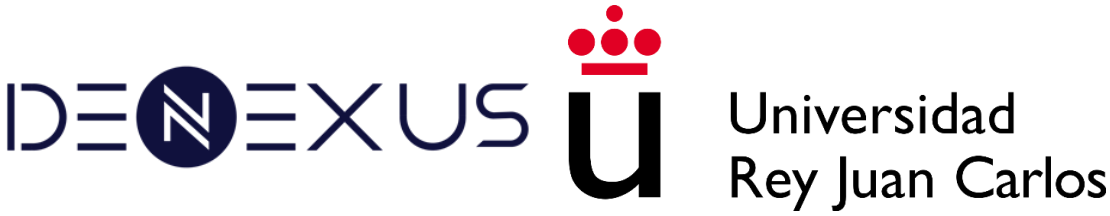
Convocatoria Colaboración Público-Privada CPP2021-009025

DICYME:

Dynamic Industrial CYberrisk Modelling based on Evidence

Iniciativa Conjunta de:



ENTREGABLE 2.4:

Documentación de los modelos y métricas para la medición/estimación de probabilidad e impacto de incidentes OT

Coordinadores:

Romy R. Ravines (DeNexus Tech)
Isaac Martín de Diego (Universidad Rey Juan Carlos)
Alberto Fernández Isabel (Universidad Rey Juan Carlos)

Contenido

1 INTRODUCCIÓN Y OBJETIVOS 3

2 ATR2ATK: Atractivo de una organización a ciberataques 3

2.1 Modelo Random Forest 4

2.2 Cálculo de la reputación online 5

2.2.1 *Engagement* de la entidad 6

2.2.2 *Engagement* de los usuarios 7

2.3 COMBINACIÓN DE VÍCTIMAS 8

3 THRACT: Actores de amenazas y víctimas 9

4 CVE2TTPs 10

5 CUANTIFICACIÓN DEL CIBERRIESGO 11

5.1 Frecuencia de eventos de pérdida (Loss Event Frequency, LEF) 12

5.2 Magnitud de la pérdida (Loss Magnitude, LM) 13

6 CONCLUSIONES Y SIGUIENTES PASOS 13

1 INTRODUCCIÓN Y OBJETIVOS

Este documento constituye el entregable final de la Actividad 2 del proyecto DICYME, centrada en la definición y documentación de los elementos clave de los modelos de estimación de la probabilidad e impacto de incidentes OT, detallados en los entregables E2.1, E2.2 y E2.3, y disponibles en *deriskGroup / DICYME Project · GitLab*.

El objetivo principal ha sido sistematizar la información técnica necesaria para el diseño de los módulos de cálculo, a partir de evidencias y datos empíricos recolectados en fases anteriores. Se ha prestado especial atención a la documentación de variables, fuentes de datos, criterios de selección, métricas relevantes y enfoques utilizados para su validación.

Este entregable sintetiza aspectos metodológicos de componentes desarrollados en los modelos, como:

- Los criterios y métricas empleados para cuantificar el atractivo de una organización como objetivo de ciberataques.
- El enfoque seguido para estimar la reputación online y su rol como factor modulador de riesgo.
- El tratamiento de datos firmográficos y su integración a través del modelo de combinación de víctimas.
- La documentación del indicador THRACT para evaluar actores de amenazas en relación con una víctima, así como del modelo CVE2TTPs para traducir vulnerabilidades a técnicas MITRE ATT&CK.
- Y, de forma transversal, la estructura cuantitativa de estimación del ciberriesgo, entendida como producto entre probabilidad de ocurrencia y severidad de impacto económico.

2 ATR2ATK: Atractivo de una organización a ciberataques

Como se ha introducido a lo largo de los entregables E2.1, E2.2 y E2.3, este modelo cuantifica la probabilidad de que una entidad sea víctima de un ciberataque. Para ello, se ha creado un conjunto de datos con diversas variables que definen a la entidad:

- Datos firmográficos: país, industria, facturación anual, ganancias anuales, número de empleados, si son lucrativas, si cotizan en bolsa.
- Reputación online.
- Victimización: dispositivos visibles en Internet y aparición en *leaks* de *ransomware*.

Todos los datos se obtienen de fuentes abiertas, aunque en algunos casos haciendo uso de herramientas de *Software-as-a-service* de pago costeadas por DeNexus que se alimentan de fuentes públicas, estructurando, normalizando y facilitando el acceso a sus datos. Es el caso de los datos firmográficos y la reputación online, aunque también se han integrado fuentes de datos gratuitas para el primer caso.

En el prototipo final, descrito en el entregable E2.3, el modelo de Machine Learning empleado para el cálculo del indicador es Random Forest, utilizándolo con todas las variables previamente mencionadas. En la *Sección 2.1 Modelo Random Forest* se aporta más documentación y detalles del proceso de entrenamiento y validación del mismo. Además, en este entregable merece especial mención el cálculo de la reputación online, pues se introdujo ya en el primer prototipo basándose en el estado del arte y publicaciones científicas del dominio del marketing y la cuantificación de la presencia de una empresa para con la ciudadanía, sus clientes, su competencia, etc. Este se describe en la *Cálculo de la reputación online*.

2.1 Modelo Random Forest

Desarrollado en el lenguaje de programación R, se ha implementado un flujo completo de entrenamiento y validación de un modelo de clasificación Random Forest para el cálculo del indicador de atractivo, aplicando una búsqueda exhaustiva de hiperparámetros y utilizando validación cruzada. Dado el dataset, que contiene empresas que han sido víctimas de ciber incidentes y otras de las que no se tiene constancia que lo hayan sufrido en el periodo recogido (diciembre 2023 – diciembre 2024), el objetivo es predecir la variable binaria `Incidente`.

En primer lugar, se cargan los datos y se convierten las variables a los tipos deseados, como factores, numéricos, etc., de manera que el código posterior trate adecuadamente cada variable, con los diferentes niveles existentes para cada una de ellas. Posteriormente, se define una cuadrícula (*grid*) con combinaciones de hiperparámetros de Random Forest:

- `mtry`: número de variables consideradas en cada división del árbol.
- `splitrule`: criterio de división.
- `min.node.size`: tamaño mínimo del nodo hoja.
- `sample.fraction`: fracción de datos usada para entrenar cada árbol (submuestreo).
- `num.trees`: número total de árboles en el bosque.

A continuación, cada combinación de esta cuadrícula se evalúa por separado utilizando validación cruzada de 5 iteraciones (`K = 5`). En este esquema, el conjunto de entrenamiento se divide en cinco partes: en cada iteración, cuatro de ellas se utilizan para entrenar el modelo y la quinta para validarlo. Este proceso se repite cinco veces, de manera que cada subconjunto actúa una vez como conjunto de validación. Así, cada combinación de hiperparámetros se somete a su propia validación cruzada, lo que permite comparar su rendimiento de forma justa y seleccionar la más adecuada. En este proceso, se buscó maximizar la métrica F1 Score, para obtener el mejor balance entre precisión (*accuracy*) y sensibilidad (*recall*). Esta estrategia busca que el modelo sea capaz de identificar correctamente los incidentes reales (alto *recall*), al tiempo que minimiza las falsas alarmas (alto *accuracy*). Como resultado de todo este proceso, se obtuvieron los siguientes parámetros:

- mtry = 2
- splitrule = "gini"
- min.node.size = 1
- sample.fraction = 0.5
- num.trees = 10

Finalmente, se evaluó el modelo, generando su matriz de confusión. Los resultados obtenidos se recogen en la *Tabla 1. Métricas de rendimiento del nuevo modelo ATR2ATK*.

Tabla 1. Métricas de rendimiento del nuevo modelo ATR2ATK

Métrica	Entrenamiento	Test
Accuracy	86%	64%
Precision	87%	73%
Recall/Sensitivity	95%	80%
F1-score	91%	76%

Además, se evaluó el modelo también en un conjunto de datos extraído *ad-hoc* de una muestra de clientes de DeNexus en 3 periodos temporales diferentes, para tratar de visualizar los cambios que se producen en el indicador en diferentes instantes. Cabe recordar que la componente temporal se introduce principalmente en la reputación online, que es el valor que cambia más frecuentemente, ya que el resto de las variables son mucho más estáticas (aunque también pueden variar, pero en menor rango). Los datos y resultados de estas entidades no se incluyen por confidencialidad, pero efectivamente se pudo validar cómo el indicador variaba según difería la reputación online. Además, se extrajeron conclusiones muy valiosas sobre el funcionamiento del indicador para diferentes industrias, países y tamaños de empresas, esenciales para el posterior desarrollo e integración del indicador en DeRISK. Asimismo, sirvió para el desarrollo del nuevo modelo de combinación de datos de víctimas, ideado para poder automatizar y recoger a aún mayor escala datos de entidades afectadas que sirvan para refinar el entrenamiento de este modelo.

2.2 Cálculo de la reputación online

La **reputación online** se refiere al nivel de aceptación y reconocimiento que tiene una entidad (como una empresa, institución o persona) entre el público en general. Este concepto se basa en la idea de que una entidad puede resultar más atractiva para un adversario dependiendo de su reputación online, que se define como el resultado de todo lo que los usuarios, clientes o empleados escriben, comunican o comparten en Internet en cualquier momento de su relación, directa o indirecta, con la entidad. Esta percepción puede influir en la atención que reciba una entidad en un momento determinado, y puede variar con el tiempo.

En este contexto, se hace una distinción entre:

- **Medios sociales (*social media*)**, que son contenidos dinámicos compartidos a través de plataformas controladas por la entidad (como su sitio web corporativo,

blogs o podcasts). Aunque es posible la interacción con los usuarios, no es frecuente ni esperada.

- **Redes sociales (*social networks*)**, que son plataformas externas a la entidad (como X -antiguo Twitter-, Facebook, Reddit o LinkedIn), donde el contenido compartido genera mayor interacción con los usuarios, y donde esta participación es esperada y habitual.

Las siguientes plataformas se han considerado para estimar la reputación online, siendo estas las diferentes fuentes de datos disponibles en la plataforma Determ, empleada para la búsqueda de menciones a las entidades del conjunto de datos:

- X (Twitter)
- Tripadvisor
- Facebook
- Reddit
- Instagram
- Sitios web
- Comentarios en sitios web
- TikTok
- YouTube
- Foros online

Para calcular la reputación online se ha utilizado la herramienta *Determ* y se ha formulado un modelo específico. Este modelo parte del concepto de $Engagement(E)$, que se define como la relación entre la $Interacción(I)$ y el $Alcance(R)$:

$$Engagement\ (E) = \frac{Interaction\ (I)}{Reach(R)}$$

Por ejemplo, si una publicación ha sido vista 100 veces y ha recibido una sola interacción (como un comentario), el $Engagement$ será:

$$Engagement\ (E) = \frac{1}{100} = 0.01$$

La **reputación online (OR)** es dinámica y dependiente del tiempo, ya que representa una captura estática de un momento concreto, influenciada por los periodos previos. Por ello, se calcula el indicador temporal OR^t para cada red y medio social, siendo t el instante temporal actual. Esta reputación se construye como una combinación ponderada del *engagement* generado por la entidad EE^t y el generado por los usuarios externos UE^t:

$$OR^t = \alpha \cdot EE^t + (1 - \alpha) \cdot UE^t$$

2.2.1 *Engagement* de la entidad

El *engagement* de la entidad EE^t se calcula como la media ponderada del compromiso en cada red social o medio donde la entidad participa:

$$EE^t = \frac{1}{E} \cdot \sum_{j=1}^{E} \left[\frac{EE_j^t}{max_{z=1\ldots t}\ EE_j^z} \right]$$

donde E es el número total de medios o redes sociales donde participa la entidad y j representa cada una de esas fuentes, de forma que EE_j^t significa el *engagement* de la entidad en el instante t y la red o medio social j, y $max_{z=1\ldots t}\ EE_j^z$ es el máximo *engagement* de la entidad alcanzado en estimaciones previas.

A su vez, EE^t se define como:

$$EE_j^t = \frac{1}{n_j^t} \cdot \sum_{i=1}^{n_j^t} \frac{IC_i^t}{RC_i^t}$$

siendo n_j^t la cantidad total de menciones en el medio o red social, mientras que IC_i^t y RC_i^t son las interacciones y alcance de la fuente j en cada mención i en el instante t

2.2.2 *Engagement* de los usuarios

El *engagement* de los usuarios se calcula de forma similar, considerando únicamente menciones externas a la entidad:

$$UE^t = \frac{1}{U} \cdot \sum_{k=1}^{U} \left[\frac{UE_k^t}{\max_{z=1\ldots t} UE_k^z} \right]$$

donde U es la cantidad de sitios donde hay menciones para la entidad en cuestión, y k representa cada una de estos sitios o redes sociales. Por tanto, UE_k^t refleja el *engagement* de los usuarios en el instante t para la fuente k, y and $\max_{z=1\ldots t} UE_k^z$ es el máximo *engagement* de los usuarios calculado hasta el momento para la entidad.

También se puede definir como sigue:

$$UE_k^t = S_k^t \cdot \frac{1}{n_k^t} \cdot \sum_{p=1}^{n_k^t} \frac{IU_p^t}{RU_p^t}$$

siendo n_k^t la cantidad total de menciones en el foro k, mientras que IU_p^t y RU_p^t son las interacciones y alcance, respectivamente, en la fuente k para cada mención p en el momento t. S_k^t Estima el sentimiento en cada mención de la fuente k y el periodo t.

La estimación del sentimiento para cada medio o red social k se realiza de la siguiente manera:

$$S_k^t = \frac{positive_k^t - negative_k^t}{positive_k^t + negative_k^t}$$

donde $positive_k^t$ es el número de menciones categorizadas como positivas para la fuente k en el periodo t, y $negative_k^t$ es el número de menciones clasificadas como negativas. Esta fórmula produce un valor entre -1 (totalmente negativo) y +1 (totalmente positivo). Cabe mencionar que no se considera el sentimiento en las publicaciones de la entidad EE^t, ya que se asume que dicho contenido es controlado y, por tanto, tiene una orientación positiva por defecto.

El valor final de la **reputación online** estará en el rango $[-1, +1]$:

- Un valor cercano a **-1** indica una **mala reputación online**, con menciones negativas y baja interacción.
- Un valor cercano a **+1** indica una **buena reputación online**, con menciones positivas y alta interacción.
- Valores cercanos a **0** reflejan neutralidad o falta de relevancia.

2.3 COMBINACIÓN DE VÍCTIMAS

En el entregable E2.3. se menciona un modelo que es capaz de obtener datos firmográficos a partir de un ciber incidente, pudiendo obtener datos sobre entidades víctimas que pueden ser de interés. El modelo consulta distintas fuentes externas de datos y realiza una combinación final de los datos, seleccionando la fuente más confiable y combinando campos de otras fuentes en caso de que la fuente seleccionada carezca de algún campo.

A los datos obtenidos de las fuentes externas se les asigna un valor de confianza numérico entre 0 y 1 que permite indicar que tan confiable es la fuente. Este valor de confianza se basa en la cantidad de campos encontrados en lugar de la calidad de los datos ya que evaluar la calidad resulta bastante difícil. Por ejemplo, una fuente de datos es Google, donde se obtienen datos de una multitud fuentes, por lo tanto, resulta difícil evaluar y tener un *ground truth* con el que comparar resultados. El cálculo de la confianza se basa en una fórmula, donde a partir de una fuente s, la confianza se calcula:

$$c_{fields_s} = \begin{cases} \frac{1}{f - m_s} \text{if } m_s < f \\ 1 \text{if } m_s = f \end{cases}$$

$$c_{responses_s} = \begin{cases} \frac{1}{n_s} \text{ if } n_s > 0 \\ 0 \text{ if } n_s = 0 \end{cases}$$

$$conf_s = \frac{1}{r_s} \cdot [\alpha \cdot c\_fields_s + (1 - \alpha) \cdot c\_responses_s]$$

La primera componente de la fórmula $cfieldss$ mide la proporción de los campos de *firmographics* extraídos correctamente ms, a partir del total de campos posibles f. La segunda componente $cresponsess$ sirve para penalizar el valor, donde si la fuente devuelve numerosos resultados ns la confianza será menor que si la fuente devuelve pocos resultados. La puntuación final de confianza, $confs$, combina estos dos componentes en un promedio ponderado utilizando un parámetro ajustable, α, y se escala según el número de reintentos rs necesarios para obtener un resultado válido.

La función seleccionará la fuente con mayor confianza $confs$ y realizará combinación de resultados si se considera oportuno. La combinación se realiza si falta alguno de los campos, donde se comprueban si la confianza de las demás fuentes supera cierto umbral, en caso afirmativo se completan los campos faltantes si dichos campos contienen valor.

Una vez se ha seleccionado la fuente con mayor confianza y se han combinado datos en caso de que algún campo faltase, la función final calcula un nuevo valor que mide la confianza del conjunto de datos finalmente extraído.

$$conf\_total = \frac{1}{2}\left(\frac{\sum_{s \epsilon S} conf_s}{|S|} + \frac{|\cup_{s \epsilon S} m_s|}{f}\right)$$

La fórmula permite recalcular la confianza, pero solo tomando en cuenta las fuentes de datos utilizadas para dar la respuesta final, que pueden ser tanto la fuente seleccionada como las que se utilizan para combinar datos. El conjunto S incluye únicamente las

fuentes utilizadas durante la combinación, donde $|S|$ representa la cardinalidad del conjunto, es decir, el número de fuentes. Para cada fuente $s \in S$ se calculan las confianzas individuales de cada una, $confs$, cuyo promedio se utiliza para obtener la confianza media de todas las fuentes. Este valor se combina con la proporción de campos no ausentes, m_s, obtenida después de la combinación (número total de campos distintos aportados por el conjunto S, $|U_{s \in S} m_s|$), en relación con el número total de campos esperados f, para producir una puntuación final que refleje tanto la completitud como la confiabilidad de los datos integrados. Esta confianza permite evaluar si es preferible combinar información de múltiples fuentes o confiar en un número menor de fuentes más confiables, dependiendo de la calidad y cobertura de los datos disponibles.

3 THRACT: Actores de amenazas y víctimas

El cálculo de este indicador ya se introdujo en el entregable E2.1 y se calcula en base a la información disponible para los actores de amenazas en la base de datos pública *Threat Group Cards: A Threat Actor Encyclopedia*, desarrollada y mantenida por la *Electronic Transactions Development Agency*. Consta de tres métricas parciales (puntuación de actividad, de capacidad y de objetivo) que se combinan para obtener el valor final:

$$Actor\ Score\ = Puntuación\ Actividad \times \frac{1}{2}(Puntuación\ Capacidad + Puntuación\ Objetivo)$$

Merece ocasión recordar cómo se produce el cálculo de cada puntuación parcial. La puntuación de actividad se otorga según los siguientes casos, basados en la última vez en que fue detectada actividad del actor:

$$Puntuación\ Actividad = \begin{cases} 1.0, & si\ días\ desde\ última\ actividad \leq 30\ días \\ 0.8, & si\ 30 < días\ desde\ última\ actividad \leq 90\ días \\ 0.6, & si\ 90 < días\ desde\ última\ actividad \leq 365\ días \\ 0.3, & si\ días\ desde\ última\ actividad\ > 365\ días \end{cases}$$

Asimismo, la puntuación de capacidad se asigna según la atribución a las capacidades del actor, según 3 niveles:

$$Puntuación\ Capacidad = \begin{cases} 0.3, & si\ capacidades\ igual\ a\ "Por\ debajo\ de\ la\ media" \\ 0.5, & si\ capacidades\ igual\ a\ "Media" \\ 0.8, & si\ capacidades\ igual\ a\ "Por\ encima\ de\ la\ media" \end{cases}$$

La puntuación de objetivo tiene en cuenta el país e industria de la víctima, calculándose como una media ponderada a partir de los siguientes valores intermedios:

$$Puntuación\ país = \begin{cases} 1.0, & si\ país\ de\ la\ víctima\ entre\ los\ países\ objetivo\ del\ actor \\ 0.6, & si\ país\ de\ la\ víctima\ es\ frontera\ de\ algún\ país\ objetivo\ del\ actor \\ 0.5, & si\ región\ de\ la\ víctima\ coincide\ con\ alguna\ región\ de\ los\ objetivos\ del\ actor \\ 0.4, & en\ otro\ caso \end{cases}$$

$$Puntuación\ industria = \begin{cases} 1.0, & si\ industrias\ objetivo\ del\ actor\ contiene\ "Oportunista" \\ 1.0, & si\ industria\ de\ la\ víctima\ coincide\ con\ alguna\ industria\ objetivo\ del\ actor \\ 0.4, & en\ otro\ caso \end{cases}$$

$$Puntuación\ motivación = \begin{cases} 1.0, & si\ motivación\ igual\ a\ "Criminal \\ 0.8, & si\ motivación\ igual\ a\ "Patrocinado\ por\ un\ estado" \\ 0.5, & si\ motivación\ igual\ a\ "Hacktivismo" \\ 0.1, & en\ otro\ caso \end{cases}$$

$$Puntuación\ Objetivo = \frac{w_{país} \times punt.país + w_{industria} \times punt.industria + w_{motivación} \times punt.motivación}{w_{país} + w_{industria} + w_{motivación}}$$

Siendo $w_{país} = 0.5,\ w_{industria} = 0.7,\ w_{motivación} = 0.4$.

4 CVE2TTPs

También introducido en el entregable E2.1, la metodología de este modelo consiste en abordar el problema multiclase y multietiqueta de mapear vulnerabilidades (CVEs) a técnicas TTPs de MITRE, utilizando un modelo BERT preentrenado que se ajusta mediante *fine-tuning*. Se añade una capa de salida correspondiente al número de TTPs, permitiendo que el modelo aprenda asociaciones complejas entre descripciones de CVEs y técnicas. Para ello, se exploran dos enfoques predictivos (véase la *Ilustración 1. Alternativas para predecir técnicas de un CVE*): uno directo, basado en la descripción y tipo de vulnerabilidad, y otro en dos etapas, donde primero se predice el tipo de vulnerabilidad antes de inferir las TTPs. Ambos enfoques permiten una modelización flexible y adaptable al nivel de información disponible, obteniendo finalmente el modelo descrito en la *Ilustración 2. Modelo de predicción CVE2TTPs*.

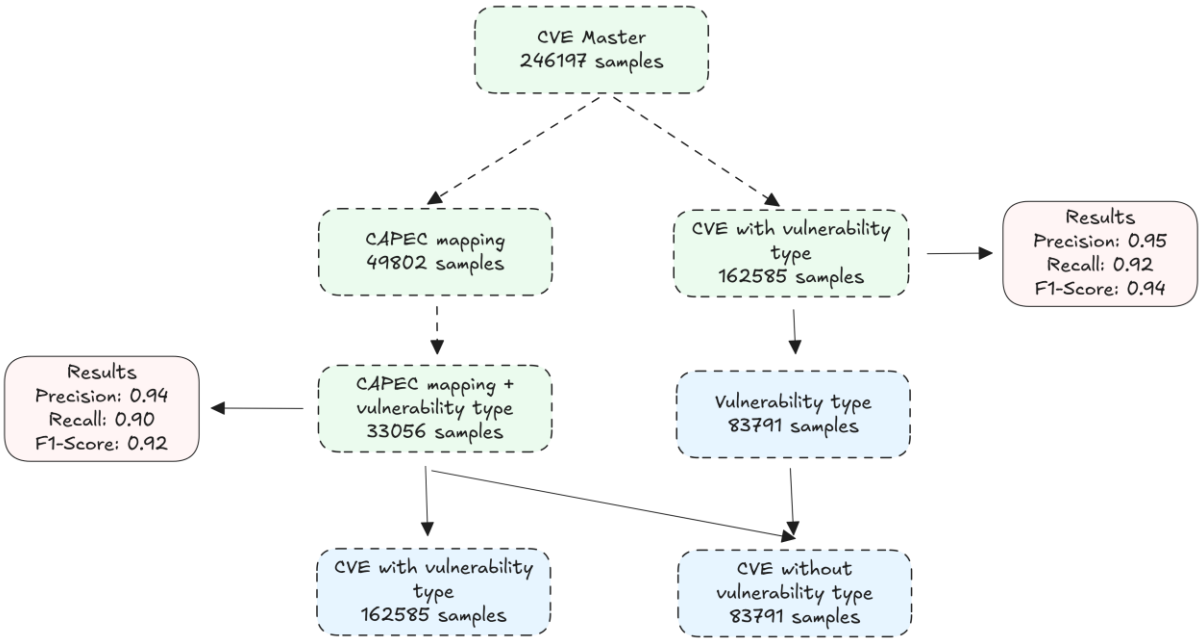


Ilustración 1. Alternativas para predecir técnicas de un CVE

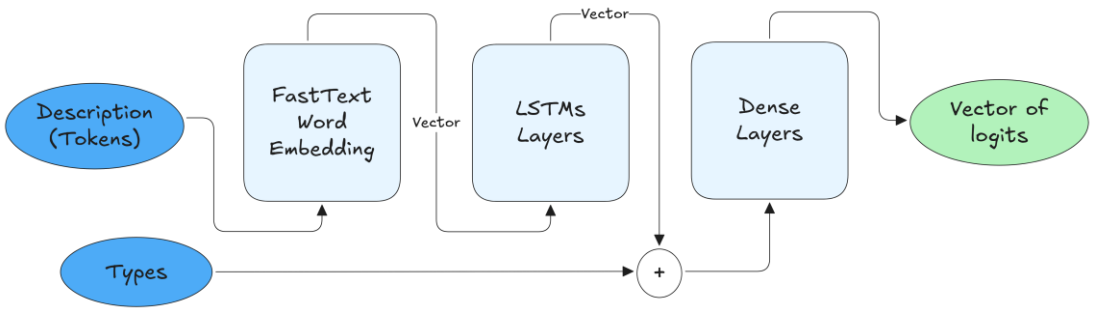


Ilustración 2. Modelo de predicción CVE2TTPs

5 CUANTIFICACIÓN DEL CIBERRIESGO

Siguiendo el enfoque propuesto en el entregable E2.2, el módulo de cuantificación del riesgo cibernético calcula la exposición al riesgo y su impacto financiero potencial como el producto de la frecuencia de eventos de pérdida y su magnitud:

$$\text{Risk} = \text{Loss Event Frequency (LEF)} \times \text{Loss Magnitude (LM)}$$

Esta métrica representa la pérdida financiera anual esperada debido a incidentes cibernéticos y se estima mediante simulaciones estocásticas de Monte Carlo. Estas simulaciones generan resultados realistas para cada componente del modelo:

- Frecuencia de eventos de pérdida (Loss Event Frequency, LEF): número esperado de incidentes cibernéticos exitosos por año.
- Magnitud de la pérdida (Loss Magnitude, LM): impacto económico promedio por incidente exitoso, incluyendo pérdidas directas e indirectas.

Ambos componentes permiten obtener una estimación final del riesgo cibernético. Para facilitar la comprensión de esta estructura, cada parte se descompone obteniendo una disposición en forma de árbol (véase la *Ilustración 3. Árbol de descomposición de la cuantificación del ciberriesgo*).

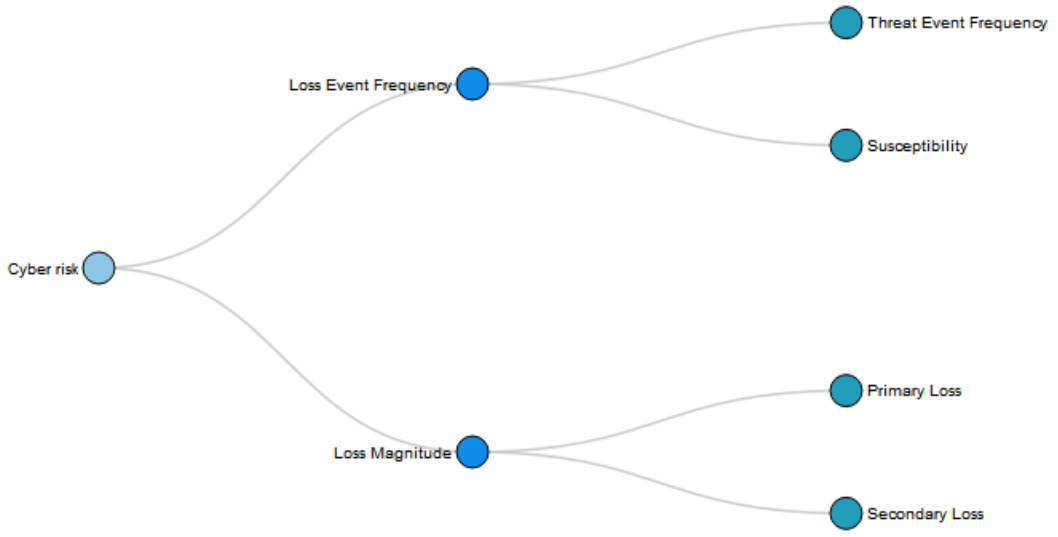


Ilustración 3. Árbol de descomposición de la cuantificación del ciberriesgo

5.1 Frecuencia de eventos de pérdida (Loss Event Frequency, LEF)

La LEF representa el número esperado de incidentes cibernéticos exitosos por año y se define como:

$$\text{LEF} = \textit{Threat Event Frequecy}\ (\text{TEF}) \times \text{Susceptibility}$$

$$\text{TEF}^{(i)} \sim \text{Poisson}(\lambda)$$

$$\text{Susceptibility}^{(i)} \sim \text{Beta}(\alpha, \beta)$$

La Frecuencia de intentos de ataque (*Threat Event Frequency*, TEF) cuantifica cuántos intentos de ciberataques se esperan contra la organización por año. Se modela con una distribución de Poisson con parámetro λ el producto de las siguientes componentes:

- Base: número promedio de ataques anuales en la industria y región de la organización, basado en informes públicos de terceros.
- Tasa de incidentes: tasa anualizada de crecimiento en la cantidad de incidentes, calculada a partir de datos históricos de las bases de datos de ciber incidentes generadas en el proyecto.
- Atractivo: características o comportamientos que aumentan el interés de posibles atacantes. Un mayor atractivo implica mayor propensión a ser atacado.

$$\text{TEF}^{(i)} \sim \text{Poisson}(\lambda)$$

$$\lambda = \text{Baseline} \times \text{Incident rate} \times \text{Attractiveness}$$

La Susceptibilidad modela la probabilidad de que un intento de ataque tenga éxito empleando una distribución Beta. Los parámetros α y β se obtienen a partir de los componentes siguientes:

- Índice de actores de amenaza: análisis de la tendencia en los últimos 3 años del indicador DICYME, que ofrece una visión del actor de amenaza en función del país y sector objetivo. Utiliza datos públicos de ETDA, ponderando actividad, capacidad y objetivos del actor.
- Exposición: cuantifica la explotabilidad de técnicas del marco MITRE ATT&CK basándose en CVEs reportadas. Se obtiene mediante el modelo CVE2TTPs, que relaciona vulnerabilidades con tácticas, técnicas y procedimientos del framework. Se identifican CVEs que permiten técnicas de la matriz ICS y se ponderan según la severidad (CVSS) y la relevancia táctica, priorizando aquellas de la táctica *Impact*.
- Perfil de seguridad: madurez defensiva de la organización, que refleja su capacidad para detectar y contener ataques.

$$\text{Susceptibility}^{(i)} \sim \text{Beta}(\alpha, \beta)$$

$$\begin{aligned} \alpha &= \lambda \times \text{Susceptibility score} \\ \beta &= \lambda \times (1 - \text{Susceptibility score}) \end{aligned}$$

$$\text{Susceptibility score} = \frac{\text{Threat Actor index}}{\text{Exposure} \times \text{Security profile}}$$

5.2 Magnitud de la pérdida (Loss Magnitude, LM)

La Magnitud de la pérdida representa el costo financiero asociado a un incidente cibernético exitoso. Se estima simulando pérdidas primarias y secundarias:

- Pérdidas primarias: costos financieros directos, como interrupciones operativas o daños a equipos.
- Pérdidas secundarias: costos indirectos, como investigaciones forenses, sanciones regulatorias o daños reputacionales.

El proceso comienza con la lista de CVEs proporcionada por el usuario. Mediante el modelo CVE2TTPs, se determina qué técnicas MITRE ATT&CK asociadas a la táctica de impacto (*Impact*) son potencialmente aplicables.

Para cada técnica identificada, se simulan:

- Una de las posibles pérdidas primarias: interrupción del negocio, daños a equipos, extorsión y daños humanos.

$$Primary\ Loss \begin{cases} Business\ Interruption_k = Cost \times duration \times Extent \\ Equipment\ Damage_k = Cost\ of\ Equipment_k \times Damage \\ Extortion_k = Revenue_k \times Proportion \\ Human\ Damage_k = Cost\ of\ Life_k \times Damage \end{cases}$$

- Todas las pérdidas secundarias asociadas: investigación forense, daño reputacional y multas o sanciones.

$$Secondary\ Loss \begin{cases} Forensic\ Investigation_k = Cost\ of\ 1\ Hour_k \times Duration\ (in\ hours) \\ Reputational\ Damage_k = Revenue_k \times Proportion \\ Penalties_k = Revenue_k \times Proportion \end{cases}$$

Finalmente, cada iteración de simulación selecciona una técnica de impacto válida, muestrea los valores de pérdida correspondientes y calcula la magnitud total de pérdida como:

$$\mathrm{LM}_k = \text{Primary Loss}_k + \sum \text{Secondary Losses}_k$$

6 CONCLUSIONES Y SIGUIENTES PASOS

La Actividad 2 del proyecto DICYME culmina con este entregable, en el que se consolidan los principales elementos técnicos y metodológicos necesarios para sustentar la estimación de probabilidad e impacto de incidentes OT en un marco riguroso de análisis de ciber riesgo.

A lo largo de esta actividad se ha documentado en detalle el uso de múltiples fuentes de información, estrategias de integración de datos, formulación de métricas clave y

enfoques de validación orientados a cuantificar tanto la exposición como el potencial impacto de ciber incidentes en organizaciones industriales. La documentación generada cubre desde componentes como la reputación online y la exposición técnica hasta el atractivo frente a actores maliciosos o el impacto económico simulado.

Este trabajo cierra una fase clave del proyecto, dotando a los siguientes módulos de base documental sólida y criterios reproducibles para la implementación del motor de cálculo de ciber riesgo. La integración progresiva de estos componentes se ha llevado a cabo de forma modular, lo que facilitará su adaptación y extensión en futuras aplicaciones o entornos organizativos distintos.

Con ello, se cierra la contribución de la Actividad 2 al objetivo general de DICYME: proporcionar una arquitectura flexible, empírica y dinámica para la medición del ciber riesgo industrial basado en evidencia real.