

Proyecto de grado

Secure Learning and Deep
Reinforcement Learning for
Android Malware detection.

Integrantes

David Huertas
Brayan vargas

Tutor

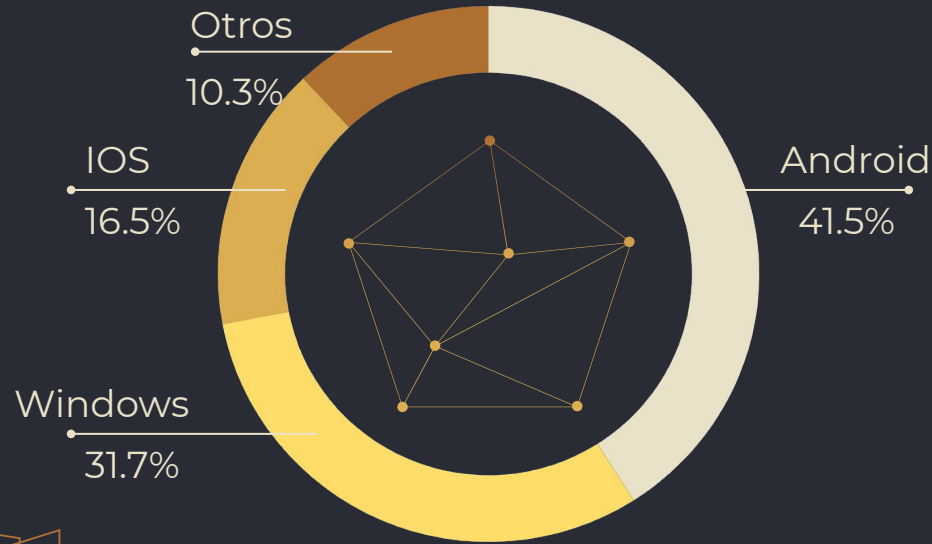
Christian Urcuqui, Msc.

TABLA DE CONTENIDOS

1. CONTEXTO
Antecedentes y justificación.
2. FORMULACIÓN DEL PROBLEMA
Síntesis del problema
3. OBJETIVOS
Principal y secundarios
4. MARCO TEORICO
Machine learning y ciberseguridad.

5. METODOLOGÍA Y RESULTADOS
CRISP-DM y Secure Learning
6. RECOMENDACIONES
7. CONCLUSIONES Y TRABAJO FUTURO
8. BIBLIOGRAFIA

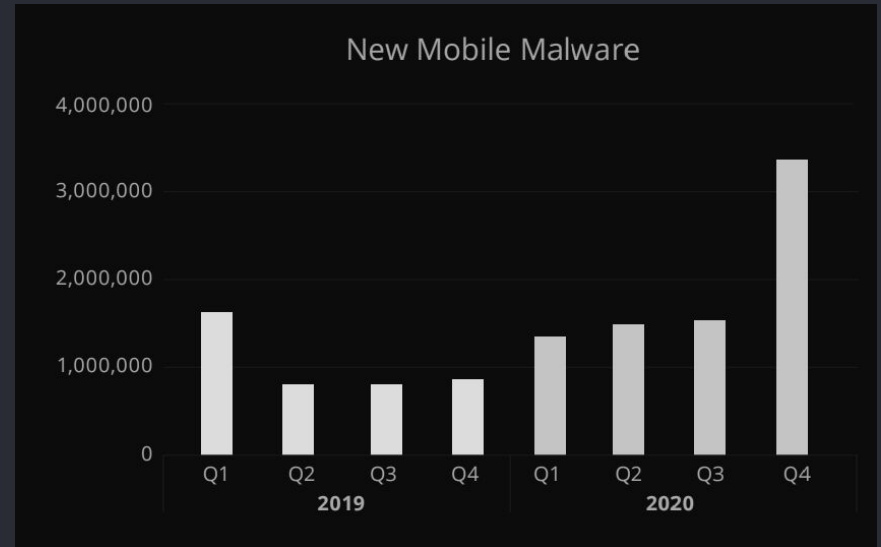
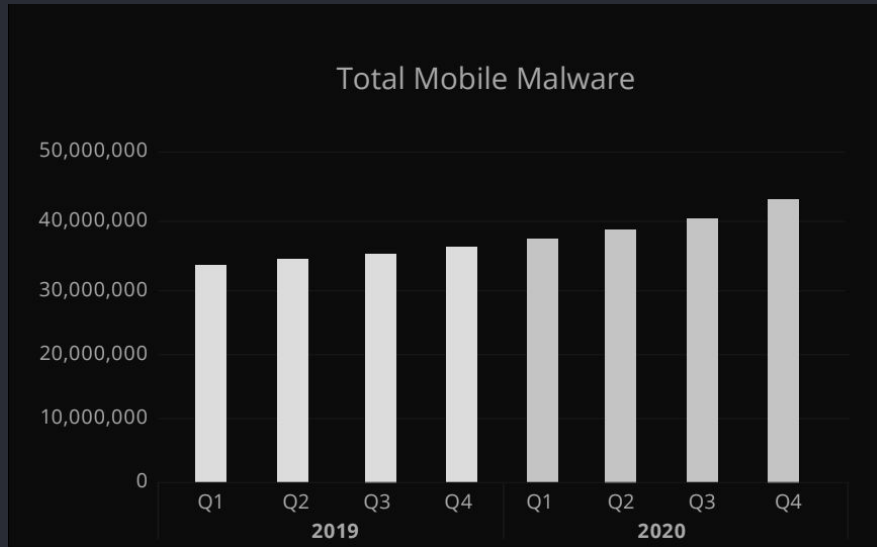
Contexto



Sistemas operativos
más usados durante el
2020.



Contexto



Antecedentes

Antecedentes

**Sesgo de los
modelos**

Secure learning

Vulnerabilidades

Antecedentes

**Conocimiento del
modelo y su entorno.**



Justificación

Contemplar escenarios de tipo gray-box para el estudio y análisis del modelo de clasificación de malware android propuesto en el proyecto antecesor.

Formulación del problema

Existen vulnerabilidades a ataques con enfoque adversario, en algoritmos de machine learning para la clasificación de malware Android donde se desconoce gran parte de la información del modelo.





OBJETIVOS

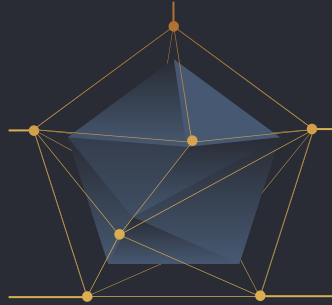


GENERAL

Reducir vulnerabilidades a ataques con enfoque adversario en algoritmos de machine learning para la clasificación de malware Android.

ESPECÍFICO 1

Propuesta de modelo adversario.



ESPECÍFICO 3

Marco de desarrollo seguro.

ESPECÍFICO 2

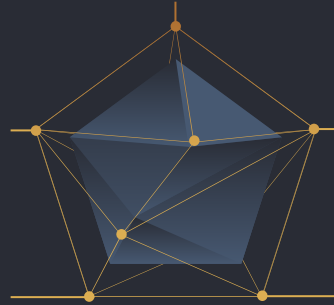
Evaluación un modelo de deep reinforcement learning.

GENERAL

Reducir vulnerabilidades a ataques con enfoque adversario en algoritmos de machine learning para la clasificación de malware android.

ESPECÍFICO 1

Proponer un modelo de ataque adversario con enfoque gray-box.



ESPECÍFICO 3

Marco de desarrollo seguro.

ESPECÍFICO 2

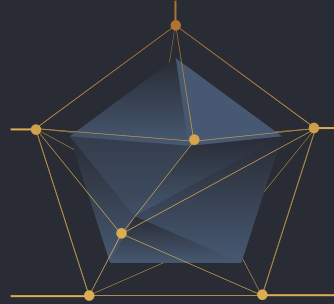
Evaluación un modelo de deep reinforcement learning.

GENERAL

Reducir vulnerabilidades a ataques con enfoque adversario en algoritmos de machine learning para la clasificación de malware android.

ESPECÍFICO 1

Propuesta de modelo adversario.



ESPECÍFICO 3

Marco de desarrollo seguro.

ESPECÍFICO 2

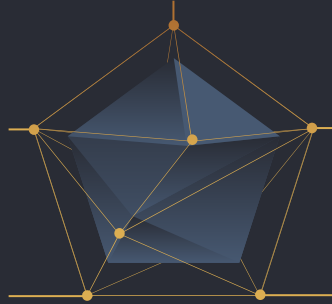
Evaluar un modelo de deep reinforcement learning para realizar perturbaciones más precisas en el proceso de aprendizaje del modelo.

GENERAL

Reducir vulnerabilidades a ataques con enfoque adversario en algoritmos de machine learning para la clasificación de malware android.

ESPECÍFICO 1

Propuesta de modelo adversario.



ESPECÍFICO 3

Proponer enfoques de desarrollo que cuenten con buenas medidas de seguridad.

ESPECÍFICO 2

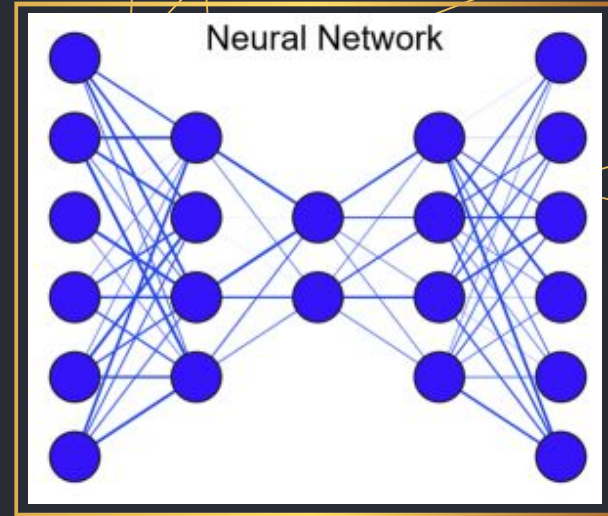
Evaluación un modelo de deep reinforcement learning.



MARCO TEÓRICO

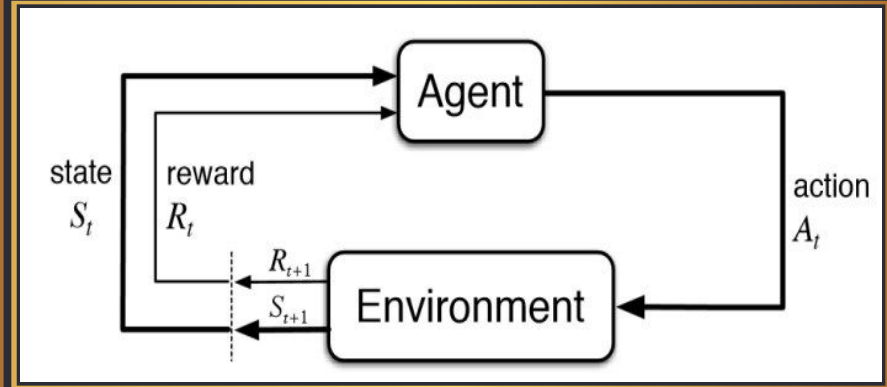
Machine learning

- **Deep learning**
- Reinforcement learning
- Deep reinforcement learning
- Generative adversarial network



Machine learning

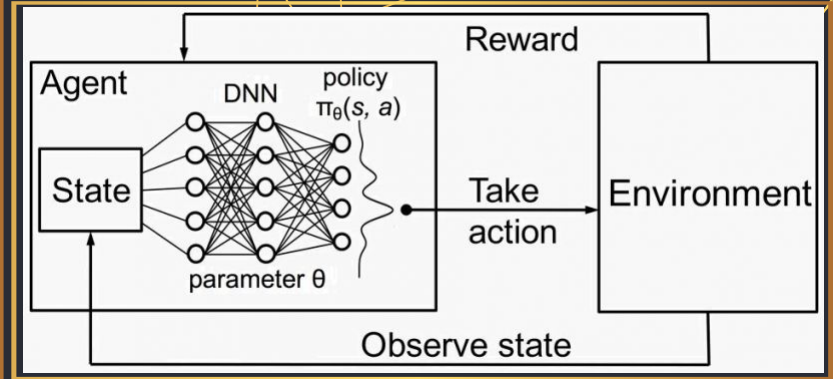
- Deep learning
- **Reinforcement learning**
- Deep reinforcement learning
- Generative adversarial network



Marco teorico

Machine learning

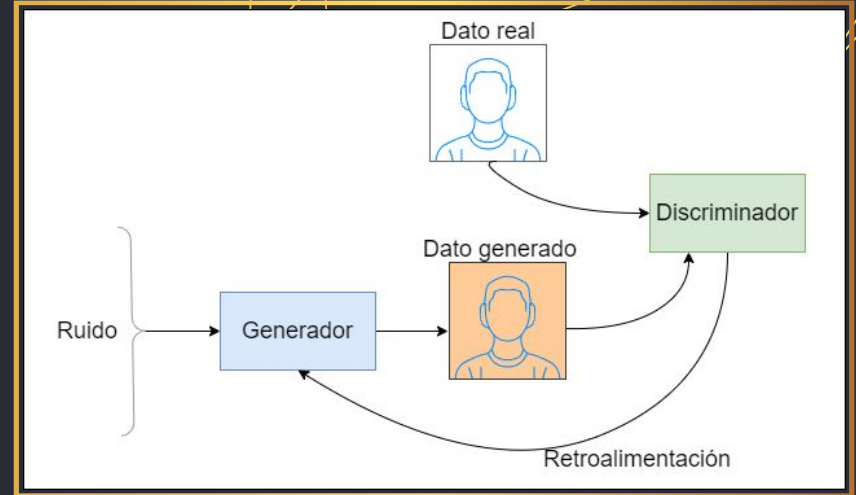
- Deep learning
- Reinforcement learning
- **Deep reinforcement learning**
- Generative adversarial network



Marco teorico

Machine learning

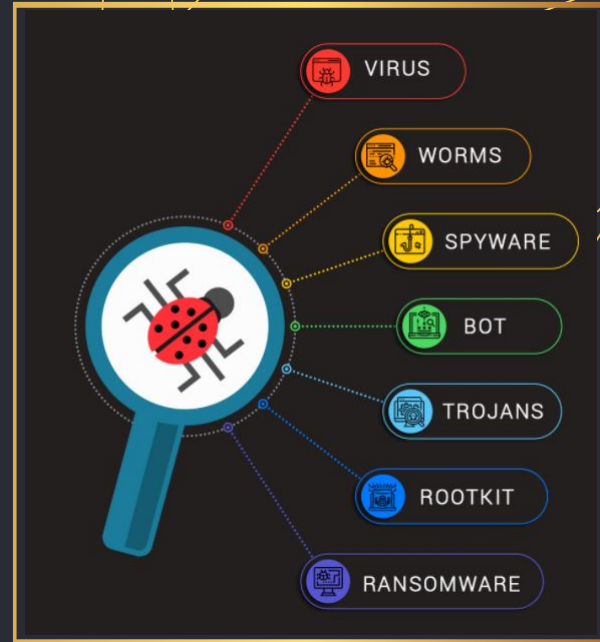
- Deep learning
- Reinforcement learning
- Deep reinforcement learning
- **Generative adversarial network**



Marco teorico

Ciberseguridad

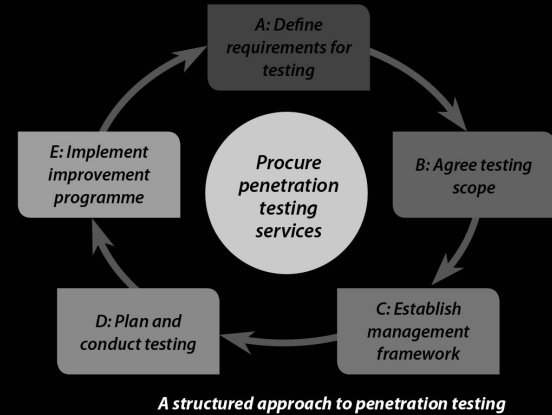
- **Malware**
- Pentesting
- Secure learning



Marco teorico

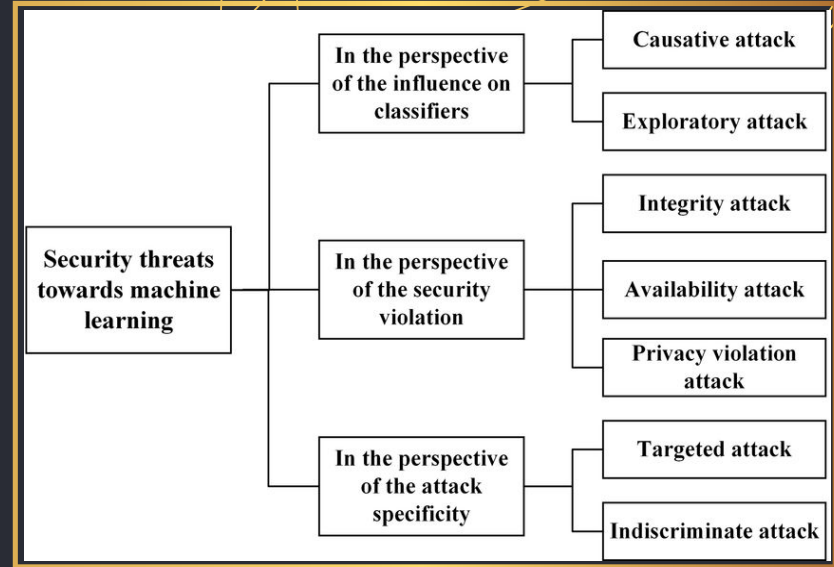
Ciberseguridad

- Malware
- **Pentesting**
- Secure learning



Ciberseguridad

- Malware
- Pentesting
- **Secure learning**



Estado del arte

Características / Papers	A	B	C	D	E	F	G
DRL		X	X				x
Gray-Box	X		X	x			x
Secure Learning	X					x	x
Malware Classifier	X		x	x	x	x	x

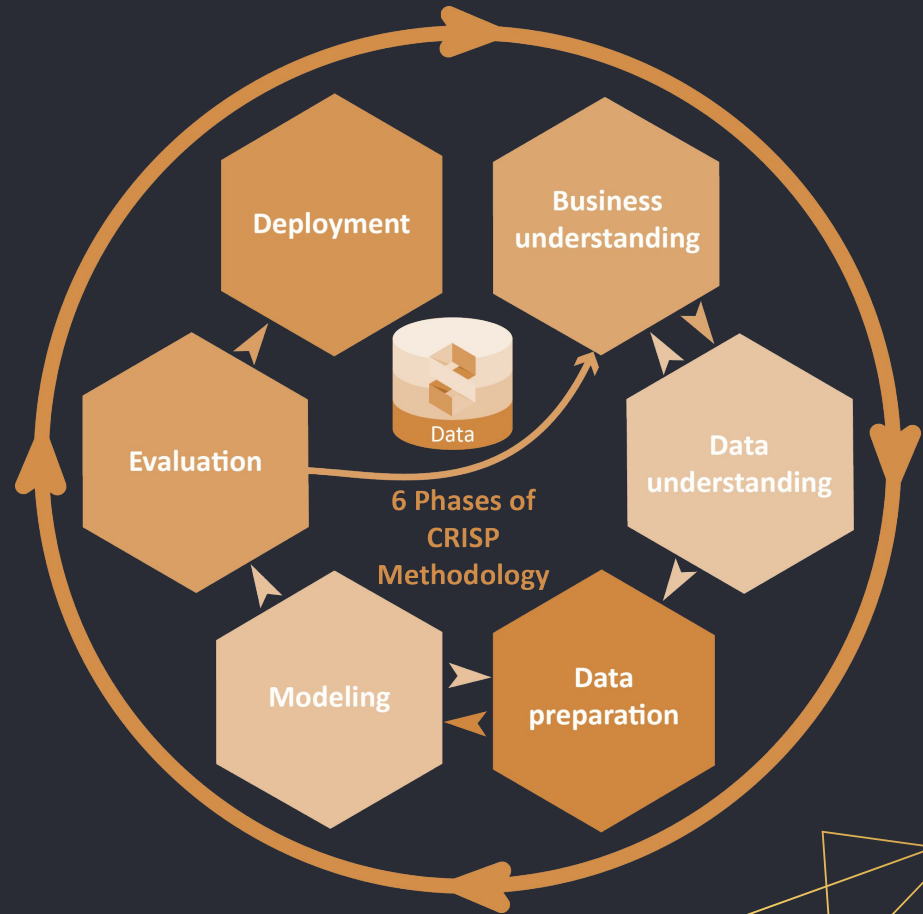
- A) Robust Android Malware Detection System Against Adversarial Attacks Using Q-Learning
- B) Deep Reinforcement Learning for Black-Box Testing of Android Apps
- C) Evading Anti-malware Engines with Deep Reinforcement Learning

- D) Adversarial-Example Attacks Toward Android Malware Detection System
- E) Black box analysis of android malware detectors
- F) Secure Learning para detección de Android Malware






METODOLOGÍA

CRISP-DM



Exploración y manejo de datos

			
DATASET	MALWARE	GOODWARE	Total
Features to detect android malware.	3141	4691	7832
Adversarial data 2019	500	0	500

C. Urcuqui, J. Delgado, A. Perez, A. Navarro, and J. Diaz, "Features to Detect Android Malware", 2018. IEEE Colombian Conference on Communications and Computing (COLCOM), 2018.

J. Delgado, "Secure Learning para detección de Android Malware," thesis, 2019.

Exploración y manejo de datos



INDICE	VARIABLES	INDICE	VARIABLES	INDICE	VARIABLES
1	Bytes enviados	4	Paquetes enviados	7	Consultas DNS
2	Bytes recibidos	5	Paquetes recibidos	8	Paquetes no TCP
3	Paquetes TCP	6	Volumen de bytes	9	Paquetes UDP
				10	IP externas

C. Urcuqui, J. Delgado, A. Perez, A. Navarro, and J. Diaz, "Features to Detect Android Malware", 2018. IEEE Colombian Conference on Communications and Computing (COLCOM), 2018.

J. Delgado, "Secure Learning para detección de Android Malware," thesis, 2019.

Modelo del clasificador

Reentrenamiento y actualización

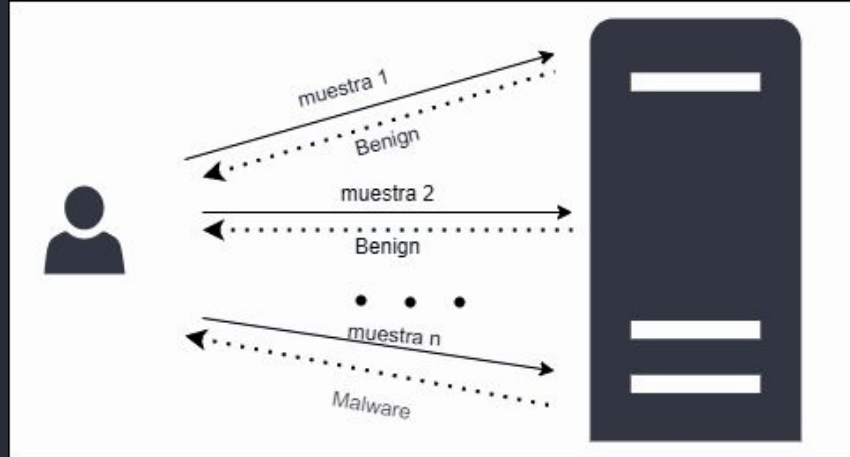
Algoritmo	Precision		Recall		F1-score		Accuracy
	benign	malign	benign	malign	benign	malign	
Naïve Bayes	0.81	0.41	0.12	0.96	0.20	0.58	0.4468
Random Forest	0.93	0.90	0.94	0.88	0.93	0.89	0.9172
KNN: K=4	0.89	0.89	0.93	0.83	0.91	0.86	0.8922
SVM	0.62	0.90	1	0.06	0.76	0.11	0.6271
Logistic regression	0.72	0.68	0.86	0.47	0.78	0.56	0.7063
Decision Tree	0.90	0.85	0.90	0.84	0.90	0.84	0.8773

Modelo del ataque

Generación de malware

Malware generado

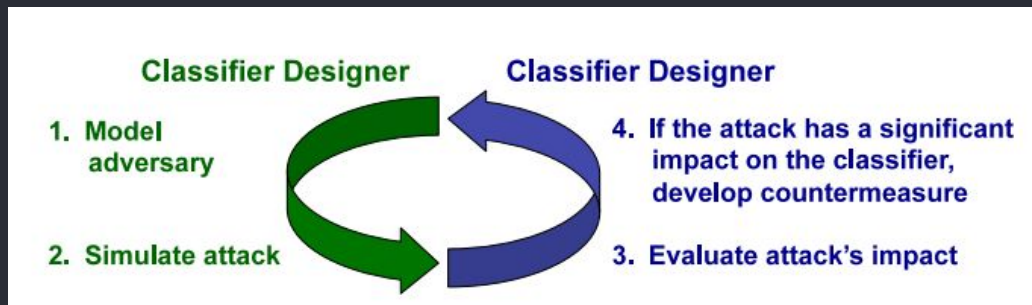
- Algoritmo de fuerza bruta
- 1500 entradas
- Consulta del modelo clasificador RF



Secure Learning

Ciclo de seguridad

- Constante revision
- Contramedidas
- Simulacion
- Clasificador

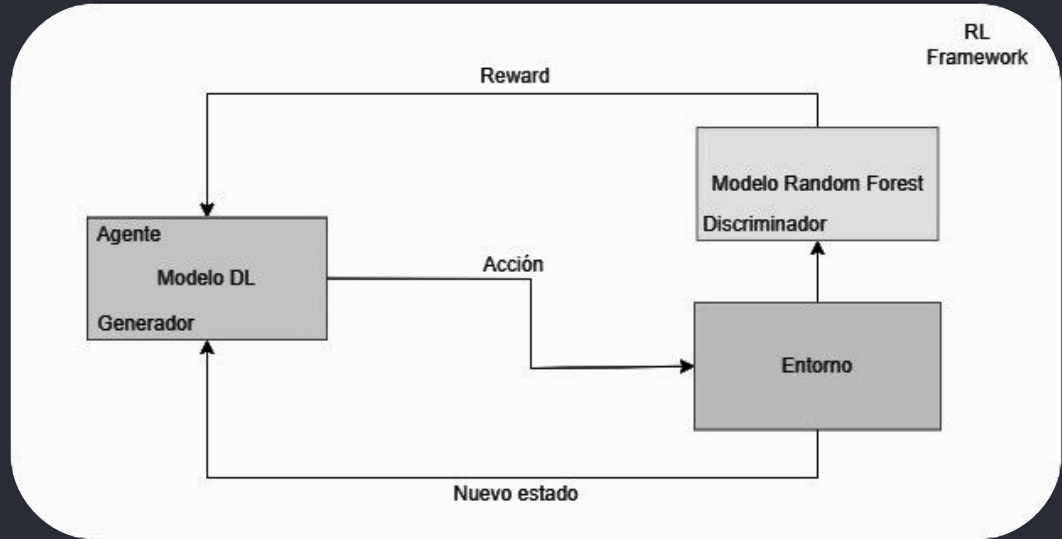


Secure Learning

DRL

Estructura

- 20 acciones
- 1000 epocas
- Learning rate de 0.001
- Gamma de 0.9
- Epsilon de 1
- 4 capas(10, 24, 72, 20)
- Relu
- Secuencial

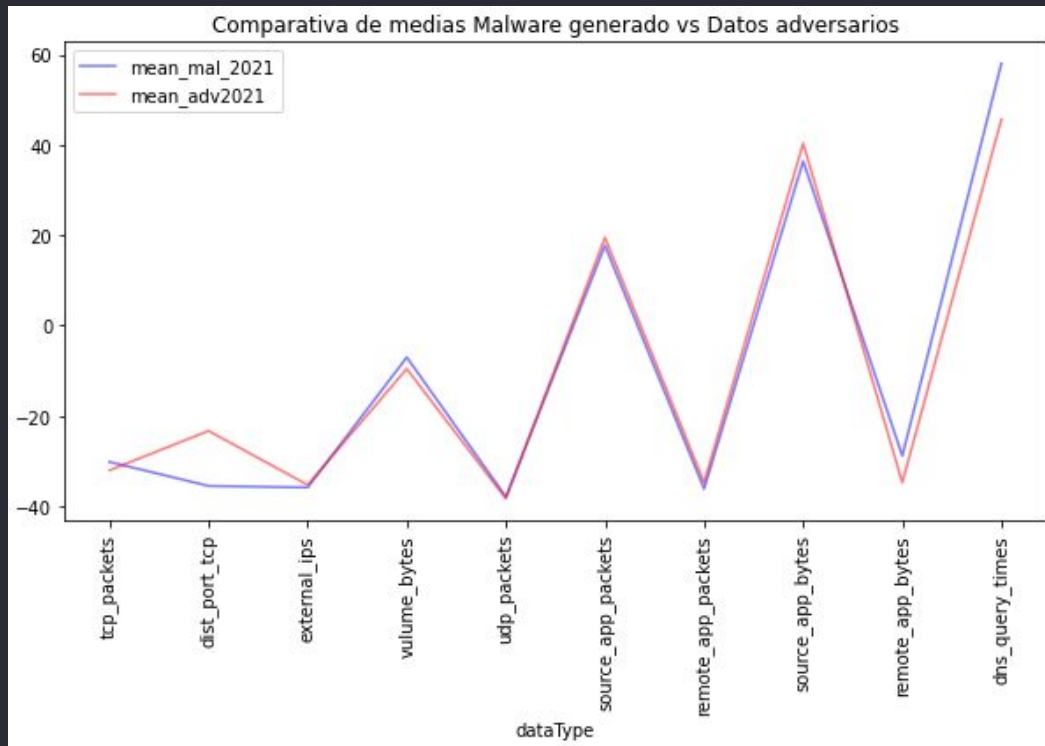


Secure Learning

Datos adversarios

Estructura

- Datos adversarios
- Variables poco relevantes
- 500 datos adversarios



Evaluación

Entrenamiento adversario

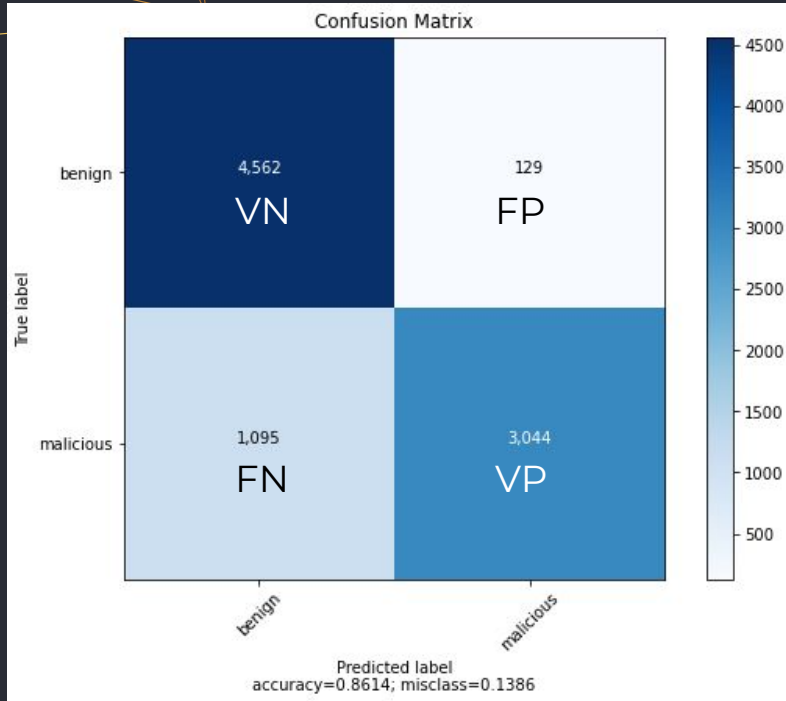
Entrenamiento

- Datos originales
- Hold-out (75%-25%)
- 1000 datos adversarios
- Accuracy de 0.97

Algoritmo	Precision		Recall		F1-score		Accuracy
	benign	malign	benign	malign	benign	malign	
Random Forest	0.98	0.97	0.97	0.98	0.98	0.97	0.9743

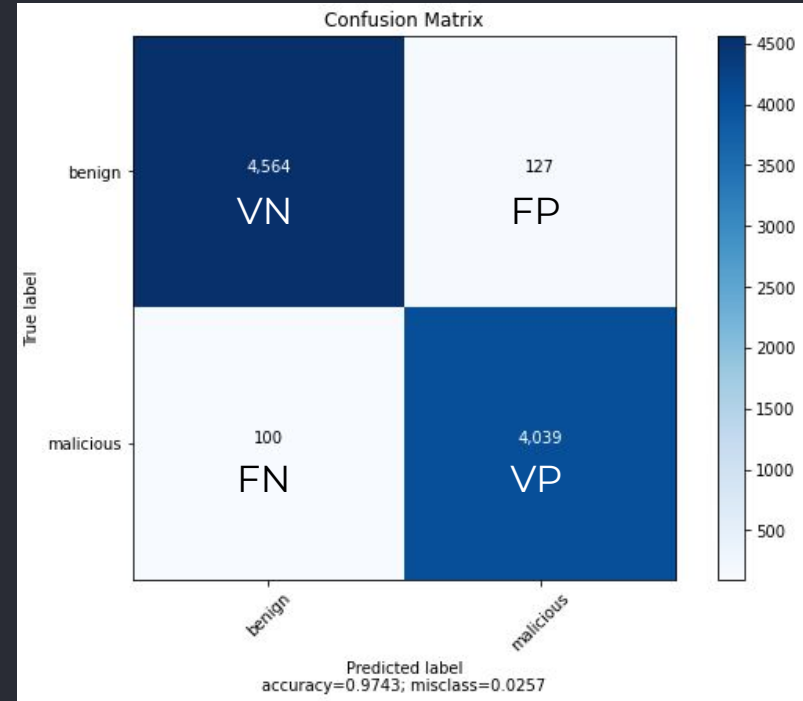
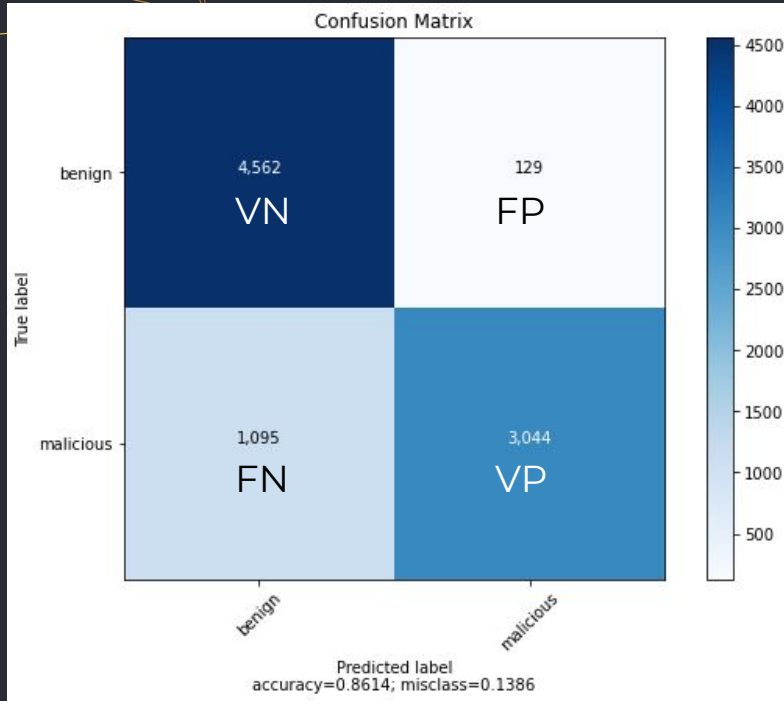
Evaluación

Entrenamiento adversario



Evaluación

Entrenamiento adversario





RECOMENDACIONES



RECOMENDACIONES

- Adicionar al modelo una capa de seguridad para contrarrestar los mensajes de alerta y error con información adicional y precisa del modelo.
- Realizar la exploración de rangos para malware dentro del modelo de forma más estructurada y precisa.
- Dificultar los intentos de reconocimiento que los atacantes pueden intentar realizar.



CONCLUSIONES Y TRABAJO FUTURO

CONCLUSIONES

- El uso de *Deep reinforcement learning* en combinación con una estructura de una red generativa adversaria es de gran utilidad al momento de realizar un proceso de entrenamiento adversario.
- Es posible realizar un ataque adversario de tipo *gray-box* a un modelo de ML de Android Malware entrenado con características de tráfico de red utilizando DRL
- Se encontró que el algoritmo de DRL alcanzó una modificación promedio de 5,96628 unidades respecto a los valores de entrada



TRABAJO FUTURO

- Llevar a cabo el proceso de modificación de un malware real con el fin de demostrar los riesgos para un modelo de clasificación poco seguro.
- Proponer un método de ataque adversario y posterior entrenamiento desde un enfoque black box.
- Continuar con el análisis de los rangos para determinar el intervalo en el que un atacante podría modificar un malware.

REFERENCIAS

- “Operating System Market Share Worldwide,” StatCounter Global Stats, Feb-2021. [Online]. Available: <https://gs.statcounter.com/os-market-share>. [Accessed: 16-Mar-2021].
- P. by S. R. Department and F. 4, “Google Play Store: number of apps 2020,” Statista, 04-Feb-2021. [Online]. Available: <https://www.statista.com/statistics/266210/number-of-available-applications-in-the-google-play-store/>. [Accessed: 16-Mar-2021].
- C. Urcuquí , M. García , J. Osorio , and A. Navarro , Ciberseguridad: un enfoque desde la ciencia de datos, 1 ed. Cali, Valle Del Cauca: Editorial Universidad Icesi, 2018.
- S. Chaieb, “Machine Learning Systems: Security,” Sahbi Chaieb, 03-Feb-2021. [Online]. Available: <https://sahbichaieb.com/mlsystems-security/>. [Accessed: 16-Mar-2021].
- J. Delgado, “Secure Learning para detección de Android Malware,” thesis, 2019.
- M. Barreno, B. Nelson, A. D. Joseph, and J. D. Tygar, “The security of machine learning,” Machine Learning, vol. 81, no. 2, pp. 121–148, 2010.
- B. Dickson, “Robust AI: Protecting Neural Networks Against Adversarial Attacks,” Experfy Insights, 18-Nov-2020. [Online]. Available: <https://www.experfy.com/blog/ai-ml/robust-ai-protecting-neural-networks-against-adversarial-attacks/>. [Accessed: 16-Mar-2021].

REFERENCIAS

- M. Barreno, B. Nelson, A. D. Joseph, and J. D. Tygar, "The security of machine learning," Machine Learning, vol. 81, no. 2, pp. 121-148, 2010.
- S. Shalev-Shwartz and S. Ben-David, Understanding Machine Learning: From Theory to Algorithms, 1st ed. New York: Cambridge University, 2014, pp. 19-25.
- J. Patterson and A. Gibson, Deep learning, 2nd ed. pp. 1-15.
- B. Brown and A. Zai, Deep reinforcement learning in action, 1st ed. NY: Manning Publications Co, 2020.
- R. Sutton and A. Barto, Reinforcement learning, 2nd ed.
- J. Graham, R. Howard and R. Olson, CYBER SECURITY ESSENTIALS, 1st ed. [Place of publication not identified]: CRC Press, 2017.
- M. Sikorski and A. Honig, Practical Malware Analysis: The Hands-On Guide to Dissecting Malicious Software, 1st ed. Manassas Park, Virginia: No Starch Press, 2012, pp. 24-30.
- G. Weidman, Penetration Testing - A hands-on introduction to Hacking, 1st ed. San Francisco: no starch press, 2014, pp. 1-10.
- M. Barreno, A. D. Joseph, and J. D. Tygar, "Can Machine Learning Be Secure?," pp. 16-25, 2006.
- Rathore H, Sahay SK, Nikam P, Sewak M. Robust Android Malware Detection System Against Adversarial Attacks Using Q-Learning. Information Systems Frontiers. 2020 Nov 15:1-6.

REFERENCIAS

- Romdhana A, Merlo A, Ceccato M, Tonella P. Deep Reinforcement Learning for Black-Box Testing of Android Apps. arXiv preprint arXiv:2101.02636. 2021 Jan.
- Fang Z, Wang J, Li B, Wu S, Zhou Y, Huang H. Evading anti-malware engines with deep reinforcement learning. IEEE Access. 2019 Mar 28;7:48867-79.
- Li H, Zhou S, Yuan W, Li J, Leung H. Adversarial-example attacks toward android malware detection system. IEEE Systems Journal. 2019 Apr 11;14(1):653-6.

The background is a dark navy blue. It features several thin, gold-colored lines that form abstract geometric shapes. On the left, there are several intersecting lines forming a series of triangles and polygons. On the right, there are more lines, some of which form a larger, more complex shape. A prominent gold-colored rectangular frame is centered on the slide, enclosing the text.

GRACIAS!