poco interesantes tipos de alertas.



*Selección de Paquete Suricata. Elaborada por Vera Córdova Ana Luisa*



*Proceso de Instalación de suricata Elaborado por Vera Córdova Ana Luisa*

## Anexo 6 Modelo de Algoritmo Meanshift

*Librería de inserción de archivos de alertas.*

```
In [260]: import pandas as pd
import re
import sys
from operator import add
import pandas as pd
from sklearn.ensemble import IsolationForest
df = pd.read_csv('/home/tesisanaug/snort5.csv', sep=";", encoding='latin1')
df
```

*Limpieza de los archivos de forma interna.*

```
In [263]: df.dtypes
Out[263]: protocolo    object
ip_ini              object
port_ini           int64
ip_fin             object
port_fin           int64
clase              object
prioridad          int64
dtype: object

In [264]: pd.unique(df['protocolo'])#se elimina no ayuda ya que existe solo 1 clase
Out[264]: array([u'TCP'], dtype=object)

In [265]: #array(['UDPv4', 'ICMPv4'], dtype=object)//protocolo
d = {'TCP' : '1'}
df['protocolo'] = df['protocolo'].map(d)
df.head()
```

*Arreglos y definición de Clases de datos*

```
In [272]: pd.unique(df['protocolo'])
Out[272]: array(['1'], dtype=object)

In [273]: pd.unique(df['ip_ini'])
Out[273]: array(['1', '2'], dtype=object)

In [274]: pd.unique(df['ip_fin'])
Out[274]: array(['1', '2', '3', '4', '5', '6', '7', '8', '9'], dtype=object)

In [275]: pd.unique(df['clase'])
Out[275]: array(['1', '2'], dtype=object)

In [276]: df[["protocolo","ip_ini","port_ini","ip_fin","port_fin","clase","prioridad"]] = df[["protocolo","ip_in
i","port_ini","ip_fin","port_fin","clase","prioridad"]].astype(int)

In [277]: df.describe()
```

*Proceso de entrenamiento de data .*

```
In [279]: from sklearn.model_selection import train_test_split
from sklearn.preprocessing import MinMaxScaler
from sklearn.neighbors import KNeighborsClassifier
from sklearn.metrics import classification_report
from sklearn.metrics import confusion_matrix
import numpy as np
from sklearn import preprocessing, neighbors
import pandas as pd

X = df[["protocolo", "ip_ini", "port_ini", "ip_fin", "port_fin"]].values
y = df['clase'].values
X_train, X_test, y_train, y_test = train_test_split(X, y, random_state=0)
scaler = MinMaxScaler()
X_train = scaler.fit_transform(X_train)
X_test = scaler.transform(X_test)

In [280]: df.hist()
```

### Umbral de detección de anomalías

```
In [282]: print(df.groupby('port_fin').size())

port_fin
53      28
80     3585
443     320
dtype: int64

In [283]: estimador = IsolationForest(n_estimators=100, #utiliza arboles... n_estimators es el numero de arboles
contamination=0.01, # umbral donde se detectan las anomalías
max_samples=256)#

estimador

Out[283]: IsolationForest(bootstrap=False, contamination=0.01, max_features=1.0,
max_samples=256, n_estimators=100, n_jobs=1, random_state=None,
verbose=0)
```

### Complementos de Matriz de confusión.

```
In [291]: n_neighbors = 7

knn = KNeighborsClassifier(n_neighbors)
knn.fit(X_train, y_train)
print('Accuracy of K-NN classifier on training set: {:.2f}'
      .format(knn.score(X_train, y_train)))
print('Accuracy of K-NN classifier on test set: {:.2f}'
      .format(knn.score(X_test, y_test)))

pred = knn.predict(X_test)
print(confusion_matrix(y_test, pred))
print(classification_report(y_test, pred))

Accuracy of K-NN classifier on training set: 1.00
Accuracy of K-NN classifier on test set: 1.00
[[977  0]
 [ 0  7]]

precision    recall  f1-score   support

1         1.00      1.00      1.00      977
2         1.00      1.00      1.00        7

avg / total         1.00      1.00      1.00     984
```



```

In [297]: from sklearn import metrics
# try K=1 through K=25 and record testing accuracy
k_range = range(1, 125)

# We can create Python dictionary using [] or dict()
scores = []

# We use a loop through the range 1 to 26
# We append the scores in the dictionary
for k in k_range:
    knn = KNeighborsClassifier(n_neighbors=k)
    knn.fit(X_train, y_train)
    y_pred = knn.predict(X_test)
    scores.append(metrics.accuracy_score(y_test, y_pred))

print(scores)
# import Matplotlib (scientific plotting library)
import matplotlib.pyplot as plt

# allow plots to appear within the notebook
%matplotlib inline

# plot the relationship between K and testing accuracy
#plt.plot(x, y)
plt.plot(k_range, scores)
plt.xlabel('Evaluacion de Normales:')
plt.ylabel('falsos Positivos')

```

### *Evaluación de falsos positivos*

## Referencia bibliográfica

- Anaconda Inc. (2020). conda-forge / packages / jupyterlab 2.2.8. Obtenido de <https://anaconda.org/conda-forge/jupyterlab>
- Acosta , N. (2017). Diseño e implementación de una Red Lan para la Empresa Palinda. Quito: USFQ. Obtenido de: <http://repositorio.usfq.edu.ec/bitstream/23000/6383/1/130874.pdf>
- Alarcon , C. (2015). Optimización del clasificador “naive bayes” usando árbol de decisión c4.5. Lima: Universidad Mayor de San Marcos. Obtenido de: <https://core.ac.uk/download/pdf/323343447.pdf>
- Arias, J. (2016). Big Data: la nueva herramienta pra el transporte. Obtenido de: <https://www.crhoy.com/tecnologia/big-data-la-nueva-herramienta-para-el-sector-transporte/>
- PowerData. (2020). Big Data: ¿En qué consiste? Su importancia, desafíos y gobernabilidad. Obtenido de IBM DeveloperWorks:. Obtenido de: <https://www.powerdata.es/big-data>
- Barrero G. (2018). Evaluación de la eficiencia de los modelos machine learning para la predicción de la calidad del software desarrollado en ibm rpg usando la matriz de confusión y las curvas roc. Obtenido de: [https://www.researchgate.net/publication/328600321\\_PREDICCION\\_DE\\_LA\\_CALIDAD\\_DE\\_SOFTWARE\\_DESARROLLADO\\_EN\\_IBM\\_RPG\\_USANDO\\_DEEP\\_LEARNING](https://www.researchgate.net/publication/328600321_PREDICCION_DE_LA_CALIDAD_DE_SOFTWARE_DESARROLLADO_EN_IBM_RPG_USANDO_DEEP_LEARNING)
- Beltran, G. (2016). *Diferencias entre geolocalizar, GPS y localizar*. Obtenido de <https://gersonbeltran.com/2012/06/14/diferencias-entre-geolocalizar-gps-y-localizar/>
- Bhattacharjee, P., & Md Fujail, A. K. (2017). Intrusion Detection System for NSL-KDD Data Set using Vectorised Fitness Function in Genetic . Cachar, Assam. Obtenido de [https://www.ripublication.com/acst17/acstv10n2\\_08.pdf](https://www.ripublication.com/acst17/acstv10n2_08.pdf)
- Bookdown. (2020). Métodos de clasificación. Obtenido de <https://bookdown.org/content/2274/metodos-de-clasificacion.html>
- Cisco. (2017). Analyzing the Cisco Enterprise Campus Architecture. Obtenido de <http://www.ciscopress.com/articles/article.asp?p=1608131&seqNum=3>

- Cisco. (2019). Cisco Annual Internet Report. Recuperado de <https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white-paper-c11-741490.html>
- Fabrega, R., Fabrega, J., y Blair, A. (2016). *La enseñanza de Lenguajes de Programación en la Escuela: ¿Por qué hay que prestarle atención?*. Obtenido de: <https://www.fundaciontelefonica.cl/wp-content/uploads/2016/12/LA-ENSEN%CC%83ANZA-DEL-LENGUAJE-DE-PROGRAMACION-ultima-version.compressed.pdf>
- Google. (Febrero de 2016). Todo lo necesario para realizar analíticas de datos predictivas. Obtenido de: <https://cloud.google.com/solutions/smart-analytics?hl=es>
- Hernández, R. (2017). Metodología de la Investigación. 6ta Edición. Obtenido de: <https://www.uca.ac.cr/wp-content/uploads/2017/10/Investigacion.pdf>
- IBM. (2016). What is cloud computing? Obtenido de IBM Cloud: <https://www.ibm.com/cloud-computing/es-es/learn-more/what-is-cloud-computing/>
- Netgate. (2020). Requerimientos para instalar pfSense. Obtenido de: <https://forum.netgate.com/topic/55419/requerimientos-para-instalar-pfsense>
- Martín, C. (2016). *BIG DATA Y HADOOP*. Obtenido de: <http://www.ctic.uni.edu.pe/index.php/tallerhadoop>
- Martinez, C (2017). Investigación Descriptiva: Tipos y Características. Obtenido de:
- Mendoza S, (2020) ¿Qué es el aprendizaje automático? Obtenido de: <https://planetachatbot.com/que-es-aprendizaje-automatico-f12afd3871cb>
- Mogollon, F. S. (2017). Desafíos de la ciberseguridad y respuestas estatales: El caso del estado Ecuatoriano en el periodo 2008-2015. Quito: Pontificia Universidad Católica del Ecuador. obtenido de: <http://repositorio.puce.edu.ec/handle/22000/14104>
- Morales, E. (2018). Integración de un IDS/IPS al controlador SDN para la prevención y detección de ataques de seguridad (DoS) en un escenario de redes definidas por software. Riobamba: Escuela Superior Politécnica de Chimborazo. obtenido de: <http://dspace.esPOCH.edu.ec/handle/123456789/10931>
- Ocampo C, (2017) Sistema de detección de intrusos en redes corporativas. Obtenido de: <https://www.redalyc.org/pdf/849/84953102008.pdf>

- Ramirez, P. (2019). Cuadro comparativo de los diferentes lenguajes de programación. Obtenido de: [https://www.academia.edu/34836420/CUADRO\\_COMPARATIVO\\_DE\\_DIFERENTES LENGUAJES\\_DE\\_PROGRAMACION](https://www.academia.edu/34836420/CUADRO_COMPARATIVO_DE_DIFERENTES LENGUAJES_DE_PROGRAMACION)
- Rodriguez, C. (2016). Diseño de un sistema de detección de intrusos con snort para la compañía Silverit SAS. Bogotá: Universidad Piloto de Colombia. Obtenido de: <http://repository.unipiloto.edu.co/handle/20.500.12277/2706>
- Rodriguez, D., & Valldeoriola, J. (2016). *Técnicas de investigación social y educativa*, Obtenido de: [https://books.google.com.ec/books?hl=en&lr=&id=ZT\\_qDQAAQBAJ&oi=fnd&pg=PT8&dq=info:aKu-kJwX8moJ:scholar.google.com&ots=\\_jQAOOBj9Z&sig=FVc8HtQw3sXCMCZO pXhJu\\_JDI5M&redir\\_esc=y#v=onepage&q&f=false](https://books.google.com.ec/books?hl=en&lr=&id=ZT_qDQAAQBAJ&oi=fnd&pg=PT8&dq=info:aKu-kJwX8moJ:scholar.google.com&ots=_jQAOOBj9Z&sig=FVc8HtQw3sXCMCZO pXhJu_JDI5M&redir_esc=y#v=onepage&q&f=false)
- Seguridad, D. d. (2016). *Cuaderno de Estrategia 185, ciberseguridad: la cooperación público-privada*. Madrid: Ministerio de Defensa. Obtenido de: [https://www.uah.es/export/sites/uah/es/investigacion/.galleries/Archivos-de-proyectos-de-Grupos-de-Investigacion/GI94-PR258887-Capacitacion\\_profesional\\_y\\_formacion\\_especializada\\_en\\_ciberseguridad.pdf](https://www.uah.es/export/sites/uah/es/investigacion/.galleries/Archivos-de-proyectos-de-Grupos-de-Investigacion/GI94-PR258887-Capacitacion_profesional_y_formacion_especializada_en_ciberseguridad.pdf)
- Sinnexus. (2016). *Datamining*. Sinergia e Inteligencia de Negocio S.L Obtenido de.: [http://www.sinnexus.com/business\\_intelligence/datamining.aspx](http://www.sinnexus.com/business_intelligence/datamining.aspx)
- Snort, (2020). ¿Nuevo en Snort?. Obtenido de: <https://www.snort.org/>
- Suricata. (2020). *Suricata*. Obtenido de <https://suricata-ids.org>.