Notebook creado por **Guillermo Grande Santi**

Imports

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import logging

from sklearn.model_selection import train_test_split
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score
from sklearn.model_selection import GridSearchCV, cross_val_score
import pickle

import tensorflow as tf
from tensorflow.keras.models import Sequential # type: ignore
from tensorflow.keras.layers import LSTM, Dense # type: ignore
from tensorflow.keras.preprocessing.sequence import pad_sequences #
type: ignore
from tensorflow.keras.preprocessing.text import Tokenizer # type:
ignore

import torch
import torch.nn as nn
import torch.optim as optim
from torch.utils.data import DataLoader, TensorDataset

from sentence_transformers import SentenceTransformer
from gensim.models import Word2Vec
from gensim.utils import simple_preprocess
import nltk
import re
import string
import spacy
import contractions

import shap
```

```
c:\Users\guigr\anaconda3\envs\tfm\Lib\site-packages\tqdm\auto.py:21:
TqdmWarning: IProgress not found. Please update jupyter and
ipywidgets. See
https://ipywidgets.readthedocs.io/en/stable/user_install.html
  from .autonotebook import tqdm as notebook_tqdm

WARNING:tensorflow:From C:\Users\guigr\AppData\Roaming\Python\
Python311\site-packages\tf_keras\src\losses.py:2976: The name
```

```
tf.losses.sparse_softmax_cross_entropy is deprecated. Please use
tf.compat.v1.losses.sparse_softmax_cross_entropy instead.
```

Carga de datos inicial

```
# Cargar datos de Kaggle
df_fake = pd.read_csv("Datasets/Fake.csv") # Noticias falsas
df_real = pd.read_csv("Datasets/True.csv") # Noticias verdaderas

# Agregar columna de etiquetas
df_fake["label"] = 0
df_real["label"] = 1

# Subject y Date no nos interesa
df_fake.drop(["subject", "date"], axis=1, inplace=True)
df_real.drop(["subject", "date"], axis=1, inplace=True)

print(df_fake.shape)
df_fake.head()
```

```
(23481, 3)
```

	title	text	label
0	Donald Trump Sends Out Embarrassing New Year’...	Donald Trump just couldn t wish all Americans ...	0
1	Drunk Bragging Trump Staffer Started Russian ...	House Intelligence Committee Chairman Devin Nu...	0
2	Sheriff David Clarke Becomes An Internet Joke...	On Friday, it was revealed that former Milwauk...	0
3	Trump Is So Obsessed He Even Has Obama’s Name...	On Christmas day, Donald Trump announced that ...	0
4	Pope Francis Just Called Out Donald Trump Dur...	Pope Francis used his annual Christmas Day mes...	0

```
print(df_real.shape)
df_real.head()
```

```
(21417, 3)
```

	title	text	label
0	As U.S. budget fight looms, Republicans flip t...	WASHINGTON (Reuters) - The head of a conservati...	1
1	U.S. military to accept transgender recruits o...	WASHINGTON (Reuters) - Transgender people will ...	1
2	Senior U.S. Republican senator: 'Let Mr. Muell...	WASHINGTON (Reuters) - The special counsel inve...	1
3	FBI Russia probe helped by Australian diplomat...	WASHINGTON (Reuters) - Trump campaign adviser G...	1
4	Trump wants Postal Service to charge 'much mor...	SEATTLE/WASHINGTON (Reuters) - President Donald...	1

```
                                                text  label
0  WASHINGTON (Reuters) - The head of a conservat...      1
1  WASHINGTON (Reuters) - Transgender people will...      1
2  WASHINGTON (Reuters) - The special counsel inv...      1
3  WASHINGTON (Reuters) - Trump campaign adviser ...      1
4  SEATTLE/WASHINGTON (Reuters) - President Donal...      1
```

```
print("Porcentaje de balanceo de clases:")
print("Fake: ", df_fake.shape[0]/(df_fake.shape[0]+df_real.shape[0]))
print("Real: ", df_real.shape[0]/(df_fake.shape[0]+df_real.shape[0]))
```

```
Porcentaje de balanceo de clases:
Fake:  0.5229854336496058
Real:  0.47701456635039424
```

```
# Mostrar la primera noticia fake
print("Primera noticia fake:")
print("Title: ", df_fake.iloc[0]['title'])
print("Text: ", df_fake.iloc[0]['text'])

# Mostrar la primera noticia real
print("Primera noticia real:")
print("Title: ", df_real.iloc[0]['title'])
print("Text: ", df_real.iloc[0]['text'])
```

```
Primera noticia fake:
Title:   Donald Trump Sends Out Embarrassing New Year’s Eve Message; This is Disturbing
Text:  Donald Trump just couldn t wish all Americans a Happy New Year and leave it at that. Instead, he had to give a shout out to his enemies, haters and  the very dishonest fake news media.  The former reality show star had just one job to do and he couldn t do it. As our Country rapidly grows stronger and smarter, I want to wish all of my friends, supporters, enemies, haters, and even the very dishonest Fake News Media, a Happy and Healthy New Year,  President Angry Pants tweeted.  2018 will be a great year for America! As our Country rapidly grows stronger and smarter, I want to wish all of my friends, supporters, enemies, haters, and even the very dishonest Fake News Media, a Happy and Healthy New Year. 2018 will be a great year for America!  Donald J. Trump (@realDonaldTrump) December 31, 2017Trump s tweet went down about as welll as you d expect.What kind of president sends a New Year s greeting like this despicable, petty, infantile gibberish? Only Trump! His lack of decency won t even allow him to rise above the gutter long enough to wish the American citizens a happy new year!  Bishop Talbert Swan (@TalbertSwan) December 31, 2017no one likes you  Calvin (@calvinstowell) December 31, 2017Your impeachment would make 2018 a great year for America, but I ll also accept regaining control of Congress.  Miranda Yaver (@mirandayaver) December 31, 2017Do you hear yourself talk? When you have to include that many people that hate you you have to wonder? Why do the they all
```

hate me? Alan Sandoval (@AlanSandoval13) December 31, 2017Who uses the word Haters in a New Years wish?? Marlene (@marlene399) December 31, 2017You can t just say happy new year? Koren pollitt (@Korencarpenter) December 31, 2017Here s Trump s New Year s Eve tweet from 2016.Happy New Year to all, including to my many enemies and those who have fought me and lost so badly they just don t know what to do. Love! Donald J. Trump (@realDonaldTrump) December 31, 2016This is nothing new for Trump. He s been doing this for years.Trump has directed messages to his enemies and haters for New Year s, Easter, Thanksgiving, and the anniversary of 9/11. pic.twitter.com/4FPAe2KypA Daniel Dale (@ddale8) December 31, 2017Trump s holiday tweets are clearly not presidential.How long did he work at Hallmark before becoming President? Steven Goodine (@SGoodine) December 31, 2017He s always been like this . . . the only difference is that in the last few years, his filter has been breaking down. Roy Schulze (@thbthttt) December 31, 2017Who, apart from a teenager uses the term haters? Wendy (@WendyWhistles) December 31, 2017he s a fucking 5 year old Who Knows (@rainyday80) December 31, 2017So, to all the people who voted for this a hole thinking he would change once he got into power, you were wrong! 70-year-old men don t change and now he s a year older.Photo by Andrew Burton/Getty Images.
Primera noticia real:
Title: As U.S. budget fight looms, Republicans flip their fiscal script
Text: WASHINGTON (Reuters) - The head of a conservative Republican faction in the U.S. Congress, who voted this month for a huge expansion of the national debt to pay for tax cuts, called himself a “fiscal conservative” on Sunday and urged budget restraint in 2018. In keeping with a sharp pivot under way among Republicans, U.S. Representative Mark Meadows, speaking on CBS’ “Face the Nation,” drew a hard line on federal spending, which lawmakers are bracing to do battle over in January. When they return from the holidays on Wednesday, lawmakers will begin trying to pass a federal budget in a fight likely to be linked to other issues, such as immigration policy, even as the November congressional election campaigns approach in which Republicans will seek to keep control of Congress. President Donald Trump and his Republicans want a big budget increase in military spending, while Democrats also want proportional increases for non-defense “discretionary” spending on programs that support education, scientific research, infrastructure, public health and environmental protection. “The (Trump) administration has already been willing to say: ‘We’re going to increase non-defense discretionary spending ... by about 7 percent,’” Meadows, chairman of the small but influential House Freedom Caucus, said on the program. “Now, Democrats are saying that’s not enough, we need to give the government a pay raise of 10 to 11 percent. For a fiscal conservative, I don’t see where the rationale is. ... Eventually you run out of other people’s money,” he said. Meadows was among Republicans who voted in late December for their party’s debt-financed tax overhaul, which is

```
expected to balloon the federal budget deficit and add about $1.5
trillion over 10 years to the $20 trillion national debt. “It’s
interesting to hear Mark talk about fiscal responsibility,” Democratic
U.S. Representative Joseph Crowley said on CBS. Crowley said the
Republican tax bill would require the  United States to borrow $1.5
trillion, to be paid off by future generations, to finance tax cuts
for corporations and the rich. “This is one of the least ... fiscally
responsible bills we’ve ever seen passed in the history of the House
of Representatives. I think we’re going to be paying for this for
many, many years to come,” Crowley said. Republicans insist the tax
package, the biggest U.S. tax overhaul in more than 30 years,  will
boost the economy and job growth. House Speaker Paul Ryan, who also
supported the tax bill, recently went further than Meadows, making
clear in a radio interview that welfare or “entitlement reform,” as
the party often calls it, would be a top Republican priority in 2018.
In Republican parlance, “entitlement” programs mean food stamps,
housing assistance, Medicare and Medicaid health insurance for the
elderly, poor and disabled, as well as other programs created by
Washington to assist the needy. Democrats seized on Ryan’s early
December remarks, saying they showed Republicans would try to pay for
their tax overhaul by seeking spending cuts for social programs. But
the goals of House Republicans may have to take a back seat to the
Senate, where the votes of some Democrats will be needed to approve a
budget and prevent a government shutdown. Democrats will use their
leverage in the Senate, which Republicans narrowly control, to defend
both discretionary non-defense programs and social spending, while
tackling the issue of the “Dreamers,” people brought illegally to the
country as children. Trump in September put a March 2018 expiration
date on the Deferred Action for Childhood Arrivals, or DACA, program,
which protects the young immigrants from deportation and provides them
with work permits. The president has said in recent Twitter messages
he wants funding for his proposed Mexican border wall and other
immigration law changes in exchange for agreeing to help the Dreamers.
Representative Debbie Dingell told CBS she did not favor linking that
issue to other policy objectives, such as wall funding. “We need to do
DACA clean,” she said.  On Wednesday, Trump aides will meet with
congressional leaders to discuss those issues. That will be followed
by a weekend of strategy sessions for Trump and Republican leaders on
Jan. 6 and 7, the White House said. Trump was also scheduled to meet
on Sunday with Florida Republican Governor Rick Scott, who wants more
emergency aid. The House has passed an $81 billion aid package after
hurricanes in Florida, Texas and Puerto Rico, and wildfires in
California. The package far exceeded the $44 billion requested by the
Trump administration. The Senate has not yet voted on the aid.

# Unir ambos datasets
df = pd.concat([df_fake, df_real])

# Mezclar datos
df = df.sample(frac=1).reset_index(drop=True)
```

```
# Ver primeras filas
print(df.head())
```

```
                                                title  \
0   Four-Year-Old Dies After Finding Loaded Gun A...
1   Anti-Trump Protestors Shut Down Major Road Le...
2  U.S. court backs Trump in battle over interim ...
3  Kazakhstan, Kyrgyzstan pledge to improve ties ...
4  Brazil's Temer sent for tests, treatment for u...

                                                text  label
0  A four-year-old Iowa boy died as the result of...      0
1  Protestors have peacefully shut down the main ...      0
2  WASHINGTON (Reuters) - A U.S. District Court j...      1
3  ALMATY (Reuters) - The leaders of Kazakhstan a...      1
4  SAO PAULO (Reuters) - Brazilian President Mich...      1
```

```
# Comprobar que los datos siguen balanceados
print(df["label"].value_counts())
```

```
0    23481
1    21417
Name: label, dtype: int64
```

```
# df.to_csv("Datasets/FakeAndRealNews.csv", index=False)
```

Preprocesado NLP

Guardamos ejemplos originales

```
import random
import json

df = pd.read_csv("../Datasets/FakeAndRealNews.csv")

# Dividimos los datos en entrenamiento y prueba
# Por ahora usaremos únicamente el texto de la noticia (omitimos el
título)
X = df["text"]
y = df["label"]
X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.2, random_state=42)

# Se usará para redes neuronales
# Usaremos un 20% del conjunto de datos para validación (16% del
total)
X_train, X_valid, y_train, y_valid = train_test_split(X_train,
y_train, test_size=0.2, random_state=42)
```

```
# Combine X_test and y_test into a list of dictionaries
news_data = [{"text": text, "label": label} for text, label in
zip(X_test, y_test)]

# Select 20 random news
random_news = random.sample(news_data, 20)

# Write the selected news to a JSON file
with open("random_news.json", "w", encoding="utf-8") as json_file:
    json.dump(random_news, json_file, ensure_ascii=False, indent=4)
```

Preprocesado

```
df = pd.read_csv("Datasets/FakeAndRealNews.csv")

# Instalar modelo de spacy
!python -m spacy download en_core_web_sm
```

```
Collecting en-core-web-sm==3.8.0
  Downloading
https://github.com/explosion/spacy-models/releases/download/en_core_we
b_sm-3.8.0/en_core_web_sm-3.8.0-py3-none-any.whl (12.8 MB)
     ---------------------------------------- 0.0/12.8 MB ? eta
-:--:--
     ---------------------------------------- 0.0/12.8 MB ? eta
-:--:--
      ---------------------------------------- 0.2/12.8 MB 2.6 MB/s eta
0:00:05
     --- ------------------------------------ 1.1/12.8 MB 10.2 MB/s
eta 0:00:02
     ------ --------------------------------- 2.1/12.8 MB 14.5 MB/s
eta 0:00:01
     --------- ------------------------------ 3.0/12.8 MB 14.8 MB/s
eta 0:00:01
     ----------- ---------------------------- 3.8/12.8 MB 16.4 MB/s
eta 0:00:01
     -------------- ------------------------- 4.8/12.8 MB 16.0 MB/s
eta 0:00:01
     ----------------- ---------------------- 5.7/12.8 MB 16.5 MB/s
eta 0:00:01
     -------------------- ------------------- 6.6/12.8 MB 16.9 MB/s
eta 0:00:01
     ----------------------- ---------------- 7.6/12.8 MB 17.3 MB/s
eta 0:00:01
     ------------------------- -------------- 8.5/12.8 MB 17.6 MB/s
eta 0:00:01
     ---------------------------- ---------- 9.5/12.8 MB 17.8 MB/s
eta 0:00:01
     ------------------------------ ------- 10.4/12.8 MB 19.9 MB/s
```

```
eta 0:00:01
     ---------------------------------- ---- 11.4/12.8 MB 20.5 MB/s
eta 0:00:01
     ---------------------------------- - 12.3/12.8 MB 20.5 MB/s
eta 0:00:01
     ---------------------------------- 12.8/12.8 MB 19.8 MB/s
eta 0:00:01
     ---------------------------------- 12.8/12.8 MB 16.0 MB/s
eta 0:00:00
✔ Download and installation successful
You can now load the package via spacy.load('en_core_web_sm')


[notice] A new release of pip is available: 24.0 -> 25.0.1
[notice] To update, run: python.exe -m pip install --upgrade pip
```

```
from nltk.corpus import stopwords

# Descargar stopwords de nltk
nltk.download("stopwords")
stop_words = set(stopwords.words("english"))

nlp = spacy.load("en_core_web_sm")
```

```
[nltk_data] Downloading package stopwords to
[nltk_data]     C:\Users\guigr\AppData\Roaming\nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
```

La limpieza del texto se realizará mediante la siguiente función:

```
# Función para limpiar el texto
def clean_function(text):
    text = text.lower() # Convertir a minúsculas para uniformidad
    text = re.sub(r"\[.*?\]", "", text) # Quitar todo contenido entre
corchetes
    text = re.sub(f"[{string.punctuation}]", "", text) # Quitar
cualquier carácter de puntuación
    text = re.sub(r"[‘’“”\"\']", "", text) # Eliminar comillas raras
    text = re.sub(r"\w*\d\w*", "", text) # Quitar palabras con números
    text = re.sub(r'https://\S+|www\.\S+', '', text) # Quitar URLs
    text = contractions.fix(text) # Expandir contracciones
    text = re.sub(r"\s.\s", " ", text) # Eliminar caracteres que miden
solo 1 (espacios a ambos lados)
    text = re.sub(r"\s+", " ", text).strip() # Quitar espacios en
blanco adicionales
    text = " ".join([word for word in text.split() if word not in
stop_words]) # Quitar stopwords
    text = " ".join([token.lemma_ for token in nlp(text)]) #
Lematización
    return text
```

```
df["clean_title"] = df["title"].apply(clean_function)
df["clean_text"] = df["text"].apply(clean_function)

# Mostrar comparaciones de la limpieza realizada en algunas filas de texto
number_new = 25000
print(df['label'][number_new])

print(df['title'][number_new])
print(df['text'][number_new])

print(df['clean_title'][number_new])
print(df['clean_text'][number_new])
```

```
0
SHE GREW UP BELIEVING BLACKS Could Only Support Democrats…Until She Took A Job With ACORN: WATCH The INCREDIBLE Story Of A Woman Who Took On Obama’s LEFTIST MACHINE [VIDEO]
Keep your eye on Anita Moncreif If knowledge is power she is the Democrat Party s worst nightmare. When you re on the left, and all of your friends are leftists, and your parents are leftists, you don t hang around with other people, and you only get the view of folks as what you see on TV, and how they present it to you. And you guys are seen as racist, angry people. Every time they get a chance, that s the image they push out there on TV. They try to find that one crazy Tea Party person and they try to get them to say something, and they make sure they play it on all the black stations. And you see that and you say,  Okay, these people are nuts.  So I didn t expect to find any kind of support from the Right. Everything Anita Moncreif believed to be true about the Left changed when she took a job with ACORN and quickly discovered the Democrat Party was not really looking out for the interests of the Black community or low income neighborhoods. When she began to understand they would do anything, including breaking the law, to grow the Democrat Party, she made the decision to expose them. She quickly found out how the mainstream media will go to any length to keep the truth about the criminal Left from the American people.Watch her amazing story here: Decades after his death, Saul Alinsky s vision has become reality. From Barack Obama to Hillary Clinton to ACORN to Black Lives Matter, Alinsky is more alive in his death now than in his four decades of community organizing.Anita is asking for the help of conservatives to make this movie a reality. She needs YOUR help to build momentum for this film.Please consider giving whatever you can today. Click HERE to donate $1, $5, $10, $20 or whatever you can afford. This is an independent fund. We have no big funders or organizations backing us  yet. That s why we need you. We need to start shooting now. Reaching our goal will allow us to begin shooting footage at the two party conventions and buy us time to raise awareness to raise the production, administrative, and promotional budgets for this much-needed film.We re going to
```

communicate with you the audience. Some of the footage we ll release before the film s debut. We ll also communicate some of our successes and our challenges along the way. Together, we can change the way films are produced and promoted.The American Left and the Right need to see this film and decide where we go from here.If the necessary funds aren t raised on Kickstarter, account funds won t be unlocked.Eight years after exposing ACORN, I have been immersed in training, speaking, and examining the effectiveness of the grassroots on both sides of the aisle. I felt that my journey was not over, and I had many more truths to tell. I am finally ready to offer a movement eye view of the legacy of Alinsky, and the rise of grassroots movements across the nation. It s a huge effort, it s expensive, and the stakes are high, so please chip in $15, $50, $500 or more to fund our efforts to film at the DNC and RNC conventions in the next few weeks.Donate now to The Children of Alinsky (Phase 1)Together, we can do great things and the possibility of a documentary filmed and funded by ordinary people determined to implement change will be a major step toward illustrating how bottom-up change is done.Your friend, Anita MonCriefHere is Part II of Anita s amazing story:
grow believe black could support democrat … until take job acorn watch incredible story woman take obamas leftist machine
keep eye anita moncreif knowledge power democrat party bad nightmare leave friend leftist parent leftist hang around people get view folk see tv present guy see racist angry people every time get chance image push tv try find one crazy tea party person try get say something make sure play black station see say okay people nut expect find kind support right everything anita moncreif believe true left change take job acorn quickly discover democrat party really look interest black community low income neighborhood begin understand would anything include break law grow democrat party make decision expose quickly find mainstream medium go length keep truth criminal leave american peoplewatch amazing story decade death saul alinsky vision become reality barack obama hillary clinton acorn black life matter alinsky alive death four decade community organizinganita ask help conservative make movie reality need help build momentum filmplease consider give whatever today click donate whatever afford independent fund big funder organization back we yet need need start shoot reach goal allow we begin shoot footage two party convention buy we time raise awareness raise production administrative promotional budget muchneede filmwe go communicate audience footage release film debut also communicate success challenge along way together change way film produce promotedthe american leave right need see film decide go hereif necessary fund raise kickstarter account fund unlockedeight year expose acorn immerse training speak examine effectiveness grassroot side aisle feel journey many truth tell finally ready offer movement eye view legacy alinsky rise grassroot movement across nation huge effort expensive stake high please chip fund effort film dnc rnc convention next weeksdonate child alinsky phase great thing possibility documentary film fund ordinary people determine implement

```
change major step toward illustrate bottomup change doneyour friend
anita moncriefhere part ii anita amazing story
```

Se puede observar que hay varias filas del DataFrame que, tras la limpieza, se quedaron en blanco (por ser URLs u otra razón). Eliminaremos dichas filas vacías.

```
# Mostrar comparaciones de la limpieza realizada en algunas filas de
texto
number_new = 83
print(df['label'][number_new])

print(df['title'][number_new])
print(df['text'][number_new])

print(df['clean_title'][number_new])
print(df['clean_text'][number_new])
```

```
0
LATINOS MAKE DISGUSTING VIDEOS Bashing TRUMP: “Make America Mexico
Again” [Video]

latinos make disgusting video bash trump make america mexico
```

```
# Eliminar filas con texto vacío
filas_antes = df.shape[0]
df = df[df['clean_text'].str.strip() != '']
filas_despues = df.shape[0]
filas_eliminadas = filas_antes - filas_despues
print(f"Se han eliminado {filas_eliminadas} filas.")

# Mostrar las primeras filas del DataFrame después de eliminar filas
vacías
df.head()
```

```
Se han eliminado 705 filas.
```

	title	\
0	WHATEVER HAPPENED To Trump’s Second Wife? [VIDEO]	
1	ABSOLUTE SUBMISSION: Trump Bows to Neocon Orth...	
2	LONDON’S MAYOR HAS HARSH WORDS For Our Communi...	
4	Trump's top defense and homeland officials to ...	
5	Support for Brazil's pension reform more organ...	

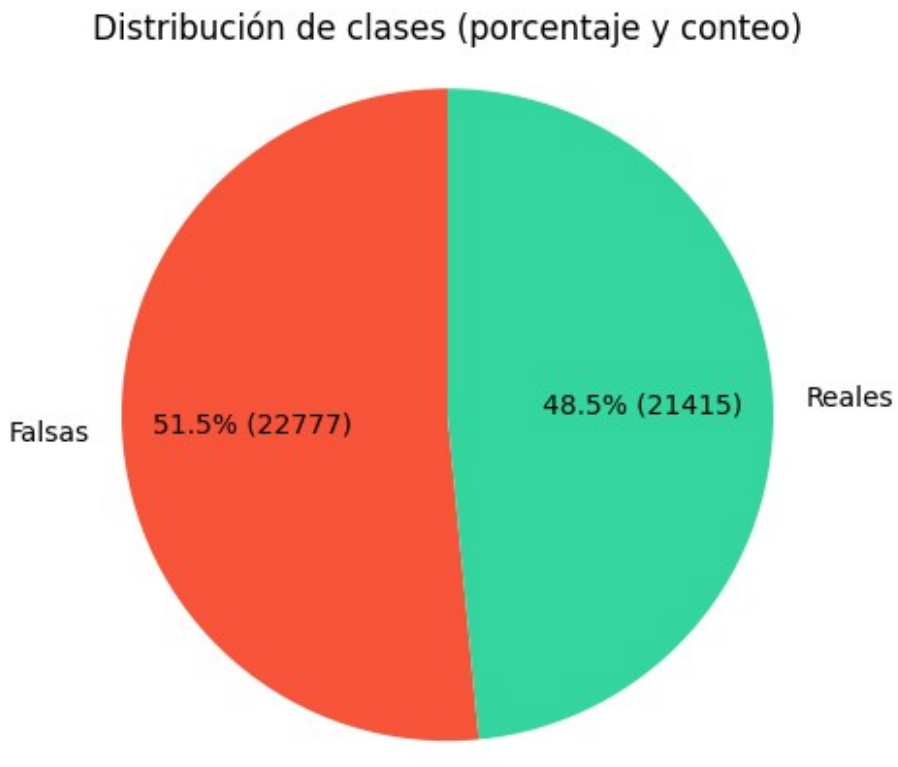
	text	label	\
0	It s a pretty safe bet that the press isn t ab...	0	
1	Consortium News Exclusive: In his Mideast trip...	0	
2	Our country is spinning out of control. Obama ...	0	
4	BERLIN (Reuters) - U.S. Secretary of Defense J...	1	
5	BRASILIA/RIO DE JANEIRO (Reuters) - The govern...	1	

```
                                          clean_title  \
0                  whatever happen trump second wife
1     absolute submission trump bow neocon orthodoxy
2  londons mayor harsh word community organizer c...
4  trump top defense homeland official attend mun...
5    support brazil pension reform organize lawmaker

                                           clean_text
0  pretty safe bet press able reveal bad blood do...
1  consortium news exclusive mideast trip saudi a...
2  country spin control obama orchestrate effort ...
4  berlin reuter us secretary defense james matti...
5  brasiliario de janeiro reuters government braz...
```

```python
# Comprobar que los datos siguen balanceados
print(df["label"].value_counts())
```

```
0    22777
1    21416
Name: label, dtype: int64
```

```python
labels = ['Falsas', 'Reales']
counts = [22777, 21416]
percent = [count / sum(counts) * 100 for count in counts]
# Colores aesthetic, suaves pero con carácter
colors = ['#f7543a', '#34d59f']  # Terracota suave y azul grisáceo

fig2, ax2 = plt.subplots()
ax2.pie(percent, labels=labels, autopct=lambda p: f'{p:.1f}% ({int(p *
sum(counts) / 100)})', startangle=90, colors=colors)
ax2.set_title('Distribución de clases (porcentaje y conteo)')
ax2.axis('equal')
plt.show()
```



Veamos la longitud media de nuestras noticias tras el preprocesado completo, lo que nos será de utilidad más tarde:

```
mean_length = df["clean_text"].apply(len).mean()
print("Mean length of clean_text:", mean_length)
```

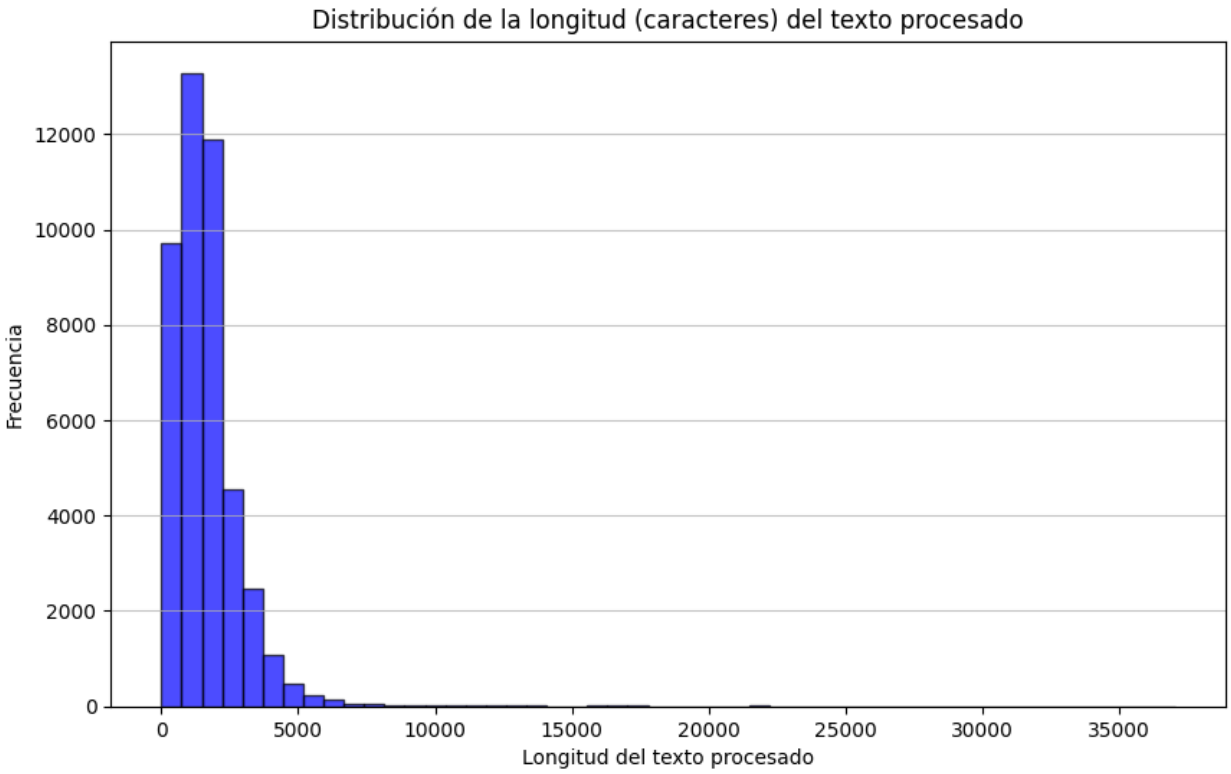
```
Mean length of clean_text: 1656.3348946665762
```

Información adicional:

```
# Calculate the lengths of clean_text
text_lengths = df["clean_text"].apply(len)

# Media de longitud
mean_length = text_lengths.mean()
print("Mean length of clean_text:", mean_length)

# Mínimo y máximo
max_length = text_lengths.max()
min_length = text_lengths.min()
print("Maximum length of clean_text:", max_length)
print("Minimum length of clean_text:", min_length)

# Histograma
plt.figure(figsize=(10, 6))
plt.hist(text_lengths, bins=50, color='blue', alpha=0.7,
```

```
edgecolor='black')
plt.title("Distribución de la longitud (caracteres) del texto
procesado")
plt.xlabel("Longitud del texto procesado")
plt.ylabel("Frecuencia")
plt.grid(axis='y', alpha=0.75)
plt.show()
```

```
Mean length of clean_text: 1656.3348946665762
Maximum length of clean_text: 37034
Minimum length of clean_text: 4
```



Guardamos en un DataFrame todos los datos por si, en un futuro, se necesita utilizarlos.

```
# Guardar el DataFrame en un archivo CSV
df.to_csv("Datasets/CleanedAllData.csv", index=False)
```

Guardaremos en otro DataFrame únicamente las columnas procesadas tras la limpieza. Este DataFrame es el que se utilizará a partir de ahora.

```
# Eliminar las columnas 'title' y 'text'
df.drop(['title', 'text'], axis=1, inplace=True)

# Mostrar las primeras filas del DataFrame
df.head()
```

```
   label                                          clean_title  \
0      0                  whatever happen trump second wife
1      0     absolute submission trump bow neocon orthodoxy
2      0  londons mayor harsh word community organizer c...
4      1  trump top defense homeland official attend mun...
5      1    support brazil pension reform organize lawmaker

                                          clean_text
0  pretty safe bet press able reveal bad blood do...
1  consortium news exclusive mideast trip saudi a...
2  country spin control obama orchestrate effort ...
4  berlin reuter us secretary defense james matti...
5  brasiliario de janeiro reuters government braz...
```

```
# Guardar el DataFrame en un archivo CSV
df.to_csv("Datasets/CleanedFakeAndRealNews.csv", index=False)
```

Vectorización TF-IDF + Clasificación mediante Random Forest (Caso Reuters)

```
# Cargar el DataFrame limpio
df = pd.read_csv("../Datasets/Cleaned-FR-News_V2.csv")

# Dividimos los datos en entrenamiento y prueba
# Por ahora usaremos únicamente el texto de la noticia (omitimos el título)
X = df["clean_text"]
y = df["label"]
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

TF-IDF

```
# Definimos y utilizamos vectorizador TF-IDF
vectorizer = TfidfVectorizer(max_features=5000)
X_train = vectorizer.fit_transform(X_train)
X_test = vectorizer.transform(X_test)

X_train.shape
```

```
(28283, 5000)
```

El **TfidfVectorizer** genera un shape de (28283, 5000) porque hay 28,283 muestras en el conjunto de entrenamiento, y cada muestra está representada por un **vector de 5,000 características**. Estas 5,000 características corresponden a las palabras más importantes del vocabulario, seleccionadas según su frecuencia en el corpus.

Random Forest

```
# Entrenar modelo
rf = RandomForestClassifier(n_estimators=100, random_state=42)
rf.fit(X_train, y_train)

# Evaluación en test
y_pred = rf.predict(X_test)
print("Precisión:", accuracy_score(y_test, y_pred))
```

```
Precisión: 0.9954746011992307
```

Obtenemos una precisión del 99.5% en test, ¡un valor extremadamente alto!

Apliquemos SHAP para ver qué palabras están afectando a la clasificación (pues es altamente probable que exista alguna palabra condicionando los resultados hacia un lado en muchos casos, lo que explicaría el valor altísimo de *accuracy* obtenido).

Explicabilidad mediante SHAP

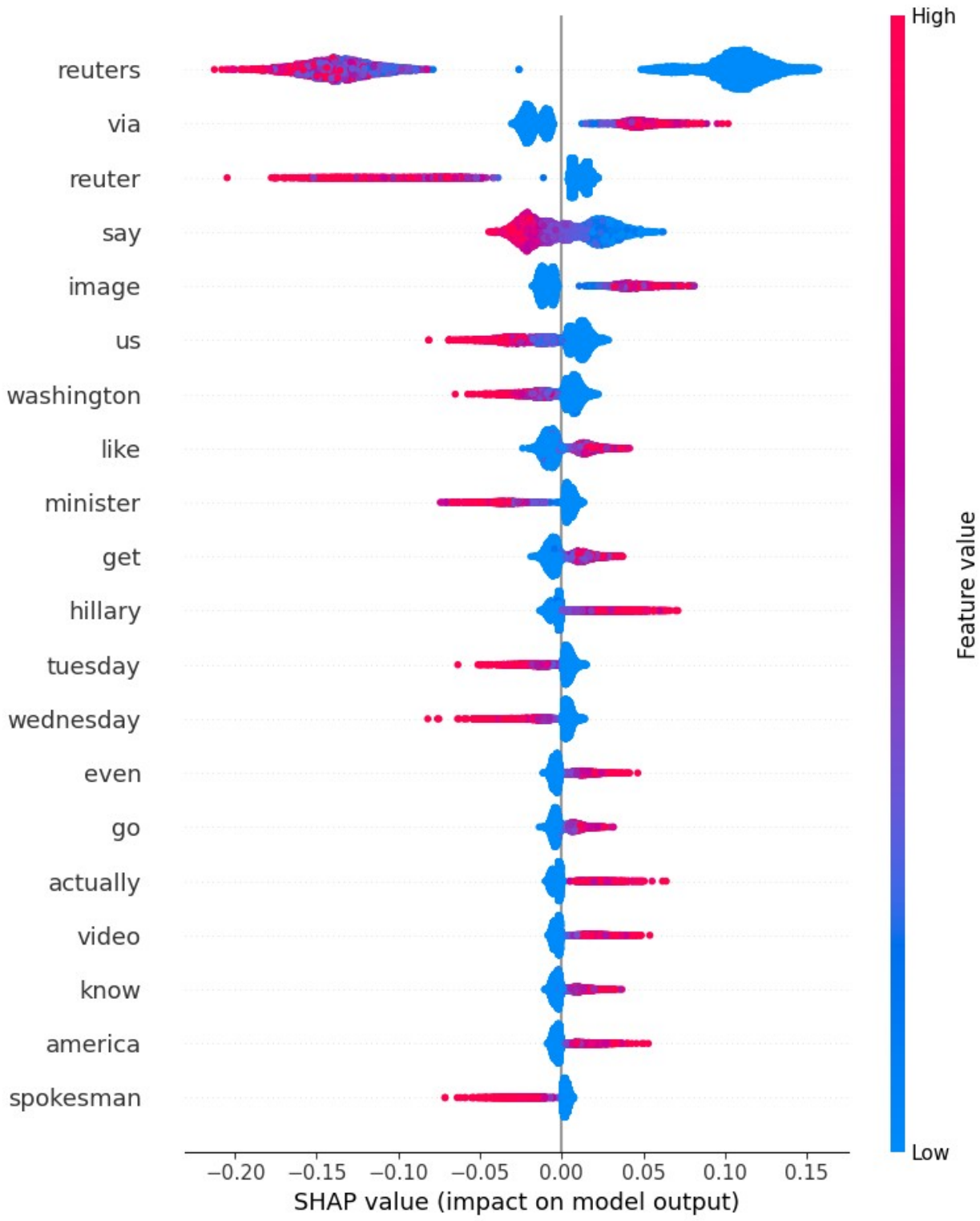
```
# Es necesario convertir las matrices dispersas a densas
X_train_dense = X_train.toarray()
X_test_dense = X_test.toarray()

# Aplicamos SHAP
explainer = shap.Explainer(rf, X_train_dense)
shap_values = explainer(X_test_dense)
```

```
c:\Users\guigr\anaconda3\envs\tfm\Lib\site-packages\tqdm\auto.py:21:
TqdmWarning: IProgress not found. Please update jupyter and
ipywidgets. See
https://ipywidgets.readthedocs.io/en/stable/user_install.html
  from .autonotebook import tqdm as notebook_tqdm
100%|==================| 17657/17678 [13:48<00:00]
```

```
shap_values.shape
```

```
(8839, 5000, 2)
```

```
# Obtenemos las palabras de los vectores TF-IDF
feature_names = vectorizer.get_feature_names_out()

# Elegimos índice a explicar
class_index = 0 # En este caso, ¿qué palabras influyen más y menos en
que una noticia sea falsa?

# Seleccionamos los valores SHAP para dicha clase
shap_values_class = shap_values[:, :, class_index]

# Mostrar valores SHAP
shap.summary_plot(shap_values_class, X_test_dense,
feature_names=feature_names)
```



Se puede observar que las palabras *Reuters*, *via* y *Reuter* son las que más influyen en nuestra clasificación.

Concretamente, **Reuters** es una de las fuentes de noticias más confiables a nivel mundial, por lo que el modelo tiende a clasificar como verdadera una noticia que contiene en su texto la marca

de la agencia (estas palabras tienen un alto impacto negativo en la clasificación como noticia falsa).

Por otro lado, el uso de la palabra *via* produce el efecto contrario, favoreciendo la clasificación de la noticia como falsa. Exploraremos esto con más detalle a continuación.

El resto de las palabras, como *say*, *image*, *US*, *Washington* o *Minister*, parecen ser más neutras a simple vista.

```
# Dividimos el DataFrame en verdadero y falso
df_real = df[df['label'] == 1]['clean_text']
df_fake = df[df['label'] == 0]['clean_text']

print("-------------------------------------------------------")
count_reuters = sum(1 for text in df_real if "reuters" in text)
print("Number of texts containing the word 'reuters' in REAL
DataFrame:", count_reuters)
count_reuters = sum(1 for text in df_fake if "reuters" in text)
print("Number of texts containing the word 'reuters' in FAKE
DataFrame:", count_reuters)
print("-------------------------------------------------------")
count_via = sum(1 for text in df_real if "via" in text)
print("Number of texts containing the word 'via' in REAL DataFrame:",
count_via)
count_via = sum(1 for text in df_fake if "via" in text)
print("Number of texts containing the word 'via' in FAKE DataFrame:",
count_via)
print("-------------------------------------------------------")
count_reuter = sum(1 for text in df_real if "reuter" in text)
print("Number of texts containing the word 'reuter' in REAL
DataFrame:", count_reuter)
count_reuter = sum(1 for text in df_fake if "reuter" in text)
print("Number of texts containing the word 'reuter' in FAKE
DataFrame:", count_reuter)
print("-------------------------------------------------------")
```

```
-------------------------------------------------------
Number of texts containing the word 'reuters' in REAL DataFrame: 18767
Number of texts containing the word 'reuters' in FAKE DataFrame: 180
-------------------------------------------------------
Number of texts containing the word 'via' in REAL DataFrame: 1116
Number of texts containing the word 'via' in FAKE DataFrame: 11447
-------------------------------------------------------
Number of texts containing the word 'reuter' in REAL DataFrame: 21378
Number of texts containing the word 'reuter' in FAKE DataFrame: 318
-------------------------------------------------------
```

```
# Search for sentences with the word "via" in them
sentences_with_via = [text for text in X if "via" in text][:5]

# Print the sentences
```

```
for sentence in sentences_with_via:
    print(sentence)
```

```
country spin control obama orchestrate effort race baiter like al
sharpton leader black life matter terrorist create divide race like
generation never know obama skips supreme court justice anton scalia
funeral funeral iconic first lady nancy reagan find time take pot shot
gop presidential frontrunner donald trump attend hipster festival
michelle barack prove time desire behave like leader go tell european
need relate eu hear baron today current president planning go britain
make case public stay european union momentarily stun happen lead
behind let country handle affairsby jade president nothing free
nothing eventual payoff sticking nose britain affairs come surprise
give track record last interminable year time officeor maybe know
deeply dislike anyone sign onto pathetic peace prize back beginning
reign maybe believe sycophantic press tell wait use kind reverse
psychology british ie know much hold contempt anything suggest would
oppose nah far remove reality grasp conceptso london mayor boris
johnson take subject far interesting follow national government figure
would via gate viennathese excerpt telegraph thing mr get wrong extent
obama undermine america sovereignty particularly southern border
subject change culture monolingual character one could say mr obama
decision drop lecture brit particular subject contradictory congruent
behavior sentiment regard america sovereigntyobviously mr johnson pay
attention american presidential campaign would do though see
unprecedented populist follow donald trump base theme sovereigntywe
want back one piecei love america believe american dream indeed hold
story past year largely america rise global greatness america helped
preserve expand democracy around world two global conflict throughout
cold war united states fight founding ideal republic government people
people people perish earthso face bit peculiar us government official
believe britain must remain within eu system democracy increasingly
underminedsome time next couple month tell president obama go arrive
country like deus ex machina pronounce matter air force one touch
lectern presidential seal erect british people tell good right thing
inform important ally interest stay eu matter flawed may feel
organisation never mind loss sovereignty never mind expense
bureaucracy uncontrolle immigrationthe american view clear whether
code en clair president tell we uk membership eu right britain right
europe right america tell way influence counsel nationsit important
argument deserve take seriously also think wholly fallacious come
uncle sam piece outrageous exorbitant hypocrisythere country world
defend sovereignty hysterical vigilance united states america nation
bear glorious refusal accept overseas control almost two half century
ago american colonist rise violently assert principle alone determine
government america george iii minister day americans refuse kneel
almost kind international jurisdiction alone western nation we decline
accept citizen subject ruling international criminal court hague even
sign convention law sea imagine americans submit democracy kind regime
euso essential britain comply system americans would reject hand
```

blatant case say entire letter go telegraph
presence kind privilege put unnecessary pressure people colour defend anger frustration fear outcome share story vajdaan tanveer rsu coordinatoryou make stuff white people experience racism firstyear journalism student trevor hewitt julia knope tell victim racialization allow stay meeting room report eventhewitt knope say make eye contact unidentified woman appeared set event approach hewitt knope ask ever racializedhewitt say tell woman want cover meeting assignment say woman tell racialize student could sit meet hewitt knope leave room feel really bad kind embarrassing knope say goal meeting end racialization need something everybody involve people cause problem need know group go accomplish anything racialize student collective part ryerson student union rsu website state group oppose form racism work towards community wellness student focus build antiracist network foster antiracist environment campuswide service campaign event knope say understand support group want other understand event list public rsu campaign seem really ironic meeting racialization prohibit certain people enter say right almost like suggest make racialization go away everyone racialize talk magically go away hewitt addedrsu coordinator vajdaan tanveer tell ryersonian phone member collective request safe space campus open conversation want racialize student feel intimidate speak mind afraid judge something say might use saidwhen ask hewitt knope incident tanveer confirm attend meet white term educate event public say use opportunity tell work get involve via whiterabbitradionet
thank president obamain democratic national convention barack obama take stage declare red state america blue state america united states america night dad tell we would president one daythen four year later fulfil dad prophecyeight year ago support hillary clinton contentious bitter primary ready first woman president think go win could notand long journey country begani look around see comfortable life live start crack along every american citizen easy life live postclinton thing pastthe economy collapse people suffer uncertainty plague every household every employee every employerit seem hope great country foundation hold lose teacher firefighter office worker nurse wait pink slip homeowner wait foreclosure notice sick senior dread medical billsthen flash screen hope changehandsome charismatic dignified junior senator illinois one really hear enter life give we hope message shine bright million americans good day ahead work together trust one another would prevail change could believe boy believe itand believe ever great man prepare leave office even face oncoming trump administrationthe fact remain deficit cut twothird stock market hit high point history still continues grow eleven million job create million americans healthcare auto industry save wage race income first time decade samesex marriage law land pentagon open door woman area expertisein word country change well hope perseverance president obama get we see around much well truth right front eye need fact tell we certainly hurt front debatenow mean perfect tpp large scale use drone failure fully communicate obamacare need strong government irk time

```
unlike leftwe rightwing ideology never look pure presidentno one
perfect never govern perfectlyso absolutely help republicans
congressional state president obama never waver commitment allow
american citizen indulge promise man create equal endow creator life
liberty pursuit happinesswe liberty believe happy nation himpresident
obama inspire pursue happiness sophomore college currently work degree
political science hope go law school specialize constitutional law
utilize resource bring change good change washington dc political
writer enter fourth year grind passion politic paper passion anger
hope sense accomplishment write share world owe president obamai burn
many bridge end couple friendship argue lot people defend president
obama policy one regret would againi come age era obama world view
politic culture shape large part publication put consequential
president modern americahis presidency mark dignity grace scandal gift
america people show we calm collect family man south side chicago
could little bit hope yearn change black man funny name put grow
without father white mother grandparent transform face nation
generation comehow profound thati miss calm cool demeanor time crisis
miss love inspire message hope miss family especially michelle first
lady pinnacle elegancethank president obama president help naive
politically inept year old grow passionate firedup year old want grow
good fellow manpresident obama build yes yes didfeature image via
white house
like mother like daughter chelsea difficult accord insider know former
clintonite hillary cuss like sailor really hammer people perhaps case
know chelsea know difficult work chelsea clinton unpleasant colleague
cause high turnover bill hillary chelsea clinton foundation source say
several top staffer leave foundation since chelsea come onboard vice
chairman lot people leave lot people leave want insider tell difficult
onetime ceo bruce lindsey push upstairs position chairman board two
year ago chelsea could bring mckinsey colleague eric braverman boy try
hire communication professional actually try run place understand
suppose say source push matt mckenna chelsea spokesman work uber ginny
ehrlich found ceo clinton health matter initiative work robert wood
johnson foundationvia ny post
rand try separate crowded pack gop presidential contender ratchet
antiwar rhetoric gain much attention dad first gop presidential debate
approach republican senator presidential candidate rand paul launch
bomb crowd field attempt differentiate candidatessenator paul tell
leftist washington post iowa weekend gop candidate want blow world
significant difference rest gop field plan make case first debate
thursdayaccording post paul say debate pit gop candidate want send
half million son daughter back iraq ask gop candidate whether want
always intervene every civil war around world via politistick
```

Acabamos de observar cómo:

- La palabra *Reuters* aparece en aproximadamente un **88%** del DataFrame de noticias **reales** y tan sólo en un **0.8%** del DataFrame de noticias **falsas**.

- La palabra *Reuter* aparece en aproximadamente un **99.8%** del DataFrame de noticias **reales** y tan sólo en un **1.4%** del DataFrame de noticias **falsas**.
- Aunque la diferencia se reduce, la palabra *via* aparece en solo un **5.2%** del DataFrame de noticias **reales**, mientras que está presente en un **50.2%** del DataFrame de noticias **falsas**. Además, al analizar ejemplos de frases que contienen esta palabra, se observa que suelen hacer referencia a fuentes de información que podrían percibirse como menos confiables que Reuters o incluso inventadas, como *via WhiteRabbitRadio.net*, via NY Post, *via Politistick* o *via White House*.

Dado que el Dataset original de noticias reales podría provenir de **Reuters**, se eliminarán las palabras *Reuter|Reuters* de los textos para evitar que influyan excesivamente en la clasificación.

Sin embargo, otras palabras más neutrales, como *via*, así como las identificadas en el análisis con **SHAP**, incluidas diversas *fuentes de información*, se mantendrán.

El objetivo es que el modelo aprenda a identificar relaciones más complejas entre palabras para determinar si una noticia es real o falsa, en lugar de basarse únicamente en la mención de un medio confiable. Por ello, repetiremos el experimento tras la eliminación.

Vectorización TF-IDF + Clasificación mediante Random Forest

Entrenamiento, test y explicabilidad

```
df = pd.read_csv("Datasets/CleanedFakeAndRealNews.csv")
# Eliminar la palabra "reuter" y "reuters" de los textos
df["clean_text"] = df["clean_text"].str.replace(r"\b(reuter|reuters)\
b", "", regex=True)
df.shape # Mantenemos el mismo número de filas
```

```
(44193, 3)
```

```
df = pd.read_csv("../Datasets/Cleaned-FR-News_V2.csv")
# Dividimos los datos en entrenamiento y prueba
X = df["clean_text"]
y = df["label"]
X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.2, random_state=42)

# Definimos y utilizamos vectorizador TF-IDF
vectorizer = TfidfVectorizer(max_features=5000)  # 5000 palabras más
importantes (ordenadas por frecuencia de aparición en el corpus)
X_train = vectorizer.fit_transform(X_train)
X_test = vectorizer.transform(X_test)

# pd.to_pickle(vectorizer, "models/tf-idf-vectorizer.pkl")
```

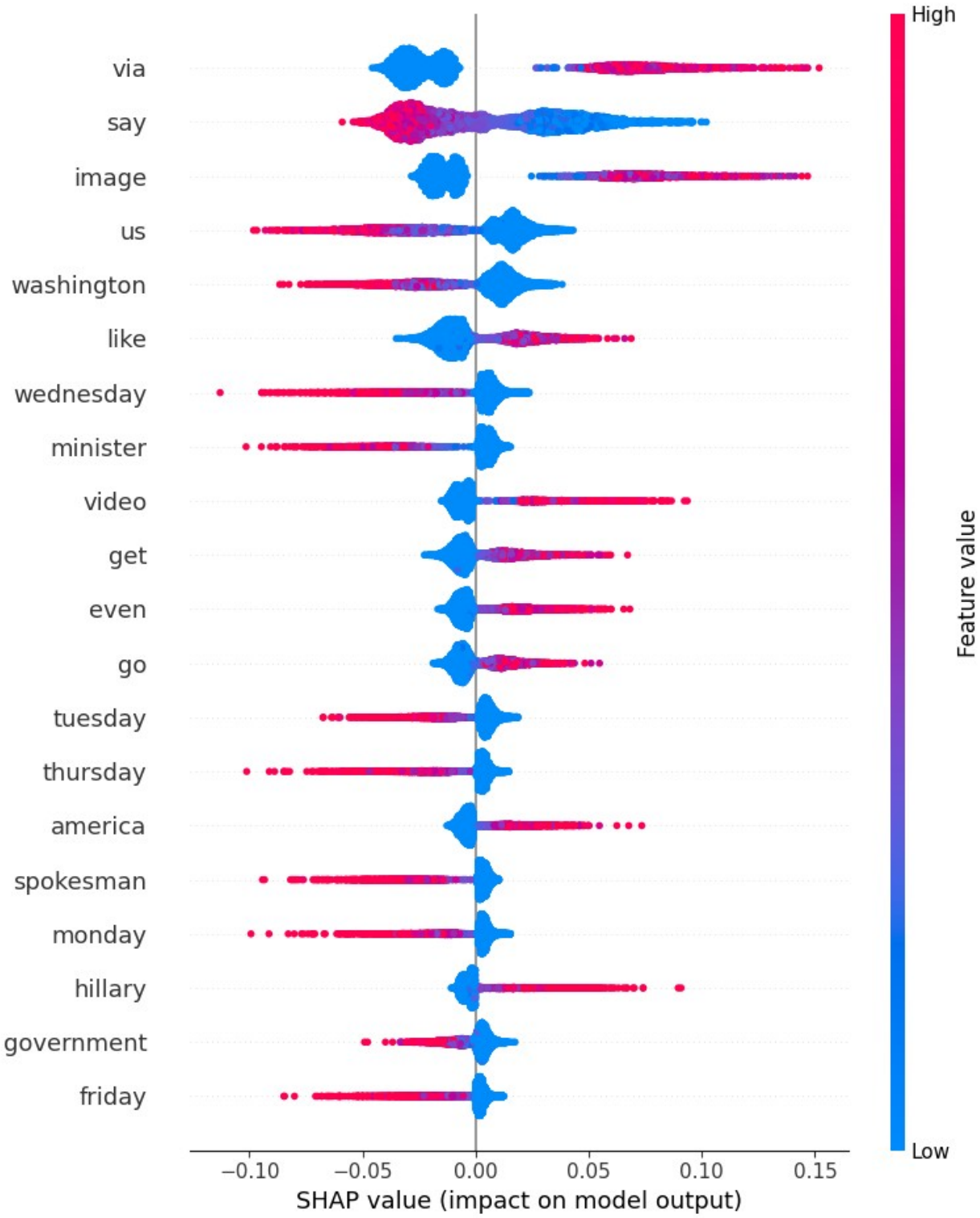
```
# Dividimos los datos en entrenamiento y prueba
X = df["clean_text"]
y = df["label"]
X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.2, random_state=42)

# Definimos y utilizamos vectorizador TF-IDF
vectorizer = TfidfVectorizer(max_features=5000)  # 5000 palabras más
importantes (ordenadas por frecuencia de aparición en el corpus)
X_train = vectorizer.fit_transform(X_train)
X_test = vectorizer.transform(X_test)

# Entrenar modelo
rf = RandomForestClassifier(n_estimators=100, random_state=42)
rf.fit(X_train, y_train)

# Evaluación en test
y_pred = rf.predict(X_test)
print("Precisión:", accuracy_score(y_test, y_pred))
```

```
Precisión: 0.982011539766942
```

La precisión ha bajado únicamente un 1%, lo cual sigue siendo un resultado muy bueno (**98,2%** en test). Veamos la explicabilidad con SHAP de nuevo.

```
# Es necesario convertir las matrices dispersas a densas
X_train_dense = X_train.toarray()
X_test_dense = X_test.toarray()

# Aplicamos SHAP
explainer = shap.Explainer(rf, X_train_dense)
shap_values = explainer(X_test_dense)

# Obtenemos las palabras de los vectores TF-IDF
feature_names = vectorizer.get_feature_names_out()

# Elegimos índice a explicar
class_index = 0 # En este caso, ¿qué palabras influyen más y menos en
que una noticia sea falsa?

# Seleccionamos los valores SHAP para dicha clase
shap_values_class = shap_values[:, :, class_index]

# Mostrar valores SHAP
shap.summary_plot(shap_values_class, X_test_dense,
feature_names=feature_names)
```

```
100%|==================| 17670/17678 [17:16<00:00]
```



A continuación, se describe por qué cada palabra podría influir en la clasificación de la noticia como Fake o Real según se observa en el gráfico SHAP:

1. **via**

   - *Motivo*: Puede emplearse para aparentar referencias externas sin verificación real, **incrementando la probabilidad de ser Fake**.
2. **say**
   - *Motivo*: Alude a declaraciones directas que, si están respaldadas por fuentes confiables, **reducen la sospecha** de información falsa.
3. **image**
   - *Motivo*: Hace referencia a imágenes que pueden ser manipuladas o sacadas de contexto, incrementando la **posibilidad de desinformación**.
4. **us**
   - *Motivo*: Al referirse a Estados Unidos, suele haber más cobertura mediática y verificación, lo que **baja la probabilidad** de que la noticia sea Fake.
5. **washington**
   - *Motivo*: Asociado a la política y prensa formal de EE. UU., a menudo ofrece fuentes oficiales, **reduciendo la posibilidad** de ser Fake.
6. **like**
   - *Motivo*: Uso informal que puede denotar lenguaje sensacionalista, lo cual tiende a **elevar la probabilidad** de ser Fake.
7. **wednesday**
   - *Motivo*: Al mencionar un día específico, suele asociarse con noticias formales y verificadas, **reduciendo la sospecha** de ser Fake.
8. **minister**
   - *Motivo*: Implica una fuente oficial o gubernamental, lo que habitualmente está ligado a noticias más **verídicas**.
9. **video**
   - *Motivo*: La mención de videos puede implicar contenido potencialmente manipulado o editado, **aumentando la desconfianza**.
10. **get**
    - *Motivo*: Verbo usado a menudo en titulares sensacionalistas o vagos, lo cual **eleva la probabilidad** de ser Fake.
11. **even**
    - *Motivo*: Suele aportar énfasis o dramatismo en el discurso, factor que puede incrementar la percepción de **contenido engañoso**.
12. **go**
    - *Motivo*: Palabra de acción que puede relacionarse con llamados a actuar de manera apresurada o **sensacionalista**.
13. **tuesday**
    - *Motivo*: Al mencionar un día específico, suele asociarse con noticias formales y verificadas, **reduciendo la sospecha** de ser Fake.
14. **thursday**
    - *Motivo*: Al mencionar un día específico, suele asociarse con noticias formales y verificadas, **reduciendo la sospecha** de ser Fake.
15. **america**
    - *Motivo*: Vincula el contenido con **asuntos geopolíticos**, un terreno fértil para la proliferación de noticias falsas.
16. **spokesman**

- *Motivo*: Hace referencia a una fuente oficial que suele **dotar de credibilidad** al texto, reduciendo la sospecha de Fake.

17. **monday**
    - *Motivo*: Al mencionar un día específico, suele asociarse con noticias formales y verificadas, **reduciendo la sospecha** de ser Fake.
18. **hillary**
    - *Motivo*: Involucra a una figura política relevante, frecuentemente **asociada a controversias** y, por ende, a posibles Fake News.
19. **government**
    - *Motivo*: Mencionar al gobierno puede implicar declaraciones oficiales o controversias políticas, un **foco común de desinformación**.
20. **friday**
    - *Motivo*: Al mencionar un día específico, suele asociarse con noticias formales y verificadas, **reduciendo la sospecha** de ser Fake.

Es particularmente notable que, de las 20 palabras con mayor impacto en la clasificación, 5 corresponden a días de la semana, representando así el 25% del total.

Verificaremos que no se trate de un formato típico de Reuters para confirmar que la mención del día de la semana efectivamente aporta credibilidad y suele mencionarse en noticias reales.

```
# Buscar frases con la palabra "monday"
sentences_with_monday = [text for text in df["clean_text"] if "monday"
in text][:5]

# Imprimir las frases
for sentence in sentences_with_monday:
    print(sentence)
```

```
brasiliario de janeiro  government brazil president michel temer far
assemble coalition need pass landmark pension reform potential
supporter measure organize key legislator say monday still enormously
far need vote party leader commit party president commit one party set
commit brazil low house speaker rodrigo maia tell journalist event rio
de janeiro pension reform cornerstone policy president temer effort
bring brazil deficit control measure widely unpopular brazilian
accustom relatively expansive welfare net order curry support congress
temer ally water original proposal november require few year
contribution private sector worker receive pension accord several
government source temer ally grow optimistic last week reform chance
however speed essential bill passage congressional recess begin dec
lawmaking thereafter hamper politic lawmaker ramp campaign election
london  foreign minister boris johnson say britain would appeal iran
humanitarian ground free jail aid worker express reservation grant
diplomatic protection would help secure release husband say wednesday
nazanin zaghariratcliffe project manager thomson  foundation sentence
five year prison convict iranian court plotting overthrow clerical
establishment deny charge johnson come pressure resign comment make
early month zaghariratcliffe teach people journalism arrest april
```

critic say comment might prompt iran extend sentence apologize remark monday thomson foundation charity organization independent thomson operate independently news say zaghariratcliffe holiday teaching journalism iran wednesday johnson meet husband richard ratcliffe tell britain would leave stone unturned bid free say british ambassador tehran early raise case iranian authority johnson also stress importance appeal humanitarian ground ratcliffe tell reporter say positive meeting official question whether would help grant wife diplomatic protection move would explicitly make zaghariratcliffe fate issue statetostate relation rather purely consular case legal opinion prepare human right charity redress zaghariratcliffe case say british government could grant diplomatic protection predominantly british citizen deny fair trial say thought would important helpful foreign secretary foreign office express reservation ratcliffe say foreign office say lawyer would meet next fortnight discuss issue iran state news agency irna also signal move could backfire cite comment unidentified international law expert iran view zaghari iranian citizen try due illegal action convict iranian court serve sentence quote expert say hence uk interference peaceful path humanitarian issue consider intervention iran naturally trigger iran severe reaction ratcliffe say johnson keen take trip iran plan end year could allow see wife threeyearold daughter care relative iran first time month important going trip stand alongside foreign secretary understand big ask reasonably unprecedented think important circumstance say ratcliffe say wife appear edge nervous breakdown due test find lump breast say think use diplomatic bargaining chip fight nothing we use vehicle fight say

mexico city mexico government monday say would work strengthen north american economy united states publish objective renegotiation nafta trade deal one mexican official describe bad fear statement mexican economy ministry say expect talks united states mexico canada renegotiate north american free trade agreement nafta able get way aug mexico would continue domestic consultation revamp accord early august add ministry say would work achieve constructive negotiation process allow trade investment flow increase consolidate cooperation economic integration strengthen north american competitiveness united states say top priority talk shrink us trade deficit mexico canada recurring complaint us president donald trump highly anticipate document send lawmaker we trade representative robert lighthizer say would seek reduce trade imbalance improve access we good export canada mexico threenation pact speak condition anonymity senior mexican official say list priority bad expect welcome united states push impose punitive tariff trump threaten official also note we wish ditch chapter dispute settlement mechanism hinder united states pursue antidumping antisubsidy case mexican canadian firm would resisted firmly canada canada fight death chapter official say

new york national security major us election issue bombing new york new jersey hillary clinton donald trump seek burnish foreignpolicy credential monday meeting world leader united nations clinton

democratic presidential nominee return role know well serve president barack obamas secretary state four year trump republican nominee newcomer global stage hurriedly try play catchup rapid succession clinton meet briefly japanese prime minister shinzo abe egyptian president abdel fattah alsisi ukrainian president petro poroshenko trump also meet sisi minute egyptian leader speak clinton manhattan hotel meeting come day start clinton suggest trump harsh rhetoric toward muslims aids islamic state militant group recruit effort trump push back argue united states less safe result obama clinton policy security question arise monday bilateral session take place world leader gather un general assembly clinton abe discuss concern north korea maritime issue involve china clinton trump speak sisi work closely egypt combat islamic state threat trump campaign release statement say trump highlight egypt we share common enemy importance work together defeat radical islamic terrorism clinton sisi also discuss goal move egypt toward new civil society new modern country uphold rule law respect human right liberty clinton poroshenko address russian incursion ukrainian territory clinton start session say ukraine face real problem threat russian aggression anxious know supportive session also resonate trump praise russian president vladimir putin early month trump call putin strong leader obama rattle democrats republicans washington evening without drama start clinton motorcade zoom pack new york street rush hour quickly rush hotel hotel trump also try bolster foreign policy credential last month go mexico meet president enrique pena nieto side end disagree whether would pay build border wall discuss clinton call episode embarrass international incident

frankfurt germany environmentalist green party poll high level year survey publish sunday overtake chancellor angela merkel wouldbe coalition partner talk form new government continue coalition agreement would see merkel extend spell germany helm help avoid collapse euro zone cement country position bloc economic powerhouse merkel must bring together green probusiness free democrats fdp conservative bloc secure majority sticking point include immigration cap whether end coal production increase defense spending poll publish german daily bild put green percent one percentage point week early fdp fell amount percent direct bearing coalition talk give green bragging right term negotiate green call abandon coal source energy germany objective share thousand demonstrator march german city bonn saturday oppose fdp germany want meet climate protection goal exit coal necessary anton hofreiter one leader green parliamentary party tell sunday merkel christian democrats cdu bavarian csu sister party stable percent social democrats say would renew rule coalition conservative slip one percentage point percent party leader slated meet monday evening large negotiating team launch detailed talk fdp leader christian lindner say interview publish sunday party fear new election negotiation fail new vote could see gain farright alternative germany afd surge parliament last month campaign channel public anger merkel decision leave germany border open migrant eurosceptic afd

```
stable sunday survey percent leftwe die linke rise one percentage
point poll percent polling firm emnid interview people october
separate survey carry pollster insa bavaria put merkel local ally csu
percent already disappointing percent general election sept csu leader
horst seehofer fend call resignation since vote bavaria main entry
point migrant germany csu want limit migrant year bavarian survey
people poll nov due appear bild newspaper monday
```

```
# Dividimos el DataFrame en verdadero y falso
df_real = df[df['label'] == 1]['clean_text']
df_fake = df[df['label'] == 0]['clean_text']

# Días de la semana
days_of_week = ["monday", "tuesday", "wednesday", "thursday",
"friday"]

# Contar ocurrencias en df_real
print("Ocurrencias en df_real:")
for day in days_of_week:
    count = sum(1 for text in df_real if day in text)
    print(f"{day.capitalize()}: {count}")

# Contar ocurrencias en df_fake
print("\nOcurrencias en df_fake:")
for day in days_of_week:
    count = sum(1 for text in df_fake if day in text)
    print(f"{day.capitalize()}: {count}")
```

```
Ocurrencias en df_real:
Monday: 4887
Tuesday: 5672
Wednesday: 5593
Thursday: 5341
Friday: 4983

Ocurrencias en df_fake:
Monday: 1877
Tuesday: 1987
Wednesday: 1795
Thursday: 1744
Friday: 1914
```

Se evidencia una mayor presencia de los días de la semana en las noticias reales en comparación con las falsas. Sin embargo, esta diferencia no es tan pronunciada como en el caso de Reuters, lo que sugiere que nuestro modelo ha capturado relaciones más complejas para realizar su clasificación de manera efectiva.

```
df.to_csv("Datasets/Cleaned-FR-News_V2.csv", index=False)
```

Grid Search y Cross Validation

En este apartado, incluiremos dos aspectos adicionales para mejorar la evaluación y optimización del modelo:

- **Grid Search**: Para explorar diferentes combinaciones de hiperparámetros y verificar si es posible mejorar el rendimiento del modelo mediante una configuración más óptima.
- **Cross Validation**: Para obtener una estimación más robusta del desempeño del modelo, calculando la media de evaluación a través de múltiples particiones de los datos. De esta manera, evitaremos basarnos únicamente en el resultado del conjunto de prueba. Esto se realiza al mismo tiempo que se exploran las diferentes combinaciones del Grid Search.

```
# Definir el rango de hiperparámetros para el Grid Search
param_grid = {
    'n_estimators': [50, 100, 200, 300, 400, 500], # Nos interesa
principalmente el número de árboles
    # El resto de hiperparámetros en *default* presentan un buen
rendimiento, por lo que no los modificamos para ahorrar tiempo de
entrenamiento.
}

# Crear un nuevo RandomForestClassifier para el Grid Search
rf_grid = RandomForestClassifier(random_state=42)

# Configurar el Grid Search
grid_search = GridSearchCV(estimator=rf_grid, param_grid=param_grid,
cv=5, scoring='accuracy', verbose=1, n_jobs=-1)

# Ejecutar el Grid Search
grid_search.fit(X_train, y_train)

# Guardar los resultados en una lista
results = grid_search.cv_results_
n_estimators = param_grid['n_estimators']
mean_accuracies = results['mean_test_score']

# Mostrar los mejores hiperparámetros
print("Mejores hiperparámetros:", grid_search.best_params_)
print("Mejor puntuación de validación cruzada:",
grid_search.best_score_)
```

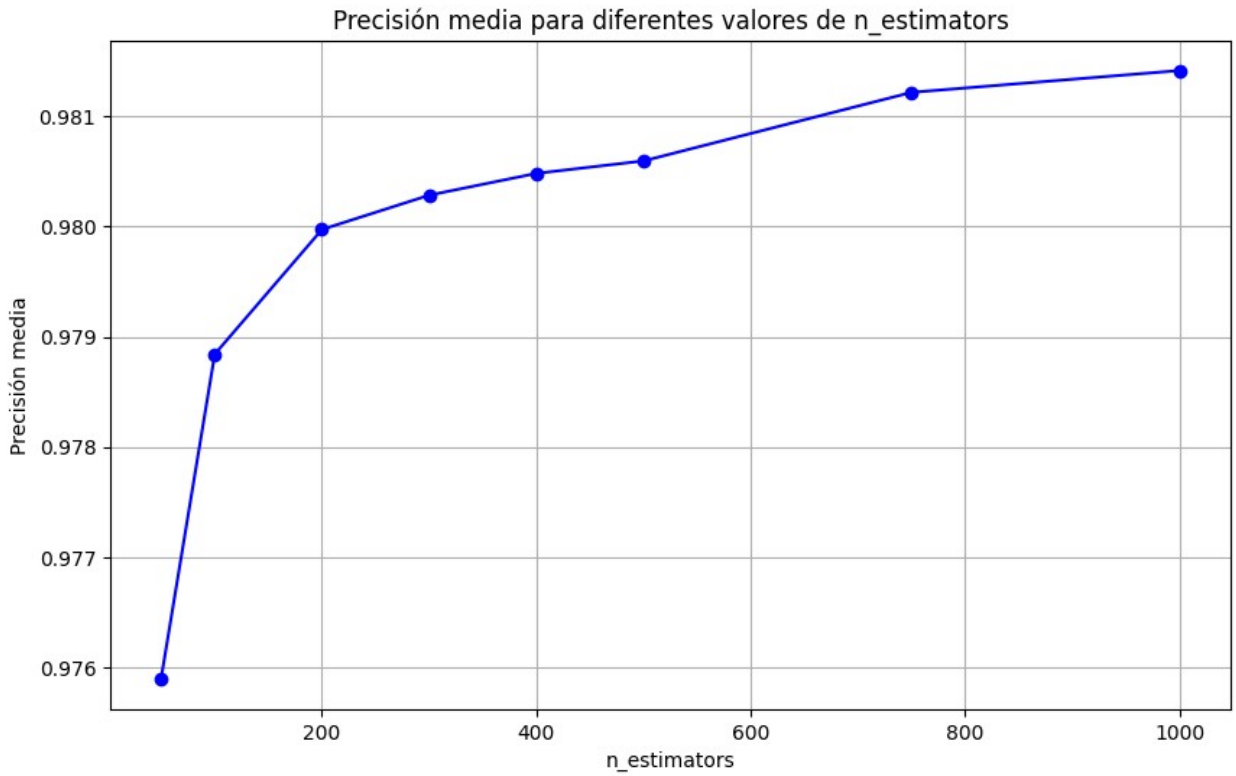
```
Fitting 5 folds for each of 6 candidates, totalling 30 fits
Mejores hiperparámetros: {'n_estimators': 500}
Mejor puntuación de validación cruzada: 0.9805963117676699
```

Dado que el mejor desempeño se obtuvo con el mayor número de estimadores evaluado, ampliaremos el rango de búsqueda agregando dos valores adicionales: *n_estimators=750* y *n_estimators=1000*.

Sin embargo, según la teoría de los árboles de decisión, es probable que la precisión ya haya alcanzado un punto de estabilización alrededor de estos valores. A partir de cierto umbral, aumentar la cantidad de estimadores suele tener un impacto marginal en el rendimiento, ya que el modelo tiende a converger. No obstante, realizar esta prueba nos permitirá confirmar si aún es posible obtener mejoras significativas o si hemos alcanzado el punto óptimo de complejidad del modelo.

```
# Ampliar el rango de búsqueda de hiperparámetros
param_grid_extended = {
    'n_estimators': [750, 1000],  # Agregar 750 y 1000
}

# Configurar el Grid Search con el rango extendido
grid_search_extended = GridSearchCV(estimator=rf_grid,
param_grid=param_grid_extended, cv=5, scoring='accuracy', verbose=1,
n_jobs=-1)

# Ejecutar el Grid Search con el rango extendido
grid_search_extended.fit(X_train, y_train)

# Mostrar los mejores hiperparámetros y la mejor puntuación
print("Mejores hiperparámetros (rango extendido):",
grid_search_extended.best_params_)
print("Mejor puntuación de validación cruzada (rango extendido):",
grid_search_extended.best_score_)

# Unir los valores de n_estimators y las precisiones medias
n_estimators = n_estimators + param_grid_extended['n_estimators']
mean_accuracies = list(mean_accuracies) +
list(grid_search_extended.cv_results_['mean_test_score'])

# Graficar las diferentes precisiones
plt.figure(figsize=(10, 6))
plt.plot(n_estimators, mean_accuracies, marker='o', linestyle='-',
color='b')
plt.title('Precisión media para diferentes valores de n_estimators')
plt.xlabel('n_estimators')
plt.ylabel('Precisión media')
plt.grid(True)
plt.show()
```

```
Fitting 5 folds for each of 2 candidates, totalling 10 fits
Mejores hiperparámetros (rango extendido): {'n_estimators': 1000}
Mejor puntuación de validación cruzada (rango extendido):
0.9814165755020255
```



El modelo con 200 árboles presenta una diferencia de solo 0.001 en precisión respecto al de 1000 árboles. Esto indica que la convergencia ocurre alrededor de 100-200 estimadores, alcanzando así la precisión óptima.

```
# Validación cruzada con 10 particiones y 200 estimadores
rf_cv = RandomForestClassifier(n_estimators=200, random_state=42)
cv_scores = cross_val_score(rf_cv, X_train, y_train, cv=10,
scoring='accuracy', n_jobs=-1)

# Mostrar resultados de la validación cruzada
print("Precisión media en validación cruzada (10 folds):",
np.mean(cv_scores))
```

```
Precisión media en validación cruzada (10 folds): 0.9820105266021107
```

```
# Entrenar el modelo con todo el conjunto de entrenamiento
rf_cv.fit(X_train, y_train)

# Evaluación en el conjunto de prueba
y_pred_test = rf_cv.predict(X_test)
print("Precisión en el conjunto de prueba:", accuracy_score(y_test,
y_pred_test))
```

```
Precisión en el conjunto de prueba: 0.9837085643172304
```

Guardamos el mejor modelo de Random Forest, con una precisión en el conjunto de prueba de un **98.37%**.

```python
# Guardar el modelo
with open('models/best_rf.pkl', 'wb') as f:
    pickle.dump(rf_cv, f)
```