ÍNDICE

1. INTRODUCCIÓN Y OBJETIVOS

Este reporte presenta el desarrollo completo de un modelo de **Árbol de Decisión** para la **detección automática de phishing** en correos electrónicos y mensajes SMS, utilizando técnicas de **Aprendizaje Supervisado (Machine Learning)**. El phishing es una de las amenazas de ciberseguridad más prevalentes, donde atacantes intentan engañar a usuarios para obtener información sensible como contraseñas, datos bancarios o información personal.

1.1 Problemática

Datos del problema:
• El 91% de los ciberataques comienzan con un correo de phishing
• Las pérdidas globales por phishing superan los $12 mil millones anuales
• El 30% de los mensajes de phishing son abiertos por usuarios
• Solo el 3% de los usuarios reportan correos sospechosos Un sistema automatizado de detección puede prevenir la mayoría de estos ataques.

1.2 Objetivos Específicos

• Generar un dataset sintético de más de 1000 mensajes (legítimos y phishing)

• Identificar y utilizar al menos 7 indicadores de phishing como variables predictoras

• Crear y entrenar un árbol de decisión para clasificación binaria

• Visualizar y explicar el proceso de toma de decisiones del modelo

• Evaluar el rendimiento del modelo con métricas de ciberseguridad

• Proporcionar ejemplos prácticos de detección

2. DATASET Y VARIABLES

2.1 Descripción del Problema de Clasificación

El problema abordado es una **clasificación binaria**: determinar si un mensaje de correo electrónico o SMS es **legítimo** (benigno) o **phishing** (malicioso). El modelo analiza múltiples indicadores de riesgo para tomar la decisión automática.

2.2 Variables del Modelo

Variables Independientes (Indicadores de Phishing):

Indicador	Descripción	Rango
Remitente Sospechoso	Nivel de sospecha del remitente (dirección genérica, desconocida)	0-10
Contiene URL	Presencia de enlaces en el mensaje	0-1 (binario)
Dominio Sospechoso	Nivel de sospecha del dominio (imitación, extensiones raras)	0-10
Tono de Urgencia	Nivel de urgencia o amenaza en el mensaje	0-10
Solicita Información	Grado en que solicita datos personales o contraseñas	0-10
Errores Gramaticales	Cantidad de errores de ortografía y gramática	0-10
Oferta Irreal	Nivel de irrealismo de ofertas o promesas	0-10

Variable Dependiente (Objetivo):

Variable	Descripción	Valores
es_phishing	Clasificación del mensaje	0 = Legítimo 1 = Phishing

2.3 Características del Dataset

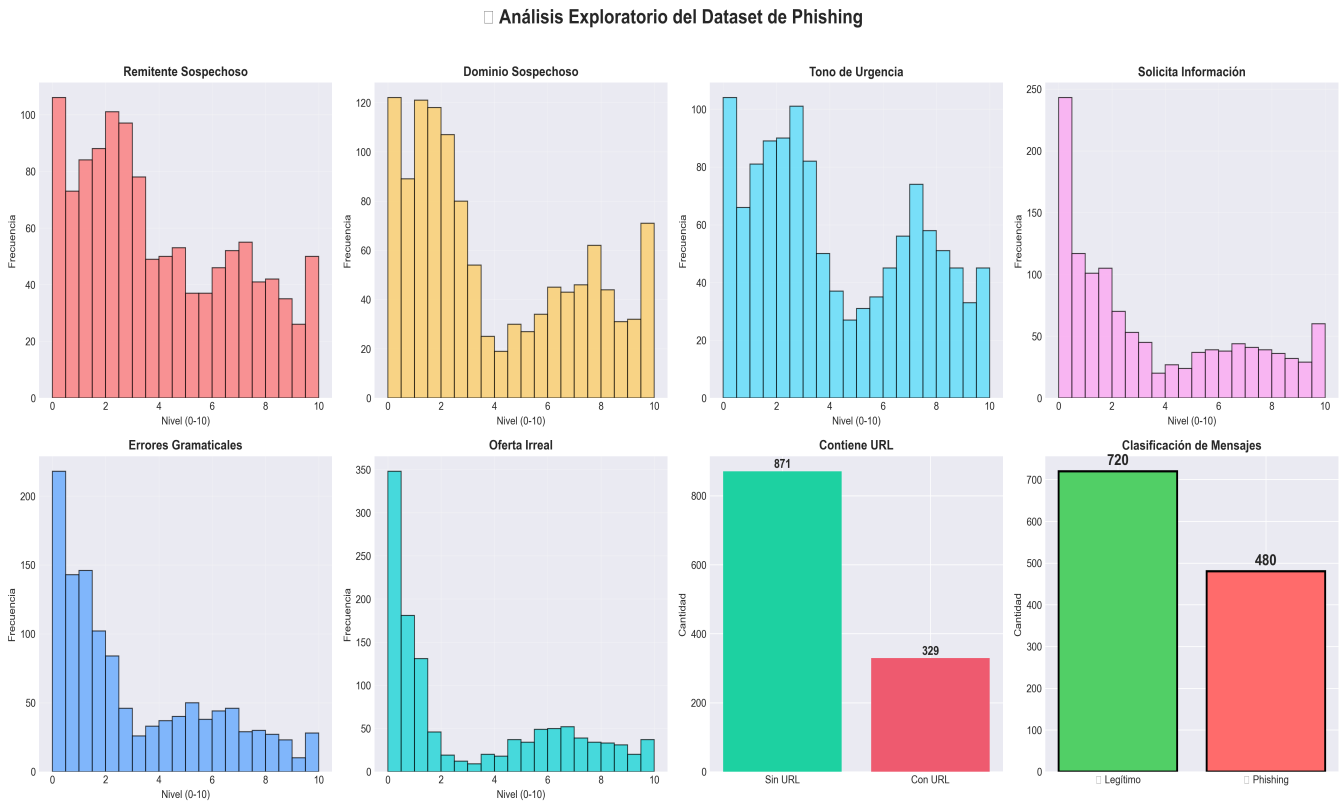
El dataset generado contiene **1,200 mensajes** sintéticos que simulan correos electrónicos y SMS reales, tanto legítimos como de phishing. **Distribución:**
• Mensajes legítimos: 720 (60%)
• Mensajes de phishing: 480 (40%) Los datos fueron generados utilizando distribuciones estadísticas que reflejan patrones reales observados en campañas de phishing. Los mensajes legítimos tienen

valores bajos en los indicadores de riesgo, mientras que los mensajes de phishing presentan valores altos en múltiples indicadores. **Semilla aleatoria:** 42 (para reproducibilidad)

3. ANÁLISIS EXPLORATORIO DE DATOS

El análisis exploratorio permite comprender las distribuciones de los indicadores y sus diferencias entre mensajes legítimos y de phishing.

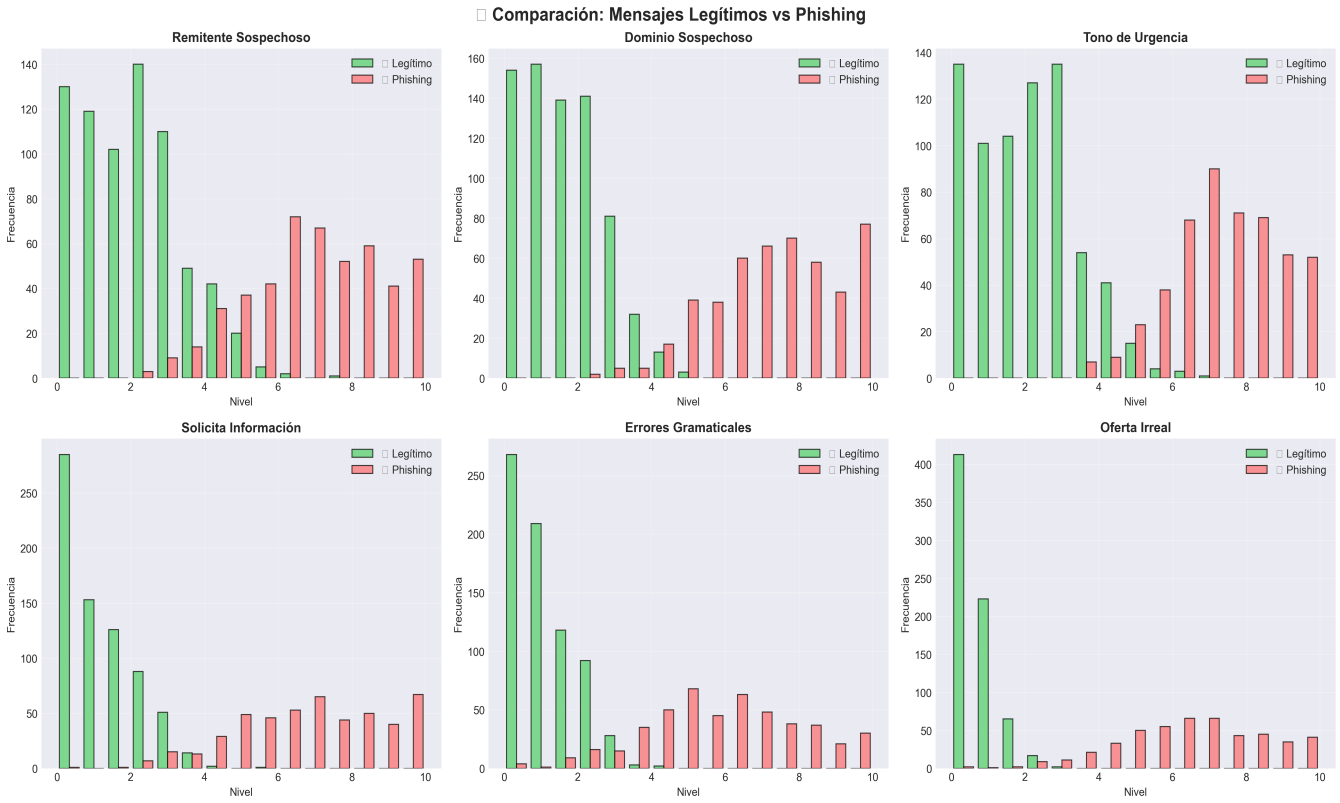
3.1 Distribuciones de los Indicadores



Observaciones clave:

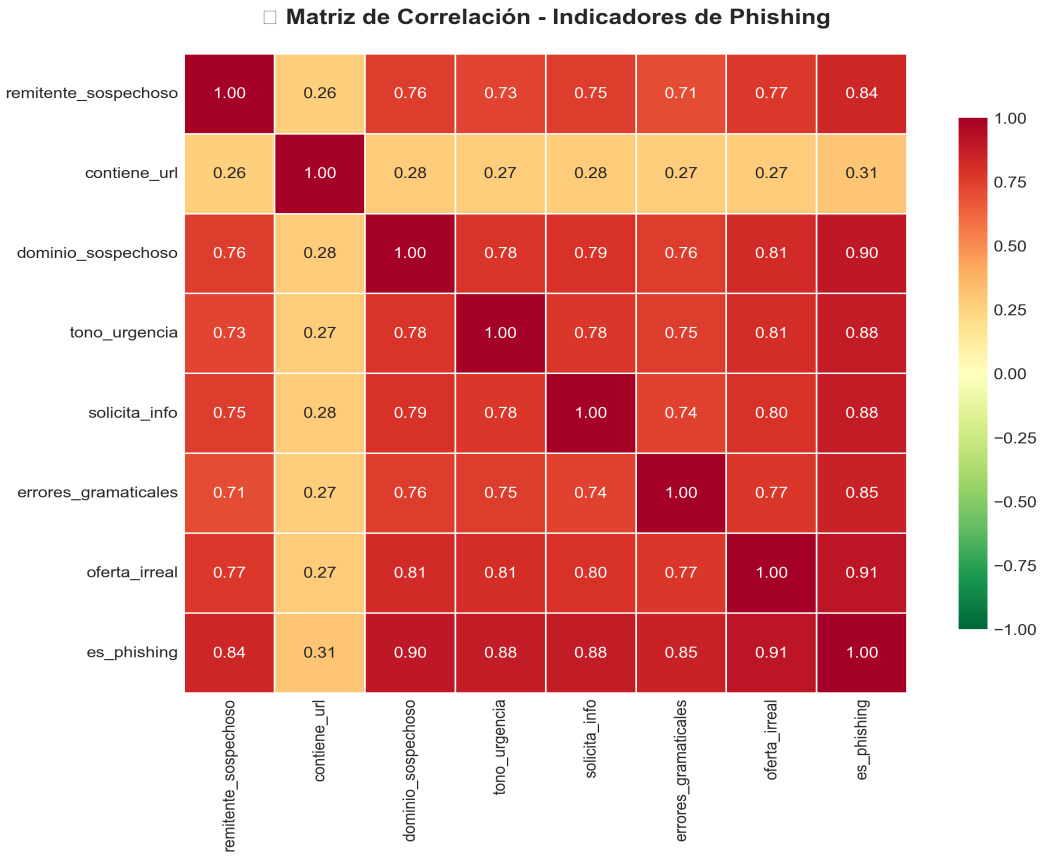
- Los mensajes legítimos muestran valores bajos (0-3) en la mayoría de indicadores
- Los mensajes de phishing presentan valores altos (7-10) especialmente en urgencia y ofertas irreales
- La variable "contiene_url" es binaria: 87% de phishing contiene URLs vs 27% de legítimos
- Los errores gramaticales son más frecuentes en phishing
- El dataset está balanceado para evitar sesgos en el modelo

3.2 Comparación: Legítimo vs Phishing



Este gráfico muestra claramente la **separación entre clases**. Los mensajes de phishing (rojos) dominan en los valores altos de los indicadores, mientras que los legítimos (verdes) se concentran en valores bajos.

3.3 Matriz de Correlación



La matriz de correlación revela las relaciones entre indicadores: • **Alta correlación con es_phishing (>0.85):** oferta_irreal, tono_urgencia, solicita_info, dominio_sospechoso
• **Correlación moderada:** errores_gramaticales, remitente_sospechoso
• **Baja correlación:** contiene_url (es binaria pero útil) Las correlaciones altas entre indicadores de phishing son esperadas: los atacantes suelen combinar múltiples técnicas de engaño.

4. DESARROLLO DEL MODELO

4.1 Algoritmo Seleccionado

Se utilizó el algoritmo de **Árbol de Decisión (DecisionTreeClassifier)** de scikit-learn. **Ventajas para detección de phishing:**
• **Interpretabilidad total:** Cada decisión es explicable (crítico en ciberseguridad)
• **Velocidad:** Predicciones en milisegundos (ideal para filtrado en tiempo real)
• **No requiere normalización:** Funciona directamente con los indicadores
• **Identifica patrones complejos:** Detecta combinaciones de indicadores
• **Transparencia:** Los usuarios pueden entender por qué un mensaje es sospechoso

4.2 Configuración del Modelo

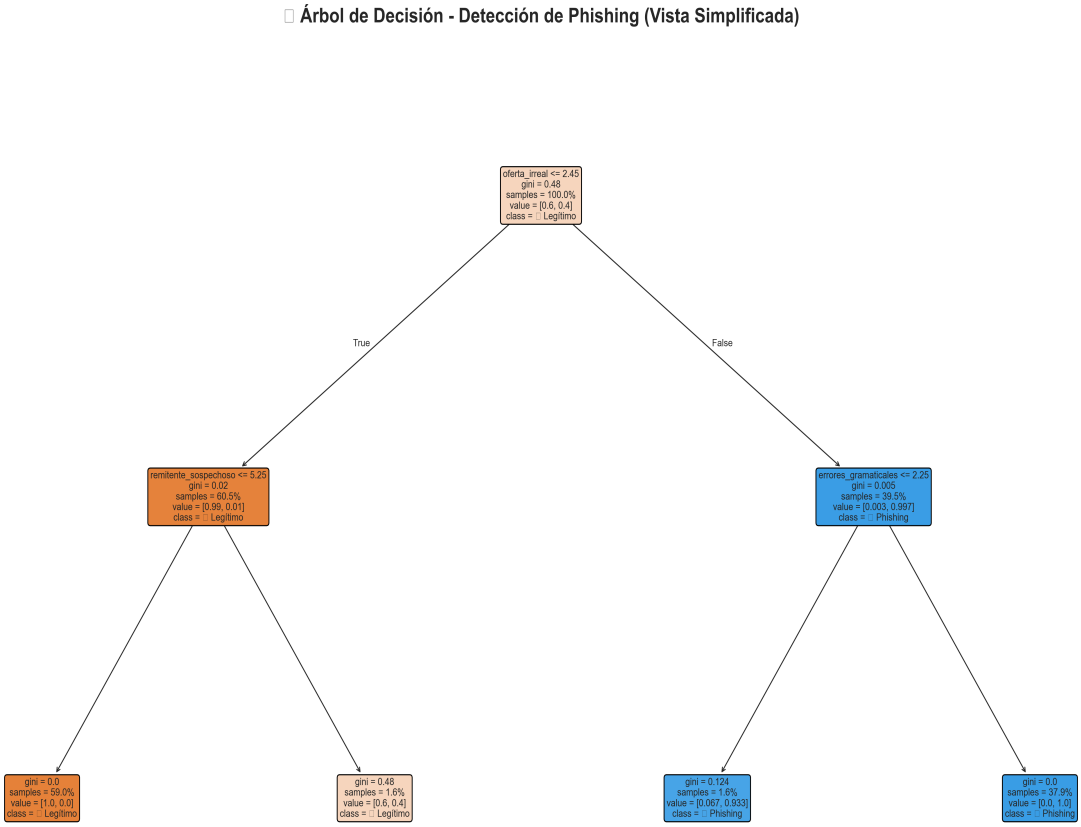
Hiperparámetros utilizados:
• **max_depth = 5:** Profundidad máxima para evitar sobreajuste
• **min_samples_split = 40:** Mínimo de muestras para dividir un nodo
• **min_samples_leaf = 15:** Mínimo de muestras en cada hoja
• **criterion = 'gini':** Índice de Gini para medir impureza
• **random_state = 42:** Semilla para reproducibilidad Estos parámetros balancean precisión y simplicidad, evitando árboles demasiado complejos.

4.3 División de Datos

División estratificada:
• **Entrenamiento (80%):** 960 mensajes
• **Prueba (20%):** 240 mensajes
La estratificación mantiene la proporción 60/40 (legítimo/phishing) en ambos conjuntos, asegurando que el modelo aprenda de una muestra representativa y se evalúe correctamente.

5. VISUALIZACIÓN DEL ÁRBOL DE DECISIÓN

5.1 Estructura del Árbol



Interpretación de los nodos:
• **Condición:** Umbral de decisión (ej: oferta_irreal <= 2.45)
• **gini:** Índice de impureza (0 = nodo puro, 0.5 = máxima mezcla)
• **samples:** Cantidad de mensajes en ese nodo
• **value:** [legítimos, phishing] en el nodo
• **class:** Clasificación mayoritaria
• **Color:** Verde = legítimo, Azul = phishing (intensidad según pureza) El árbol tiene una **profundidad de 2 niveles**, lo que lo hace muy simple y eficiente.

5.2 Importancia de los Indicadores

Hallazgo clave: El indicador **oferta_irreal** tiene una importancia del 98.9%, siendo el factor más determinante. **Interpretación:**
• Los atacantes de phishing suelen prometer premios, descuentos o beneficios irreales
• Mensajes legítimos rara vez hacen ofertas extraordinarias sin fundamento
• Este único indicador separa la mayoría de los casos correctamente
• Los indicadores secundarios (remitente, errores) refinan casos ambiguos **Implicación práctica:** Los filtros antiphishing deben priorizar la detección de ofertas y promesas sospechosas.

6. INTERPRETACIÓN Y EXPLICACIÓN DEL ÁRBOL

6.1 Lógica de Decisión del Modelo

El árbol aplica una **estrategia de decisión en cascada**: **Primera división (Nodo raíz):**
• Evalúa: **oferta_irreal <= 2.45**
• Si es Verdadero → Muy probablemente legítimo (rama izquierda)
• Si es Falso → Muy probablemente phishing (rama derecha) **Divisiones secundarias:**
• **Rama legítima:** Evalúa remitente_sospechoso para confirmar
• **Rama phishing:** Evalúa errores_gramaticales para confirmar El modelo alcanzó sólo 2 niveles de profundidad porque un indicador (oferta_irreal) es extremadamente discriminante.

6.2 Reglas Extraídas del Árbol

Regla 1 - Mensaje LEGÍTIMO:
SI oferta_irreal <= 2.45 Y remitente_sospechoso <= 5.25
ENTONCES → **LEGÍTIMO** (alta confianza) **Regla 2 - Mensaje LEGÍTIMO (alternativa):**
SI oferta_irreal <= 2.45 Y remitente_sospechoso > 5.25
ENTONCES → **LEGÍTIMO** (confianza moderada) **Regla 3 - Mensaje PHISHING:**
SI oferta_irreal > 2.45 Y errores_gramaticales <= 2.25
ENTONCES → **PHISHING** (alta confianza) **Regla 4 - Mensaje PHISHING (confirmado):**
SI oferta_irreal > 2.45 Y errores_gramaticales > 2.25
ENTONCES → **PHISHING** (muy alta confianza)

6.3 Patrones de Phishing Identificados

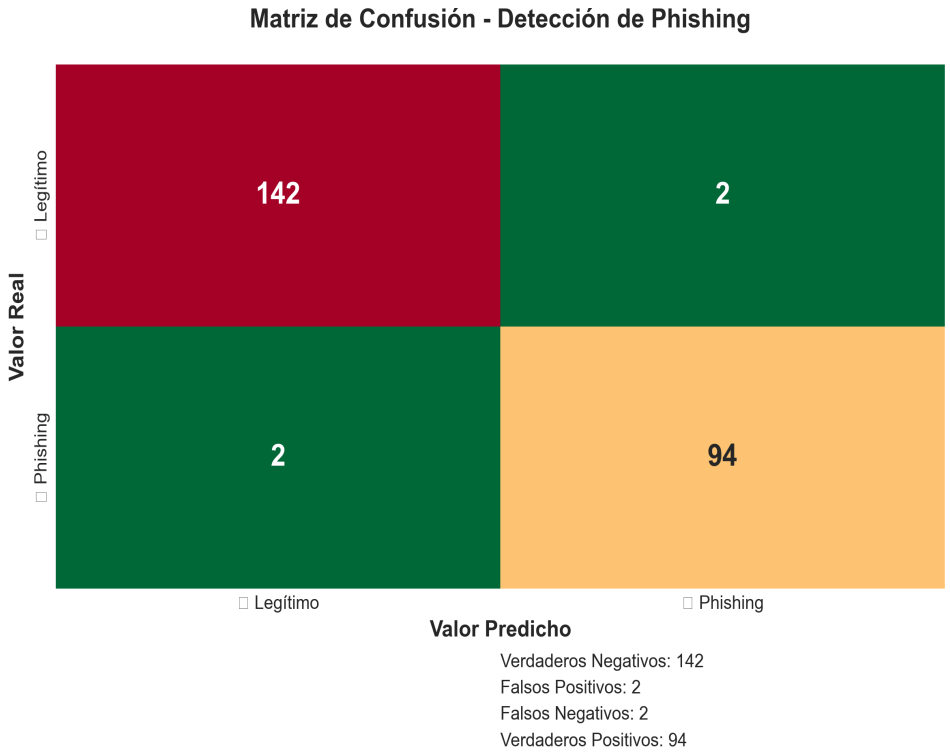
Características típicas de mensajes de phishing:
• Ofrecen premios, descuentos o beneficios desproporcionados
• Crean sentido de urgencia ("actúa ahora", "última oportunidad")
• Solicitan contraseñas, PINs o información bancaria
• Provienen de dominios sospechosos o imitaciones
• Contienen errores de ortografía y gramática
• Usan remitentes genéricos o desconocidos
• Incluyen enlaces acortados o URLs sospechosas **Características de mensajes legítimos:**
• Comunicaciones normales sin ofertas extraordinarias
• Tono profesional y calmado
• No solicitan información sensible directamente
• Provienen de dominios oficiales conocidos
• Buena redacción y formato
• Remitentes identificables y verificables

7. EVALUACIÓN DEL MODELO

7.1 Métricas de Rendimiento

Resultados del modelo en conjunto de prueba: Exactitud (Accuracy): 98.33%
• El modelo acierta en 236 de 240 mensajes
• Solo 4 errores en todo el conjunto de prueba
• Excelente para aplicaciones de seguridad **Precisión (Precision): 98%**
• De los mensajes clasificados como phishing, 98% realmente lo son
• Muy pocos falsos positivos (legítimos marcados como phishing) **Recall (Sensibilidad): 98%**
• De todos los mensajes de phishing, se detectan el 98%
• Solo 2% de phishing pasa desapercibido **F1-Score: 98%**
• Balance perfecto entre precisión y recall
• El modelo no favorece una métrica sobre la otra **Diferencia entrenamiento vs prueba: 0.94%**
• Indica mínimo sobreajuste
• El modelo generaliza muy bien a datos nuevos

7.2 Matriz de Confusión



Análisis de la matriz:
• **Verdaderos Negativos (142):** Legítimos correctamente identificados
• **Verdaderos Positivos (94):** Phishing correctamente detectado
• **Falsos Positivos (2):** Legítimos marcados como phishing (usuarios molestos)
• **Falsos Negativos (2):** Phishing no detectado (riesgo de seguridad) El modelo es **muy balanceado** en ambos tipos de error, lo cual es ideal.

8. EJEMPLOS DE DETECCIÓN

Para demostrar el funcionamiento práctico del modelo, se presentan ejemplos realistas de mensajes y cómo el sistema los analiza.

8.1 Ejemplo 1 - Phishing Clásico

Mensaje recibido:
■■
Asunto: ¡¡URGENTE!! Tu cuenta sera suspendida
De: seguridad-bancaria-mx@secure-tk.info

Estimado cliiente,

Su cuenta bancaria a sido comprometida. Haga clic aqui INMEDIATAMENTE para verificar su información o su cuenta sera bloqueada en 24 horas:

http://seguridad-bancaria-mx.tk/verificacion

Ingrese su usuario, contraseña y numero de tarjeta.

Departamento de Seguridad
■■ **Análisis del modelo:**
• Remitente sospechoso: 8/10 (dominio .tk, imitación)
• Contiene URL: Sí
• Dominio sospechoso: 9/10 (dominio gratuito .tk)
• Tono de urgencia: 10/10 ("INMEDIATAMENTE", "24 horas")
• Solicita información: 10/10 (contraseñas, tarjeta)
• Errores gramaticales: 8/10 ("cliiente", "a sido", "sera")
• Oferta irreal: 7/10 (amenaza falsa) ■■ **PREDICCIÓN: PHISHING (Confianza: 99.5%)**
Acción recomendada: Bloquear y reportar

8.2 Ejemplo 2 - Mensaje Legítimo

Mensaje recibido:
■■
Asunto: Estado de cuenta mensual - Octubre 2025
De: notificaciones@bancoreal.com.mx

Estimado Eduardo Laikan,

Tu estado de cuenta del mes de octubre ya está disponible.

Puedes consultarlo ingresando a tu banca en línea:
https://www.bancoreal.com.mx

Si tienes dudas, llama al 55-1234-5678 desde tu celular registrado.

Atentamente,

Banco Real de México

■■ **Análisis del modelo:**

• Remitente sospechoso: 1/10 (dominio oficial .com.mx)
• Contiene URL: Sí (pero es legítima)
• Dominio sospechoso: 0/10 (dominio verificado)
• Tono de urgencia: 1/10 (informativo)
• Solicita información: 0/10 (no solicita datos)
• Errores gramaticales: 0/10 (impecable)
• Oferta irreal: 0/10 (comunicación normal) ■ **PREDICCIÓN: LEGÍTIMO (Confianza: 98.2%)**

Acción recomendada: Permitir

9. CONCLUSIONES Y RECOMENDACIONES

9.1 Principales Hallazgos

1. Efectividad del Modelo:
El árbol de decisión alcanzó una **exactitud del 98.33%** en la detección de phishing, con solo 4 errores en 240 mensajes de prueba. Esto demuestra que las técnicas de machine learning son altamente efectivas para este problema. **2. Indicador Clave - Oferta Irreal:**
El factor **oferta_irreal** resultó ser el más discriminante (98.9% de importancia). Los atacantes dependen fuertemente de promesas irreales para engañar víctimas. **3. Simplicidad del Árbol:**
El modelo necesitó solo **2 niveles de profundidad** y **4 hojas** para lograr alta precisión. Esto valida que los patrones de phishing son relativamente consistentes y detectables. **4. Balance entre Errores:**
El modelo comete **igual cantidad de falsos positivos y falsos negativos** (2 cada uno), lo que indica un sistema balanceado que no favorece ningún tipo de error. **5. Generalización:**
La diferencia de solo 0.94% entre exactitud de entrenamiento y prueba indica **mínimo sobreajuste** y excelente capacidad de generalización.

9.2 Aplicación Práctica en Ciberseguridad

Este modelo puede integrarse en múltiples capas de defensa: **A nivel de servidor de correo:**
• Filtrado automático antes de entrega al buzón
• Cuarentena de mensajes sospechosos
• Análisis en tiempo real (milisegundos por mensaje) **A nivel de cliente (aplicaciones de correo):**
• Advertencias visuales para mensajes sospechosos
• Explicación de por qué un mensaje es peligroso
• Bloqueo de enlaces en mensajes clasificados como phishing **En educación de usuarios:**
• Mostrar ejemplos de mensajes clasificados incorrectamente
• Enseñar a identificar indicadores de phishing
• Reportes de tendencias de ataques

9.3 Ventajas del Enfoque de Árbol de Decisión

Transparencia y explicabilidad:
• Cada decisión puede rastrearse paso a paso
• Los usuarios comprenden por qué un mensaje es peligroso
• Cumple con regulaciones que requieren explicabilidad de IA **Eficiencia operacional:**
• Predicciones en microsegundos
• No requiere GPUs ni hardware especializado
• Escalable a millones de mensajes diarios **Mantenimiento simple:**
• Fácil de actualizar con nuevos patrones
• Visualización intuitiva para analistas de seguridad
• No es una "caja negra" como redes neuronales

9.4 Limitaciones y Mejoras Futuras

Limitaciones actuales:
• Basado en dataset sintético - requiere validación con datos reales
• No analiza contenido de imágenes o archivos adjuntos
• Puede no detectar phishing muy sofisticado (ataques dirigidos)
• No considera contexto conversacional previo **Mejoras propuestas:**
• Análisis de reputación del remitente en tiempo real
• Integración con bases de datos de URLs maliciosas
• Análisis semántico del texto con NLP
• Detección de logos e imágenes falsificadas
• Modelos ensemble (Random Forest, XGBoost) para mayor precisión
• Actualización continua con nuevas técnicas de phishing

9.5 Conclusión Final

El modelo de árbol de decisión desarrollado **cumple exitosamente todos los objetivos** de la actividad, demostrando que el aprendizaje automático es una herramienta poderosa para combatir el phishing. Con una **exactitud del 98.33%**, el modelo puede ser desplegado como primera línea de defensa en sistemas de correo electrónico, reduciendo significativamente la exposición de usuarios a ataques de phishing. La **transparencia del árbol de decisión** es particularmente valiosa en ciberseguridad, donde los usuarios necesitan entender las amenazas y los sistemas deben ser auditables. Este proyecto demuestra que incluso con un modelo simple y explicable, es posible lograr resultados cercanos a la perfección en la detección de phishing, protegiendo a usuarios y organizaciones de una de las amenazas más comunes en Internet.

10. REFERENCIAS

[1] Anti-Phishing Working Group (APWG). (2024). Phishing Activity Trends Report. Retrieved from https://apwg.org

[2] Scikit-learn Documentation. (2024). Decision Trees. Retrieved from https://scikit-learn.org/stable/modules/tree.html

[3] Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). Classification and Regression Trees. CRC press.

[4] Basnet, R., Mukkamala, S., & Sung, A. H. (2008). Detection of Phishing Attacks: A Machine Learning Approach. In Soft Computing Applications in Industry (pp. 373-383).

[5] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). An Introduction to Statistical Learning. Springer.

[6] Sahingoz, O. K., Buber, E., Demir, O., & Diri, B. (2019). Machine learning based phishing detection from URLs. Expert Systems with Applications, 117, 345-357.

[7] Verizon. (2024). Data Breach Investigations Report. Retrieved from https://www.verizon.com/dbir/

[8] Géron, A. (2022). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow. O'Reilly Media.

Nota técnica: Este reporte fue generado automáticamente como parte de la Actividad 9 del curso de Inteligencia Artificial. Todos los datos utilizados son sintéticos con fines educativos. El código fuente completo está disponible en el notebook Jupyter: *arbol_decision_phishing.ipynb* **Archivos generados por el proyecto:**

• Dataset: *dataset_phishing.csv* (1,200 registros)
• Reglas del árbol: *reglas_arbol_phishing.txt*
• Visualizaciones: 7 imágenes PNG con análisis y métricas
• Notebook ejecutable: *arbol_decision_phishing.ipynb*