Entrega 2

Responsables:

DAVID LONDONO LÓPEZ

ISAAC JIMÉNEZ FERNADEZ

Profesor:

Raul Ramos Pollan

Universidad de Antioquia

Introducción a la inteligencia artificial

2023/1

1. Planteamiento del problema

El problema que se está planteando es un problema de clasificación, que tiene como objetivo detectar y clasificar páginas con phishing. El término anti-phishing se refiere a las medidas preventivas para bloquear los ataques de phishing. El phishing es un delito cibernético en el que los atacantes se hacen pasar por entidades confiables o conocidas y contactan a las personas a través de diferentes medios, como correo electrónico, mensajes de texto o teléfono, con el fin de obtener información confidencial.

Por lo general, en un ataque de phishing por correo electrónico, el mensaje engañoso sugerirá que hay un problema con una factura, que ha habido actividad sospechosa en una cuenta o que el usuario debe iniciar sesión para verificar una cuenta o contraseña. Además, los atacantes pueden solicitar a los usuarios que ingresen información de la tarjeta de crédito, detalles bancarios y otros datos personales confidenciales.

Una vez que los atacantes recopilan esta información, pueden utilizarla para acceder a cuentas, robar datos e identidades, así como descargar malware en la computadora del usuario. Por lo tanto, la implementación de medidas de seguridad efectivas para prevenir los ataques de phishing es esencial para proteger la privacidad y seguridad en línea de los usuarios.

1.2.Dataset

El dataset seleccionado es Phishing Dataset for Machine Learning (https://www.kaggle.com/datasets/shashwatwork/phishing-dataset-for-machine-learning) Este conjunto de datos contiene 48 atributos extraídos de 5000 páginas web de phishing y 5000 páginas web legítimas, que se descargaron de enero a mayo de 2015 y de mayo a junio de 2017. Se emplea una técnica mejorada de extracción de funciones aprovechando el marco de automatización del navegador (es decir, Selenium WebDriver), que es más preciso y robusto en comparación con el enfoque de análisis basado en expresiones regulares.

Algunos de los atributos más significativos son:

- NumDots, variable continua numérica
- UrlLength, variable continua numérica
- NumDash, variable continua numérica
- NoHttps, variable categórica 0 y 1
- IpAddress, variable categórica 0 y 1
- RandomString, variable categórica 0 y 1
- HostnameLength, variable continua numérica
- PopUpWindow, variable categórica 0 y 1

1.3. Métricas

Para la evaluación del sistema se emplearán dos métricas de evaluación principales: el accuracy y el f1 score, ya que ambos se enfocan en la precisión. Además de estas métricas técnicas, se tiene en cuenta la métrica de negocio, la fiabilidad de las

predicciones para determinar si una página tiene phishing o no. Es crucial que estas predicciones sean confiables para que el navegador web pueda evitar que sus usuarios accedan a páginas maliciosas.

1.4. Desempeño

En un modelo de este tipo, se espera que la precisión de las predicciones sea alta, superando el 80%, además será importante evitar un gran número de falsos positivos.

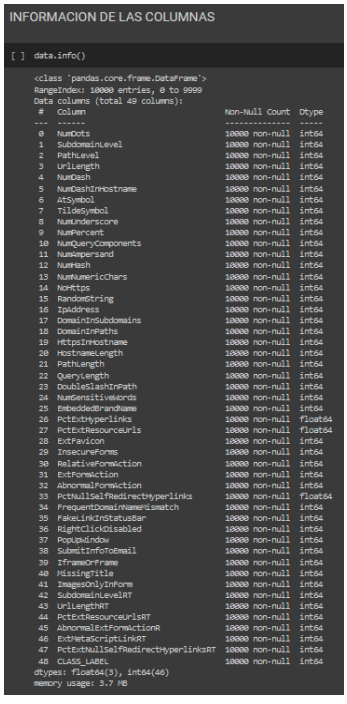
En un ambiente productivo, el modelo sería utilizado como un filtro para prevenir que los usuarios accedan a páginas sospechosas y, de esta manera, garantizar la seguridad de los usuarios. Por lo tanto, es fundamental que el modelo pueda proporcionar predicciones confiables y precisas para lograr este objetivo.

2. Exploración descriptiva del dataset

La base de datos utilizada en este proyecto es el "Phishing Dataset for Machine Learning" y consta de 10,000 muestras y 48 variables. De estas variables, 47 se utilizan para describir las características de una URL de un sitio web, mientras que la otra variable es la salida "CLASS_LABEL", que indica si el sitio web es o no un sitio de phishing.

La información contenida en las 47 variables descriptivas se utiliza para identificar patrones y características comunes entre los sitios web de phishing. Al analizar estas variables, se pueden crear modelos de aprendizaje automático que puedan detectar de manera efectiva la presencia de sitios web de phishing y proteger a los usuarios contra posibles ataques.

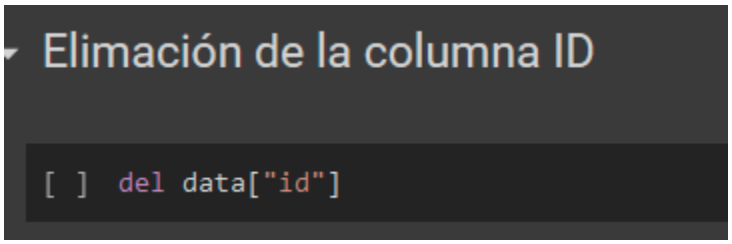
El nombre de las columnas presentes en el dataset son las siguientes:



Además, se optó por eliminar la columna "ID" durante el análisis, ya que esta variable no es relevante para el propósito del estudio.

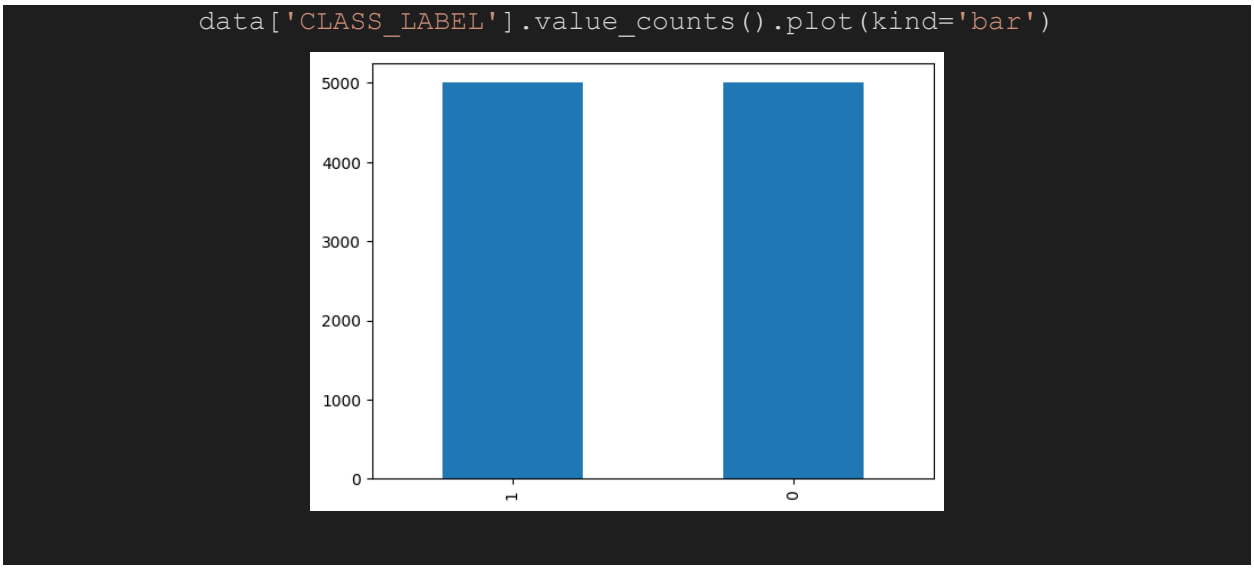
La columna "ID" suele ser un identificador único asignado a cada muestra en un conjunto de datos. En la mayoría de los casos, esta variable no tiene ningún impacto en la predicción o en el análisis de los datos. Es solo un número de referencia que se utiliza para identificar la muestra en la base de datos.

Por lo tanto, eliminar la columna "ID" no afectará la calidad de los resultados y puede simplificar el análisis, ya que se reducirá la cantidad de variables en el conjunto de datos. Esto también puede ayudar a mejorar el rendimiento y la eficiencia de los modelos de aprendizaje automático, ya que tendrán menos variables que analizar. En general, la eliminación de variables irrelevantes puede ser una práctica común en el preprocesamiento de datos para mejorar la precisión y la eficiencia en el análisis de los datos.



2.1 Normalización, balanceo de los datos y train test split

Después de explorar el conjunto de datos, se observó que no hay desequilibrio en los datos. El dataset consta de un total de 10000 muestras, de las cuales 5000 pertenecen a la clase 1 (no phishing) y 5000 pertenecen a la clase 0 (phishing).



Este es un resultado muy importante en el análisis de datos, ya que el desequilibrio en la distribución de las muestras puede llevar a problemas en la construcción de modelos de aprendizaje automático. El desbalance en los datos puede hacer que el modelo esté sesgado hacia la clase dominante, lo que puede llevar a una baja precisión en la clasificación de la clase minoritaria.

Sin embargo, en este caso, el hecho de que no exista un desequilibrio en los datos facilita la construcción de modelos precisos para detectar sitios web de phishing. Al tener un conjunto de datos equilibrado, se puede entrenar el modelo de manera justa y precisa, lo que se traduce en mejores resultados y una mayor seguridad en línea para los usuarios.

Se dividió él dataset en dos bloques destinados al entrenamiento de los datos y validación del modelo.

Resultados del entrenamiento

```
[ ] resultados_dt = experimentar_dt([3,10,50,100],MinMaxScaler().fit_transform(X), Y)
    resultados_dt
```

	profundidad del arbol	eficiencia de entrenamiento	desviacion estandar entrenamiento	eficiencia de prueba	desviacion estandar prueba	accuracy
0	3.0	0.942922	0.002025	0.9367	0.017607	0.9367
1	10.0	0.987400	0.000788	0.9549	0.007543	0.9549
2	50.0	1.000000	0.000000	0.9480	0.014471	0.9480
3	100.0	1.000000	0.000000	0.9491	0.013671	0.9491

Se están realizando pruebas eliminando datos random del dataset

```
[ ] remove_n = 3000
    drop_indices = np.random.choice(data.index, remove_n, replace=False)
    df_subset = data.drop(drop_indices)

    X = df_subset.drop('CLASS_LABEL', axis=1).values
    Y = df_subset['CLASS_LABEL'].values
    print (X.shape , Y.shape)

    (7000, 48) (7000,)

[ ] resultados_dt = experimentar_dt([3,10,20,100],MinMaxScaler().fit_transform(X), Y)
    resultados_dt
```

	profundidad del arbol	eficiencia de entrenamiento	desviacion estandar entrenamiento	eficiencia de prueba	desviacion estandar prueba	accuracy
0	3.0	0.943016	0.002393	0.929000	0.029166	0.929000
1	10.0	0.987587	0.000954	0.944000	0.027978	0.944000
2	20.0	0.999730	0.000355	0.938429	0.027334	0.938429
3	100.0	1.000000	0.000000	0.939143	0.029062	0.939143

Referencias

https://www.kaggle.com/datasets/shashwatwork/phishing-dataset-for-machine-learning

https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html

https://support.minitab.com/es-mx/minitab/21/help-and-how-to/statistical-modeling/regression/supporting-topics/basics/what-are-categorical-discrete-and-continuous-variables/

https://scikit-learn.org/stable/modules/generated/sklearn.metrics.accuracy_score.html

https://scikit-learn.org/stable/modules/tree.html#tree