UNIVERSIDAD DEL VALLE DE GUATEMALA

Facultad de Ingeniería

Laboratorio 01

Detección de Pishing

Edwin Andrés Ortega Kou – 22305

Esteban Zambrano Garoz – 22119

Security Data Science

Guatemala, febrero 2026

Parte 1 – Ingeniería de características

Exploración de Datos

El conjunto de datos fue cargado en un DataFrame utilizando la librería pandas. Posteriormente, se visualizaron las primeras cinco observaciones con el objetivo de comprender la estructura del dataset y verificar el formato de las variables disponibles. El dataset cuenta con dos columnas principales: url, que contiene la dirección web a analizar, y status, que indica si la URL corresponde a un sitio legítimo o a un sitio de phishing.

La inspección inicial permitió confirmar que las URLs se encuentran en formato de texto crudo y que la variable status es de tipo categórico, con valores asociados a las clases de interés para el problema de clasificación.

```
== Cargando dataset ==
Shape: (11430, 2)

Head:
                                                 url      status
0             http://www.crestonwood.com/router.php  legitimate
1  http://shadetreetechnology.com/V4/validation/a...    phishing
2  https://support-appleld.com.secureupdate.duila...    phishing
3                                http://rgipt.ac.in  legitimate
4  http://www.iracing.com/tracks/gateway-motorspo...  legitimate

Class counts:
status
legitimate    5715
phishing      5715
Name: count, dtype: int64

Class %:
status
legitimate    50.0
phishing      50.0
Name: proportion, dtype: float64
```

Como parte de la exploración inicial, se analizó la distribución de las observaciones según la etiqueta status. El conjunto de datos contiene un total de 11,430 observaciones, de las cuales 5,715 corresponden a URLs legítimas y 5,715 a URLs de phishing.

Esto representa una distribución 50% legítimas y 50% phishing, lo que indica que el dataset se encuentra perfectamente balanceado. Debido a este balance, no fue necesario aplicar técnicas adicionales de re-muestreo, para corregir posibles sesgos entre clases. Este balance facilita el entrenamiento de los modelos de clasificación y permite una evaluación más justa de su desempeño.

Derivación de Características

¿Qué ventajas tiene el análisis de una URL contra el análisis de otros datos, cómo el tiempo de vida del dominio, o las características de la página Web?

R/ Ya que la URL está disponible inmediatamente al recibir un enlace y no requiere consultas externas ni cargar el contenido de la página web. A diferencia de datos como la antigüedad del dominio o las características visuales del sitio, este enfoque no depende de información histórica ni de la ejecución del sitio, lo que reduce el costo computacional y los riesgos de seguridad.

¿Qué características de una URL son más prometedoras para la detección de phishing?

R/ Las características más prometedoras para detectar phishing a partir de una URL son las que están relacionadas con la estructura y complejidad, como la longitud total, la cantidad de subdominios, el número de parámetros y el uso de caracteres especiales. Las URLs de phishing suelen ser más largas y complejas, con el objetivo de confundir al usuario y simular legitimidad. También, indicadores como la codificación hexadecimal, la presencia de palabras clave asociadas a servicios legítimos y las medidas de entropía permiten identificar URLs con patrones poco naturales o generados automáticamente, lo cual es común en ataques de phishing.

Con base en el análisis anterior, se definieron e implementaron veintiocho funciones para derivar características numéricas a partir de cada URL, las cuales fueron añadidas al dataset original. Estas funciones permiten transformar las URLs, originalmente en formato de texto, en representaciones numéricas que pueden ser utilizadas por modelos de aprendizaje automático.

Las características derivadas incluyen:

- Longitud total de la URL.
- Longitud del dominio (host).
- Longitud del path.
- Longitud del query string.
- Número de puntos en la URL.
- Número de subdominios.
- Número de guiones y guiones bajos.
- Número de caracteres especiales como /, ?, =, & y %.
- Proporción de caracteres numéricos, alfabéticos y no alfanuméricos.
- Frecuencia máxima de repetición de un mismo carácter.
- Presencia de símbolos sospechosos como @.
- Uso de direcciones IP en lugar de nombres de dominio.
- Presencia de puertos explícitos.
- Uso de acortadores de URL.

- Uso de dominios de nivel superior poco comunes.
- Presencia de codificación hexadecimal.
- Número de palabras clave comúnmente asociadas a ataques de phishing.

Preprocesamiento

La variable categórica status fue transformada a una variable binaria, asignando 0 a las URLs legítimas y 1 a las URLs de phishing, con el fin de facilitar el entrenamiento de los modelos de clasificación. También, se evitó el uso de la URL o del dominio como variable de entrada para prevenir posibles problemas de data leakage. Las características derivadas se encuentran en formato numérico, por lo que no fue necesario aplicar técnicas adicionales de codificación en esta etapa.

Selección de Características

Se analizaron las características derivadas para eliminar aquellas constantes, redundantes o poco informativas. Para ello, se evaluó la varianza de las variables y la correlación entre ellas, descartando características altamente correlacionadas. Asimismo, se verificó que el dataset no contuviera observaciones duplicadas. Este proceso permitió seleccionar un conjunto de variables relevantes y no redundantes para la clasificación de URLs legítimas y de phishing.

¿Qué columnas o características fueron seleccionadas y por qué?

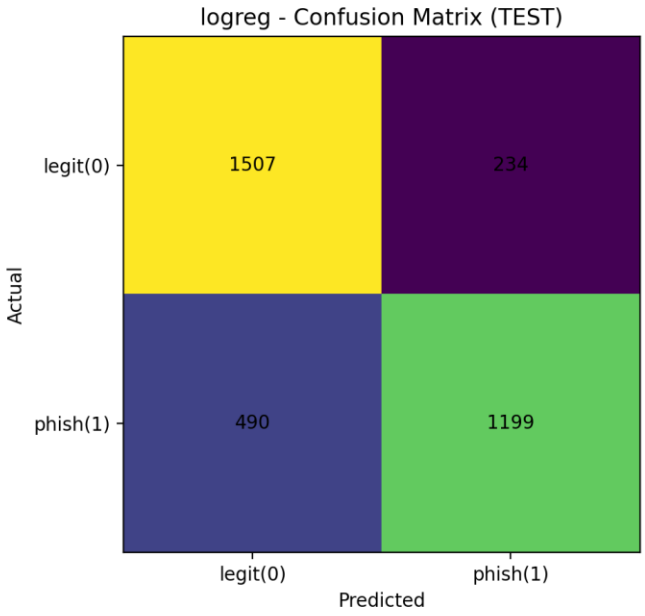
R/ Se seleccionaron 28 características, las cuales describen la longitud, estructura y complejidad de las URLs, el uso de caracteres especiales, patrones comúnmente asociados al phishing y medidas de entropía. Estas características fueron elegidas porque presentan mayor capacidad para diferenciar entre URLs legítimas y URLs de phishing sin introducir redundancia ni ruido en el modelo.

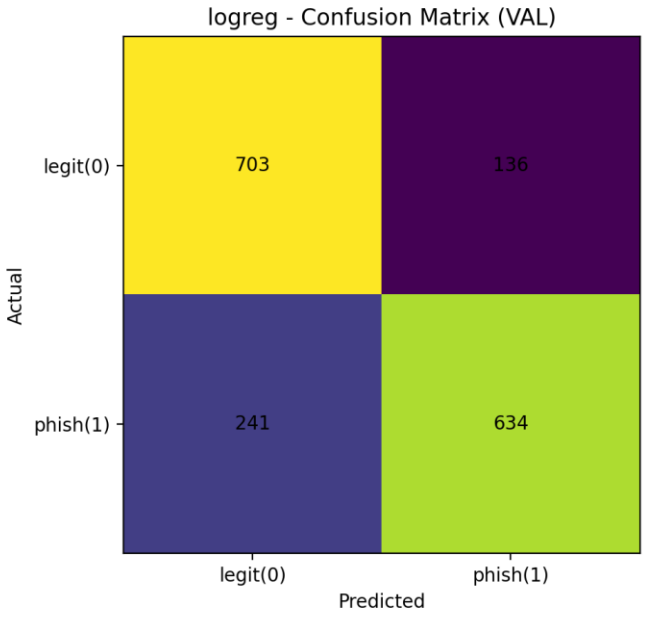
Parte 2 – Implementación

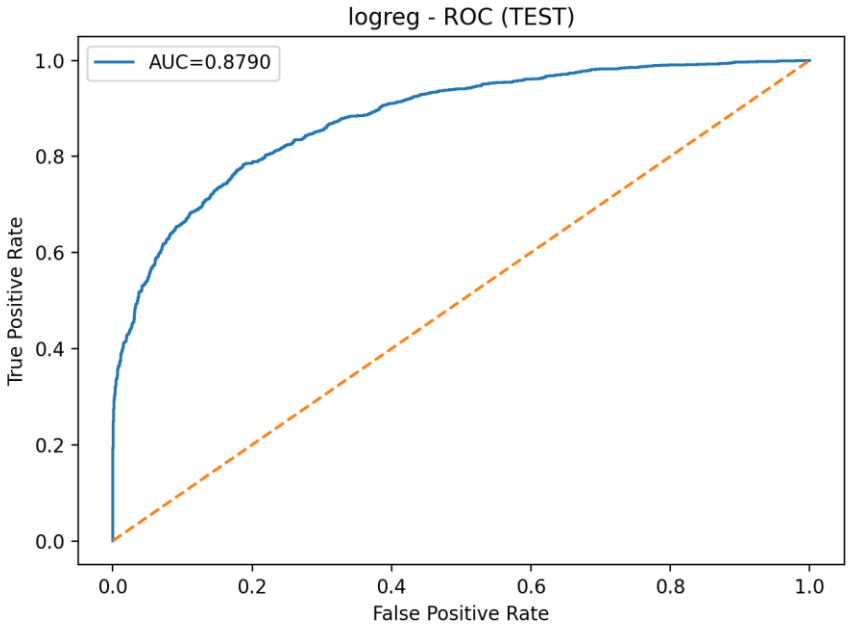
Implementación y evaluación de modelos

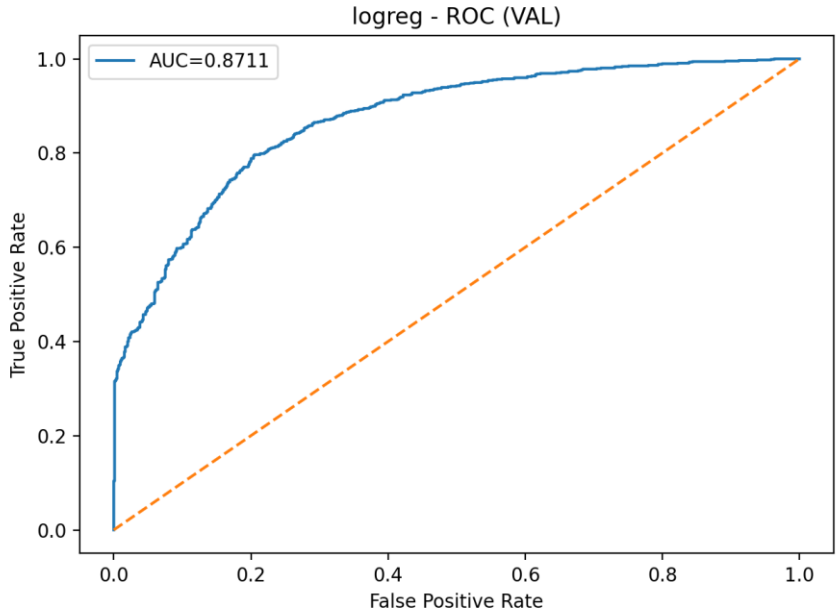
Para la evaluación comparativa de los modelos, se han medido las métricas de precision, recall y AUC-ROC, el cual el problema de detección de phishing presenta unas aunque el dataset de entrenamiento es balanceado, el problema real presenta desbalance que afectan los errores de clasificación, lo cual conlleva un coste elevado. La métrica de precision hace referencia a cuántas de las alertas emitidas realmente correspondían a ataques de phishing, algo de suma importancia para evitar que las alertas tengan un elevado número de falsos positivos, que podrían saturar las alertas. La métrica de recall corresponde a la capacidad del modelo de detectar correos electrónicos maliciosos reales, mediante el cual se podrá minimizar los falsos negativos, es decir, los ataques que fueron desestimados por el sistema. Finalmente, el AUC-ROC corresponde a la capacidad que tiene el propio modelo de discriminar entre ambas clases y a la vez la explica independientemente del umbral de decisión, lo cual permite comparar entre los distintos clasificadores.

El modelo de regresión logística presentó un desempeño estable al entre validación y prueba validando, obteniéndose un AUC de 0.871 en validación y 0.879 en prueba, lo que indica una buena capacidad separadora entre correo legítimo y de phishing. En el conjunto de prueba, la precisión es de 0.8367 con un recall de 0.7099, indicando que el modelo intenta reducir los falsos positivos a costa de perder algunos ataques reales. Esto se puede observar en la matriz de confusión donde aún quedan falsos negativos de relevancia. En un escenario de base rate del 15% el modelo tiene 5,324 verdaderos positivos frente a 5,712 falsos positivos, que evidencian un compromiso aceptable entre detección y generación de alertas.


logreg - Confusion Matrix (TEST)
1507
234
490
1199
legit(0)
phish(1)
Actual
legit(0)
phish(1)
Predicted

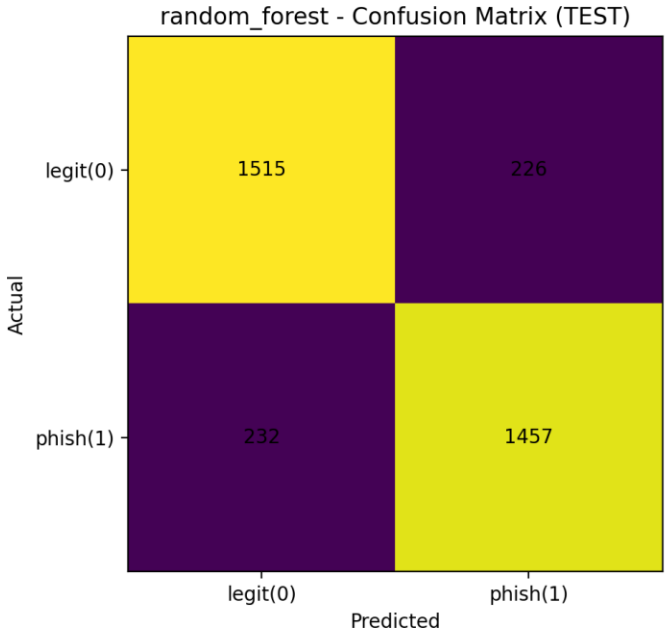

logreg - Confusion Matrix (VAL)
703
136
241
634
legit(0)
phish(1)
Actual
legit(0)
phish(1)
Predicted

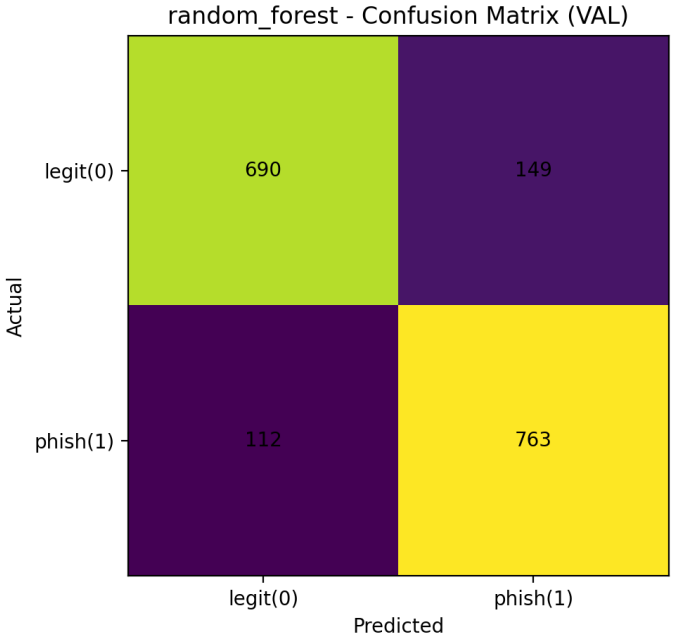


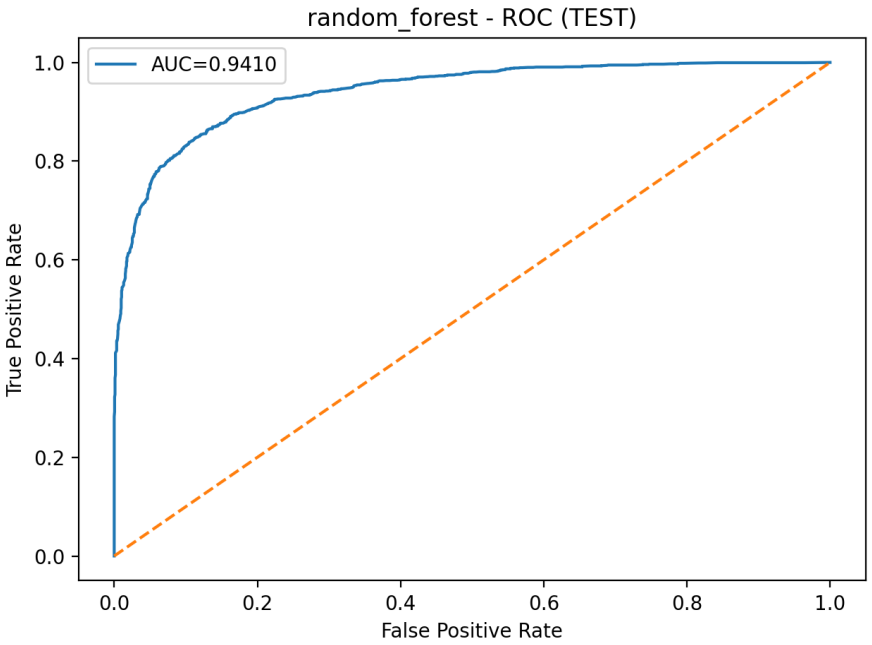


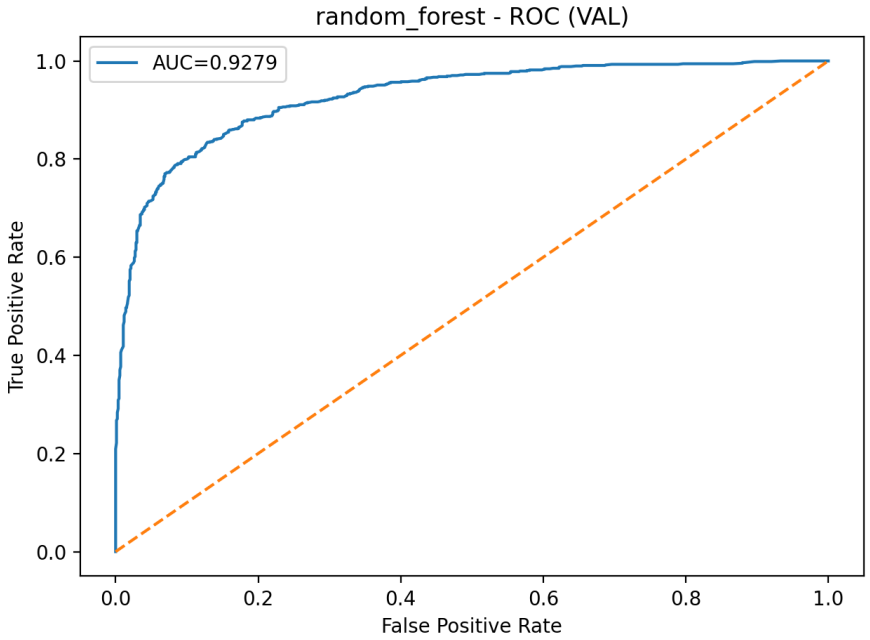
El modelo de Random Forest mostró un rendimiento superior en comparación con el modelo de regresión logística para todos los parámetros añadidos. En el conjunto de prueba, el modelo alcanzó valores de precisión de 0.8657, recall de 0.8626 y AUC de 0.9410, lo que sugiere que esta técnica proporciona un nivel de detección de ataques bastante alto sin aumentar él número de falsos positivos. También se puede observar que la matriz de confusión se comporta mucho mejor,

mostrando un comportamiento importante de los falsos negativos en comparación con la matriz confusa del modelo linear. En el escenario base rate del 15% se detectan 6,470 ataques reales con 5,517 falsos positivos por lo que Random Forest es un modelo bastante adecuado para un entorno de seguridad en el cual es importante detectar ataques lo antes posible.









En conclusión, los resultados indican que Random Forest es el que ofrece el mejor compromiso entre precisión y recall, algo que lo convierte en un método más efectivo para la detección de phishing en los escenarios realistas. Aun así, la regresión logística sigue siendo una opción interpretable y más simple computacionalmente.

4. **¿Cuál es el impacto de clasificar un sitio legítimo como phishing?**
   La detección de un falso positivo puede hacer que un sitio totalmente legítimo sea bloqueado, lo que a su vez suele implicar la frustración de los usuarios ante interrupciones en el acceso a recursos válidos y con pérdida de productividad. Es más, si además se dan altos niveles de falsos positivos esto puede dar lugar a la fatiga de las alertas disminuyendo la confianza en el sistema de detección.

5. **¿Cuál es el impacto de clasificar un sitio de phishing como legítimo?**
   Un falso negativo genera la posibilidad de acceder a un sitio malicioso al que los usuarios están expuestos a un robo de credenciales, fraude, eventual acceso al sistema de la organización víctima o al compromiso de seguridad. Suele tener un impacto más fuerte que un falso positivo, normalmente puede derivar en incidentes de seguridad con costes económicos y repercusiones reputacionales.

6. **En base a las respuestas anteriores, ¿Qué métrica elegiría para comparar modelos similares de clasificación de phishing?**
   La métrica dominante que se usa para llevar a cabo la comparación de modelos principales es el recall, dado que en la temática de phishing, se prefiere poder identificar la mayor cantidad posible de sitios maliciosos; además el AUC se utiliza para determinar la capacidad del modelo como discriminador de URLs legítimas y de phishing sin tenerse en cuenta el umbral de decisión utilizado.

7. **¿Qué modelo funcionó mejor para la clasificación de phishing? ¿Por qué?**
   El modelo Random Forest mostró un mejor rendimiento frente a la Regresión Logística. En el conjunto de pruebas se obtuvieron mayores valores de recall y AUC, lo cual indica una mayor capacidad de detectar URLs de phishing sin aumentar considerablemente la cantidad de falsos positivos. Por esta razón, se considera este modelo como el que mejor se adecúa a este problema.

8. **Una empresa desea utilizar su mejor modelo, debido a que sus empleados sufren constantes ataques de phishing mediante e-mail. La empresa estima que, de un total de 50,000 emails, un 15% son phishing. ¿Qué cantidad de alarmas generaría su modelo? ¿Cuántas positivas y cuantas negativas? ¿Funciona el modelo para el BR propuesto? En caso negativo, ¿qué propone para reducir la cantidad de falsas alarmas?**
   En el supuesto de contar con 50.000 correos electrónicos y una tasa real de phishing de un 15%, el modelo Random Forest identificaría correctamente aproximadamente 6.470 emails de phishing y generaría cerca de 5.517 falsos positivos. De esta forma, el número de alertas llegaría a la cifra cercana a 11.987.

Aunque el modelo muestra buen comportamiento en la detección de ataques, el volumen de falsos positivos indica que el sistema podría beneficiarse de algunas mejoras como la variación del umbral de decisión, la priorización de alertas de alto riesgo o bien la incorporación de revisión humana para los casos dudosos, con el fin de disminuir la carga de trabajo.