Part I: Selecció del conjunt de dades

Autor: Josep Rau

Assignatura: M2.989 - Arquitectures de bases de dades no relacionals - Aula 1

Índex

Descripció del conjunt de dades

El conjunt de dades escollit per a la pràctica final és “Global Cybersecurity Threads (2015-2024)”. Aquest conjunt de dades proporciona informació de ciberatacs ocurreguts entre 2015 i 2024 en l’ àmbit empresarial. Conté dades sobre els països afectats, les indústries afectades, el tipus d’ atac utilitzat, etc. El conjunt de dades está compost per 3000 files i 10 columnes.

Nom	Global Cybersecurity Threads
Files	3000
Columnes	10
Format	Csv
Font	Kaggle
URL	https://www.kaggle.com/datasets/atharvasoundankar/global-cybersecurity-threats-2015-2024?resource=download
Llicència	CC0 - Public Domain
Última actualització	Març 2025

Justificació

M’ he decantat per aquest conjunt de dades perquè tracta d’ un tema actual, en aquest cas la ciberseguretat i les seves amenaces. En el meu àmbit professional, constantment rebem correus electrònics maliciosos amb els que s’ intenta poder accedir als nostres sistemes i atacar. Donat que ho veig en el meu dia a dia, m’ interessa analitzar l’ evolució dels atacs cibernètics en els últims anys, si hi ha hagut alguna indústria més propensa a ser víctima d’atacs, quins mètodes de defensa són més efectius, etc. Finalment, un altre motiu per el qual em motiva aquest tema, és que mentre jo estudiava a la UAB, hi va haver un ciber atac que va fer caure tot el sistema de la universitat durant un parell de dies fins que van complir les demandes econòmiques.

Rellevància

Amb l’auge de les tecnologies i la digitalització de les dades, les amenaces de ciberatacs incrementen també. Aquest conjunt de dades engloba el període de 2015 fins a 2024 a nivell mundial, per tant, són dades força actuals. Tot i que aquest conjunt de dades només contempla atacs a empreses, és un tema que afecta a empreses, institucions públiques i ciutadans, de manera directa o indirecta. Depenent de l’objectiu de l’atac poden haver implicacions econòmiques o de privadesa de les dades, com ja ha succeït en més d’una ocasió.

Complexitat

És un conjunt de dades que compte amb 3000 registres i 10 columnes, les que combinen atributs quantitatius i categòrics. A continuació es detalla la composició del conjunt de dades:

Variable	Tipus	Descripció	Exemple
Country	Categòric	País de l'atac	China
Year	Quantitatiu	Any de l'atac	2019
Attack Type	Categòric	Tipus d'atac (malware, phishing, DDoS...)	Phishing
Target Industry	Categòric	Indústria atacada (Retail, IT, Banking...)	Education
Financial Loss (in Million $)	Quantitatiu	Cost econòmic associat a l'atac	80.53
Number of affected users	Quantitatiu	Nombre d'usuaris afectats per l'atac	773169
Attack Source	Categòric	Origen de l'atac (Hacker Group, Insider, Unknown...)	Hacker Group
Security Vulnerability Type	Categòric	Vulnerabilitat del sistema (Social Engineering, Zero-day...)	Unpatched Software
Defense Mechanism Used	Categòric	Mecanisme de Defensa Utilitzat (VPN, Firewall, Encryption...)	VPN
Incident Resolution TIme (in Hours)	Quantitatiu	Temps de resolució en hores	63

Encara que sigui un conjunt de dades simple, es pot observar com és bastant complet en quant a atributs per contextualitzar. Hi ha tres mètriques disponibles, com són el cost econòmic, el nombre d'usuaris afectats i el temps de resolució, que es poden analitzar des de diferents perspectives (per any, per indústria, per tipus d'atac, etc.)

Originalitat

És un conjunt de dades extret de Kaggle i que ha estat objecte d'anàlisi per usuaris de la plataforma. També és un tema que es tracta sobretot en àmbits laborals amb formacions i també en àmbit periodístic. No obstant, no és un tema que hagi estudiat ni analitzat fins ara, i veig aquestes dades amb potencial per treballar les diferents tècniques de visualització que hem vist durant el curs.

Tot i tenir un nombre just d' atributs, són suficients per realitzar un anàlisi profund i poder extreure'n conclusions. Per a completar el conjunt d'atributs es pot obtenir una mètrica com *Pèrdues financeres per minut*, ja que ens permetria analitzar quins atacs penalitzen econòmicament. Un altre atribut podria ser la severitat segons el nombre d' usuaris afectats, la qual es podria categoritzar en baixa, mitjana, alta o crítica.

Per altra banda, aquest conjunt es podria complimentar amb altres datasets que continguin informació sobre inversions en ciberseguretat de cada país i analitzar quins resultats estant tenint.

Qüestions

Amb aquest conjunt de dades em plantejo respondre les preguntes següents:

- Quina és la tendència dels ciberatacs en els últims 10 anys ?
- Quins són els països més afectats? I quin tipus d'atac és més comú a cada país?
- Quin impacte econòmic tenen els atacs ? Quin implica més diners?
- Quin sector és el més atacat ? I quina és la vulnerabilitat més comuna en cada sector?
- Quina vulnerabilitat és més perillosa en quant a nombre d'usuaris afectats?
- Quin tipus d'atac requereix més temps per solucionar ?
- Quines són les solucions més efectives en base al temps de resolució?

La meva idea a dia d' avui és presentar tota aquesta informació en un format que permeti a l'usuari interactuar i extreure'n les seves pròpies conclusions. He mostrat algunes de les preguntes que es podran resoldre amb la visualització, però al fer-ho interactiu, aquestes dependran també de les intencions i creativitat de l'usuari.