Clasificación del tráfico de red mediante técnicas de aprendizaje automático en Redes Definidas por Software

X. Ordóñez, L. Pisco

Universidad de las Fuerzas Armadas ESPE
Tutor: Ing. Daniel Nuñez, Mgtr.
4 de septiembre de 2023

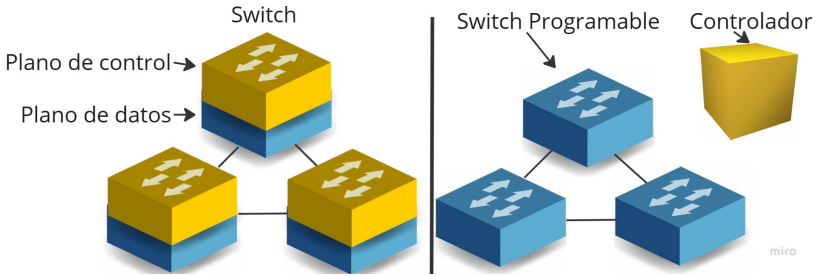
Trabajo de Integración Curricular, previo a la obtención de título de Ingeniero/a en Tecnologías de la Información

Contenido

Introducción/Antecedentes

- Interconexión e intercambio de información.

- Redes Definidas por Software.



- Clasificación Basada en Puertos, DPI y Estadística.

Introducción/Planteamiento del Problema

- **Creciente variedad de aplicaciones y servicios en línea**, dificultad de clasificar el tráfico de manera confiable y eficiente.
- Aún existen desafíos por abordar, como la **clasificación de nuevas clases de tráfico** y la optimización de enfoques de clasificación.
- El problema se centra en la clasificación precisa y eficiente del **tráfico de red en entornos SDN** mediante el uso de **técnicas de ML**.

Introducción/Justificación

- La combinación de **SDN e IA**.
- Modelo de **clasificación de tráfico de red** utilizando **algoritmos de ML**.
- Análisis para evaluar su precisión en la **clasificación de tráfico en una SDN**.

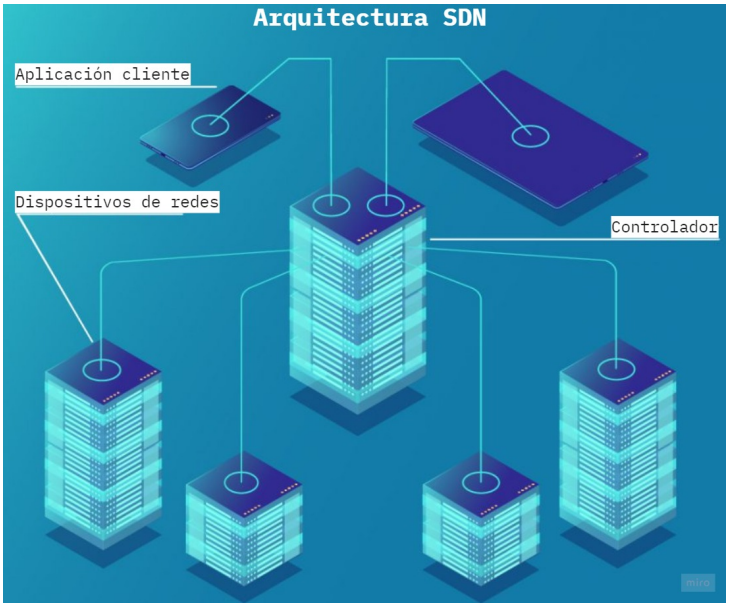
Introducción/Objectivos

- Determinar los enfoques de clasificación de tráfico con técnicas de aprendizaje automático en SDN.
- Realizar la selección de métodos de clasificación, para detectar el tráfico de aplicaciones en un entorno fuera de línea.
- Evaluar los clasificadores seleccionados en un entorno de banco de pruebas de red.

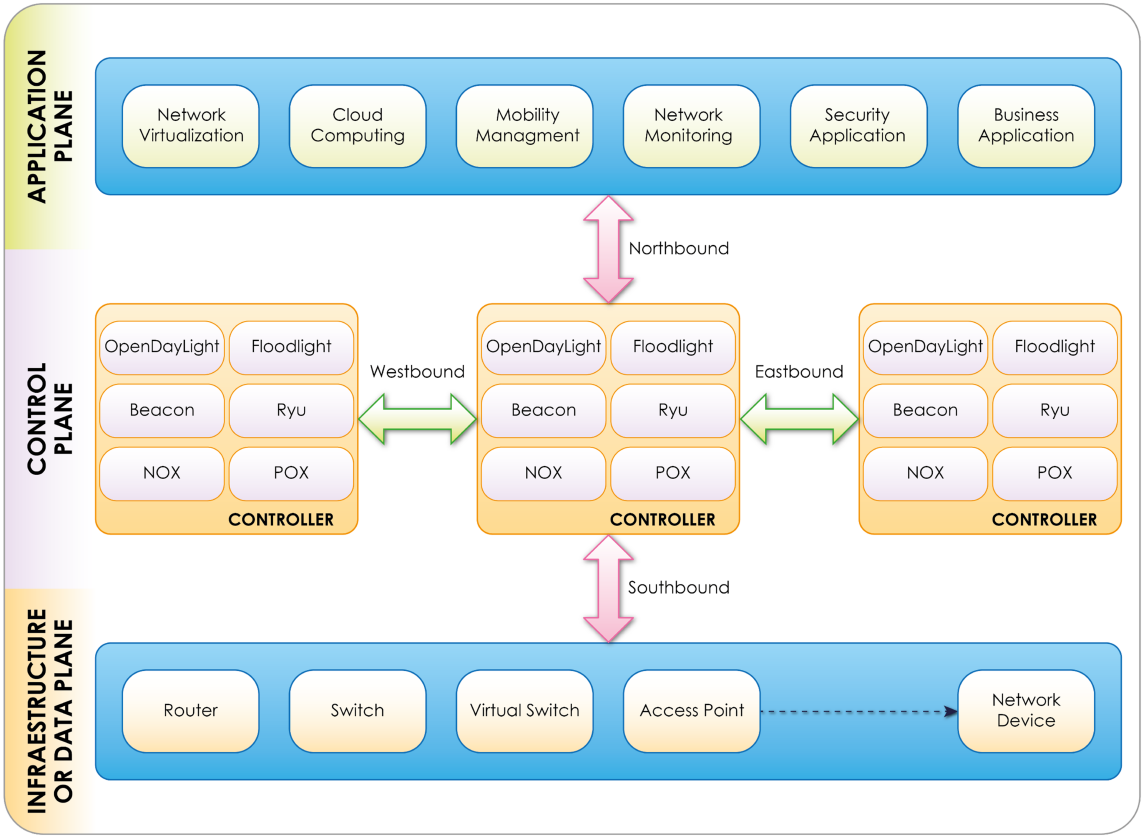
Introducción/Alcance

Modelo de clasificación de tráfico por aplicación.

- **Redes Definidas por Software:**
  - Enfocado en la **programabilidad y la automatización.**
  - Reemplaza redes tradicionales al **centralizar la configuración de toda la infraestructura de red**.



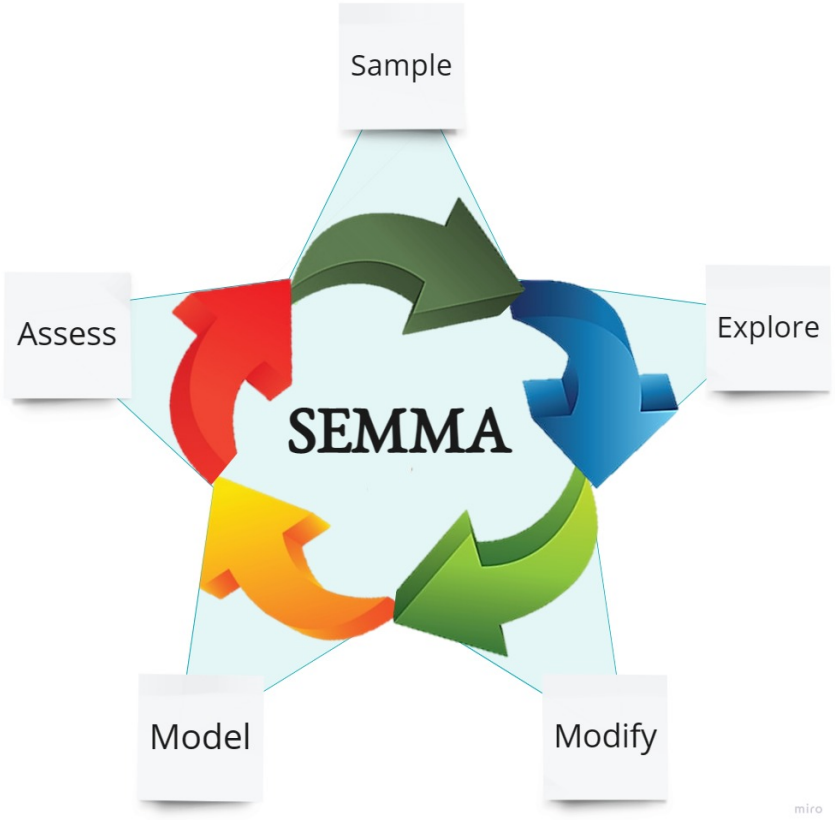
Arquitectura funcional de SDN.



- **Aprendizaje Automático:**
  - Rama de la **IA**, en él se centra el desarrollo de algoritmos y modelos, los tipos de aprendizajes automáticos: **Supervisado, No Supervisado y Por Refuerzo**.
  - Existen diferentes métodos de clasificación:

Metodología

Fases de la metodología SEMMA.



Metodología/ Fase de Muestreo *(Sample)*

- Selección del conjunto de datos.

Referencia	Nombre	Clases	Tamaño	Formato
Video Streaming Service Identification on Software-Defined Networking	Datavideo1 y Datavideo2 (2021)	3	7.8GB	PCAP
Characterization of Encrypted and VPN Traffic using Time-related Features	ISCX-VPN-NonVPN (2016)	14	28GB	PCAP
A novel dataset for encrypted virtual private network traffic analysis	Encrypted VPN conjunto de datos (2022)	5	18GB	JSON
Discriminators for use in flow-based classification	Moore (2005)	10	No publicado	ARFF

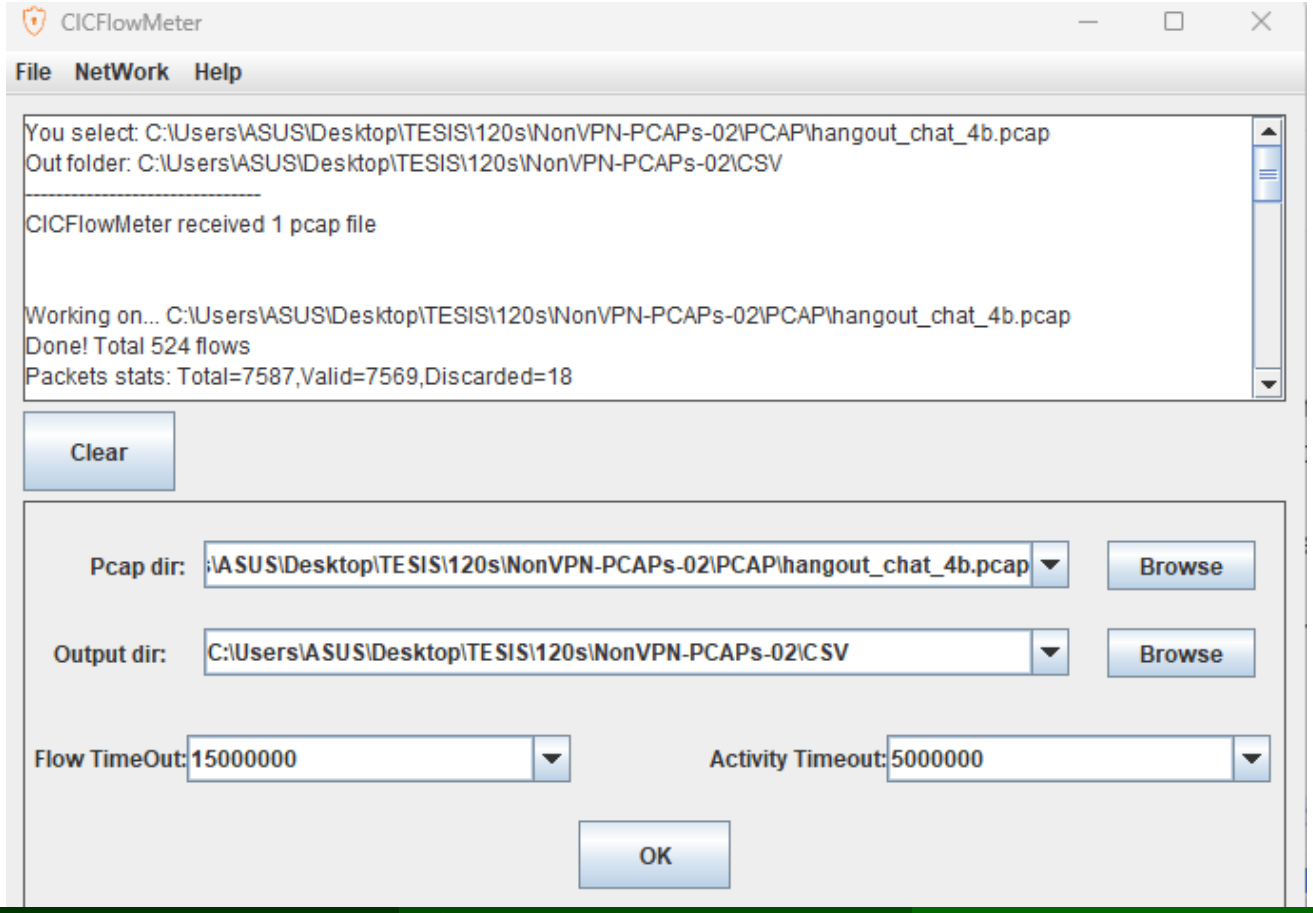
Metodología/ Fase de Muestreo *(Sample)*

- Protocolos y aplicaciones capturadas del conjunto de datos ICX-VPN-NonVPN.

Tráfico	Contenido
Navegación Web	Firefox y Chrome
Correo Elétrónico	SMPTS, POP3S e IMAPS
Chat	ICQ, AIM, Skype, Facebook y Hangouts
Streaming	Vimeo y Youtube
Transferencia de Archivos	Skype, FTPS y SFTP mediante Filezilla y un servicio externo
VoIP	Llamadas de voz de Facebook, Skype y Hangouts (1h de duración)
P2P	Torrent y Transmisión (Bittorrent)

Metodología/ Fase de Muestreo *(Sample)*

- Extracción de archivos PCAP.



Metodología/ Fase de Exploración *(Explore)*

- Granularidad en las etiquetas.

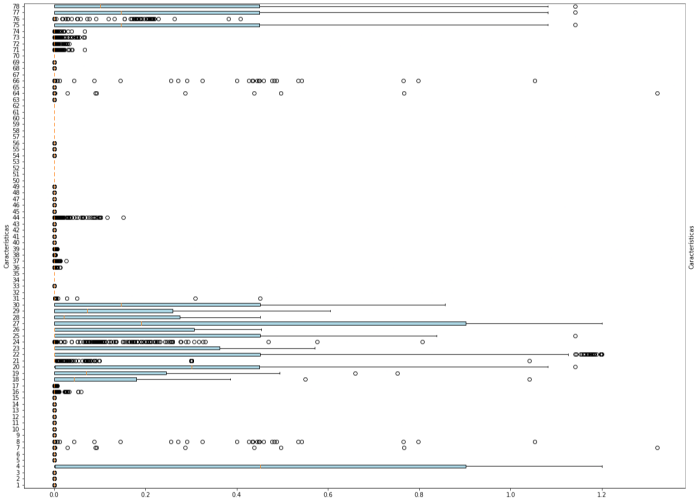
Aplicación	Tipo	Instancias en 120s	Instancias en 15s
AIM Chat	Texto en tiempo real	596	1033
Email Client	Texto	5640	6976
Facebook Audio	Voz sobre IP	54191	62544
Facebook Chat	Texto en tiempo real	1885	2285
Facebook Video	Video en tiempo real	633	1226
FTPS	Archivo de transferencia	1501	2045
Gmail Chat	Texto en tiempo real	615	1053
Hangouts Audio	Voz sobre IP	61899	72397
Hangouts Chat	Texto en tiempo real	3509	4033
Hangouts Video	Video en tiempo real	1921	2882
ICQ	Texto en tiempo real	601	1131
Netflix	Video	1059	2084
SCP	Archivo de tranferencia	8046	9546
SFTP	Archivo de transferencia	303	449
Skype Audio	Voz sobre IP	24037	27756
Skype Chat	Texto en tiempo real	4883	6703
Skype File	Archivo de transferecia	36792	42640
Skype Video	Video en tiempo real	1062	1594
Spotify	Música	898	1498
BitTorrent	P2P	709	777
Vimeo	Video	1038	1752
VoIP Buster	Voz sobre IP	4408	6217
YouTube	Video	1845	2962

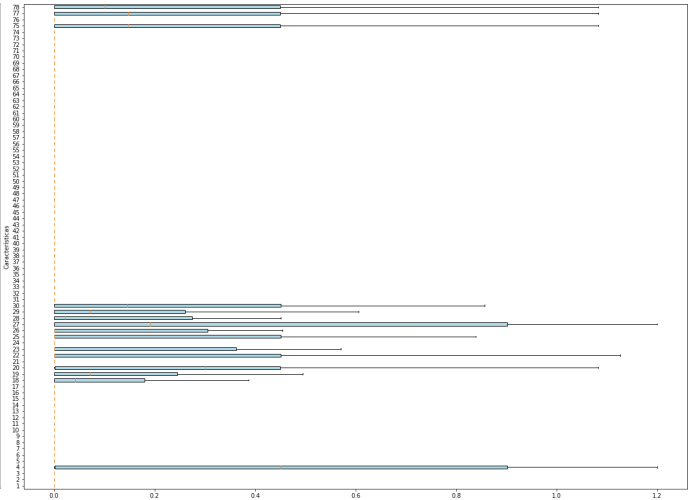
Metodología/ Fase de Modificación *(Modify)*

- **Limpieza de datos**
  - Datos Nulos
  - Datos duplicados
  - Datos erróneos
  - Datos Infinitos
  - Eliminación

Metodología/ Fase de Modificación *(Modify)*

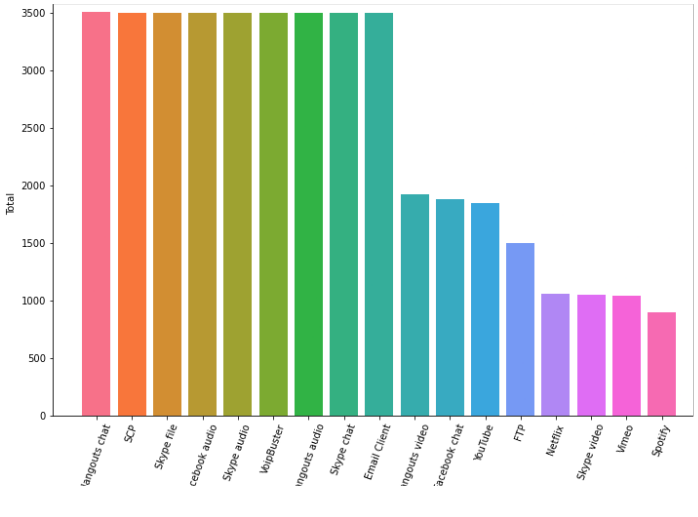
- **Valores Atipicos**
  - Técnica de imputación media
  - EL proceso se aplicó para cada uno de los archivos correspondientes a las aplicaciones.
  - Control de valores atípicos es más específico y adaptado a las características de cada clase.

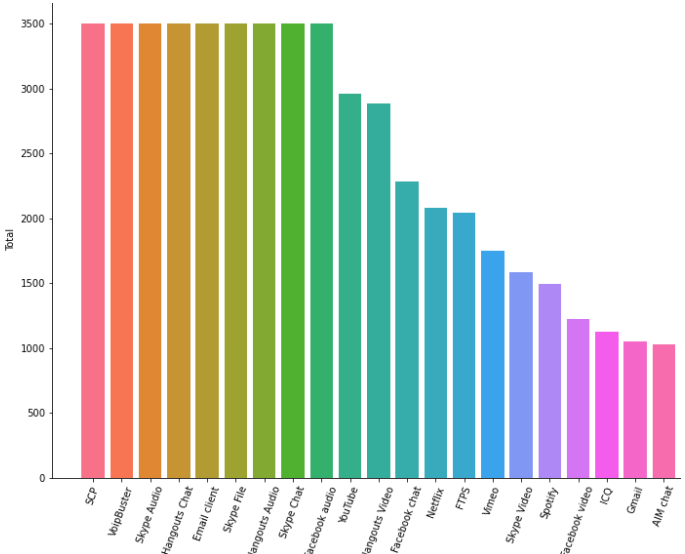




Metodología/ Fase de Modificación *(Modify)*

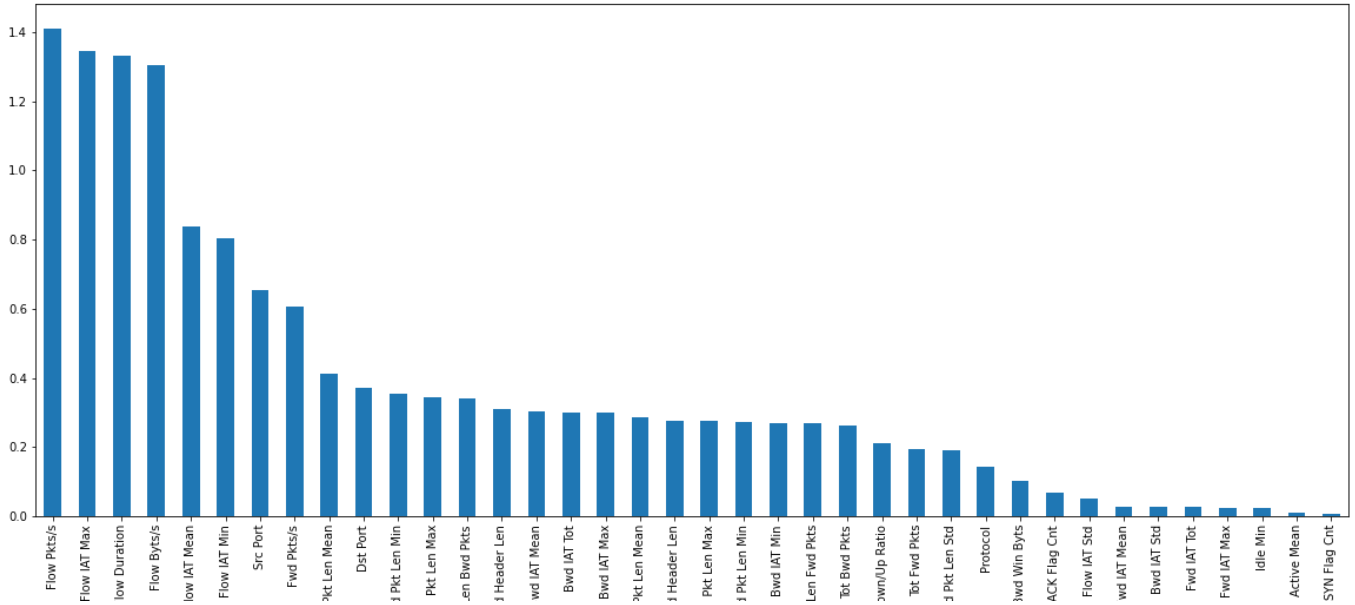
- **Desviación Estandar:**
  - Evaluar la dispersión o variabilidad de los valores en relación con la media.
  - Reducir la dimensionalidad y complejidad del conjunto de datos.
- **Balanceo de Datos:**
  - Abordar el problema de desequilibrio en un conjunto de datos.
  - Lograr el equilibrio en la cantidad de ejemplos en casi todas las clases.





Metodología/ Fase de Modificación *(Modify)*

- **Ingenieria de Caracteríticas:**
  - Identificar el número de características que influyen en los resultados de precisión.



Metodología/ Fase de Modelado *(Model)*

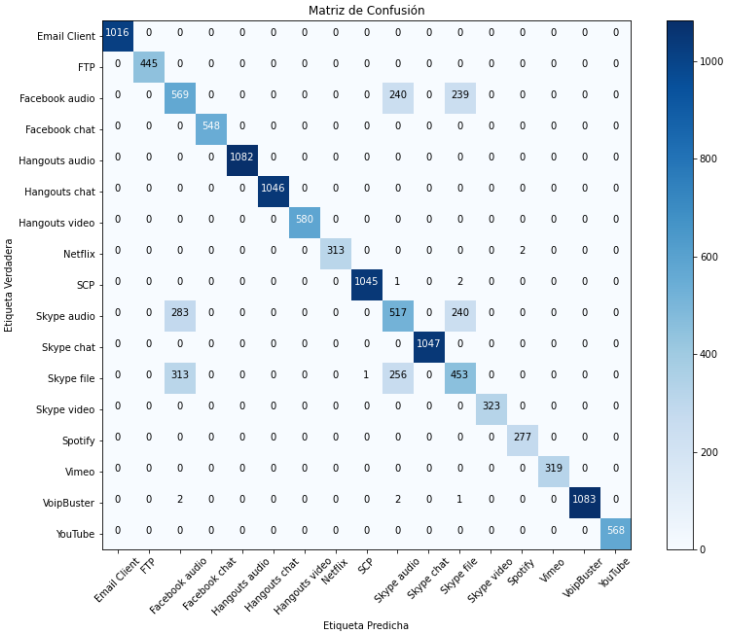
- Investigación basada en artículos relevantes que aplican algoritmos ML.

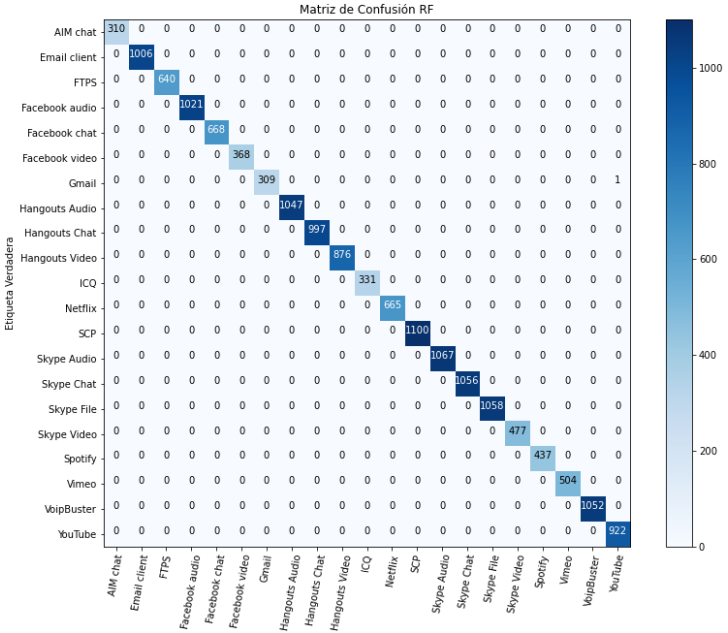
Referencia	Técnica ML	Características	Etiquetas	Presición
A Deep Learning Based System for Traffic Engineering in Software Defined Networks	k-NN, SVM, DT, RF, DNN y CNN	13	3	kNN:98% RF:97% CNN:95% DNN:94% DT:88% SVM:86%
Investigation of machine learning based network traffic classification	SVM y K-Means	30	8	SVM:98.7% K-M:88%
FWFS: Selecting robust features towards reliable and stable traffic classifier in SDN	DT, k-NN, NB y SVM	Seleccionadas por el algoritmo	12	DT:99.39% kNN:98.34% NB:96.71% SVM:94.65%
Deep neural networks for application awareness in SDN-based network	DNN, SVM, k-NN y DT	9	4	DNN:88% DT:85% SVM:80% kNN:79%
An intelligent traffic classification in sdn-iot: A machine learning approach	k-NN, RF y DT	6	4	DT:87.2% RF: 85.1% kNN: 79.5%
A dynamic network traffic classifier using supervised ML for a Docker-based SDN network	k-NN, RF y DT	7	10	kNN:97.14% RF:96.69% DT:95.8%
Fine-grained Traffic Classification Based on Improved Residual Convolutional Network in Software Defined Networks	ResNet Mejorado	Seleccionadas por el algoritmo	16	ResNet:99.93%

Metodología/ Fase de Evaluación *(Assess)*

Se evalúan los modelos construidos para medir su calidad y el rendimiento, considerando los dos grupos de características previamente establecidos.

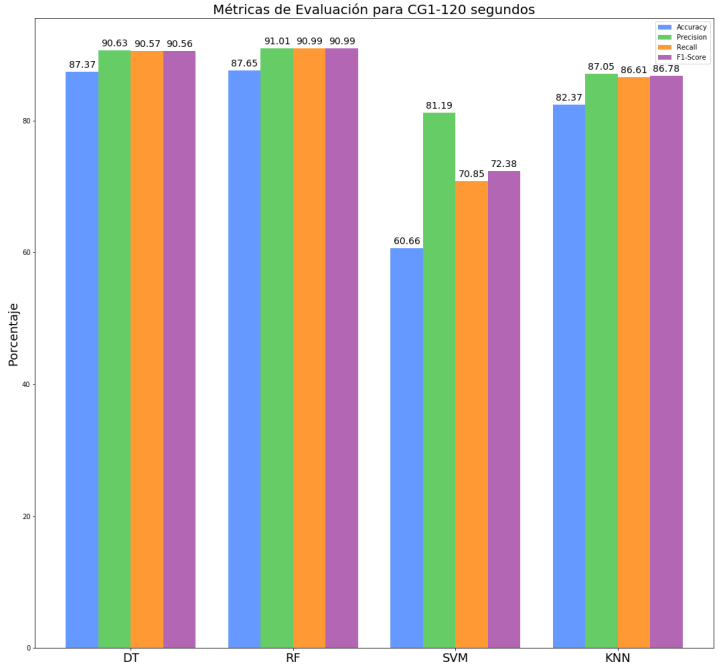
- **Matriz de confusión:** Se evalua el rendimiento de un modelo en función de las predicciones realizadas.

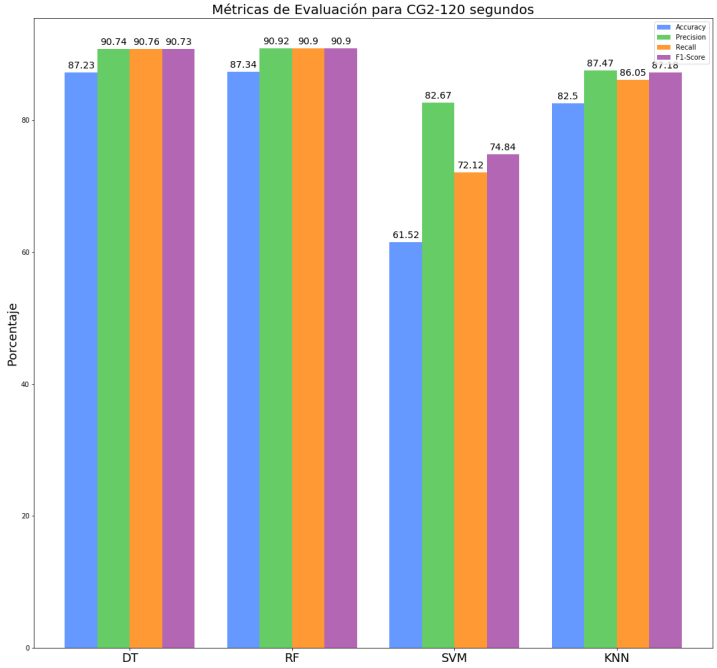




Metodología/ Fase de Evaluación *(Assess)*

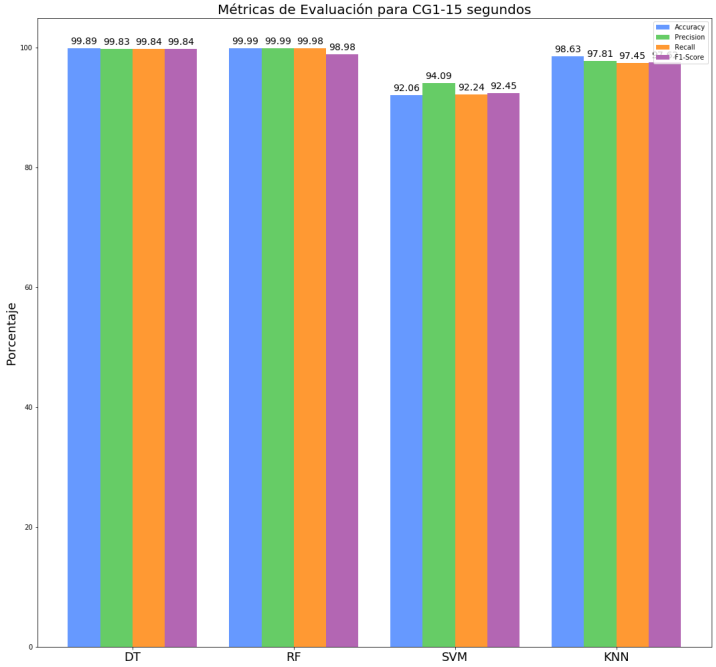
- **Métricas de Evaluación:** Se mide el desempeño y se evalua la capacidad predictiva.
  Métricas de evaluación para el conjunto de datos con tiempo de espera en flujo de 120 segundos.

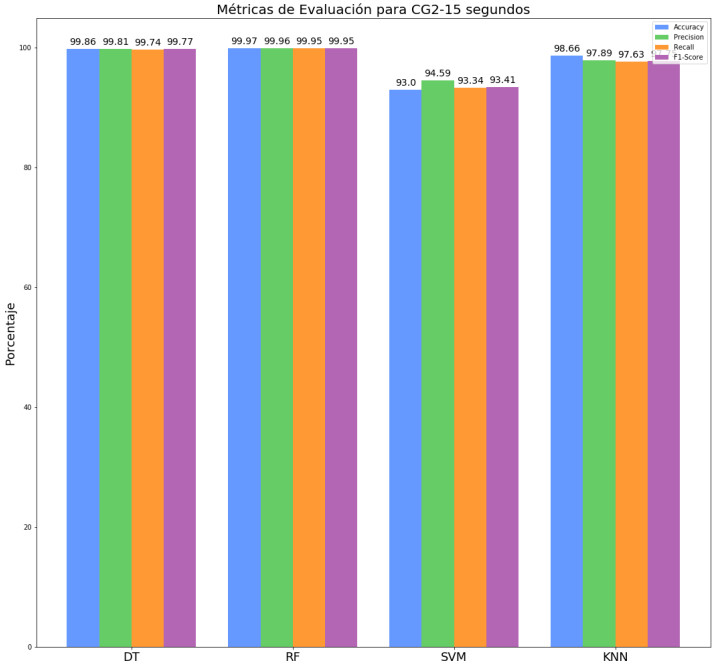




Metodología/ Fase de Evaluación *(Assess)*

- Métricas de Evaluación para el conjunto de datos con tiempo de espera en flujo de 15 segundos.





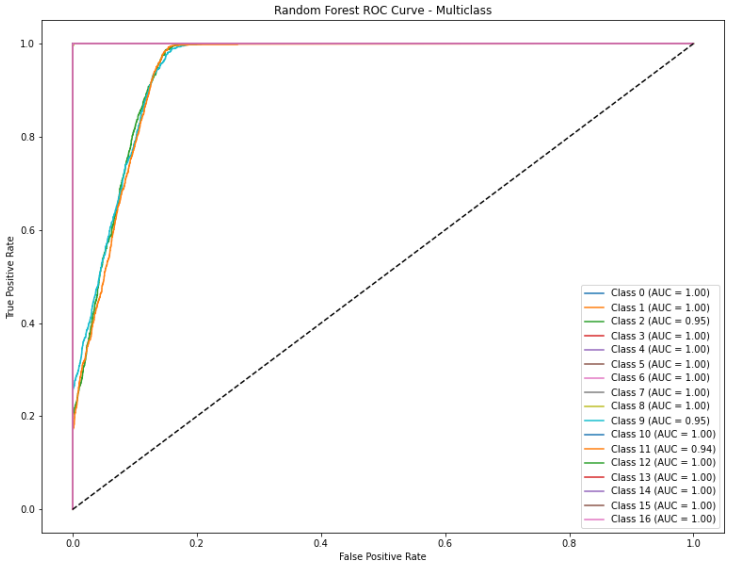
Metodología/ Fase de Evaluación *(Assess)*

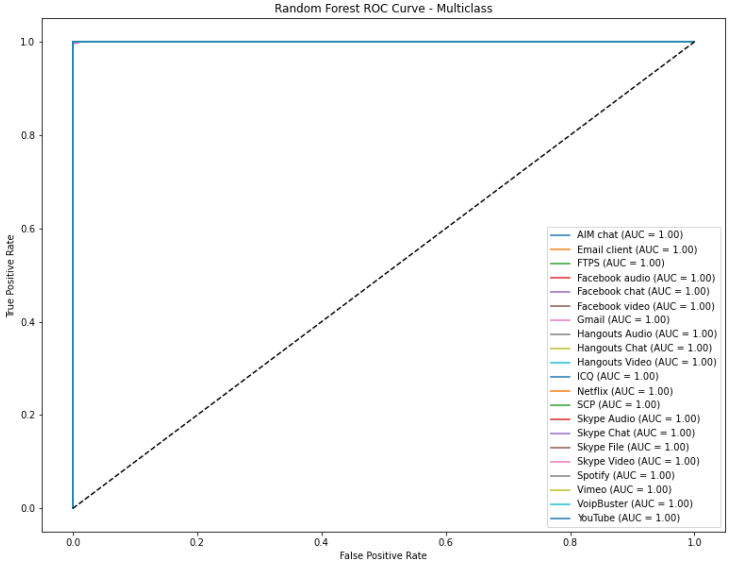
- **Validación cruzada:** Utilizada para evaluar el rendimiento de un modelo de aprendizaje automático de manera más robusta y confiable.

Grupo	Modelo	Accuracy	Presicion	Recall	F1-Score
CG1 - 120 segundos	DT	87.89%	90.98%	90.97%	90.96%
	RF	87.89%	91.33%	91.31%	91.31%
	SVM	55%	81%	45%x	49%
	KNN	83.8%	88.21%	87.88%	88.01%
CG2 - 120 segundos	DT	87%	91%	91%	91%
	RF	88%	91%	91%	91%
	SVM	56%	55%	54%	54%
	KNN	82%	85%	84%	84%
CG1 - 15 segundos	DT	99.88%	99.82%	99.82%	99.82%
	RF	**99.99%**	**99.98%**	**99.98%**	**99.98%**
	SVM	92.11%	9372%	92.41%	92.44%
	KNN	99.93%	99.91%	99.89%	99.90%
CG2 - 15 segundos	DT	99.91%	99.86%	99.84%	99.85%
	RF	99.98%	99.97%	99.97%	99.97%
	SVM	91.92%	93.01%	92.47%	91.92%
	KNN	98.74%	98.01%	97.68%	97.83%

Metodología/ Fase de Evaluación *(Assess)*

- **Curva ROC:** Se evalúa y compara la capacidad de discriminación de los modelos.
  Curva Roc para el modelo RF.





Implementación

- Modelos de ML pre-entrenados con la mayor precisión: **RF** y **DT**.
- Implementación en el controlador RYU.
- En entornos reales los datos tienden a ser desnormalizados.

Algoritmo 1 Clasificación de flujos

```
procedure CLASIFICACIÓN EN EL CONTROLADOR
Require: Cargar modelo ML
    procedure GESTIÓN PARA CARGAR EL MODELO PRE-ENTRENADO
Require: Metodo para predecir el trafico a partir de un archivo pcap
        LeerPCAP ← PCAP
        for packet in LeerPCAP : do
            Calcular_Caracteristicas ← LeerPCAP
            Agregar_a_Lista ← Calcular_Caracteristicas
        end for
        Guardar_Dataframe ← Agregar_a_Lista
        Normalizar_Datos ← Dataframe
        Predecir ← Normalizar_Datos
Require: Decorador para manejar eventos de paquetes entrantes
        Cargar ← PCAP
        Predicciones ← MetodoPredecir
        for i, prediction in enumerate(Predicciones) do
            predicted_class ← prediction
            if predicted_class in class_counts then
                class_counts[predicted_class]+ = 1
            end if
        end for
        for class_name, count in class_counts.items do
            Imprimirconteoclases
        end for
        Imprimir Total de Ejemplos, Predicciones Correctas y Exactitud del Modelo
    end procedure
end procedure
```

Implementación

Clasificación del tráfico de tipo AIM chat, modelo DT y RF.

```
Package 1236 - Prediction: AIM chat (Class 0)
Package 1237 - Prediction: Facebook chat (Class 4)
Package 1238 - Prediction: Facebook chat (Class 4)
Package 1239 - Prediction: AIM chat (Class 0)
Package 1240 - Prediction: Facebook chat (Class 4)
Package 1241 - Prediction: AIM chat (Class 0)
Package 1242 - Prediction: Facebook chat (Class 4)
Package 1243 - Prediction: AIM chat (Class 0)
Predicted class count:
AIM chat: 495 times
Facebook chat: 453 times
Hangouts Chat: 295 times
Total number of samples: 1243
Correct predictions: 495
Model accuracy: 39.82%
```

```
Package 1237 - Prediction: Facebook video (Class 5)
Package 1238 - Prediction: Facebook video (Class 5)
Package 1239 - Prediction: Facebook video (Class 5)
Package 1240 - Prediction: Facebook video (Class 5)
Package 1241 - Prediction: Facebook video (Class 5)
Package 1242 - Prediction: Facebook video (Class 5)
Package 1243 - Prediction: Facebook video (Class 5)
Predicted class count:
Facebook video: 888 times
Vimeo: 25 times
VoipBuster: 11 times
AIM chat: 35 times
Facebook chat: 284 times
Total number of samples: 1243
Correct predictions: 35
Model accuracy: 2.82%
```

Implementación

Clasificación del tráfico de tipo Facebook chat, modelo DT y RF.

```
Package 5522 - Prediction: Facebook chat (Class 4)
Package 5523 - Prediction: AIM chat (Class 0)
Package 5524 - Prediction: AIM chat (Class 0)
Package 5525 - Prediction: AIM chat (Class 0)
Package 5526 - Prediction: AIM chat (Class 0)
Package 5527 - Prediction: AIM chat (Class 0)
Predicted class count:
Facebook chat: 1801 times
AIM chat: 2134 times
Hangouts Chat: 1592 times
Total number of samples: 5527
Correct predictions: 1801
Model accuracy: 32.59%
```

```
Package 5522 - Prediction: Facebook video (Class 5)
Package 5523 - Prediction: Facebook video (Class 5)
Package 5524 - Prediction: Facebook video (Class 5)
Package 5525 - Prediction: Facebook video (Class 5)
Package 5526 - Prediction: Facebook video (Class 5)
Package 5527 - Prediction: Facebook video (Class 5)
Predicted class count:
Facebook video: 3457 times
AIM chat: 175 times
Vimeo: 287 times
Spotify: 17 times
VoipBuster: 86 times
Gmail: 36 times
Facebook chat: 1469 times
Total number of samples: 5527
Correct predictions: 1469
Model accuracy: 26.58%
```

Conclusiones

- Se determinó que los enfoques para la clasificación de tráfico en un entorno SDN, se basan en la implementación de los algoritmos de aprendizaje supervidos.
- En cada uno de los modelos, se realizó una clasificación en grano fino utilizando el conjunto de datos VPN-NonVPN con 53037 y 42710 instancias. Teniendo como resultado que, DT obtuvo un 99.89 %, RF alcanzó el 99.99 %, SVM logró un 92.06 % y KNN presentó un 98.63 %.
- En base a la evaluación de los modelos, se estableció que RF y DT son los mas óptimos para tener una mejor precisión al clasificar el tráfico en una red SDN.

Recomendaciones

- Valores Atípicos.
- El ajuste de hiperparámetros.
- Pruebas adicionales donde se utilicen otros conjuntos de datos.
- Proceso de evaluación de lo modelos en el controlador.

Trabajos Futuros

- Aplicar los cuatro modelos de ML desarrollados en un entorno real.
- Controladores OpenDaylight y ONOS.
- Técnicas de aprendizaje profundo.

Clasificación del tráfico de red mediante técnicas de aprendizaje automático en Redes Definidas por Software

X. Ordóñez, L. Pisco

Universidad de las Fuerzas Armadas ESPE
Tutor: Ing. Daniel Nuñez, Msg.
4 de septiembre de 2023

ESPE
UNIVERSIDAD DE LAS FUERZAS ARMADAS
INNOVACIÓN PARA LA EXCELENCIA

ITIN
INGENIERÍA EN TECNOLOGÍAS DE LA INFORMACIÓN

Gracias, alguna pregunta?