Extracción de variables para caracterización multi-clase de la severidad de IPs

David Escudero García
RIASC. Universidad de León
Campus de Vegazana s/n 24071
descg@unileon.es

Noemí DeCastro-García
Dpto. de Matemáticas. Universidad de León.
Campus de Vegazana s/n 24071
ncasg@unileon.es

Miguel V. Carriegos
Dpto. de Matemáticas. Universidad de León.
Campus de Vegazana s/n 24071
miguel.carriegos@unileon.es

***Resumen*—Determinar la severidad de un incidente de ciberseguridad es fundamental para establecer medidas efectivas contra el mismo. En este contexto, el aprendizaje automático es utilizado para crear modelos capaces de clasificar y predecir la peligrosidad de los eventos de ciberseguridad. Uno de los aspectos más importantes en el uso de este tipo de técnicas es la extracción de variables que permitan obtener modelos eficientes.**

El objetivo de este trabajo es construir un conjunto de variables o *features* que caracterice la maliciosidad de una dirección IP de manera multi-clase. La configuración final son 23 variables: 18 de ellas obtenidas mediante series temporales y listas de reputación, y 5 relacionadas con la geolocalización de la IP. No solo se han extraído las features, sino que se ha realizado un análisis estadístico para estudiar su adecuación y optimización. En el caso de las variables de geolocalización, por los posibles cambios que pueden sufrir en el tiempo. En el caso de las series temporales, por los hiper-parámetros inherentes a la construcción de las variables.

***Index Terms*—Severidad, aprendizaje automático, selección de variables, direcciones IP**

Tipo de contribución: *Investigación original*

I. Introducción

El objetivo de este trabajo es extraer un conjunto de variables o *features* estadísticamente significativo que caracterice, mediante aprendizaje automático, la peligrosidad o maliciosidad de una dirección IP de manera multi-clase. Se propone un conjunto de variables conformado por características que podamos encontrar en herramientas externas sobre localización geográfica de una dirección IP, y características construidas mediante series temporales y consulta a listas negras o de reputación de direcciones IP.

El conjunto final tiene 23 variables, 18 relacionadas con información temporal, y 5 con información geográfica. El análisis de la significancia de las mismas se basa en las limitaciones que presentan. Por un lado, las variables de geolocalización pueden sufrir cambios con el paso del tiempo. En este caso, las preguntas de investigación van dirigidas a determinar si existe deriva conceptual y, en caso afirmativo, si tiene efecto sobre los resultados de los modelos obtenidos. Por otro lado, las características temporales dependerán de dos hiper-parámetros que definen las ventanas temporales. Las preguntas de investigación se dirigen a determinar el efecto de los mismos en el ajuste de los modelos de clasificación.

El estudio se ha realizado sobre un conjunto de 99720 IPs proporcionado por INCIBE [1]. Se han llevado a cabo diferentes análisis descriptivos e inferenciales. Los resultados muestran que uno de los hiper-parámetros de los que depende la extracción de las variables temporales tiene un efecto significativo sobre los resultados alcanzados. Por otra parte, los cambios en la geolocalización no parecen implicar una degradación de los modelos.

Este artículo está organizado de la siguiente manera: en la sección 2, se desarrolla el trabajo relacionado. En la sección 3, se describe el proceso de construcción y obtención de las variables de caracterización. En la sección 4, se incluyen todos los detalles experimentales del estudio. En la sección 5, se desglosan y discuten los resultados obtenidos. Finalmente, se incluyen las conclusiones, los agradecimientos y las referencias.

II. Trabajo relacionado

Caracterizar la severidad de un evento de ciberseguridad, entendida como una medida del riesgo que supone, es fundamental para poder reaccionar de una manera eficiente ante el mismo. Actualmente, existen diferentes metodologías y estándares que asignan una puntuación para evaluar la severidad de eventos de ciberseguridad, y que se basan en taxonomías, o aplicación de consultas en informes internacionales (*Microsoft Security Bulletin Vulnerability Rating*, [1],*Common Vulnerability Scoring System* (CVSS)[2],*Open Web Application Security Project* (OWASP) *Risk Rating Methodology*, [3], *Cyber Incident Scoring System* , [4], entre otros). Recientemente, también se han utilizado técnicas de aprendizaje automático para esta tarea ([5]). En este último trabajo, la severidad de eventos de ciberseguridad de diferente naturaleza es caracterizada, de forma multi-clase, mediante 113 variables recogidas por un *Computer Emergency Response Team* (CERT).

En particular, si hablamos de la severidad de una dirección IP, suele ser habitual entender la misma como su reputación. El enfoque tradicional para determinar la maliciosidad de una IP se basa en el uso de *blacklists* que contienen conjuntos de IPs que han sido detectadas llevando a cabo comportamientos maliciosos. Muchas herramientas para la gestión de *firewalls*, como FireHOL [6], agrupan información de varias *blacklists* para bloquear eficientemente IPs sospechosas. Existen varios trabajos centrados en la extracción de información de *blacklists* para la predicción de la maliciosidad de una IP. En [7] se agregan las IPs en subredes usando un prefijo CIDR seleccionado y se construye una serie temporal usando como magnitud el número de IPs de la subred presentes en *blacklists* en diferentes intervalos de tiempo. Estas series temporales se dividen en 3 franjas ”buena”, ”normal”, y ”mala” según si el

[1]La publicación del *dataset* se encuentra en estudio legal en la fecha de envío de este trabajo

valor de la serie temporal está por debajo, en o por encima de la media. De estas series temporales se extraen características como el valor medio en cada franja, o la proporción de tiempo en la que la serie temporal permanece en una franja. Se alcanza un ratio de verdaderos positivos de 0.7. Otro enfoque se presenta en [8], en el que usando técnicas de clustering se determina qué partes del espacio de direcciones IP contienen una mayor frecuencia de IPs maliciosas. Se obtiene una tasa de acierto de 0.776. En [9] también se usa clustering, agrupando las IPs a clasificar de forma que maximice la dependencia entre la pertenencia de un IP a un cluster y su presencia en una *blacklist*. Este esquema alcanza precisiones de en torno a 0.8, pero requiere que se repita el proceso de clustering para actualizarse a cambios en la blacklist. En estos casos, el proceso se basa en la hipótesis de que direcciones IP de una misma subred o de redes contiguas tienen una mayor probabilidad de compartir un grado de maliciosidad, por lo que es necesario disponer de un volumen relativamente alto de IPs para poder crear un modelo que no sea demasiado local y sirva para extender las predicciones a un espectro amplio de IPs.

Por otro lado, existen enfoques basados en el uso de información adicional sobre la IP y su comportamiento como la geolocalización o registros DNS asociados para obtener predicciones más generalizables. Servicios como Maxmind [10] permiten obtener información sobre la geolocalización de la IP; otros como IPQualityScore [11] proporcionan un nivel de maliciosidad basándose en ciertas características de la IP como el contenido, registros DNS, etc. El trabajo en [12] propone una herramienta que utiliza información de geolocalización, además de la propia IP o dominio tratado, para predecir su maliciosidad. Se obtiene una tasa de acierto de 0.75 con el mejor modelo, que supera a las tasas de acierto obtenidas por otras fuentes como VirusTotal que se evalúan en el artículo. Otros trabajos como [13] se centran en la detección de IPs maliciosas a partir del tráfico web y de correo eléctrónico usando features como el volumen de peticiones, el número de correos de spam recibidos de una IP, etc. Las tasas de acierto alcanzadas son más altas que otros métodos basados en información contextual de la IP, pero el proceso de monitorización y análisis del tráfico es costoso. En [14] se propone un esquema más complejo que combina información de fuentes externas como una medida de fiabilidad de las predicciones, análisis de muestras de malware asociadas y análisis de las ocurrencias de cada IP en el tiempo. La tasa de acierto llega a alcanzar un 93 % en el mejor modelo, pero el procedimiento de análisis es complejo y la obtención de la información asociada conlleva un importante despliegue de recursos.

En general, métodos más simples alacanzan una tasa de acierto limitada, inferior a 0.8. Métodos como el propuesto en [14] son más eficaces pero todo el proceso de obtención y análisis de muestras de malware y análisis del tráfico requiere una infraestructura que limita su aplicabilidad.

Otra de las posibles limitaciones existentes en la caracterización de una dirección IP es el cambio a lo largo del tiempo de la misma. Su geolocalización puede cambiar, el dominio asociado puede ser diferente y, aunque una IP pueda estar asociada a actividad maliciosa en un instante de tiempo, puede no estarlo más tarde: quizás el equipo original fuese infectado pero se ha solventado el problema. En muchos trabajos se sugiere que es necesario mantener actualizados los modelos para ajustarse a estos cambios, pero esto conlleva un cierto consumo de recursos, así que sería deseable poder estimar el impacto que los cambios en la caracterización de las IPs tienen sobre los modelos. Estos cambios en los datos se pueden analizar bajo el marco teórico de deriva conceptual. La deriva conceptual es el cambio en la distribución de los datos en escenarios dinámicos de aprendizaje. Sea X el espacio de vectores de features o variables de una muestra de datos, y $P(X)$ la distribución de probabilidad marginal. Además, sea Y el espacio de etiquetas de X. En términos matemáticos, se define un *concepto* como la distribución conjunta de X e Y, $P(X, Y)$ ([15]). Si denotamos la distribución marginal de los datos en un instante t como $P_t(X)$, y la distribución condicional (posterior) de las etiquetas de los mismos mediante $P_t(Y \mid X)$, entonces la deriva conceptual ocurre cuando $P_t(y \mid X) \neq P_{t+\triangle t}(y \mid X)$ y/o $P_t(X) \neq P_{t+\triangle t}(X)$. Este puede ser el caso, por ejemplo, de la geolocalización de una dirección IP si, por ejemplo, esta cambia porque se asigna a un lugar diferente ($P_t(X)$ cambia) o se reciben muchas alertas de IPs maliciosas procedentes de un país particular ($P_t(Y \mid X)$ cambia). Aunque existen caracterizaciones más generales de la deriva conceptual [16], en términos generales esta puede ser de tres tipos ([15], [17]): real ($P_t(Y \mid X) \neq P_{t+\triangle t}(Y \mid X)$ pero $P_t(X) = P_{t+\triangle t}(X)$), virtual ($P_t(X) \neq P_{t+\triangle t}(X)$ pero $P_t(Y \mid X) = P_{t+\triangle t}(Y \mid X)$), o ambas ($P_t(Y \mid X) \neq P_{t+\triangle t}(Y \mid X)$ y $P_t(X) \neq P_{t+\triangle t}(X)$). Una de las líneas de investigación actuales sobre deriva conceptual está dirigida a encontrar técnicas y algoritmos que puedan detectarla. En este trabajo, se destaca el cálculo de la distancia entre los conceptos de los periodos t y $t + \triangle t$ dada en [18] mediante el concepto de Magnitud usando como distancia la variación total de Levin [19] y su versión corregida para el cálculo de la deriva condicional:

$$\sigma_{t,t+\triangle t}(Z) = \frac{1}{2} \sum_{\overline{z} \in Dom(Z)} \mid P_t(\overline{z}) - P_{t+\triangle t}(\overline{z}) \mid \text{ y} \qquad (1)$$

$$\sigma^{Y|X}_{t,t+\triangle t} = \sum [\frac{P_t(\overline{x}) + P_{t+\triangle t}(\overline{x})}{2} \frac{1}{2} \sum \mid P_t(y \mid \overline{x}) - P_{t+\triangle t}(y \mid \overline{x}) \mid] \qquad (2)$$

Otra limitación existente en la caracterización de la maliciosidad de las IPs está en que la mayoría de los trabajos tratan un problema biclase: distinguir entre IPs maliciosas y no maliciosas. En este trabajo, tratamos el problema de asignar un nivel de maliciosidad asociado a la IP.

En este trabajo, seguimos un enfoque mixto para la caracterización de las IPs por su severidad. Usamos features derivadas de *blacklists* como en [7], pero añadimos *features* relacionadas con la geolocalización de la IP. El conjunto de *features* resultantes es ligero y no requiere todo el procesado de apoyo de herramientas como la presentada en [14]. Además, realizamos un estudio del impacto de la deriva conceptual en la predicción de la severidad asociadas a IPs, así como de la posible optimización de los parámetros que afectan a la construcción de las variables extraídas mediante series temporales.

III. Variables de caracterización

Las variables de caracterización que se proponen en este trabajo son de diferente naturaleza. Por un lado, ciertas propiedades de una serie temporal que puede crearse a través de consultas a listas de reputación o listas negras. Por otro lado, la geolocalización de la IP. En total, tendremos 23 *features* o variables, que denotaremos mediante F_i con $i = 1, \ldots, 23$.

El esquema de extracción de las variables de series temporales está basado en el trabajo presentado en [7]. Sea una red $r = A.B.C.0$ de direcciones IP. Se puede crear una serie temporal $\mathfrak{X}_r(t)$ que asigna a cada instante de tiempo t, el número de direcciones IP pertenecientes a r contenidas en las listas de referencia (listas negras o de reputación) en ese instante. Este es un proceso que puede realizarse sobre cualquier lista de IPs con un *time stamp* asociado.

A partir de un fragmento de la serie temporal $\{X_r(t_1), X_r(t_2), \ldots, X_r(t_h)\}$ de tamaño *h*, vamos a extraer un vector de 9 *features* $(F_1, \ldots, F_9)$. Las primeras tres coordenadas del vector responden a la intensidad de la ventana temporal. Las tres siguientes a la duración, y las tres últimas a la frecuencia.

Para construirlas, seguiremos el siguiente procedimiento:

1. Fijamos un número real $\delta > 0$.
2. Se calcula el número medio de IPs de la red que aparecen en las listas negras o de reputación durante el fragmento de la serie temporal elegido:
$$\mu = \frac{\sum_{i=1}^{h} X_r(t_i)}{h} \tag{3}$$
3. Asignamos un nivel o rango a cada instante t_i del fragmento de la ventana temporal: diremos que un instante t_k estará en un nivel si cumple las siguientes condiciones:
$$\text{Nivel} = \begin{cases} \text{Bajo} & \text{si } t_k : X_r(t_k) \leq (1-\delta)\mu \\ \text{Medio} & \text{si } t_k : (1-\delta)\mu < X_r(t_k) < (1+\delta)\mu \\ \text{Alto} & \text{si } t_k : (1+\delta)\mu \leq X_r(t_k) \end{cases} \tag{4}$$
4. El primer trío de *features* se relaciona con la intensidad: para cada nivel (bajos, medios, altos), la intensidad es el valor medio de la serie temporal en los instantes de tiempo de esa nivel. Esto es,
$$i(nivel) = \frac{\sum_{t_k \in nivel} X_r(t_k)}{|nivel|} \tag{5}$$
donde $|nivel|$ es el número total de instantes del fragmento temporal que han sido asignados a ese nivel. Así, $F_1 = i(bajos), F_2 = i(medios), F_3 = i(altos)$.
5. El segundo trío de *features* se relaciona con la duración: para cada franja de valores la duración es el número medio de instantes k en los que la serie temporal permanece en el nivel concreta (varios k consecutivos permanecen en el mismo nivel). Así, $F_4 = d(bajos), F_5 = d(medios), F_6 = d(altos)$.
6. El tercer trío de *features* se relaciona con la frecuencia: para cada nivel, la frecuencia es la proporción de instantes que pertenece a ese nivel:. Esto es,
$$f(nivel) = \frac{|nivel|}{|h|} \tag{6}$$
Así, $F_7 = f(bajos), F_8 = f(medios), F_9 = f(altos)$.

En la figura 1 puede encontrarse un ejemplo de una serie temporal construida a lo largo de 20 días. Fijemos $h = 5, \delta = 0.001$.

En este caso $X_r(t_1) = 3, X_r(t_2) = 6, X_r(t_3) = 1, X_r(t_4) = 9, X_r(t_5) = 2$. El número medio de IPs de la red que aparecen en la lista durante esta ventana temporal de cinco días es $\mu = 4.2$. Por lo tanto, t_1, t_3 y t_5 son instantes del nivel bajo, mientras que t_2 y t_4 son instantes del nivel alto. Así, $(F_1, F_2, F_3, F_4, F_5, F_6, F_7, F_8, F_9) = (2, 0, 7.5, 0, 0, 0, 0.6, 0, 0.4)$.

Puede observarse que las features $F_1, \ldots, F_9$ se asocian a redes de IPs. Por lo que dos IPs que pertenezcan a la misma red, tendrán los mismos valores en estas *features*. Este hecho también puede verse como una medida de cómo de cercanas son las IPs.

Cabe destacar que h es el tamaño de la ventana temporal que elegimos. Cuanto más pequeña sea, menos recursos , en datos y computacionales, requerirá la extracción de *features*. Si necesitmos una h muy elevada, entonces es posible que la extracción de las variables no resulte eficiente.

En lo anterior se han agrupado las IPs en redes de la forma A.B.C,0 esto es redes que tienen un sentido físico como la red local de un hogar. Para añadir otras nueve caracteristicas, agrupamos en redes que no tienen un sentido real pero consituyen una relación de equivalencia como otra cualquiera. A saber, las redes de la forma A.B,0.D esto es redes donde A, B, D se mantienen fijos y 0 puede ser cualquier número. Lo que nos da otras nueve características. Por lo tanto, F_i de $i = 1, \ldots, 18$.

Por otro lado, las *features* de geolocalización son las siguientes (F_i de $i = 19, \ldots, 23$):

- Latitud y longitud se conservan como números decimales (F_{19} y F_{20}).
- El código de país se categoriza en forma de número entero, cuyo valor va de 0 a al número de países representados según el código ISO 3166-1 (F_{21}).
- La IP se transforma a un valor numérico entero con la siguiente fórmula:
$$A.B.C.D \rightarrow A * 256^3 + B * 256^2 + C * 256^1 + D * 256^0$$
(F_{22}).
- La fecha de ocurrencia se transforma al tiempo UNIX (número de segundos transcurridos desde el 01/01/1970 a las 00:00) (F_{23}).

Algunas de las cuestiones que debemos analizar al utilizar las *features* seleccionadas son las siguientes:

1. Las características temporales que se han extraído dependen de dos hiper-parámetros, h y δ. ¿Tienen influencia en los resultados obtenidos?
2. Cabe esperar que los datos de la geolocalización de la IP cambien en el tiempo, ya sea por la reasignación de las direcciones o por imprecisiones en la geolocalización. Este cambio podría causar que el modelo pierda capacidad predictiva. Esta cuestión es de particular importancia en el caso de tratar con datos de listas públicas, ya que nuevas IPs tienden a incorporarse o eliminarse de listas negras. Uno de los objetivos de este estudio es evaluar el cambio en la geolocalización implica un

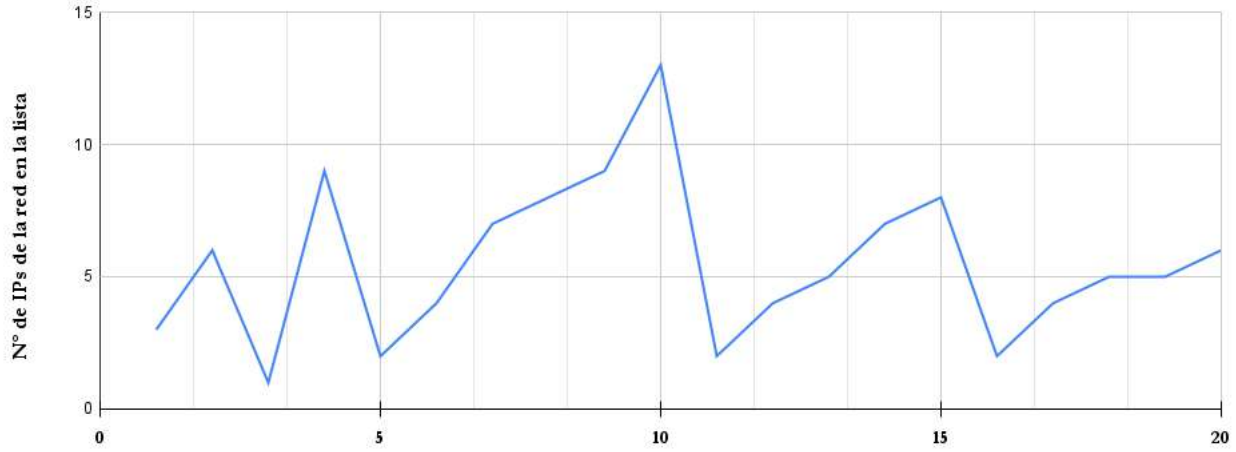

Figura 1. Ejemplo de serie temporal.Eje X: días. Eje Y: número de IPs de la red en la lista negra a cada instante

escenario de aprendizaje con deriva conceptual y, en ese caso, determinar si la capacidad predictiva disminuye . En caso positivo, habría que analizar en qué medida se degrada el modelo.

3. Si las cuestiones anteriores tienen efecto en la capacidad de los modelos de aprendizaje automático para clasificar IPs según su maliciosidad, habrá que determinar la configuración que nos aporte las mejores *features.*

IV. Sección experimental

En esta sección se describen los conjuntos de datos, las preguntas de investigación y los análisis realizados.

IV-A. Conjuntos de datos

Para la realización de este trabajo se ha utilizado un conjunto de datos de eventos de ciberseguridad aportado por INCIBE que corresponde al mes de mayo de 2021 [2]. Lo denotaremos por D. Se trata de un archivo CSV que contiene 99720 IPs. De cada IP se tiene la siguiente información:

1. IP.
2. Fecha del evento.
3. Latitud: extraída en el momento de recepción del evento.
4. Longitud: extraída en el momento de recepción del evento.
5. Código del país: extraída en el momento de recepción del evento.
6. Severidad asociada: el valor de severidad viene asignado por un experto de INCIBE. A su vez, se utiliza un modelo de aprendizaje automático para generar esta severidad [5], pero utilizando información de 113 variables que INCIBE recoge en su modelo de inteligencia. Es un valor multi-clase con 4 niveles (1, 3, 6 y 9, ordenados de menor a mayor severidad).

Trataremos este conjunto D como una lista de reputación o lista negra de IPs. El primer paso realizado ha sido curar el conjunto de datos. Para evitar el exceso de notación, lo volveremos a denotar D. Se eliminaron aquellos datos que contenían valores inválidos o incorrectos. El siguiente paso fue recalcular la variable Severidad para tener cuatro etiquetas o clases:

$$\text{Severidad} = \begin{cases} 1 & \text{si } Severidad_{antigua} \in 0, 1 \\ 3 & \text{si } Severidad_{antigua} \in 2, 3, 4, Low \\ 6 & \text{si } Severidad_{antigua} \in 5, 6, 7, Warning \\ 9 & \text{si } Severidad_{antigua} \in 8, 9, 10, High \end{cases} \tag{7}$$

A partir de D se van a extraer las variables $F_1, \ldots, F_{18}$ de las 99720 IPs. Para poder dar respuesta a las preguntas de investigación planteadas, para cada una de las IPs, realizamos una consulta a MaxMind para obtener la información de las columnas *latitud, longitud*, y *código del país*. La consulta se ha realizado en octubre de 2021 para comprobar si ha habido cambios y, por lo tanto, deriva conceptual. De estas 99720 IPs, 42677 de ellas tienen una geolocalización distinta de acuerdo con MaxMind [3]. Por lo tanto, se tendrán finalmente dos conjuntos de datos para crear los modelos de clasificación, D_1 y D_2. Ambos estarán formados por las 42677 IPs en las que ha habido cambios de geolocalización. Las features $F_1, \ldots, F_{18}$ serán iguales en ambos conjuntos de datos, así como la etiqueta asignada en la variable *Severidad*. Las variables $F_{19}, \ldots, F_{23}$ serán calculadas con la información contenida en las variables de geolocalización dadas por INCIBE en el conjunto D_1. Para D_2, las variables de geolocalización se calcularán con la información aportada por MaxMind. Cabe destacar, en términos de estudiar la deriva conceptual, que tomaremos $D_1 = D_t$ y $D_2 = D_{t+\triangle t}$.

La proporción de las clases (valores de severidad) en cada uno de los conjuntos se presenta en la tabla I.

IV-B. Preguntas de investigación

Las preguntas se dividen en los análisis sobre el conjunto de variables de geolocalización y las de series temporales.

PI1 ¿Hay deriva conceptual entre D_1 y D_2?

PI2 ¿Existen diferencias significativas entre los resultados obtenidos al utilizar D_1 y D_2? En caso positivo, ¿en qué conjunto se obtienen mejores resultados? ¿Hay degradación en los resultados?¿Es relevante?

[2]La publicación del *dataset* se encuentra en estudio legal en la fecha de envío de este trabajo

[3]https://www.maxmind.com/en/home

Tabla I
FRECUENCIAS DE LAS DIFERENTES CLASES EN LOS CONJUNTOS DE DATOS D_1, D_2 Y D.

Conjunto de datos	Severidad	Frecuencia	Proporción
D	1	8402	8.4255 %
	3	24943	25.0130 %
	6	54437	54.5898 %
	9	11938	11.9715 %
D_1 y D_2	1	2898	6.7907 %
	3	12317	28.8616 %
	6	22868	53.5851 %
	9	4593	10.7624 %

PI3 ¿Existen diferencias significativas entre los resultados obtenidos en las variables respuesta al variar el parámetro h en F_i con $i = 1, \ldots, 18$? En caso positivo, ¿con qué h se obtienen mejores resultados? ¿El efecto de h es relevante?

PI4 ¿Existen diferencias significativas entre los resultados obtenidos en las variables respuesta al variar el parámetro δ en F_i con $i = 1, \ldots, 18$? En caso positivo, ¿con qué δ se obtienen mejores resultados? ¿El efecto de δ es relevante?

IV-C. Análisis

Los modelos de clasificación se obtienen usando la herramienta *AutoSklearn* [20]. Se reserva un 75 % de los datos para el proceso de optimización de hiperparámetros y el 25 % restante se usa para evaluar el modelo con los hiperparámetros ya optimizados. Para optimizar los hiperparámetros utilizamos el método SMAC (Sequential Model-based Algorithm Configuration) [21]. Los experimentos se han realizado considerando los mejores resultados tras hacer *10-fold cross-validation.*

Los experimentos se llevan a cabo probando diferentes valores para los hiperparámetros h y d. Se fijan los valores de $h \in \mathbb{N}$ en el intervalo $[6, 10]$ y de $\delta = 0.001, 0.002, 0.003, 0.004]$. La muestra D es de 30 días, por lo que se toma un h máxima de 10.

Las variables respuesta analizadas son dos: el ajuste o *Accuracy*, y el coeficiente de correlación de Matthews o MCC [22] que se define de la siguiente manera:

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}, \tag{8}$$

donde TP, TN, FP, y FN denotan verdaderos positivos, verdaderos negativos, falsos positivos y falsos negativos, respectivamente. Utilizamos esta métrica debido a que se considera más adecuada en el caso de conjuntos de datos no balanceados [23].

El resto de análisis llevados a cabo se enumeran a continuación:

1. Para analizar si las variables de geolocalización presentan deriva conceptual en los conjuntos de datos, se ha utilizado la aproximación dada en [18]. El grado de deriva se ha calculado para $P_{t,t+\triangle t}(X)$ y $P_{t,t+\triangle t}(Y \mid X)$, mediante el cálculo de $\sigma_{t,t+\triangle t}(Z)$ y $\sigma^{Y|X}_{t,t+\triangle t}$, véase Eq. (1) y Eq (2). Ambas toman valores entre 0 y 1, siendo más altas cuanta más deriva haya.
2. El primer análisis estadístico realizado es la prueba de normalidad de Kolmogorov–Smirnov con la corrección de Lilliefors. Al salir $\rho_{K-S}(D_1) = 0.002, \rho_{K-S}(D_2) = 0.000, \rho_{K-S}(D_1 \cup D_2) = 0.000$, los análisis inferenciales serán no paramétricos.
3. Para PI2, y debido a que los conjuntos de features D_1 y D_2 tienen valores diferentes en las variables de geolocalización, se ha aplicado el test U de Mann-Whitney para 2 muestras independientes. De esta manera, podremos determinar si existen diferencias estadísticamente significativas entre los resultados obtenidos. Si han existido diferencias, utilizamos un estudio descriptivo para determinar con qué conjunto obtenemos mejores resultados de clasificación de severidad.
4. En relación con la PI3, queremos determinar si existen diferencias significativas en las variables respuesta obtenidas cuando variamos el parámetro h en la construcción de las *features* correspondientes al bloque de series temporales. Al variar h, el valor de las features F_i con $i = 1, \ldots, 18$ cambia, por lo que lo conjuntos de datos con los que se construyen los modelos son diferentes y, por lo tanto, los análisis inferenciales utilizados han de ser para muestras independientes. Por otra parte, al utilizar cinco categorías para $h = 6, 7, 8, 9, 10$, se utilizará el test de Kruskall-Wallis para comparar los cinco grupos. En caso afirmativo, se realizarán comparaciones *Post Hoc* dos a dos y se utilizará la corrección de Bonferroni $\left(\overline{\alpha} = \frac{\alpha}{\text{número de combinaciones posibles}}\right)$.
5. Para PI4, se realizará un análisis similar al anterior comparando los resultados obtenidos cuando variamos el hiper-parámetro $\delta = 0.001, \ldots, 0.004$.
6. En las preguntas de investigación planteadas, no sólo se trata de analizar si hay diferencias estadísticamente significativas (ρ-valor < 0.05 sin corrección de Bonferroni), sino que también es necesario determinar la relevancia de las mismas. Para estudiar el efecto que tienen los grupos en aquellos casos en los que si aparecen diferencias, se ha utilizado la d de Cohen mediante la estandarización de las diferencias de medias. Este test se utiliza para medir la asociación entre variables cuantitativas (medidas en escala continua) y cualitativas (variables dicotómicas). Para interpretar el índice, se utiliza la escala descrita en Eq. 9 ([24]):

$$\text{Efecto} = \begin{cases} \text{Pequeño} & \text{si } d \in [0, 0.3] \\ \text{Medio} & \text{si } d \in [0.5, 0.8] \\ \text{Grande} & \text{si } d > 0.8 \end{cases} \tag{9}$$

7. Todos los análisis inferenciales se han hecho con $\alpha = 0.05$.

V. RESULTADOS

La sección de resultados está organizada en función de las preguntas de investigación planteadas.

V-A. PI1

En la tabla II, podemos encontrar los estadísticos descriptivos del cálculo de $\sigma_{t,t+\triangle t}(Z)$ y $\sigma^{Y|X}_{t,t+\triangle t}$. Los valores obtenidos no son muy elevados, por lo que podemos concluir que la deriva que se produce en las variables de geolocalización es baja. Además, hay más deriva virtual que real.

Tabla II
RESULTADOS DEL ANÁLISIS DE LA DERIVA CONCEPTUAL

Descriptivo	**Marginal** $\sigma_{t,u}(Z)$	**Posterior** $\sigma_{t,u}^{Y\|X}$
Media	0.2724	0.1362
Mediana	0.2746	0.1373
Desviación típica	0.3164	0.1582
Mínimo	0.2053	0.1027
Máximo	0.3371	0.1686

V-B. PI2

En la tabla III, podemos encontrar los resultados en las variables respuesta cuando comparamos los resultados de los modelos creados con D_1 y D_2.

Tabla III
RESULTADOS DEL TEST U DE MANN- WHITNEY PARA COMPARAR D_1 CON D_2

Variable respuesta	**Z**	ρ**-valor**
MCC	-2.929	0.003
Accuracy	-2.170	0.030

En ambos casos, hay diferencias significativas. Podemos ver en la tabla IV, los estadísticos descriptivos de ambas muestras.

Tabla IV
ESTADÍSTICOS DESCRIPTIVOS DE D_1 Y D_2

Variable respuesta		$\overline{X}$	σ	$Mediana$
MCC	D_1	0.7776	0.1168	0.7788
	D_2	0.7770	0.2732	0.7889
Accuracy	D_1	0.8640	0.0072	0.8648
	D_2	0.8626	0.1664	0.8696

La mediana en D_2 es mayor que en D_1 para ambos casos, en MCC y en *Accuracy*. Bien es cierto que la media en ambos casos es ligeramente superior para D_1, pero la desviación es menor. En general, se puede concluir que la presencia de desviación en la geolocalización de las IPs influye en el rendimiento de los modelos. En este experimento, los modelos no se ven demasiado perjudicados, probablemente porque la diferencia en la geolocalización no es muy elevada, solo 374 de las muestras ven modificadas su país de origen. Si estudiamos la relevancia de las diferencias, el efecto es pequeño ($d_{Cohen}(MCC) = 0.039, d_{Cohen}(Accuracy) = 0.10889$). Es probable que con cambios más bruscos de geolocalización, la capacidad predictiva del modelo varíe en un grado mayor, así que la bondad de ajuste de los resultados dependerá de cómo de representativa sea la magnitud de los cambios de la deriva conceptual de la geolocalización.

Por otra parte, podemos observar que el MCC en ambos casos es de, aproximadamente, 0.77, y el *Accuracy* se aproxima a 0.86. Aunque no sean tasas de ajuste superiores al $90\,\%$, sí que superan a las tasas conseguidas en otros trabajos de investigación. Cabe destacar que sería adecuado realizar los experimentos sobre los mismos conjuntos de datos para poder extraer conclusiones generales.

V-C. PI3

En la tabla V, podemos observar que existen diferencias significativas en el MCC cuando varíamos el parámetro h, tanto para D_1 como para D_2. Sin embargo, en la variable *Accuracy*, estas diferencias únicamente aparecen para D_1.

Tabla V
RESULTADOS TEST DE KRUSKALL-WALLIS ENTRE GRUPOS DADOS POR h

Variable respuesta		**H de Kruskall**	ρ**-valor**
MCC	D_1	16.433	0.002
	D_2	10.481	0.033
Accuracy	D_1	17.869	0.001
	D_2	9.441	0.051

Pasamos entonces a estudiar en detalle estas diferencias para extraer alguna conclusión sobre qué valores de h son los más óptimos. En la tabla VI, vemos las comparaciones *Post Hoc.*

Tabla VI
COMPARATIVA POST HOC CON CORRECCIÓN DE BONFERRONI

Variable respuesta	**Comparación h's**	ρ**-valor**
MCC(D_1)	$6-7$	0.136
	$6-8$	0.096
	$6-9$	1
	$6-10$	0.003
	$7-8$	1
	$7-9$	1
	$7-10$	1
	$8-9$	1
	$8-10$	1
	$9-10$	0.081
Accuracy(D_1)	$6-7$	0.540
	$6-8$	0.017
	$6-9$	1
	$6-10$	0.003
	$7-8$	1
	$7-9$	1
	$7-10$	0.918
	$8-9$	0.302
	$8-10$	1
	$9-10$	0.081
MCC(D_2)	$6-7$	1
	$6-8$	1
	$6-9$	0.155
	$6-10$	1
	$7-8$	11
	$7-9$	0.294
	$7-10$	1
	$8-9$	0.397
	$8-10$	1
	$9-10$	0.025

Como podemos observar, las diferencias no se dan entre todos los pares comparados. En el caso de D_1, las diferencias más significativas aparecen entre $h = 6$ y $h = 10$, tanto para el MCC como para el *Accuracy*. Y para el *Accuracy,* también para $h = 6$ versus $h = 8$. Los mejores ajustes de esas comparaciones se alcanzan con $h = 6$ (Mediana[MCC D_1]=0.7871, Mediana[*Accuracy* D_1]=0.8702 para $h = 6$.). Para D_2, las únicas diferencias aparecen cuando comparamos $h = 9$ y $h = 10$, debiéndose las mismas a que el MCC que se alcanza en D_2 con $h = 10$ es el más bajo de todos (Mediana[MCC D_2]=0.73226 para $h = 10$). Si estudiamos la relevancia del efecto de h, esta es elevada, véase tabla VII.

Tabla VII
VALOR DE d DE COHEN

Variable respuesta	**Comparación h's**	d**- Cohen**
MCC(D_1)	$6-7$	1.4158
Accuracy(D_1)	$6-8$	1.7586
	$6-10$	1.4586
MCC(D_2)	$9-10$	1.7057

Habría que profundizar entonces en el estudio de los resultados al combinar h con diferentes valores de δ, y de los recursos computacionales consumidos que requiere cada combinación.

V-D. PI4

En la tabla VIII, podemos observar que no existen diferencias significativas cuando varíamos el parámetro δ.

Tabla VIII
RESULTADOS TEST DE KRUSKALL-WALLIS ENTRE GRUPOS DADOS POR δ

Variable respuesta		H de Kruskall	ρ-valor
MCC	D_1	0.241	0.971
	D_2	0.439	0.637
Accuracy	D_1	0.241	0.971
	D_2	0.932	0.888

V-E. Discusión

Como conclusión, el mejor ajuste se alcanza con la siguiente combinación de hiper-paramétros:

1. MCC(D_1) = 0.7933 con $h = 9$, y $\delta = 0.001, 0.002, 0.003, 0.004$.
2. *Accuracy*(D_1) = 0.8729 con $h = 9$ y $\delta = 0.001, 0.002, 0.003, 0.004$ ó $h = 6$ y $\delta = 0.001$.
3. MCC(D_2) = 0.7871 con $h = 6$ y $\delta = .001, 0.002, 0.003, 0.004$.
4. *Accuracy*(D_2) = 0.8702 con $h = 6$ y $\delta = .001, 0.002, 0.003, 0.004$.

Para el caso de D_1, es evidente que el mejor resultado se consigue con $h = 9$, ya que obtiene el MCC y la *Accuracy* más altos para cualquier valor de δ . Sin embargo, habría que determinar si la pérdida de ajuste en el MCC es demasiada en relación con $h = 6$, ya que el *Accuracy* es el mismo para ambos valores de h, pero el consumo de recursos es menor para $h = 6$ por tener que gestionar una menor cantidad de datos al construir la serie temporal.

Para el caso de D_2, los mejores resultados se alcanzan claramente con $h = 6, \delta = 0.001, 0.002, 0.003, 0.004$ para ambas variables respuesta.

Por lo tanto, en general, el mejor resultado se obtiene con el valor de $h = 6$, el más bajo que se ha probado. Una posible pregunta de investigación futura sería determinar cuánto puede disminuir el valor de este hiper-parámetro manteniendo un buen ajuste.

VI. CONCLUSIONES

En este trabajo se ha extraído un conjunto de *features* que es significativo para categorizar, de manera multi-clase, la maliciosidad de una IP. Las variables son de doble naturaleza: temporal y de geolocalización. Los resultados alanzados tienen tasas de ajuste cercanas a 0.77 para el MCC, y a 0.86 para el *Accuracy.* Aunque estos valores no sean muy elevados, si que superan a otros trabajos de caracterización de reputación de IPs, teniendo en cuenta, además, que estos son para categorización bi-clase y suelen requerir una extracción de variables costosa.

Además, en el estudio hemos analizado el impacto de la deriva conceptual en las *features* relacionadas con la geolocalización, así como la influencia de los hiper-parámetros necesarios para calcular las variables que se extraen de las características temporales.

En los resultados se observa que no existe demasiado impacto en la capacidad predictiva de los modelos a causa de los cambios en la información de geolocalización; esto puede deberse a que la magnitud de los cambios es relativamente baja: solo 374 de las muestras tienen diferencias, por ejemplo, en el país de origen. Así, podría concluirse que los cambios en las *features* contextuales podría no ser tan relevante para los modelos como el tratamiento de las IPs cuya clasificación cambia, por ejemplo, al cesar la actividad maliciosa. Desde el punto de vista de las variables temporales, sí que encontramos que el efecto de uno de los hiper-parámetros es elevado, por lo que habrá que estudiar cuáles son los valores más óptimos del mismo.

Como trabajo futuro o de extensión, la investigación va dirigida a determinar si se puede realizar una selección de *features* del conjunto que aporte mejores resultados que los obtenidos. Además, trataremos de determinar el valor más óptimo de los hiper-parámetros que tienen influencia sobre los modelos. Por último, se estudiará la aplicación de más algoritmos de aprendizaje automático, y se estudiará la posible deriva conceptual con períodos más largos de tiempo. Todos los objetivos propuestos, se intentarán llevar a cabo sobre conjuntos de datos públicos para que puedan ser replicados y comparados con otras investigaciones.

AGRADECIMIENTOS

Este trabajo se enmarca dentro de os contratos art. 83 *Adenda 3: Machine learning para la calidad de los datos del modelo de inteligencia de INCIBE* y *Adenda 7: Prórroga de la Adenda 3* entre la Universidad de León e INCIBE en el periodo 2018-2022. Además, queremos agradecer a Diego Asterio de Zaballa el trabajo realizado en RIASC desde septiembre de 2020 a septiembre de 2021 y que forma parte de esta investigación.

REFERENCIAS

[1] Microsoft, "Security update severity rating system," Recuperado de https://www.microsoft.com/en-us/msrc/security-update-severity-rating-system.
[2] Forum of Incident Response and Security Teams (FIRST), "Common vulnerability scoring system," Recuperado de https://www.first.org/cvss/calculator/3.0.
[3] OWASP Foundation, "Owasp testing guide v4: Owasp risk rating methodology," Recuperado de https://www.owasp.org/index.php/OWASP_Risk_Rating_Methodology.
[4] Cybersecurity and Infrastructure Security Agency (CISA), "Nciss cyber incident scoring system," Recuperado de https://www.us-cert.gov/NCCIC-Cyber-Incident-Scoring-System.
[5] N. DeCastro-García, Á. L. Muñoz Castañeda, and M. Fernández-Rodríguez, "Machine learning for automatic assignment of the severity of cybersecurity events," *Computational and Mathematical Methods*, vol. 2, no. 1, p. e1072, 2020.
[6] "Firehol - linux firewalling and traffic shaping for humans," Recuperado de https://firehol.org/.
[7] Y. Liu, J. Zhang, A. Sarabi, M. Liu, M. Karir, and M. Bailey, "Predicting cyber security incidents using feature-based characterization of network-level malicious activities," in *Proceedings of the 2015 ACM International Workshop on International Workshop on Security and Privacy Analytics*, ser. IWSPA '15. New York, NY, USA: Association for Computing Machinery, 2015, p. 3–9. [Online]. Available: https://doi.org/10.1145/2713579.2713582
[8] D. Likhomanov and V. Poliukh, "Predicting malicious hosts by blacklisted ipv4 address density estimation," in *2020 IEEE 11th International Conference on Dependable Systems, Services and Technologies (DESSERT)*, 2020, pp. 102–109.

[9] B. Coskun, "(Un)wisdom of crowds: Accurately spotting malicious ip clusters using not-so-accurate ip blacklists," *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 6, p. 1406–1417, jun 2017. [Online]. Available: https://doi.org/10.1109/TIFS.2017.2663333
[10] "Maxmind," Recuperado de https://www.maxmind.com/en/home.
[11] "Ipqualityscore," Recuperado de https://www.ipqualityscore.com/.
[12] J. L. Lewis, G. F. Tambaliuc, H. S. Narman, and W.-S. Yoo, "Ip reputation analysis of public databases and machine learning techniques," in *2020 International Conference on Computing, Networking and Communications (ICNC)*, 2020, pp. 181–186.
[13] Y. Huang, J. Negrete, A. Wosotowsky, J. Wagener, E. Peterson, A. Rodriguez, and C. Fralick, "Detect malicious ip addresses using cross-protocol analysis," in *2019 IEEE Symposium Series on Computational Intelligence (SSCI)*, 2019, pp. 664–672.
[14] N. Usman, S. Usman, F. Khan, M. A. Jan, A. Sajid, M. Alazab, and P. Watters, "Intelligent dynamic malware detection using machine learning in ip reputation for forensics data analytics," *Future Generation Computer Systems*, vol. 118, pp. 124–141, 2021.
[15] J. a. Gama, I. Žliobaite, A. Bifet, M. Pechenizkiy, and A. Bouchachia, "A survey on concept drift adaptation," *ACM Computing Surveys*, vol. 46, no. 4, 2014.
[16] G. Webb, R. Hyde, H. Cao, H. Nguyen, and F.Petitjean, "Characterizing concept drift," *Data Mining and Knowledge Discovery*, vol. 30, no. 4, pp. 964–994, 2016.
[17] J. Lu, A. Liu, F. Dong, F. Gu, J. Gama, and G. Zhang, "Learning under concept drift: A review," *IEEE Transactions on Knowledge and Data Engineering*, vol. 31, no. 12, pp. 2346–2363, 2019.
[18] G. Webb, L. Lee, and B. Goethals, "Analyzing concept drift and shift from sample data," *Data Mining and Knowledge Discovery*, vol. 32, pp. 1179 – 1199, 2018.
[19] D. Levin, Y. Peres, and E. Wilmer, *Markov chains and mixing times*. American Mathematical Society, Providence, 2008.
[20] M. Feurer, A. Klein, K. Eggensperger, J. Springenberg, M. Blum, and F. Hutter, "Efficient and robust automated machine learning," in *Proceedings of 28 Conference in Advances in Neural Information Processing Systems*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds., 2015, pp. 2962–2970.
[21] F. Hutter, H. H. Hoos, and K. Leyton-Brown, "Sequential model-based optimization for general algorithm configuration," in *Proceedings of 5th Conference in Learning and Intellingent Optimization.*, C. A. C. Coello, Ed., 2011, pp. 507–523.
[22] B. Matthews, "Comparison of the predicted and observed secondary structure of t4 phage lysozyme," *Biochimica et Biophys. Acta (BBA) - Protein Struct.*, vol. 405, no. 2, pp. 442–451, 1975.
[23] D. Chicco and G. Jurman, "The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, no. 1, 2020.
[24] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences (2nd ed.).* Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers., 1988.