FACOLTÀ DI SCIENZE TECNOLOGICHE E DELL'INNOVAZIONE

CORSO DI LAUREA IN INGEGNERIA INFORMATICA

Prova Finale in
SICUREZZA INFORMATICA

Intelligenza Artificiale e Algoritmi di Machine Learning applicati alla sicurezza informatica

RELATORE
Chiar.mo
Prof LEONARDO GALTIERI

CANDIDATO
ALBERTO GENNARO
MATR. 0082200836

Anno Accademico 2023/24

Sommario

INTRODUZIONE

Con il progredire delle tecnologie ICT (Information and Communication Technology), il mondo del digitale si è trovato di fronte a nuovi pericoli e criticità da difendere nei vari layer di complessità che hanno raggiunto queste tecnologie; Di pari passo il compito della Sicurezza Informatica è diventato sempre più complesso e pressante per la tutela dei sistemi informatici. Nell'ultimo decennio è stato possibile notare un rapido incremento sia delle minacce alla sicurezza che dell'evoluzione delle tecnologie di Intelligenza Artificiale; Nel 2010 potevano essere identificati meno di 50 milioni di malware univoci, Nel 2012 sono raddoppiati a 100 milioni e nel 2019 si è arrivati a più di 900 milioni di soli Malware noti; I benefici di questa tecnologia stanno diventando sempre più utili come integrazione alle pratiche e metodologie per la prevenzione delle minacce; Questa criticità è sottolineata dal fatto che anche gli attori malevoli sfruttano algoritmi di Machine Learning per ottimizzare e rendere ancora più efficaci i loro attacchi; Già solamente l'incremento esponenziale del numero di attacchi giornalieri ha reso chiaro che era necessario trovare metodi per difendersi alternativi a quelli già esistenti; Metodi che non fossero statici e che necessitassero della costante supervisione di analisti e professionisti, ma tecnologie integrate in grado di adattarsi alla velocità e mutabilità degli attacchi attuali, Tecnologie che imparavano dai dati conosciuti per estrapolare metodi di risposta adeguati.

Algoritmi per il riconoscimento delle minacce in rete, piattaforme di risposta agli attacchi automatizzate, monitoraggio in tempo reale del traffico di rete, riconoscimento di pattern nei comportamenti, tutte necessità nate con l'evoluzione della tecnologia e delle possibilità del digitale.

L'obiettivo di questo lavoro è analizzare le potenzialità e i rischi che L'intelligenza Artificiale offre al settore della Cyber Security; andremo ad analizzare prima i concetti fondamentali di sicurezza e di algoritmi di Machine Learning, per poi passare ad un'analisi di come questi vengono utilizzati sia dai difensori che dagli attaccanti nel panorama della Cyber Security, concludendo infine con un caso di studi riguardante la piattaforma Vectra.

Capitolo I: CONCETTI FONDAMENTALI

CYBERSECURITY

Il termine Cybersecurity è inteso come la parte di difesa e prevenzione delle informazioni processate e scambiate attraverso la rete. Con questo termine si identificano l'insieme di mezzi, procedure e tecnologie volte alla protezione di sistemi informatici. Il NIST (National Institute of science and technology) definisce il termine come segue: *"Computer Security: Measures and controls that ensure confidentiality, integrity, and availability of information system assets including hardware, software, firmware, and information being processed, stored, and communicated."* [1]

I tre obiettivi cardine della sicurezza informatica, denominati modello CIA Triad, sono la Riservatezza (Confidentiality), l'Integrità (Integrity) e la Disponibilità (Availability).

Riservatezza: è la protezione delle informazioni da accessi o divulgazioni non autorizzate da parte di individui o sistemi. Essa può essere mantenuta applicando pratiche quali il controllo sugli accessi, la crittografia o l'autenticazione degli utenti, metodi più recenti includono anche security token o autenticazioni biometriche per gli utenti.

Integrità: è la garanzia che i dati o i sistemi siano affidabili, inalterati e non compromessi da agenti esterni non autorizzati. Una minaccia all'integrità può essere data, ancor prima di pensare ad elementi malevoli, semplicemente da un

[1] NIST Internal/Interagency Report NISTIR 7298 (*Glossary of Key Information Security Terms*, May 2013)

malfunzionamento di sistema, da errori nell'inserimento di dati o da metodi di backup inadeguati. Si può garantire l'integrità delle informazioni tramite pratiche di autenticazione utente, firme digitali e funzioni di hash.

Disponibilità: è la garanzia che dati e sistemi siano sempre accessibili in un determinato momento per ogni utente autorizzato. Per assicurarsi la disponibilità di dati o servizi è buona norma mantenere un corretto funzionamento sia del lato Hardware che del lato Software. La compromissione di questo aspetto può portare a casi di Denial of Service.

VULNERABILITA' RISCHIO E IMPATTO

Il concetto di *gestione del rischio* nella sicurezza informatica è il processo di identificazione e prevenzione delle minacce informatiche. Minacce per le quali l'attività o impresa legata al sistema informatico possano avere degli impatti, siano essi negativi o positivi, sul rendimento della stessa.

Rischio: Il NIST definisce il rischio come "una misura della probabile entità della perdita o del danno derivante dallo sfruttamento di una vulnerabilità". In cybersecurity, il rischio rappresenta la combinazione tra la probabilità che una minaccia sfrutti una vulnerabilità e l'impatto che ne deriva. L'analisi del rischio è un processo continuo che comprende l'identificazione, la valutazione e la prioritizzazione delle minacce, insieme all'implementazione di misure di sicurezza per mitigare tali rischi. Un approccio comune per gestire il rischio è l'utilizzo di un

modello basato su probabilità e impatti. La combinazione della probabilità che un evento accada e delle sue conseguenze potenziali.

Vulnerabilità

Secondo il NIST, una vulnerabilità è definita come una "debolezza in un sistema informatico, nelle procedure di sicurezza, nei controlli interni o nell'implementazione che può essere sfruttata da una minaccia per ottenere accesso non autorizzato a informazioni o interrompere servizi critici"[2]. In termini tecnici, essa rappresenta una falla nel sistema che, se non corretta, può essere utilizzata da attaccanti per compromettere la sicurezza del sistema, causando potenziali danni. La gestione delle vulnerabilità prevede il monitoraggio continuo e l'applicazione di patch per ridurre al minimo il rischio.

L'Impatto

E' il risultato potenziale di una compromissione nei criteri di riservatezza, integrità o disponibilità in un sistema, esso viene definito con un valore di basso, moderato o alto dipendentemente dalle conseguenze [3].

Nello specifico, riferendoci alla triade CIA, una compromissione della *riservatezza* può rendere in sistema "fallato", ad esempio può permettere accesso ad informazioni senza predisporre dei privilegi adeguati.

Compromissione dell'*integrità* può rendere un sistema corrotto, ovvero che i dati al suo interno o i suoi comportamenti sono stati manipolati da fattori esterni.

Compromissione della *disponibilità* porta invece ad un sistema che non riesce a svolgere il compito richiesto, o a causa di tempi di risposta lunghi o per un blocco dei processi.

[2] NIST SP 1800-17b
[3] CNSSI 4009-2015 under impact value from NIST SP 800-30 Rev. 1

ARTIFICIAL INTELLIGENCE

"*An AI system is a machine-based system that, for* ***explicit or implicit*** *objectives,* ***infers, from the input it receives, how to generate outputs such as*** *predictions,* ***content****, recommendations, or decisions* ***that can*** *influenc****e physical*** *or virtual environments.* ***Different*** *AI systems vary* ***in their*** *levels of autonomy* ***and adaptiveness after deployment.****"*[4]

Un sistema di intelligenza artificiale, come definito dall'OCSE (AIGO), è "un sistema basato sull'utilizzo di una macchina" che, per una determinata serie di obiettivi definiti esternamente, è in grado di: fare previsioni, stime o prendere decisioni che hanno un impatto sugli ambienti reali o virtuali. Esso utilizza input provenienti da altre sorgenti-macchina o direttamente dall'utente, per percepire ambienti reali e/o virtuali; trasforma tali percezioni in modelli astratti (in modo automatizzato, per esempio con un trattamento-processo ML o manualmente) e utilizza modelli di deduzione logica per formulare scelte, al fine di ottenere un'informazione o definire un'azione. I sistemi d'intelligenza artificiale sono progettati per operare con diversi livelli di autonomia"

Le fasi del ciclo di vita del sistema di intelligenza artificiale sono:

- pianificazione e progettazione, raccolta ed elaborazione dei dati e costruzione del modello ed interpretazione;
- verifica e convalida;
- dispiegamento;
- funzionamento e monitoraggio.

Nello specifico, sotto il termine AI ricadono diverse tipologie di applicazioni quali:

Machine Learning (ML): modelli che apprendono da dataset per fare previsioni

[4] Definizione aggiornata OECD

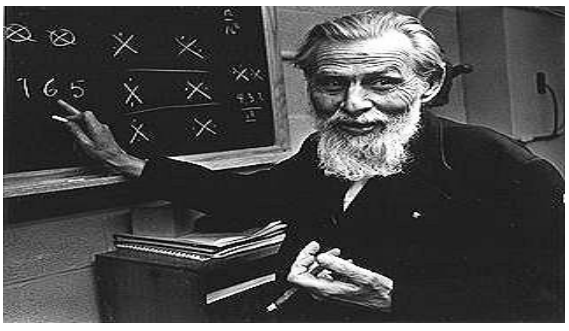
senza essere programmati direttamente, include sottocategorie di tecniche di apprendimento quali *Supervised, Unsupervised e Reinforced Learning***.**

Natural Language Processing (NLP): che racchiude le tecnologie di comprensione del testo, traduzione e generazione di linguaggio naturale.

Neural Networks (NN): Modelli ispirati alla struttura del cervello umano, utilizzati per riconoscere pattern complessi nei dati. Possono essere generalizzati come aventi tre strati principali: lo strato di input, lo strato nascosto, e lo strato di output. I dati fluiscono dallo strato di input, attraverso quello nascosto, che è composto da più layer e nodi, fino a raggiungere lo strato di output dove viene prodotto il risultato. Le reti neurali sono costituite da nodi strutturati e interconnessi, i cui valori sono determinati dai pesi di tutte le connessioni che arrivano a ciascun nodo.

EVOLUZIONE DELL'AI

L'IA come è oggi è nata nel secondo dopoguerra, grazie soprattutto all'avvento dei calcolatori, fin da subito considerati come il perfetto strumento per riprodurre il ragionamento umano. Tra i pionieri spiccano *Warren McCulloch* e *Walter Pitts*, che nel 1943 proposero un modello di rete neurale ispirato al cervello umano. Questa rete, composta da "neuroni" binari, poteva implementare i principi della logica booleana e ipotizzava capacità di apprendimento, grazie alla regola di modifica sinaptica ideata da Donald Hebb nel 1949. Successivamente, Marvin Minsky sviluppò il primo computer a rete neurale, SNARC, nel 1950, simulando una rete di 40 neuroni.



Negli anni '50, figure come John McCarthy e Alan Turing gettarono le basi teoriche dell'AI. Turing, con il suo articolo "Computing Machinery and Intelligence", introdusse il celebre Test di Turing, un criterio per valutare se una

macchina potesse dimostrare un comportamento intelligente. McCarthy invece fu il primo a coniare il termine Intelligenza artificiale durante la Conferenza di Dartmouth ad Hanover nel 1956. La conferenza si riferisce al **Dartmouth Summer Research Project on Artificial Intelligence**, svoltosi nel 1956, e considerato come l'evento ufficiale che segna la nascita del campo di ricerca.[5]

Dopo un periodo di entusiasmo iniziale, seguito da delusioni e riduzioni dei finanziamenti negli anni '70 e '80;

Dagli anni ’90 c’è stata una nuova crescita esponenziale che arriva fino ad oggi grazie all’introduzione di nuovi algoritmi e una maggiore potenza di calcolo. Si arriva al riconoscimento delle immagini, c’è un miglioramento del Natural Language Processing, il sistema di traduzione delle lingue che fino ad allora aveva avuto scarsi risultati. Inoltre la disponibilità di grandi quantità di dati ha permesso l’addestramento di modelli più sofisticati e complessi. Oggi, l'AI è integrata in molti aspetti della vita quotidiana, dai motori di ricerca ai sistemi di riconoscimento vocale e visivo, continuando a evolversi e a rivoluzionare numerosi settori.

MACHINE LEARNING

“L'apprendimento automatico (Machine Learning, abbreviato in ML) è una branca

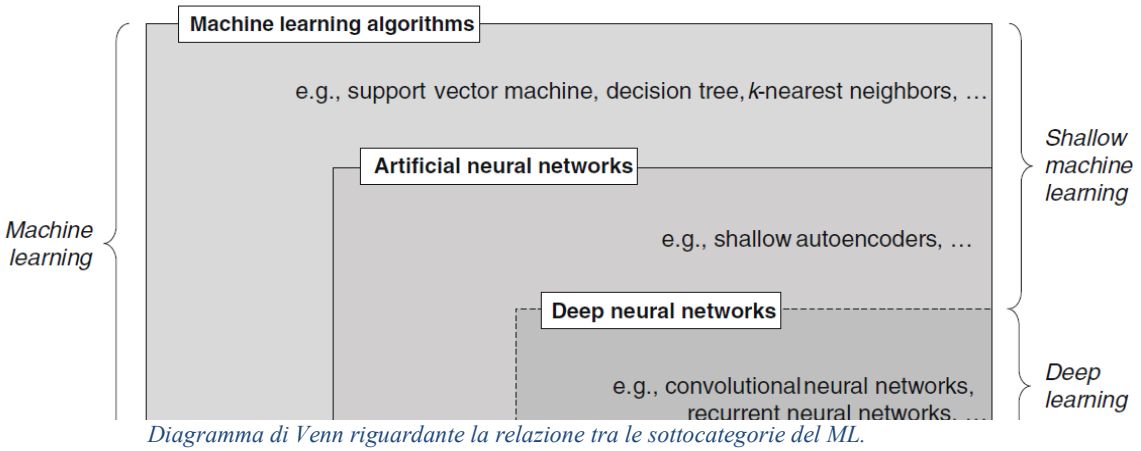


Diagramma di Venn riguardante la relazione tra le sottocategorie del ML.

[5] https://it.wikipedia.org/wiki/Conferenza_di_Dartmouth

dell'intelligenza artificiale che raccoglie metodi sviluppati negli ultimi decenni del XX secolo in varie comunità scientifiche, sotto diversi nomi quali: statistica computazionale, riconoscimento di pattern, reti neurali artificiali, filtraggio adattivo, teoria dei sistemi dinamici, elaborazione delle immagini, data mining, algoritmi adattivi, ecc; che *utilizza metodi statistici per migliorare la performance di un algoritmo* nell'identificare pattern nei dati."[6]

Fondamentalmente i modelli di Machine Learning si concentrano nel fare apprendere al programma attraverso grandi quantità di dati. Questi modelli sono composti da un insieme di regole, metodi e funzioni che possono essere applicate per riconoscere dei pattern di dati, fare predizioni di comportamenti e identificare in maniera veloce l'obiettivo dei modelli.

Per questo è giusto dire che tramite il ML le capacità e precisione di un programma aumentano con "*l'esperienza*", più precisamente tramite le iterazioni di model building effettuato dai vari algoritmi che imparano ricorsivamente da training data riguardanti problemi specifici, portando il programma a identificare pattern complessi senza essere stati scritti in maniera esplicita.

Si apprezza un incremento delle capacità in situazioni per cui vengono processate grandi quantità di dati, ad esempio in compiti di *Classificazione, Regressione e Clustering*. Per questo motivo, gli algoritmi di Machine Learning trovano efficienti applicazioni in vari ambiti della sicurezza informatica, tra cui ad esempio il rilevamento delle frodi, di malware, sistemi di Intrusion Detection e Prevention, ecc. Andando più nello specifico analizziamo la gerarchia tra i le varie metodologie di Machine learning e la distinzione tra 3 tipologie di Machine Learning, Supervised, Unsupervised e Reinforcement.

- *Apprendimento supervisionato*

  È un approccio che si focalizza su task e richieste specifiche; esso regola il training degli algoritmi su dei dataset "labeled", che includono esempi riguardanti gli input previsti dal programma da correlare a risposte in output.

[6] https://it.wikipedia.org/wiki/Apprendimento_automatico

Questo metodo è utilizzato principalmente per problemi di Classificazione e Regressione. (Classificazione dove il risultato predetto è una categorizzazione in base a delle classi stabilite, Regressione dove il risultato è un valore numerico). Nell'ambito della sicurezza informatica queste tecniche sono utili per fare predizioni riguardo attacchi DDoS, identificare tipologie di attacchi di rete come scanning e spoofing; tecniche di Regressione vengono utilizzate invece per calcolare la quantità di attacchi di Phishing durante un dato periodo di tempo o controllare i parametri di pacchetti di dati.[7]

- *Apprendimento non supervisionato*

I metodi di apprendimento non supervisionato vengono utilizzati quando c'è la necessità che il sistema identifichi Pattern, strutture e correlazioni all'interno di dati non strutturati (*Unlabeled*), non etichettati, applicando algoritmi di Clustering su gruppi di elementi che condividono proprietà comuni o proiezioni di dati da dataset con enormi volumi di informazioni, *High-Dimensional spaces*, a spazi semplificati (*Dimensionality Reduction*).

In questo caso il sistema durante la fase di training tenta di emulare i dati ricevuti e iterativamente sfrutta l'errore in output per autocorreggersi. Gli algoritmi per l'apprendimento non supervisionato possono raggrupparsi *in 3 tipologie*:

- Clustering: per raggruppare gli input in base alle similitudini delle loro caratteristiche disponendoli in uno spazio multidimensionale e calcolando la distanza tra i vari dati.
- Algoritmi associativi: Per identificare correlazioni tra variabili e trovare pattern o regole associative.
- Algoritmi di Dimensionality Reduction: Utilizzati con grandi quantità di Dati in input, dove estrapolano caratteristiche ricorrenti o importanti, si effettua una compressione delle informazioni, ricombinando i dati in componenti indipendenti tra loro ma di difficile interpretazione.

Di solito i compiti di tipo "generativo" utilizzano metodi di Unsupervised learning,

[7] Cybersecurity data science, Iqbal H. Sarker (2020)

ad esempio per la generazione di immagini, video e LLM (Large Language Models).

Nel campo della Cybersecurity, alcune minacce come ad esempio i Malware possono essere difficili da identificare, avendo la capacità di modificare il loro comportamento, in questo caso i metodi di Apprendimento non supervisionato possono essere utilizzati per rilevare queste minacce.

- *Reinforcement Learning*

  I metodi di apprendimento per rinforzo comportano principi di "Trial and error" per affinare in maniera iterativa i risultati ottenuti.

Diversamente dai due metodi precedenti, questo paradigma si occupa di problemi di decisioni sequenziali, in cui l'azione da compiere dipende dallo stato attuale del sistema e ne determina quello futuro.

"La qualità di un'azione è data da un valore numerico di "ricompensa", ispirata al concetto di rinforzo, che ha lo scopo di incoraggiare comportamenti corretti dell'agente. Questo tipo di apprendimento è solitamente modellizzato tramite i processi decisionali di Markov e può essere effettuato con diverse tipologie di algoritmi, classificabili in base all'utilizzo di un modello che descriva l'ambiente, alle modalità di raccolta dell'esperienza (in prima persona o da parte di terzi), al tipo di rappresentazione degli stati del sistema e delle azioni da compiere (discreti o continui)."[8]

Tra gli algoritmi per di apprendimento con rinforzo troviamo il Q-learning, Monte Carlo learning e il Deep Q Networks.

- *Altri metodi di apprendimento*

  L'apprendimento semi-supervisionato può essere descritto come un'ibrido

[8] https://it.wikipedia.org/wiki/Apprendimento_per_rinforzo

delle tecniche supervisionato e non supervisionato discusse prima poiché utilizza dataset etichettati o meno.

Nell'ambito della sicurezza informatica questo può dimostrarsi utile per etichettare dati senza il bisogno della supervisione umana per incrementare l'efficienza dei modelli di cybersecurity.

GENERATIVE AI

Con Generative AI si intende una branca dell'Intelligenza artificiale con la capacità di generare contenuti in risposta ad un dato prompt; essi di solito sono di tipo multimediali, come testo, immagini, video o audio. Degli esempi possono essere ChatGPT e Stable Diffusion.

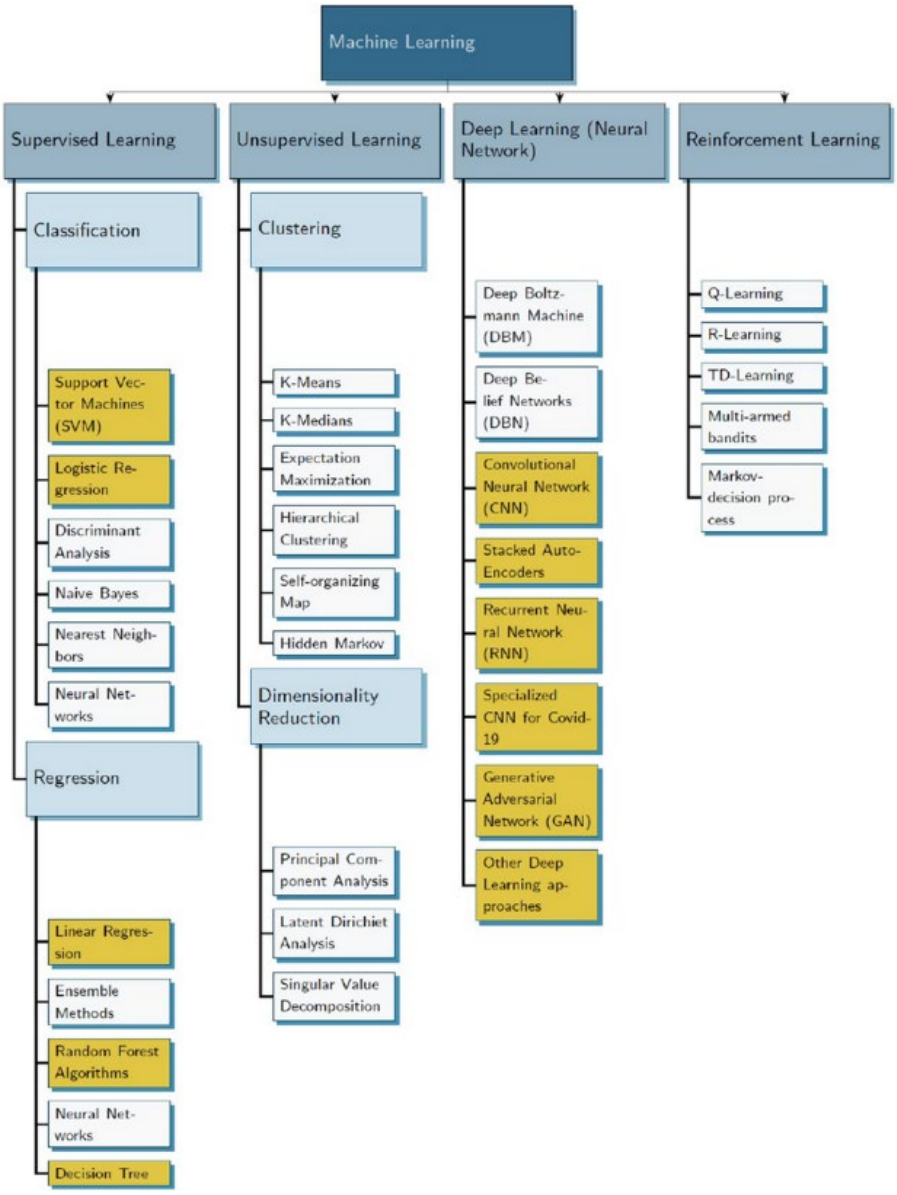
Dall'inizio dell'ultimo decennio c'è stata un' enorme attenzione nei riguardi di questa tecnologia, che ha avuto il suo boom con l'avvento di ChatGPT, chatbot di OpenAI, verso la fine del 2022.

La particolarità di questo sistema è che a differenza delle AI tradizionali che generalmente sono progettate per operare in ambiti più specifici, mentre la generative AI creano contenuti "originali", che sono simili ma non corrispondono ai dataset di training. Altra differenza è che le AI tradizionali sono sempre addestrate tramite algoritmi di "supervised" learning, mentre quest'ultima deve essere necessariamente addestrata utilizzando metodi di apprendimento "unsupervised"

La Gen AI utilizza algoritmi di deep learning e modelli generativi per produrre output partendo da un grande volume di dati utilizzato per fare training.

Capitolo II: ALGORITMI DI MACHINE LEARNING

Nel corso degli anni sono stati sviluppati moltissimi algoritmi di machine learning, ognuno con le proprie caratteristiche e i propri obiettivi analitici. Tra quelli comunemente più diffusi ritroviamo:



Algoritmi di Regressione

Sono algoritmi che creano previsioni numeriche continue sui dati in input. Ad esempio, un algoritmo di regressione può predire andamenti finanziari o tempistiche su progetti.

Regressione Lineare

La regressione lineare è l'algoritmo introduttivo al machine learning. L'algoritmo classico si chiama OLS (Ordinary Least Squares), che adatta un modello lineare con coefficienti $w=(w1,...,wn)$w=(w1,...,wn) per ridurre al minimo la somma residua dei quadrati tra gli n dati osservati e le previsioni fatte dall'approssimazione lineare.

Regressione logistica

La regressione logistica è un algoritmo di **Classificazione** e fa parte dei metodi **Supervisionati**. viene utilizzata per prevedere un output binario, ad esempio "Si" o "No" in base ad altre variabili di input. L'algoritmo genera una funzione logistica o Sigmoide.

Tra le applicazioni possibili, la regressione logistica può prevedere se un paziente svilupperà una malattia in base ai fattori di rischio documentati (dati di input).

TRAINING DATASET

examples	x1	x2	x3	x4	x5	label: y
1	0	1	2	1	0	1
2	1	1	4	0	1	0
3	1	0	1	1	0	0
4	0	1	3	0	0	1
...						
n	0	1	4	0	1	1

L'algoritmo inizialmente prende in input un dataset di training composto da N dati. Ogni esempio è composto da m attributi X e da un'etichetta y che indica la corretta classificazione.

L'algoritmo individua un vettore dei pesi W da associare al vettore degli attributi Xm degli esempi, in modo tale da massimizzare la percentuale di risposte corrette (o minimizzare quelle sbagliate).

La combinazione lineare z dei pesi L per gli attributi X fornisce una risposta del sistema per ogni esempio del training dataset.

$$z = W \cdot X = w_1 x_1 + \ldots + w_m x_m$$

Nella regressione logistica la combinazione lineare z è l'argomento della

funzione logistica che lo traduce in un valore compreso tra 0 e 1.

$$f(z) = [0, 1]$$

Per scegliere la distribuzione W* migliore l'algoritmo utilizza un'altra funzione di massimizzazione della probabilità L.

$$max\ L(W) = \sum_{i=1}^{N} y_i \log f(z_{(i)})(x) + (1 - y_i) \log(1 - f(z_{(i)}))$$

La sommatoria confronta tutte le risposte del modello con le risposte corrette calcolando la log-probabilità.

Al termine dell'addestramento l'algoritmo produce un modello, utilizzabile per classificare qualsiasi altro esempio non compreso nel training set.

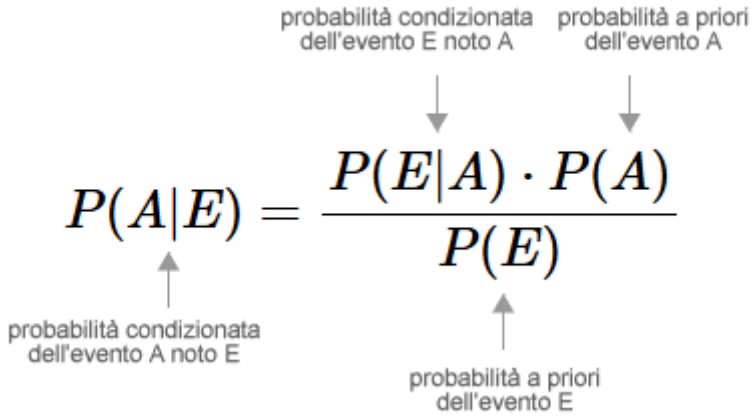
Naive Bayes

Altro algoritmo di Classificazione statistica Supervisionato, Per usare l'algoritmo Naive Bayes occorre conoscere o stimare le probabilità a priori e condizionali del problema.

Ad esempio si può classificare una mail come Spam se all'interno del messaggio troviamo termini come "soldi" o "IBAN", se contiene un url o un file allegato e se il mittente non ha mai scritto al nostro indirizzo. Si tratta di 3 caratteristiche differenti che l'algoritmo Naive Bayes considera come contributi indipendenti al fatto che la mail sia Spam.

Le caratteristiche del dataset sono di tipo dicotomico, nel senso che possono assumere due valori, come: "sì o no"; termine presente o assente nel testo. Tutte le caratteristiche del dataset sono importanti e quindi indipendenti tra di loro, non c'è nessuna probabilità condizionata tra le features e quindi è possibile calcolare la probabilità della classe andando a calcolare la probabilità condizionata. Infatti data

una combinazione di features se sono presenti o assenti/Yes o no è possibile calcolare la probabilità della classe.



Teorema di Bayes

La formula utilizzata dall'algoritmo naive bayes invece è la seguente:

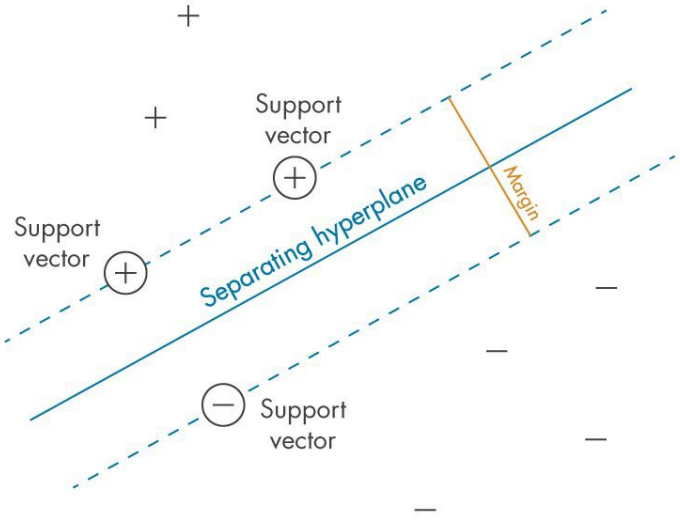
$$P(C_L \mid F_1,\ldots,F_n) = \frac{1}{Z} p(C_L) \prod_{i=1}^{n} p(F_i \mid C_L)$$

Questo tipo di classificatore è ritenuto veloce ed efficace per il fatto che il training viene eseguito in maniera veloce, perché si tratta di stimare solo le frequenze dei termini; Tuttavia alcuni svantaggi sono il fatto di non essere affidabile per dataset con grandi quantità di caratteristiche.

Support Vector Machine

Questo sistema di algoritmi viene utilizzato per risolvere problemi prevalentemente di classificazione, ma anche per quelli di regressione. Le macchine a vettori di supporto sono modelli di Supervised Learning che hanno come scopo trovare l'iperpiano ottimale in uno spazio N-dimensionale, dove N corrisponde al numero di features interessate, in grado di separare i punti dati in diverse classi nello spazio delle feature. L'iperpiano cerca di fare in modo che il margine tra i punti più vicini delle diverse classi sia il più ampio possibile.

Per margine si intende la larghezza massima della linea parallela all'iperpiano che non ha punti di dati interni. L'algoritmo può identificare un iperpiano nello spazio solo per classi separabili in modo lineare.



1.Definizione del "margine" tra le classi

L'idea principale della SVM è quella di "allargare" lo spazio delle feature usando funzioni quadratiche, cubiche o di dimensionalità più elevata dei predittori e dei termini di interazione, o anche funzioni più complicate, per descrivere soglie di separazione non lineari tra le classi. I mattoni fondamentali del SVM sono i "prodotti interni" delle osservazioni (piuttosto che le osservazioni stesse), che entrano nel classificatore dopo che sono stati trasformati tramite un kernel (o nucleo): grossolanamente una funzione delle osservazioni delle variabili indipendenti. Ci sono molte possibili scelte di kernel, con le più popolari che sono i cosiddetti kernel polinomiali o radiali.

in generale maggiore è il margine fra questi punti, minore è l'errore di generalizzazione commesso dal classificatore.

Spesso gli insiemi che devono essere distinti non sono linearmente separabili nello spazio. Per questo motivo è stato proposto che lo spazio originale di dimensioni finite venisse mappato in uno spazio con un numero di dimensioni maggiore, rendendo meno complesso trovare una distinzione in questo nuovo spazio.

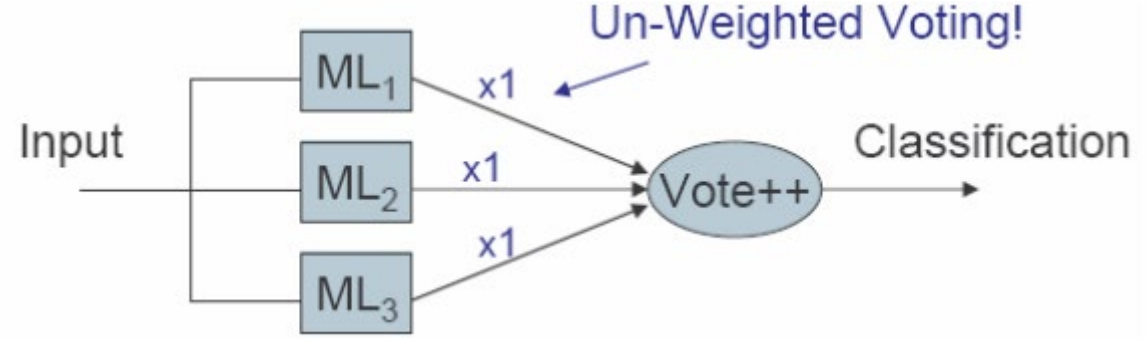
"Per mantenere il carico computazionale accettabile, le mappature utilizzate dalle SVM sono fatte in modo tale che i prodotti scalari dei vettori delle coppie di punti in ingresso siano calcolati facilmente in termini delle variabili dello spazio originale, attraverso la loro definizione in termini di una funzione kernel k(x,y) scelta in base al problema da risolvere."[9]

[9] https://it.wikipedia.org/wiki/Macchine_a_vettori_di_supporto

Gli algoritmi di SVM trovano applicazione nei campi del Natural Language, riconoscimento di immagini, nella Bioinformatica e nell'ambito della cybersecurity può essere utilizzata per il rilevamento di spam e negli IDS.

Random Forest

E' un algoritmo volto alla risoluzione di problemi di classificazione e regressione.
Esso è composto da diversi alberi decisionali che sono classificatori condizionali con lo scopo di "pesare" le features dei dati. In seguito effettua una aggregazione dei risultati, dando un risultato o per votazione nel caso dei problemi di



classificazione, o per media nei problemi di regressione.

Il funzionamento consiste nell'utilizzo delle tecniche di bootstrapping, utilizzate per ridurre la varianza di un metodo di apprendimento; Si estraggono diversi sottoinsiemi dal training set, per simulare la variazione dei data.
In seguito si effettua l'ensemble delle predizioni sulle istanze *x* di test.

$$f_{bag}(\boldsymbol{x}) = \frac{1}{B}\sum_{i=1}^{B} f_i(\boldsymbol{x})$$

Gli algoritmi di Random Forest hanno una stima dell'errore nel test (Out of Bag), che rappresenta una media degli errori di predizione per ogni istanza di training.

Questo tipo di algoritmi vengono utilizzati nella cybersecurity per l'identificazione di attacchi DOS e il rafforzamento dei modelli di Intrusion Detection System.

Hidden Markov Models

Sono modelli che possono identificare un processo sequenziale, essi definiscono in maniera probabilistica l'andamento o dinamica di un sistema senza saperne le cause, il loro compito è identificare in maniera probabilistica gli stati che hanno prodotto un risultato.

Un esempio esplicativo può essere il seguente: Devo identificare lo stato di un semaforo stradale senza avere la possibilità di vederlo, ma ho delle telecamere che monitorano il traffico nelle vicinanze, in base al flusso di autovetture rilevate per strada posso effettuare una predizione sullo stato "nascosto" del semaforo.

Gli HMM hanno alla base i processi di Markov, essi modellano un sistema che si evolve attraverso N stati e dichiarano che la probabilità che il sistema raggiunga uno stato è determinata solamente dallo stato precedente.

La formula base è la seguente:

$$P(Y) = \sum_{X} P(Y \mid X)P(X)$$

In cui $Y = y(0), y(1), \ldots, y(L-1)$ è l'insieme ordinato degli eventi e $X = x(0), x(1), \ldots, x(L-1)$ è l'insieme ordinato degli stati nascosti.

K-Nearest Neighbor

Sono utilizzati per risolvere problemi di regressione e classificazione; Misurano quanto sono vicine o distanti i dati nel training set, comparandoli con i K dati vicini. In parole semplici, questo algoritmo pensa che più le cose sono vicine tra loro, più sono simili.

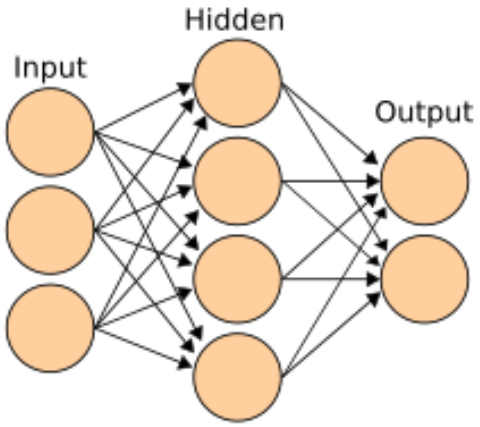
La classificazione di un record **z** è ottenuta con un processo di majority voting tra i k elementi Dz del training set D più vicini (o simili) a z.[10]

$$\bar{y} = \arg\max_{y \in \mathbf{Y}} \sum_{(\mathbf{x}_i, y_i) \in D_z} I(y_i = y)$$

Riguardo il valore K, aumentarlo o diminuirlo permette di calibrare l'efficienza dell' algoritmo, con un K troppo piccolo, l'algoritmo è sensibile ai dati anomali, con un K troppo grande al contrario si può "sforare" in altre classi.

Neural networks

La famiglia delle reti neurali artificiali consiste in una trasposizione matematica del processo di connessione neurale biologico. Si suddivide in shallow Neural network e Deep Neural network.



Come si può vedere, una ANN è un'interconnessione di nodi denominati Neuroni che, come per le sinapsi del cervello, dialogano a vicenda ed elaborano input. La differenza tra Shallow e Deep dipende dalla complessità e quantità dei layer comunicanti. Di solito le DNN vengono utilizzate in maniera unsupervised, quindi dove il training dei dati è unlabeled, visto che sono indicati per processare enormi volumi di dati, al contrario delle SNN che sono utilizzate per compiti più specifici.

[10] Classificatori K-NN, Matteo Golfarelli. Alma Mater Studiorum - Università di Bologna

Capitolo III: MINACCE E ATTACCHI INFORMATICI

In questo capitolo analizzeremo brevemente le varie tipologie di minacce informatiche; sotto questo termine ricadono tutti i tentativi di attacco volti a compromettere la triade già citata nel capitolo precedente di Riservatezza, Integrità e Disponibilità di un sistema informatico. Data la velocità con cui queste minacce evolvono, è importante definire una classificazione per essere in grado di identificare il loro comportamento e applicare le misure di sicurezza adeguate.
Nella tassonomia delle minacce informatiche più comuni troviamo i Malware, il Phishing e la Social Engineering, gli Attacchi di Rete tra cui i DOS e i Man in the Middle, i Zero Day Exploit e le SQL injection.

Attributi di un attacco informatico

Malware

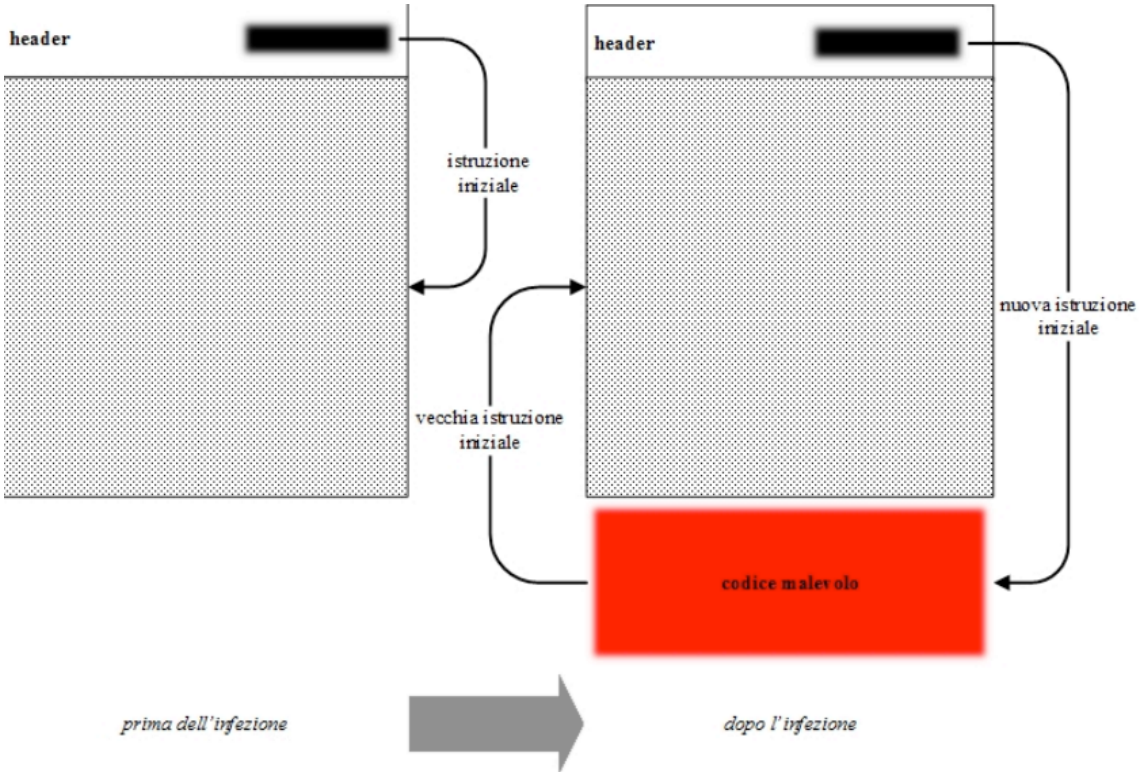
E' l'abbreviazione di Malicious Software, questi sono tra i più comuni elementi di cyber threat presenti; sotto questa categoria ricadono tutti i programmi scritti con l'obiettivo di danneggiare o compiere azioni non autorizzate su sistemi informatici.
Il NIST SP 800-83[11] definisce il termine Malware come: "A program that is inserted into a system, usually covertly, with the intent of compromising the confidentiality, integrity, or availability of the victim's data, applications, or operating system or otherwise annoying or disrupting the victim."

[11] *Guide to Malware Incident Prevention and Handling for Desktops and Laptops,* July 2013

Secondo un'analisi recente di Tang et al. (2020), i malware possono essere classificati in diverse categorie basate sui loro comportamenti (PAYLOAD), eventi scatenanti(TRIGGER) e metodi di distribuzione (PROPAGATION); trai le categorie di malware più conosciute troviamo *Virus, Worm, Trojan, Ransomware, Spyware, Adware e Rootkit*; Ogni categoria con caratteristiche e strategie di difesa adeguate.

I **virus** sono programmi che si auto replicano, collegandosi a file presenti nel sistema e propagandosi quando questi vengono eseguiti. Si può effettuare un'altra classificazione riguardo i Virus, in base alla modalità di parassitismo nei confronti del file legittimo:

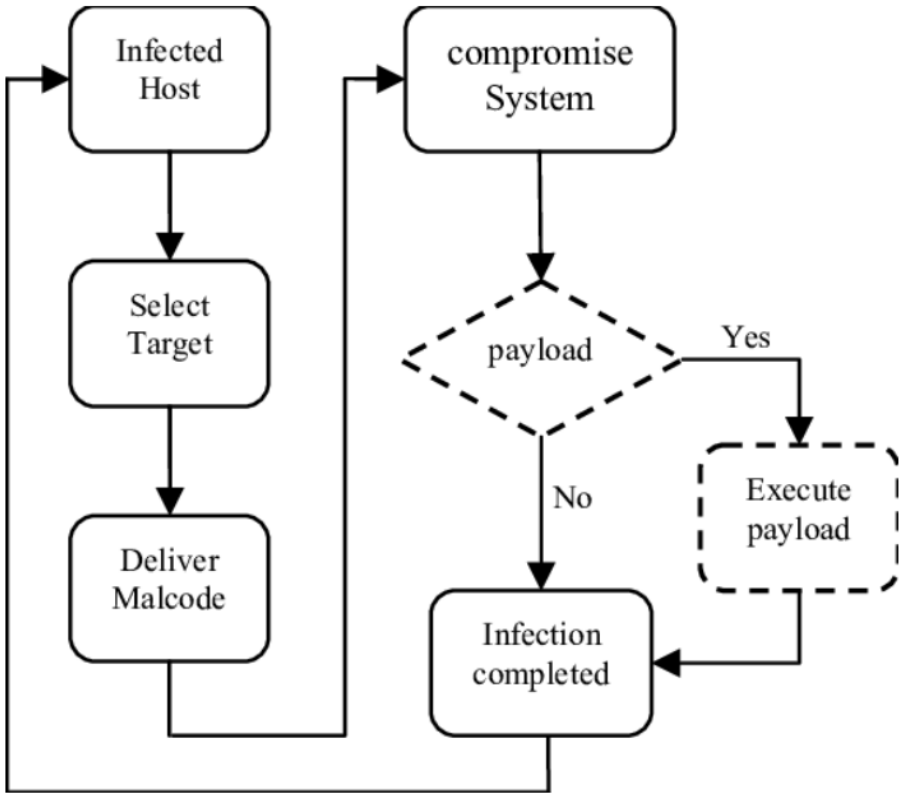
- Sovrascrittura: quando il virus sostituisce il codice del file aggredito con il proprio, eliminando il programma originale.
- Accompagnamento: modifica l'ordine di esecuzione dei file da parte del sistema operativo, posizionandosi prima del file eseguibile infetto.
- Parassitismo: modifica solamente la prima riga di codice del file infetto, impostando come prima istruzione un reindirizzamento all'esecuzione del virus, per poi ritornare all'esecuzione del codice infetto.



A causa della continua evoluzione nel panorama della cybersecurity, i virus spesso sono sviluppati in modo da essere difficilmente riconoscibili, come ad esempio i virus cifrati, che si autocriptano e generano al loro interno una chiave di decrittazione, per utilizzarne una diversa per ogni istanza di virus durante la propagazione; I virus polimorfici che durante la propagazione eseguono la stessa funzione ma al loro interno hanno righe di codice diverse, spesso inutili, per avere dimensione o disposizione di bit differenti ad ogni istanza.

I **Worm**, a differenza dei Virus, sono una tipologia di malware che riesce a propagarsi senza avere bisogno di file "Host" da infettare; hanno la capacità intrinseca di effettuare una copia di loro stessi e diffondersi ad altri dispositivi connessi, utilizzando email, condivisione di file o servizi di rete non sicuri. Il worm copia se stesso in diverse posizioni sul sistema infetto e in seguito si assicura che venga eseguito all'avvio del sistema modificando le chiavi di registro o utilizzando tecniche di persistenza nelle impostazioni del sistema.

Contrariamente ai Virus questi non hanno bisogno di azioni compiute dall'utente per propagarsi, diffondendosi rapidamente attraverso le reti a cui è collegato il sistema.



Alcuni esempi dei metodi di diffusione dei Worm sono i seguenti;

- Posta elettronica o applicazioni di messaggistica: In questo caso il worm invia una copia di se stesso tramite allegato di un messggio.
- Condivisione dati: individuano periferiche come dischi o chiavi usb, per copiarsi al loro interno o infettare altri file come i virus, per poi replicarsi su un altro sistema tramite esecuzione automatica.
- Accesso da Remoto: dove il worm utilizza prevalentemente falle nel sistema o errori nei protocolli di rete per accedere al suo interno.

Un esempio da citare è l'SQL Slammer o Sapphire, worm che ha causato un'epidemia globale nel 2003. Questo worm ha sfruttato una vulnerabilità nel servizio Microsoft SQL Server per diffondersi rapidamente attraverso Internet.

Questo Worm ha avuto effetti devastanti a causa della sua estrema velocità di propagazione per la quale, in soli dieci minuti, ha infettato circa 75.000 sistemi. La sua diffusione ha saturato la larghezza di banda di molte reti, causando interruzioni di servizio e rallentamenti a livello globale.

I **Trojan**, o Trojan horse, sono una tipologia di malware che si presenta come un software legittimo per ingannare gli utenti a scaricarlo e installarlo sui propri sistemi. A differenza dei virus e dei worm, i trojan non si replicano autonomamente; necessitano invece dell'azione dell'utente per diffondersi, sfruttando spesso tecniche di ingegneria sociale. Una volta installati, i trojan possono eseguire una serie di attività dannose, come rubare informazioni sensibili (ad esempio, credenziali bancarie), registrare le attività degli utenti tramite keylogging, o creare backdoor nel sistema per consentire l'accesso remoto da parte degli attaccanti.
I trojan possono anche trasformare il sistema infetto in uno zombie per attacchi DDoS o per far parte di una botnet. Un esempio famoso di trojan è Zeus, noto per il furto di informazioni finanziarie. Una volta installato, Zeus poteva eseguire una serie di attività dannose, tra cui il keylogging per registrare le sequenze di tasti dell'utente, il furto di cookie di sessione per accedere a conti protetti e la creazione di falsi campi di input nei siti bancari per catturare informazioni sensibili.

I **Rootkit** è un insieme di programmi malevoli che hanno come scopo quello di acquisire e mantenere i privilegi di sistema, o root. Essi restano nascosti agli strumenti di sicurezza alterando il comportamento di tool diagnostici e software antivirus, oppure nascondendo la propria presenza utilizzando tecniche avanzate di stealth come la sostituzione di file di sistema e la cifratura del traffico di rete. Oltre alla loro capacità di occultamento, la pericolosità dei rootkit è data anche dal fatto che questi possono agire a diversi livelli di sistema, tra cui il Kernel, dove possono modificare chiamate di sistema, processi, intaccare i firmware ed infiltrarsi in BIOS e UEFI. Uno dei casi di attacco Rootkit più noto è stato quello della Sony BGM nel 2005, quando l'etichetta discografica, per prevenire la copia

illegale dei loro CD, inserì un XCP (Extended Copy Protection) autoinstallante sulle copie vendute. Il programma, oltre a non notificare l'installazione agli utenti, aveva comportamenti di camuffamento tipici dei rootkit, non poteva essere rimosso dal sistema, veniva eseguito in background con consumo di risorse eccessivo e creava molteplici falle nella sicurezza dei computer infettati che potevano essere sfruttati da malware di terze parti. Lo scandalo portò a diverse azioni legali contro l'etichetta discografica e mise il sistema giudiziario americano davanti alla necessità di regolamentare le policy di DRM (Digital Rights Management) delle aziende private.

Spyware

Gli spyware sono una categoria di malware volta a raccogliere informazioni sulle attività degli utenti senza il loro consenso. Questo tipo di malware può monitorare e registrare diverse tipologie di dati, inclusi i tasti premuti (keylogging), la cronologia di navigazione, le credenziali di accesso e altri dati personali sensibili. Altri metodi di attacco sono lo Screen Grabbing, che consisteva nell'acquisire screenshot dello schermo dell'utente; Il Camera Control, tra le minacce più invasive degli Spyware, che ottengono accesso alla webcam del computer e registrano.

Gli spyware possono infiltrarsi nei sistemi attraverso varie modalità, come download nascosti in pacchetti software legittimi, vulnerabilità del browser, o tramite email di phishing. Una volta installati, operano in modo furtivo, raccogliendo e trasmettendo informazioni a terze parti senza lasciare tracce evidenti.

Tra gli spyware più minacciosi troviamo Hawkeye e Agent Tesla, che con molta probabilità appartengono allo stesso gruppo hacker; questi malware, oltre ad essere stati dormienti per un tempo di circa 7 anni, Agiscono come Remote Access Trojan, ottenendo quindi il controllo quasi totale sul sistema.

Phishing

E' una tipologia di attacco volta a trarre gli utenti in inganno per acquisire informazioni sensibili quali Password, Credenziali d'accesso, codici e Pin di carte di credito o altri dati personali. Questa categoria di minaccia informatica si basa su comunicazioni fraudolente con la vittima, mail che possono sembrare vere, messaggi di testo o link ingannevoli, che simulano fonti legittime per ottenere la fiducia dell'utente.

Esistono diverse tipologie di Phishing:

- Email Phishing: È la forma più comune, in cui gli attaccanti inviano email che sembrano provenire da fonti fidate, con link a siti web fraudolenti progettati per rubare informazioni.

- Spear Phishing: Un tipo più mirato di phishing, in cui gli attaccanti personalizzano i messaggi utilizzando informazioni specifiche sulla vittima. Questo tipo di attacco è particolarmente pericoloso nelle aziende, come indicato da Abroshan et al. (2019), poiché può portare alla compromissione di informazioni sensibili aziendali.[12]

- Whaling: Una variante dello spear phishing che attacca individui di alto profilo, come dirigenti aziendali, con messaggi su misura per ottenere informazioni preziose o accesso a risorse critiche.
- Pharming: Coinvolge la manipolazione del sistema DNS per reindirizzare gli utenti a siti web fraudolenti, anche se inseriscono l'indirizzo corretto. Questo attacco sfrutta vulnerabilità di sistema.

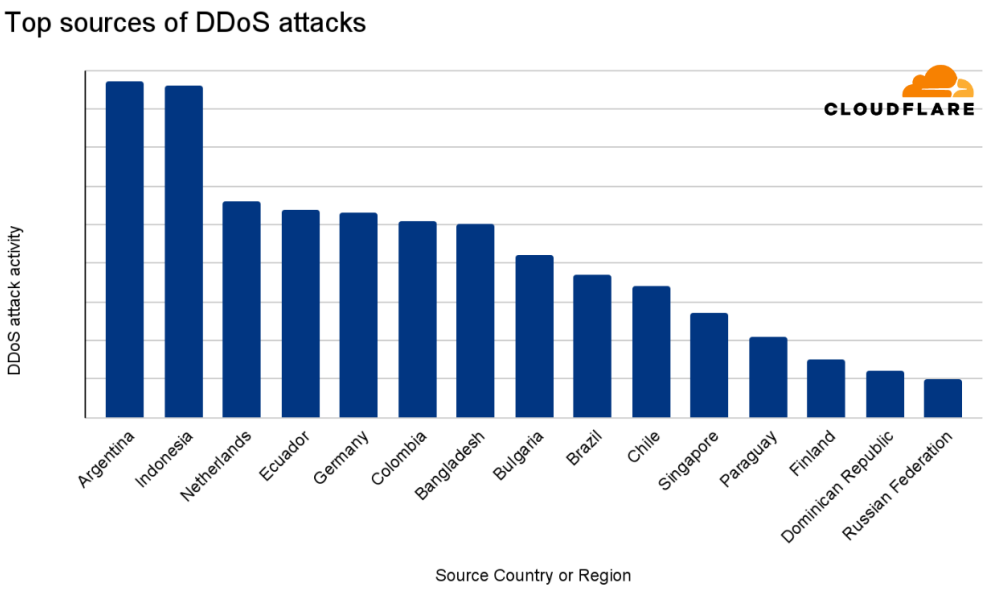
Tra gli attacchi di phishing più noti c'è quello alla Sony Pictures Entertainment; Nel 2014 il gruppo hacker "Guardians of Peace" effettuò un vero e proprio attacco di Whaling che includeva l'allora CEO dell'azienda Michael Lynton, inviando mail apparentemente provenienti da apple che chiedevano ai

[12] 3. Abroshan, H., Dehghantanha, A., Mahmoud, R., & Karimipour, H. (2019). Phishing Attacks Detection through Machine Learning Models and Optimization. In Cyber Defence in Industry 4.0 and the Industrial Internet of Things (pp. 101-123). Springer.

dipendenti Sony di fornire le credenziali di accesso tramite un link di reindirizzamento falso. Grazie a queste informazioni il gruppo Hacker acquisì l'accesso a oltre 100 Terabyte di dati cancellandoli. In seguito vennero rese note le motivazioni del gruppo, associato al governo Nord Coreano, che aveva sferrato l'attacco per impedire la pubblicazione del film "The Interview", commedia in cui veniva orchestrato un attentato al dittatore Kim Jong-un.

Denial Of Service

L'obiettivo di questi attacchi è, contrariamente alle altre minacce, non l'intrusione o l'acquisizione di dati, ma la disabilitazione di un sistema, sito o applicazione o l'incapacità dello stesso a fornire un servizio. Gli attacchi DOS sfruttano la capacità fisica limitata di una macchina di processare molteplici richieste, per portarla al collasso o al crash del sistema; Nello specifico si effettua un sovraccarico del sistema inviando, tramite Botnet, numerose istanze di chiamate o sfruttando vulnerabilità di rete per interrompere il servizio. La capacità di carico di questi attacchi va dai 50 Gbps ai 600 Gbps come l'attacco alla BBC del 2016. Questo tipo di attacchi può essere lanciato contro qualsiasi tipo di risorsa collegata alla rete.



Anche questo tipo di attacchi può essere diviso secondo una tassonomia specifica:

- Flood Attacks:
    1. SYN Flood: Questo tipo di attacco sfrutta il processo di handshake

TCP, sovraccaricando il server con richieste di connessione incomplete (SYN) lasciando in attesa il processo. Il server rimane bloccato aspettando di completare queste connessioni, esaurendo le sue risorse e impedendo nuove connessioni legittime.

2. UDP Flood: Attacco che invia un grande volume di pacchetti UDP (User Datagram Protocol) a porte casuali su un server bersaglio. il server risponde con pacchetti ICMP (internet control message protocol) di tipo "Destination Unreachable", esaurendo le sue risorse.
3. ICMP Flood (Ping Flood): In questo attacco, l'aggressore invia un numero elevato di richieste ICMP Echo (ping) al bersaglio, che deve rispondere a ciascuna, esaurendo la capacità di elaborazione del sistema.

- Attacchi basati su vulnerabilità:
    1. Ping of Death: Consiste nell'invio di pacchetti IP frammentati in maniera maligna o che superano la lunghezza massima del campo di offset del pacchetto (7 byte) al sistema bersaglio, che non riesce a gestirli correttamente, andando in contro a casi di Buffer Overflow e causando un crash o un comportamento anomalo del sistema.[13]
    2. Teardrop Attack: Sfrutta un bug nella frammentazione dei pacchetti IP in sistemi operativi datati (Windows 3.1 o 95 e versioni obsolete di Linux dello stesso periodo), inviando frammenti di pacchetti che il bersaglio non riesce a riassemblare correttamente, portando a un crash del sistema operativo.
- Amplification Attacks:
    1. DNS Amplification: L'attaccante invia richieste DNS con l'indirizzo IP della vittima come sorgente a server DNS aperti, i quali rispondono con grandi quantità di dati, sovraccaricando il bersaglio. Un singolo pacchetto inviato dall'attaccante può generare una

[13] https://it.wikipedia.org/wiki/Ping_of_Death

risposta molto più grande, amplificando l'effetto dell'attacco.

2. NTP Amplification: Simile al DNS amplification, questo attacco sfrutta server NTP (Network Time Protocolconfigurati in maniera errata per inviare un volume di dati molto più grande rispetto alla richiesta originale.

- Distributed Denial of Service (DDoS):
  Distributed poichè gli attaccanti hanno a disposizione una grande quantità di dispositivi infettati, spesso una botnet, per colpire simultaneamente il bersaglio, inoltre in questo modo diventa anche difficile risalire alla fonte dell'attacco.

Uno degli attacchi DoS più famosi e pericolosi è stato l'attacco DDoS contro Dyn avvenuto il 21 ottobre 2016. Dyn è un importante fornitore di servizi DNS (Domain Name System), e l'attacco ha sfruttato una botnet chiamata **Mirai**, composta principalmente da dispositivi IoT compromessi, come telecamere di sicurezza e router. Questo attacco ha generato un volume di traffico massiccio che ha sovraccaricato i server DNS di Dyn, causando interruzioni significative a molti dei più grandi siti web e servizi online, tra cui Twitter, Reddit, Netflix, PayPal, e molti altri.

Capitolo IV: MACHINE LEARNING UTILIZZATO NELLE MINACCE INFORMATICHE

Il cambio di paradigma avvenuto con l'esplosione delle intelligenze artificiali e degli algoritmi di machine learning ha portato non solo ad un incremento delle performance da parte degli agenti a difesa della sicurezza informatica, ma, come era prevedibile, anche ad un' evoluzione delle minacce e degli approcci degli hackers al modo in cui gli attacchi vengono concepiti e implementati; Le potenzialità di queste tecnologie open source vengono sfruttate per creare attacchi dinamici, adattabili e difficili da rilevare, possono proteggere le loro reti botnet o ottimizzare le strategie di profilazione per il phishing. Secondo uno studio del 2020 di Cybersecurity Ventures gli attacchi informatici coadiuvati da AI e ML incrementeranno del 300% entro il 2025.[14] Andiamo ad analizzare nello specifico le possibili applicazioni da parte di agenti malevoli:

Polimorfismo Malware

La prima applicazione critica riguarda la generazione di Malware Polimorfici e Metamorfici. In condizioni normali i malware vengono rilevati attraverso l'identificazione di firme specifiche, ossia righe di codice nel malware note agli agenti di prevenzione, che restavano costanti; Un Malware polimorfico utilizza un "**Mutation Engine**"[15], che riscrive porzioni di codice senza mutare la sua funzione, insieme ad una chiave di cifratura. La mutazione può avvenire attraverso vari metodi, quali Dead-Code insertion, Subroutine reordering o Register Reassignment. I Malware Metamorfici al contrario non sono composti da chiavi di

[14] Cybersecurity Ventures (2020). Cybersecurity Almanac: 100 Facts, Figures, Predictions and Statistics for 2020.

[15] https://en.wikipedia.org/wiki/Polymorphic_engine

cifratura, quindi tutto il corpo del malware cambia invece di generare nuovi Decryptor.

A seguire un esempio di codice Python semplificato che mostra il comportamento di un Malware Polimorfico.

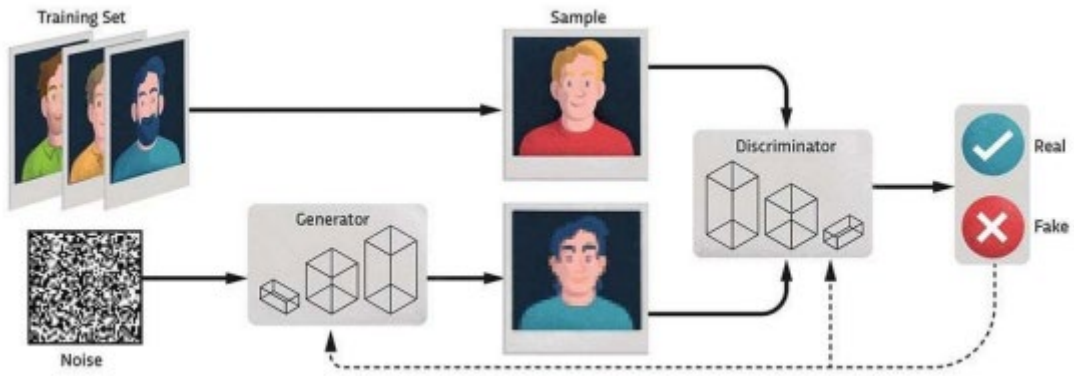
```python
import random
import string

def generate_obfuscated_code():
    # Genera una stringa di caratteri casuali
    random_code = ''.join(random.choices(string.ascii_letters + string.digits, k=10))

    # Codice che potrebbe essere mutato
    code = f'''
def polymorphic_function():
    obfuscated_variable = "{random_code}"
    print("This is a polymorphic function with variable:", obfuscated_variable)
    '''

    return code

def execute_polymorphic_code():
    # Genera il codice
    new_code = generate_obfuscated_code()

    # Scrive il codice in un file temporaneo (simulazione)
    with open("temp_code.py", "w") as file:
        file.write(new_code)

    # Esegue il codice generato
    exec(new_code)

if __name__ == "__main__":
    execute_polymorphic_code()
```

*In cui **generate_obfuscated_code():** genera una stringa casuale di caratteri e numeri che simula il payload malevolo cambiante all'interno della funzione principale.*

***execute_polymorphic_code():** Genera il nuovo codice e lo esegue. Ogni volta che questa funzione viene chiamata, il contenuto della funzione*

polymorphic_function cambia leggermente, mantenendo la stessa logica, ma con un valore variabile diverso.

Lo sviluppo di malware *Polimorfici e Metamorfici* sfrutta tecniche quali *il Reinforced Learning*, tramite il quale il malware riesce ad apprendere quali modifiche al codice siano più efficaci per sfuggire al rilevamento e continuare a propagarsi, oppure le *GAN(Generative Adversarial Networks)* in cui due reti neurali, Generatore e Discriminatore producono e sconfutano varianti del codice malware, nello specifico il generatore tenta di ingannare il discriminatore con codice che sembra non riconoscibile come Malware, Il discriminatore a sua volta diventa più preciso nel rilevare il malware generato. Questo processo continua iterativamente fino a quando il generatore non produce codice malevolo che superi costantemente il discriminatore.



GAN

Tra i malware polimorfici comosciuti troviamo il worm **WannaCry** e il virus **Cryptolocker**.

Blackmamba[16]è un malware polimorfico che genera dinamicamente codice malevolo sfruttando un eseguibile benigno che ha accesso alle API di OpenAI, utilizzato per scrivere il malware direttamente in memoria e quindi rendendolo

[16] https://www.hyas.com/blog/blackmamba-using-ai-to-generate-polymorphic-malware

invisibili ai sistemi di sicurezza. In seguito tira fuori i dati rubati attraverso canali di comunicazione legittimi come Discord o Microsoft Teams.

Phishing e Social Engineering Potenziati dal ML

Gli attacchi di phishing, già tra i più diffusi nel panorama delle minacce informatiche, sono resi ancora più pericolosi dall'uso del machine learning. Algoritmi di Natural Language Processing (NLP) possono analizzare milioni di conversazioni e creare email di phishing su misura per individui specifici, aumentando le possibilità che la vittima cada nella trappola.

Large Language Models possono migliorare la grammatica di una mail di phishing o elaborare messaggi basandosi su informazioni in tempo reale provenienti dal panorama di interesse della vittima.

Nel 2021 la Singapore's Government Technology Agency ha mostrato I risultati di uno studio che ha utilizzato API di OpenAI e AIaaS (Artificial Intelligence as a Service) di analisi della personalità per simulare attacchi di Spear Phishing più efficienti di quelli compiuti in maniera classica.[17]

Anche il **Vishing** o **Voice Phishing**, dove si utilizzano chiamate o registrazioni vocali per ingannare i recipienti, ha beneficiato dell'applicazione di Large Language Models per raccogliere informazioni utili alla social engineering dietro l'attacco, o l'utilizzo di Generative AI per clonare la voce di una persona nota all'attaccante, creando degli audio **Deepfake.**

Secondo un report di Symantec, il 76% delle organizzazioni ha subito almeno un attacco di phishing nel 2021, con una crescente sofisticazione nei contenuti grazie al machine learning.

[17] Turing in a Box: Applying Artificial Intelligence as a Service to Targeted Phishing and Defending against AI-generated Attacks. Eugene Lim et al., GovTech Singapore

Scoperta di vulnerabilità Zero-Day

Le vulnerabilità Zero-Day sono quella tipologia di falle nella sicurezza di un sistema ancora non note o "patchate" da parte degli sviluppatori. La loro minacciosità consiste proprio nel fatto di essere vulnerabilità non identificate, quindi che possono rendere potenzialmente esposti i sistemi per giorni o addirittura settimane. Gli hacker utilizzano il machine learning per automatizzare la scoperta di vulnerabilità zero-day, analizzando grandi volumi di codice sorgente e comportamenti delle applicazioni. I modelli di apprendimento supervisionato possono essere addestrati su dataset di vulnerabilità note per predirre o identificare pattern corrispondenti. Tecniche di Deep Learning possono non solo rilevare le vulnerabilità, ma anche generare automaticamente del codice che sfrutti quella determinata debolezza, rendendo gli attacchi più efficaci e veloci. Inoltre, l'utilizzo del machine learning per creare input avversari (adversarial attacks) consente agli attaccanti di manipolare i sistemi di rilevamento basati su AI, inducendoli a classificare il malware come legittimo.

Un esempio può essere **HPTSA** (Hierarchical Planning and Task-Specific Agents) il team di LLM Agents[18] in grado di riconoscere le vulnerabilità Zero-Day e sfruttarle. In questa ricerca dell'università dell'Illinois viene dimostrato come diversi LLM, ognuno concentrato su una singola task e governati da un manager agent che indirizzava gli attacchi in maniera iterativa verso parti del sistema non ancora aggredite dagli agenti.

Un altro caso di studio rilevante è quello dell'algoritmo **Mayhem**[19], sviluppato durante la DARPA Cyber Grand Challenge, che utilizza tecniche di ML per individuare e sfruttare vulnerabilità in tempo reale, dimostrando il potenziale devastante di queste tecnologie se utilizzate da attaccanti.

[18] arXiv:2406.01637v1 [cs.MA] 2 Jun 2024
[19] https://www.darpa.mil/news-events/2016-08-04

Ottimizzazione attacchi DoS

Nel campo degli attacchi Denial of Service, gli attori malevoli sfruttano algoritmi avanzati di ML e AI per rendere questo tipo di attacchi più efficaci, gestire le botnet per aumentare durata e impatto dell'attacco o adattarsi alle contromisure di sicurezza. Gli Hacker forniscono i loro dataset contenenti informazioni sulla sicurezza dei vari sistemi in pasto ai modelli di machine learning; Ad esempio, un modello di classificazione potrebbe essere addestrato su caratteristiche di rete, come la velocità di risposta e il throughput, per individuare i server più suscettibili a un attacco DDoS. tra i vari modelli possiamo analizzare:

- **Generative Adversarial Networks (GAN)**

  Le GAN possono essere utilizzate per generare un traffico dati realistico che copia un comportamento legittimo aspettato dal Sistema, rendendo difficile il riconoscimento da parte delle contromisure, esaurendo le risorse della rete vittima in maniera più efficiente.
- **Reinforcement Learning (RL) per ottimizzare strategie d'attacco**

  Il reinforcement learning può essere impiegato per ottimizzare la quantità di traffico generato da un attacco DoS. Un agente RL potrebbe apprendere ad esempio, tramite simulazioni, quale sia il volume di traffico ottimale per arrivare ad ottenere l'interruzione del servizio senza superare soglie che potrebbero attivare contromisure automatiche.
- **IoT botnets**

  I dispositivi IoT sono spesso molto suscettibili alle intrusioni esterne, rendendoli il bersaglio "Zombie" ideale per costituire delle botnet. Gli attaccanti possono impiegare algoritmi di apprendimento supervisionato per selezionare dispositivi vulnerabili da compromettere, aumentando così la dimensione della botnet. Algoritmi di apprendimento rinforzato possono essere utilizzati per ottimizzare la distribuzione del traffico generato dagli attacchi, adattando dinamicamente l'intensità e il pattern d'attacco per massimizzare il danno e minimizzare la rilevabilità.

"Recent trends point to the involvement of powerful botnets, such as those powered by Mirai-variant malware, which have been responsible for some of the largest attacks to date. In late 2023, an attack peaked at 350 million packets per second, which was once considered record-breaking but is now becoming more common."[20]

Un'analisi di Nexusguard mostra che gli attacchi DDoS basati su AI hanno una probabilità di successo del 40% superiore rispetto agli attacchi tradizionali.[21]

Evasione degli IDS

L'evasione dei sistemi di rilevamento, come IDS/IPS, è facilitata da tecniche di machine learning come le Generative Adversarial Networks (GANs). Le GANs possono generare traffico di rete che imita perfettamente il traffico legittimo, rendendo estremamente difficile per i sistemi di sicurezza distinguere tra attività malevole e normale. Uno studio condotto da MIT Lincoln Laboratory ha dimostrato che le GANs possono ridurre il tasso di rilevamento degli attacchi di oltre il 30%.

Adversarial Perturbations

Le adversarial perturbations rappresentano una tecnica avanzata utilizzata per evadere i sistemi di machine learning, inclusi quelli implementati per la sicurezza informatica. Si tratta di piccole modifiche intenzionali e quasi impercettibili applicate agli input dei modelli di machine learning che causano errori significativi nella previsione o classificazione del modello.

Perturbazioni Minime: Le perturbazioni vengono progettate per essere minime, ovvero abbastanza piccole da non alterare visivamente o funzionalmente l'input, ma sufficienti a confondere il modello. Ad esempio, una leggera variazione nei

[20] https://socradar.io/global-ddos-attack-landscape-insights-from-q1-2024/
[21] https://www.nexusguard.com/blog/what-is-an-ai-powered-ddos-attack

pixel di un'immagine o nelle caratteristiche di un pacchetto di rete può essere sufficiente a indurre un errore nel modello di rilevamento.

Ottimizzazione dell'Attacco: Gli attaccanti utilizzano tecniche come il Fast Gradient Sign Method (FGSM) o Projected Gradient Descent (PGD) per calcolare le perturbazioni ottimali che massimizzano l'errore del modello. Queste tecniche si basano sul calcolo delle derivate parziali della funzione di perdita rispetto all'input, permettendo di determinare la direzione ottimale per la perturbazione.

Evasione dei Sistemi di Difesa*:* Le perturbazioni avversarie possono essere applicate a input di reti neurali utilizzate per la rilevazione di malware, spam o altre minacce, rendendo queste minacce invisibili ai sistemi di rilevamento. Ad esempio, un malware potrebbe essere leggermente modificato per evitare di essere rilevato come tale, pur mantenendo la sua funzionalità malevola.

Capitolo V: IMPLEMENTAZIONI E VANTAGGI DEL ML NELLA CYBERSECURITY

Nell'ultima decade la branca dell'intelligenza artificiale ha apportato cambiamenti significativi nel campo della sicurezza informatica, permettendo di fronteggiare le minacce che abbiamo analizzato in maniera sempre più rapida ed efficace; Come visto nel capitolo precedente, anche gli agenti malevoli adottano tecniche di machine learning negli attacchi perpetrati, per questo motivo le necessità di sviluppare sistemi adattabili e che apprendano con il tempo diventa cruciale. I modelli di Deep learning, nello specifico le convolutional neural networks (CNNs) e le recurrent neural networks (RNNs), si sono dimostrate in grado di apprendere pattern complessi e relazioni tra dati per identificare in maniera precisa minacce e anomalie; Algoritmi di Machine Learning possono creare modelli di quello che è

un normale tipo di traffico su una rete per poi identificare comportamenti anomali che possono indicare potenziali attacchi; Si possono analizzare pattern comportamentali degli utenti, come tempi di login, modalità di accesso, localizzazione o traffico di rete per identificare anomalie sintomo di account compromessi o attacchi in corso; Per effettuare il training di solito vengono utilizzati dataset noti, raccolta di diversi enti e aziende di cybersecurity, tra questi possiamo trovare il CICIDS2017 per le analisi di rete, MITRE ATT&CK, oppure l' IMPACT, curato da una community (PREDICT) che produce dati riguardanti l'attivita di security sulla rete.

Analizziamo alcuni metodi di machine learning e come vengono applicati contro le minacce alla sicurezza:

Phishing

Essendo gli attacchi di phishing tra quelli che hanno più successo, rafforzare la supervisione grazie a tecniche di machine learning, oltre che educare gli utenti ad una buona pratica di sicurezza, è diventato fondamentale negli ultimi anni. Prendiamo come esempio il sistema di Jain e Gupta22 di rilevamento phishing basato su hyperlink. Il modello rileva URL di phishing grazie a 12 features distinte che analizzano i dati e determinano se sia una minaccia o meno; Gli autori hanno utilizzato algoritmi di Regressione Logistica come classificatore binario per la loro accuratezza nel compito. Per la creazione del training dataset sono stati inseriti 2544 indirizzi web, di cui 1428 erano True Positives. Grazie a questo metodo di apprendimento supervisionato, il modello ha avuto un efficacia del 98.4%.

Sahingoz [23] invece propose allo stesso modo un modello di identificazione phishing basato su URL, composto da un dataset di training di 36.300 siti legittimi e 37.170 indirizzi malevoli; questo modello includeva algoritmi di NLP, vettori di

[22] Jain, A.K., Gupta, B.B.: A machine learning based approach for phishing detection using hyperlinks information. J. Amb. Intell. Human. Comput. 10, 5 (2019)
[23] Sahingoz, O.K., Buber, E., Demir, O., Diri, B.: Machine learning based phishing detection from URLs. Exp. Syst. Appl. 117, 345–357 (2019)

parole e features ibride. In questo caso sono stati testati diversi algoritmi di classificazione: Adaboost, K-Star, Naive Bayes, Random Forest, Decision Tree e K-Nearest (n=3). Il modello ha raggiunto un'accuratezza predittiva del 97.9%, riuscendo a identificare anche gli zero-day phishing URL.

Malware

L'applicazione di algoritmi di machine learning (ML) per la rilevazione e prevenzione dei malware è diventata anch'essa fondamentale nella sicurezza informatica. I modelli di apprendimento automatico, come le reti neurali convoluzionali (CNN) e gli algoritmi di clustering non supervisionato, sono impiegati per identificare pattern anomali nel comportamento del software, distinguendo i malware dai programmi legittimi.

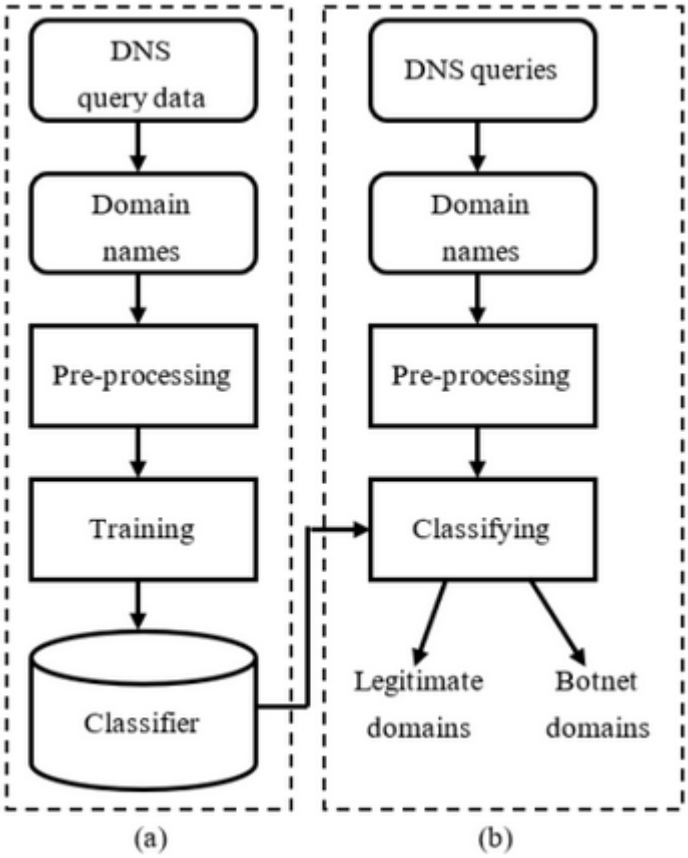
Nel modello di Le[24] il problema dell'identificazione dei malware è stato trasposto in un problema di classificazione, raggiungendo un'accuratezza del 98% con i software malevoli divisi in 9 classi; come training dataset è stato utilizzato il Microsoft Malware Classification Challenge, successivamente hanno scalato il file riducendolo ad una stringa di 10.000 byte, che è stata data in pasto a 3 modelli distinti. La velocità di rilevamento da parte del modello è stata di 0.02 secondi per per sample.

[24] Le, Q., Boydell, O., Namee, B.M., Scanlon, M.: Deep learning at the shallow end: Malware classification for non-domain experts (2018).

Botnet Detection

Xuan Dau Hoang e Quyng Chi Nguyen[25] descrivono le Botnet come un ambiente sempre mutabile. Queste inviano query per ottenere l'indirizzo IP del Command & Control Server (C&C) nei sistemi DNS; Che in contrasto cambiano spesso DNS e indirizzo IP per prevenire queste intrusioni. I due ricercatori hanno sviluppato un modello di machine learning a due fasi per il riconoscimento di queste minacce;

Come si può osservare dallo schema, il modello nella prima fase, quella di apprendimento, durante il pre-processing utilizza metodi di labeling e classificazione per raccogliere dati sulle query DNS dei bot ed estrarre i nomi dei domini che vengono utilizzati dai botmaster per controllare i dispositivi. In seguito, durante il Training, utilizza queste informazioni per apprendere i pattern di azione e identificare le botnet.



Durante la seconda fase, quella di identificazione, vengono analizzati i risultati precedenti per produrre dei classificatori utilizzati per determinare la veridicità o meno dei nomi di dominio. In questo modello sono stati utilizzati algoritmi di Machine Learning supervisionati quali kNN, Naive Bayes, Alberi decisionali e Random Forest; con gli algoritmi kNN e RF che hanno ottenuto rispettivamente il 90% e 88% di successo nel riconoscimento delle botnet.

[25] Xuan, D. H., & Nguyen, Q. C. (2018). Botnet detection based on machine learning techniques using DNS query data. *Future Internet, 10*(5), 43.

Rilevazione DoS

Gaur e Kumar[26] nella loro ricerca hanno proposto una metodologia ibrida che integrava ML e algoritmi di selezione delle caratteristiche (FS) per ridurre il tempo di addestramento e migliorare le prestazioni del sistema. Utilizzando il dataset CICDDoS2019, sono state applicate tecniche di selezione delle caratteristiche, come il test chi-quadrato, l'ANOVA e l'extra tree, su quattro classificatori ML: RF, KNN, DT e XGBoost. XGBoost ha mostrato la massima accuratezza, indipendentemente dagli algoritmi di selezione delle caratteristiche applicati. Iterazioni approfondite di metodi FS hanno ridotto la dimensionalità dei dati, portando XGBoost, insieme al metodo ANOVA FS, a ottenere un'accuratezza del 98,34% con 15 caratteristiche. È stato inoltre suggerito come futuro lavoro l'ottimizzazione degli iperparametri attraverso metodi come grid search, Bayesian optimization e random search, per affrontare problemi di overfitting e underfitting.

Capitolo VI: CASO DI STUDIO

Tra i vari sistemi di Cybersecurity che utilizzano l'intelligenza artificiale troviamo:

- **Sophos' Intercept X Tool**

  Sophos è un'azienda Britannica di cybersecurity. Il loro Sistema, Intercept X, utilizza una Deep neural Network per rilevare minacce informatiche. La particolarità di questo tool è che utilizza l'algoritmo che ha alla base il Cyber Genome Program, progetto della DARPA del 2010 volto a scoprire e catalogare firme e tracce di ogni malware o minaccia esistente.
- **Darktrace Antigena**

[26] Gaur, V.; Kumar, R. Analysis of Machine Learning Classifiers for Early Detection of DDoS Attacks on IoT Devices. Arab. J. Sci. Eng. **2022**, 47, 1353–1374.

Darktrace Antigena è una piattaforma di risposta autonoma alle minacce informatiche, basata su intelligenza artificiale. Utilizzando machine learning, Antigena identifica comportamenti anomali in rete in tempo reale e reagisce automaticamente per contenere minacce prima che si diffondano. La piattaforma si integra con i sistemi di sicurezza esistenti, riducendo drasticamente i tempi di risposta agli incidenti..

- **IBM QRadar Advisor**
  integra l'intelligenza artificiale di Watson per automatizzare il processo di analisi degli eventi di sicurezza, utilizzando algoritmi di machine learning e analisi comportamentale. Tra i benefici, migliora la capacità di correlare minacce complesse in tempo reale e riduce il tempo di risposta agli incidenti. Un caso applicativo rilevante è l'identificazione e la prevenzione di attacchi avanzati come APT (Advanced Persistent Threats) in ambienti aziendali. Gli algoritmi utilizzati includono reti neurali e modelli di apprendimento supervisionato per l'analisi del comportamento delle minacce.

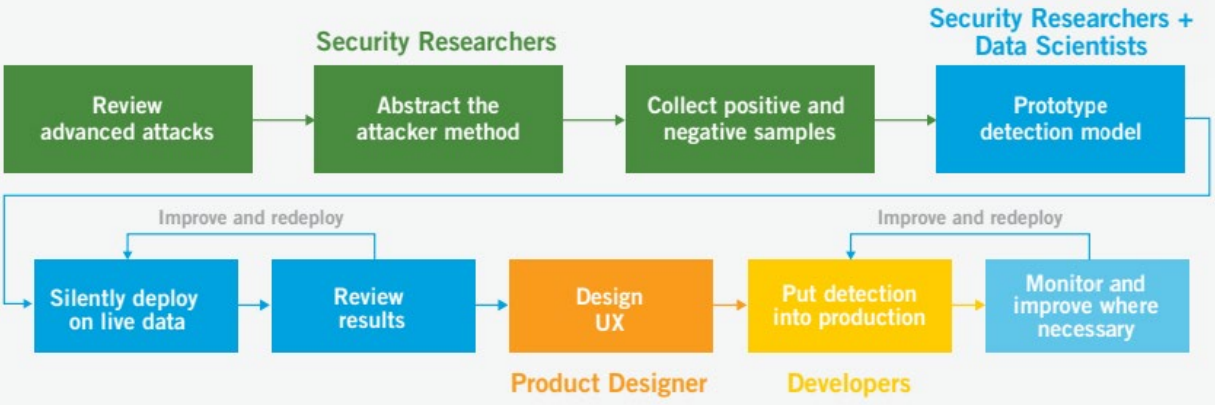
Questo caso di studi prenderà in considerazione la piattaforma *Vectra*.

La *Vectra Cognito Threat Detection and Response Platform* sfrutta algoritmi di machine learning per la gerarchizzazione delle priorità e la prevenzione delle minacce.
La piattaforma spazia in maniera ibrida tra gli ambienti *Azure* e *AWS*[27], comprendendo applicazioni SaaS, IoT e infrastrutture per medie e grandi imprese. La piattaforma include i servizi *Cognito Detect for Network*, volta a identificare minacce nascoste in datacenter e infrastrutture delle imprese; *Cognito Detect for Microsoft Office*, mirato alla sicurezza in ambienti Microsoft Office 365 e Azure AD, ora Microsoft Entra ID; *Vectra Detect e Recall* che collaborano insieme, Detect identifica gli host compromessi in tempo reale come punto di partenza dell'indagine, Vectra Recall trova le minacce che il rilevamento non ha identificato

[27] Amazon Web Services

analizzando i metadati storici. *Vectra Stream*, per il rilevamento e risposta alle minacce in tempo reale che fornisce una visibilità continua del traffico e delle attività di rete inviando metadati ai *SIEM* (Security information and event management) per identificazioni mirate. [28]

Di solito le piattaforme di sicurezza informatica tentano di identificare le minacce grazie ad archivi di firme e comportamenti già esistenti, in questo modo un attore



How Vectra develops security-led AI algorithms.

malevolo può sfruttare questa metodica, come abbiamo già visto, per aggirare le difese e attaccare; La piattaforma Cognito invece effettua un processo di apprendimento sul traffico di rete che può durare da giorni a mesi, arrivando a rilevare minacce comparandole con il normale comportamento previsto su quella rete. Nel caso di attacco coordinato, gli admin sono in grado di riallocare risorse verso la minaccia più importante durante l'intrusione.

l'Intelligenza Artificiale di Vectra utilizza reti neurali e algoritmi di deep learning per avere una panoramica dettagliata in tempo reale del network, con lo scopo di identificare, categorizzare e prioritizzare gli attacchi in corso. Questo processo è dato dal fatto che gli agenti malevoli eseguono spesso più azioni su diversi domini per raggiungere i loro obiettivi. Un algoritmo di correlazione analizza i comportamenti tra account, host, reti e cloud, attribuendo tali comportamenti a punti fissi come account o macchine host. Ad esempio, in ambienti di rete e cloud

[28] https://www.vectra.ai/

ibrido, gli IP temporanei vengono attribuiti a macchine host stabili basate su artefatti[29] osservati tramite un algoritmo chiamato **host-id**. Gli artefatti provengono dai metadati di rete, che includono informazioni come i principali host Kerberos, gli indirizzi MAC DHCP e i cookie, nonché integrazioni API come EDR, vCenter, Azure e AWS. Una volta attribuiti gli artefatti a una determinata macchina host, qualsiasi IP associato può essere correlato al comportamento dell'attaccante, garantendo una tracciabilità più precisa delle minacce.

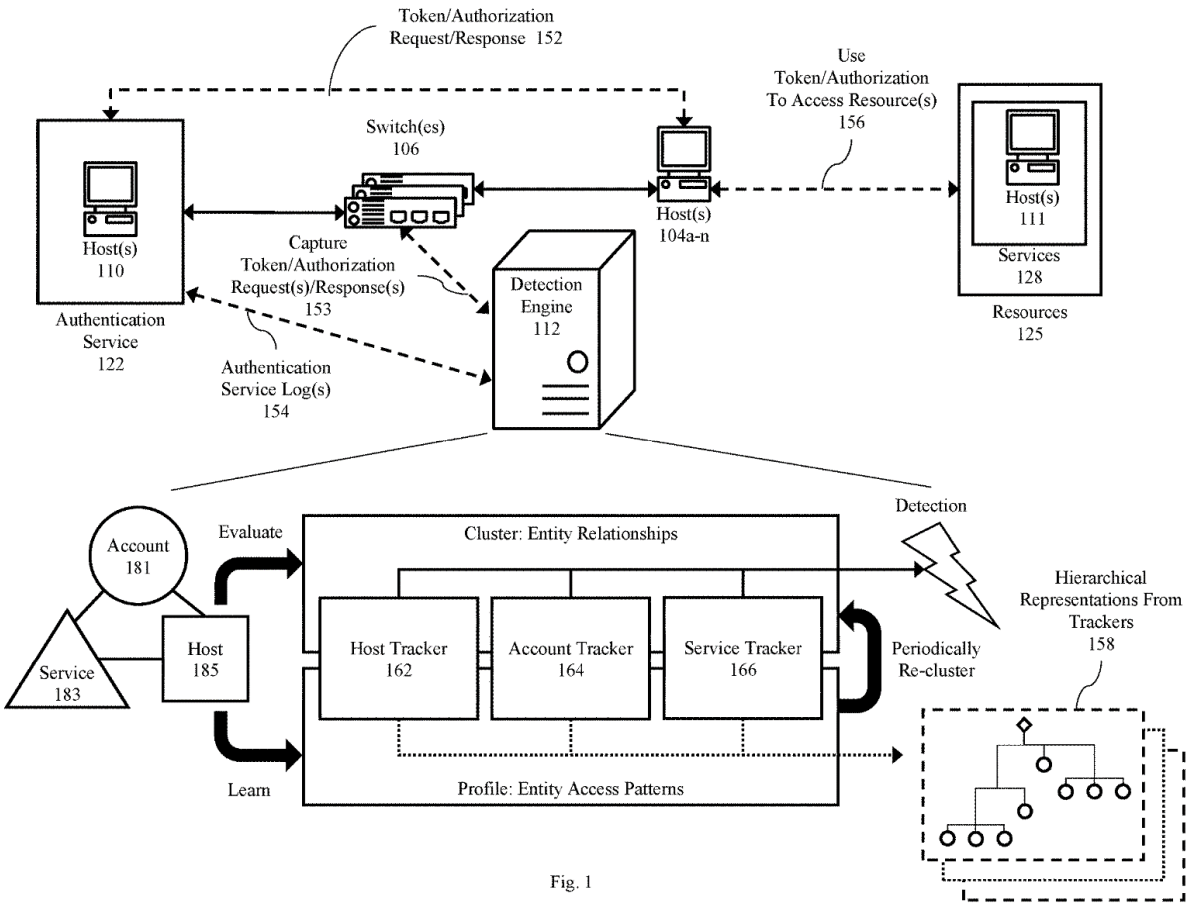


Fig. 1

Vectra Patent US11595416B2, ensemble of hierarchical machine learning models for detection of security risks and breaches in a network.

[29] il termine artefatto si riferisce a qualsiasi dato o informazione estratto da una rete o un sistema che può essere utilizzato per identificare o attribuire attività sospette o malevole.

Nel contesto di AWS, un problema particolare legato all'attribuzione degli attacchi riguarda il modo in cui gli eventi vengono registrati nel pannello di controllo di AWS, dove sono associati a "Assumed Roles" piuttosto che agli account utente reali.

Diversi account possono assumere lo stesso ruolo, rendendo fondamentale identificare l'utente IAM[30] o SAML[31] specifico che ha assunto il ruolo per rispondere efficacemente a un attacco. Gli attaccanti avanzati possono complicare ulteriormente la difesa concatenando i ruoli per nascondere l'origine dell'attacco. Vectra, attraverso una tecnologia personalizzata chiamata **Kingpin**, riesce a risalire lungo questa catena di ruoli per attribuire gli attacchi osservati all'utente reale, piuttosto che a un ruolo ambiguo. Una volta che i comportamenti dell'attaccante vengono attribuiti a un indicatore stabile, essi vengono correlati per identificare il profilo comportamentale sottostante del sistema, il che consente di etichettare e prioritizzare le minacce in corso. L'algoritmo di correlazione è stato progettato per replicare le azioni degli analisti e dei ricercatori di sicurezza di Vectra, fornendo la capacità di classificare scenari di attacco avanzati, come attori esterni o minacce interne a livello amministrativo, per una revisione immediata.

Caso di applicazione: Abuso di privilegio in una rete[32]

- **Metodologie dell'Attaccante**

  Gli agenti malevoli che ottengono credenziali privilegiate acquisiscono accesso a risorse di rete e cloud. Queste credenziali permettono l'intrusione nel sistema senza dover utilizzare malware che possa lasciare tracce o attivare allarmi preventivi. Sebbene imporre requisiti relativi al livello minimo di privilegio per gli utenti possa mitigare alcune attività malevole, non sempre questa è una soluzione percorribile. Il rilevamento di un attaccante che ruba e sfrutta un account spesso diventa una problematica ostica, dato che ogni azione eseguita dall'attaccante è

[30] Identity and Access Management
[31] Security Assertion Markup Language
[32] https://it.vectra.ai/platform

esplicitamente permessa dalle sue autorizzazioni. Un attaccante che ha acquisito familiarità con l'ambiente cercherà di mimetizzarsi, eseguendo azioni che non sono nuove per un account specifico, per evitare di destare sospetti. Per identificare efficacemente l'abuso delle credenziali privilegiate, è necessario un approccio basato sulla sicurezza che consideri ciò che l'attaccante tenta di ottenere con le credenziali rubate.

Observed privilege is a zero-trust view of the normal privilege a user needs to do their job. Use of privilege beyond what is normally necessary warrants additional scrutiny.

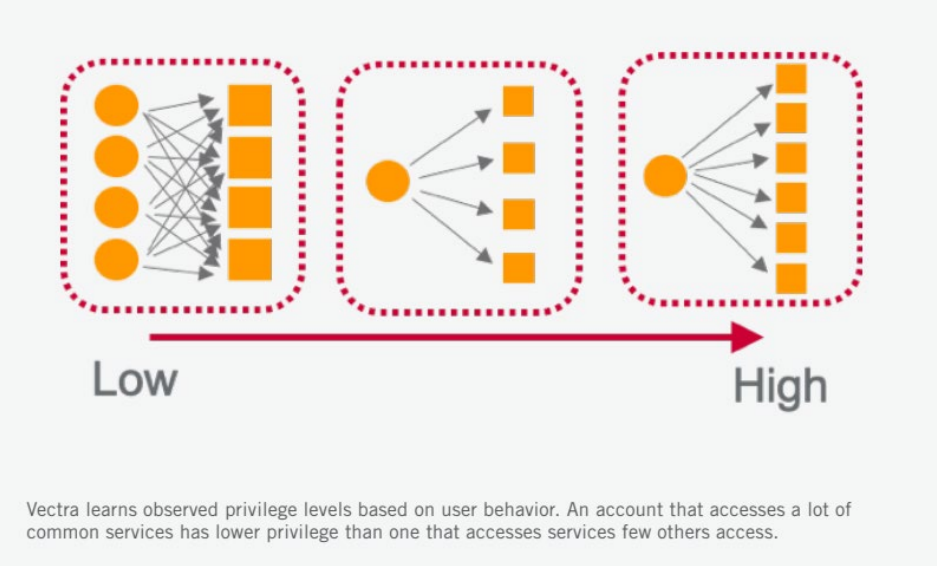
- **Metodologia di Rilevamento**

  . Il valore delle credenziali privilegiate per un attaccante risiede nella capacità di accedere a servizi e funzionalità considerati di alto valore e privilegio nell'ambiente. Il sistema Vectra AI utilizza un metodo differente di attribuzione
  I ricercatori di sicurezza di Vectra hanno identificato che, se si conosce il privilegio effettivo di ogni account, macchina host, servizio e operazione cloud, si può mappare tutte le risorse di alto valore esistenti. Sebbene i concetti di privilegio concesso siano ben consolidati, questa rappresentazione fornisce un limite superiore a ciò che è il vero privilegio rispetto al minimo necessario. Invece, il team di ricerca e il team di data science di Vectra hanno identificato un nuovo modo di rappresentare il valore dei sistemi in un ambiente, basato su ciò che è stato osservato nel tempo. Questa visione dinamica e concreta del valore è chiamata

privilegio osservato. Questa visione basata sui dati del privilegio fornisce un efficace approccio zero-trust all'uso delle credenziali senza configurazioni manuali.

L'IA di Vectra calcola il ***privilegio osservato***[33], considerando le interazioni storiche tra le entità monitorate, anziché basarsi sul privilegio definito da un amministratore IT. La variazione tra accessi e uso contribuiscono fortemente al punteggio assegnato al privilegio osservato. Un sistema che accede a molti altri nodi, normalmente non acceduti da altri, otterrà un punteggio di privilegio alto, mentre un sistema con accessi comuni avrà un punteggio basso. Questo approccio permette di differenziare tra account amministrativi e account utente normali nel dominio.

Una volta calcolati i punteggi di privilegio osservato, tutte le interazioni tra account, servizi, host e operazioni cloud vengono mappate per comprendere le interazioni storiche normali tra i sistemi. Successivamente, una suite di algoritmi di apprendimento non supervisionato che considerano i punteggi di privilegio viene utilizzata per identificare casi anomali di abuso di privilegio, con l'impiego di algoritmi di rilevamento anomalie personalizzati e implementazioni del **Clustering Gerarchico Basato sulla Densità con Rumore** (HDBSCAN).



Vectra learns observed privilege levels based on user behavior. An account that accesses a lot of common services has lower privilege than one that accesses services few others access.

La metrica del privilegio osservato focalizza il rilevamento sulle azioni anomale rilevanti, garantendo una maggiore precisione e recall rispetto a un approccio che ignora questa prospettiva critica.

[33] un metodo di monitoraggio dinamico che mette in relazione i privilegi conferiti dall'admin con i nodi di risorse a cui quel livello di privilegio può accedere

CONCLUSIONE

Le minacce informatiche sono in continua evoluzione, rendendo i modelli di machine learning addestrati su dati storici potenzialmente obsoleti. L'implementazione di meccanismi di apprendimento continuo permette al modello di adattarsi e aggiornare le proprie conoscenze con nuove informazioni sulle minacce in tempo reale, migliorando così l'accuratezza nell'identificazione delle minacce emergenti. Gli attacchi adversariali sono tecniche utilizzate dagli attori malevoli per eludere il rilevamento da parte dei modelli di machine learning. L'apprendimento automatico adversariale si concentra sullo sviluppo di modelli robusti capaci di resistere a tali attacchi. Tecniche come l'adversarial training, la distillazione difensiva e la sanitizzazione degli input possono essere impiegate per migliorare la resilienza dei modelli.

Mentre i modelli di machine learning possono automatizzare e potenziare l'identificazione delle minacce, l'esperienza umana rimane fondamentale. Integrare input umani e conoscenze di dominio può aiutare a validare e interpretare le previsioni dei modelli, migliorando l'accuratezza e riducendo i falsi positivi/negativi. La valutazione periodica delle prestazioni dei modelli di machine learning è essenziale per identificare aree di miglioramento. Il feedback di analisti ed esperti di cybersecurity può essere utilizzato per affinare e perfezionare i modelli, garantendo un'accuratezza crescente nel tempo.

È importante sottolineare che, sebbene il machine learning possa migliorare significativamente l'identificazione delle minacce informatiche, esso dovrebbe essere parte di un framework di sicurezza informatica comprensivo, che includa altre tecniche come il monitoraggio della rete, i sistemi di rilevamento delle intrusioni, le pratiche di codifica sicura e la formazione sulla consapevolezza degli utenti. L'uso del machine learning nei meccanismi di difesa proattiva segna un cambio di paradigma nella sicurezza informatica. Sfruttando la potenza delle

analisi avanzate e del riconoscimento dei pattern, il machine learning contribuisce a una postura difensiva più intelligente, adattativa ed efficiente. La capacità di rilevare anomalie sottili, automatizzare le risposte e apprendere continuamente dalle minacce in evoluzione posiziona il machine learning come un pilastro nelle strategie di sicurezza informatica moderne.

In conclusione, l'integrazione del machine learning nei meccanismi di difesa proattiva rappresenta una necessità per le organizzazioni che vogliono restare in vantaggio nei confronti delle minacce alla sicurezza. La sfida principale sarà quella di mantenere un equilibrio tra l'evoluzione delle tecnologie di difesa e le sempre più sofisticate tecniche di attacco supportate dall'intelligenza artificiale.

In sintesi, il machine learning non è più solo un'arma a favore dei difensori, ma anche un potente strumento nelle mani degli attaccanti, capace di amplificare la portata e l'efficacia delle minacce informatiche. Questa nuova realtà richiede un ripensamento delle strategie di sicurezza e una continua innovazione per mantenere la resilienza delle infrastrutture digitali.

Bibliografia

- Tang, Q., Zhang, H., Li, Y., & Liu, J. (2020). "A Survey of Malware Detection Techniques".
- Jones, D., Smith, K., & Lee, R. (2022). *Spyware Threats and Privacy Implications*. Journal of Cybersecurity, 17(2), 99-113.
- Jain, A.K., Gupta, B.B.: A machine learning based approach for phishing detection using hyperlinks information. J. Amb. Intell. Human. Comput.
- Xuan, D. H., & Nguyen, Q. C. (2018). Botnet detection based on machine learning techniques using DNS query data. *Future Internet, 10*(5), 43.
- Cybersecurity fundamentals study guide -- Information Systems Audit and Control Association – 2017
- Computer Security_ Principles and Practice, 4th, Global -- Lawrie Brown, William Stallings -- 4th, 2018 – Pearson
- Investigating the applications of artificial intelligence in -- Abbas, Naveed Naeem; Ahmed, Tanveer; Shah, Syed Habib Ullah
- Machine learning for computer and cyber security_ -- Gupta, Brij; Sheng, Quan Z -- 2019 -- CRC Press
- Handbook of Research on Machine and Deep Learning -- Padmavathi Ganapathi (editor), D. Shanmugapriya (editor) -- 2020 -- Information Science Reference

Sitografia

https://it.stemkb.com/machine-learning

https://it.stemkb.com/machine-learning/reti-neurali/

https://www.itgovernance.eu/blog/en/the-5-biggest-phishing-scams-of-all-time

https://demakistech.com/how-hackers-use-ai-and-machine-learning-to-target-enterprises/

https://www.techtarget.com/searchsecurity/tip/Generative-AI-is-making-phishing-attacks-more-dangerous

https://www.nexusguard.com/blog/what-is-an-ai-powered-ddos-attack

https://it.vectra.ai/platform

https://ieeexplore.ieee.org/document/9254703

https://www.techtarget.com/searchsecurity/definition/metamorphic-and-polymorphic-malware