

# PROYECTO 1 ETAPA 1

## **Reporte del Análisis de datos para la detección de noticias falsas**

INTELIGENCIA DE NEGOCIO  
Equipo 7

Anderson Arévalo Mendoza, 202014997  
Juan José Hurtado Cantin, 202224472  
Nicolás Pérez Ramos, 202221292

FECHA: FEBRERO, 2025

## TABLA DE CONTENIDO






Contexto del problema .....	3
CANVAS .....	3
Entendimiento y preparación de los datos.....	4
Modelado y evaluación .....	6
Modelo 1 – NAIVE BAYES – Juan Jose .....	6
Modelo 2 – Super Vector Machine – Nicolas .....	6
Modelo 3 – Random Forest Classifier – Anderson Arevalo .....	7
Resultados y conclusiones .....	7
Trabajo en equipo .....	7
Contenido	

# Contexto del problema

La proliferación de noticias falsas (**fake news**) en plataformas digitales representa un desafío crítico para las organizaciones que buscan mantener la integridad de la información y proteger a sus usuarios de la desinformación. Estas noticias engañosas, diseñadas para manipular la opinión pública o generar clics masivos, pueden tener consecuencias negativas como la propagación de rumores, pérdida de credibilidad y desconfianza entre los usuarios. Ante este contexto, surge la necesidad de implementar sistemas automáticos de detección de fake news que permitan identificar y filtrar contenido fraudulento de manera eficiente, garantizando así la difusión de información verídica y protegiendo la reputación de la plataforma.

## CANVAS

<div>TAREA DE APRENDIZAJE</div> <div></div> <div>Se realiza un aprendizaje supervisado ya que se cuenta con datos etiquetados en el dataset indicando si la noticia es falsa o verdadera. Es necesario predecir si una noticia es falsa o verdadera. Cuando el modelo ya ha sido entrenado, se toma muy pocos segundos para realizar la predicción cuando este se ejecuta sobre un nuevo conjunto de datos, sin embargo, el reentrenamiento del modelo se podría demorar bastantes minutos dependiendo de la cantidad de datos y el algoritmo elegido.</div>	<div>DECISIONES</div> <div>Para transformar los resultados del modelo en recomendaciones para el usuario se pueden realizar varias acciones. Como mostrar mensajes de alerta si se encuentra una noticia falsa o si hay sospechas de la veracidad de la noticia se pueden mostrar enlaces confiables relacionados con el tema. Ya que el modelo no es 100% acertado, cuando se encuentre una noticia falsa o con posibilidades de que sea falsa, se enviará a revisión manual para lograr identificar con mayor certeza si la noticia es falsa o verídica.</div>	<div>PROPUESTA DE VALOR</div> <div>Los que se benefician de la implementación del modelo son los periodistas, medios de comunicación, plataformas de redes sociales y en general empresas que trabajen con la publicación y uso de la información que se publica. Algunos de los problemas a los que toca enfrentarse son la desinformación y falsas noticias en áreas como la política, ya que se puede manipular la opinión pública e influir en procesos electorales y dañar la imagen de algunas instituciones. Otro problema es la dificultad para verificar las noticias falsas debido a la cantidad de información incorrecta que está circulando por los medios de comunicación. El uso de este modelo puede traer algunos riesgos como los falsos positivos o falsos negativos, es decir noticias que pueden ser clasificadas de manera incorrecta, ya sea como verdaderas o falsas. Otro de los riesgos sería el sesgo en los datos, ya que si se entrena el modelo con datos inclinados hacia un sector de la política se puede favorecer este sector por encima de los demás.</div>	<div>RECOLECCIÓN DE DATOS – NO SE DEBE DILIGENCIAR</div> <div>¿Cómo se obtiene el conjunto inicial de entidades y resultados (por ejemplo, extractos de bases de datos, extracciones de API, etiquetado manual)? ¿Qué estrategias se aplican para actualizar los datos continuamente, controlando los costos y manteniendo la vigencia?</div>	<div>FUENTES DE DATOS</div> <div>Los datos provienen de documentos csv con toda la información de noticias organizadas por columnas como ID, fecha, título y descripción. Esta información fue obtenida a partir de fuentes externas como periódicos, redes sociales y demás medios de comunicación que publican noticias. Los datos de los archivos pueden ser usados para realizar un análisis, sin embargo, primero es necesario realizar una limpieza y un procesamiento de estos.</div>
--	---	--	--	---

 <p><b>SIMULACIÓN DE IMPACTO</b></p> <p>Algunos de los beneficios dados por las decisiones tomadas son, la automatización del análisis de datos, la clasificación de los textos, la toma de decisiones fundamentada en estadísticas y la escalabilidad. Sin embargo, los costos asociados son los falsos positivos o falsos negativos, los recursos computacionales que se consumen dependiendo del modelo y la dependencia de los datos de entrada, ya que si estos están desbalanceados pueden afectar la toma de decisiones.</p>	 <p><b>APRENDIZAJE (USO DEL MODELO)</b></p> <p>Elegimos el modelo Random Forest, el cual principalmente se ejecuta por lotes, ya que requiere múltiples evaluaciones, de manera que procesa grandes volúmenes de datos en intervalos definidos. La frecuencia de actualización depende de la variabilidad de los datos, si se cambian los datos, es recomendable actualizar el modelo cada cierto tiempo.</p>	 <p><b>CONSTRUCCIÓN DE MODELOS</b></p> <p>Se realizaron 3 modelos, sin embargo, seleccionamos el modelo de Random Forest, para quedarse solamente con este es necesario monitorear su desempeño. Para actualizar el modelo es necesario esperar a que los datos cambien, aunque es recomendable actualizarlo cada cierto tiempo. Además, hay que reentrenar el modelo si la tendencia de los datos cambia o su desempeño disminuye.</p>	 <p><b>INGENIERÍA DE CARACTERÍSTICAS</b></p> <p>El modelo de Random Forest utiliza una combinación de variables clave seleccionadas a partir del análisis de datos, incluyendo características numéricas y categóricas relevantes para la predicción. Se aplican transformaciones como normalización o codificación de variables categóricas, y se pueden generar variables derivadas para mejorar la precisión. Además, se utilizan técnicas de manejo de valores nulos y selección de características para optimizar el desempeño del modelo.</p>
	<p><b>MONITOREO NO SE DEBE DILIGENCIAR</b></p>  <p>¿Qué métricas y KPI se utilizan para hacer un seguimiento del impacto de la solución de ML una vez desplegada, tanto para los usuarios finales como para la empresa? ¿Con qué frecuencia deben revisarse?</p>		

## Entendimiento y preparación de los datos

El conjunto de datos analizado contiene 57,063 registros y está compuesto por las columnas ID, Label, Titulo, descripción y Fecha. La columna Label indica si la noticia es real o falsa, mientras que Titulo y descripción contienen el contenido de la noticia. Al explorar los datos, se identificaron 16 noticias que no tenían títulos o sus valores eran nulos, por lo que se realiza la eliminación de estos datos. Además, se encontraron 5,458 noticias duplicadas por el Titulo y 7,425 noticias duplicadas al tener la misma descripción. los registros que tenían título o descripción duplicada se eliminaron. Para extraer información relevante del texto, se calcularon varias características, como el conteo de palabras en los títulos y descripciones, la palabra más frecuente (moda) y la longitud mínima y máxima de palabras. Los histogramas generados muestran que la mayoría de los títulos tienen entre 50 y 150 caracteres, con palabras que varían entre 1 y 40 caracteres de longitud según lo mostrado en la Fig. 1.

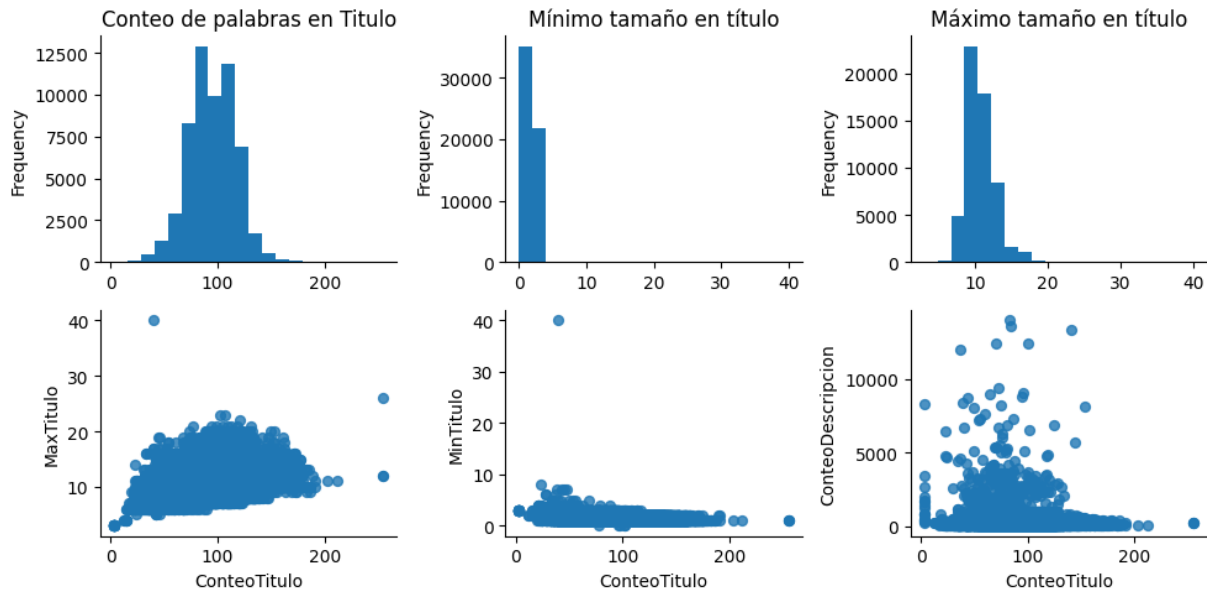


Figura 1. Estadística descriptiva para Título

En el caso de las descripciones, la distribución es mucho más dispersa, con algunas noticias alcanzando hasta 10,000 caracteres, lo que podría un desbalance en el tamaño de palabras por descripción. Sin embargo, la gran mayoría de descripciones se agrupan por debajo de 5000, como se observa en la Fig. 2.

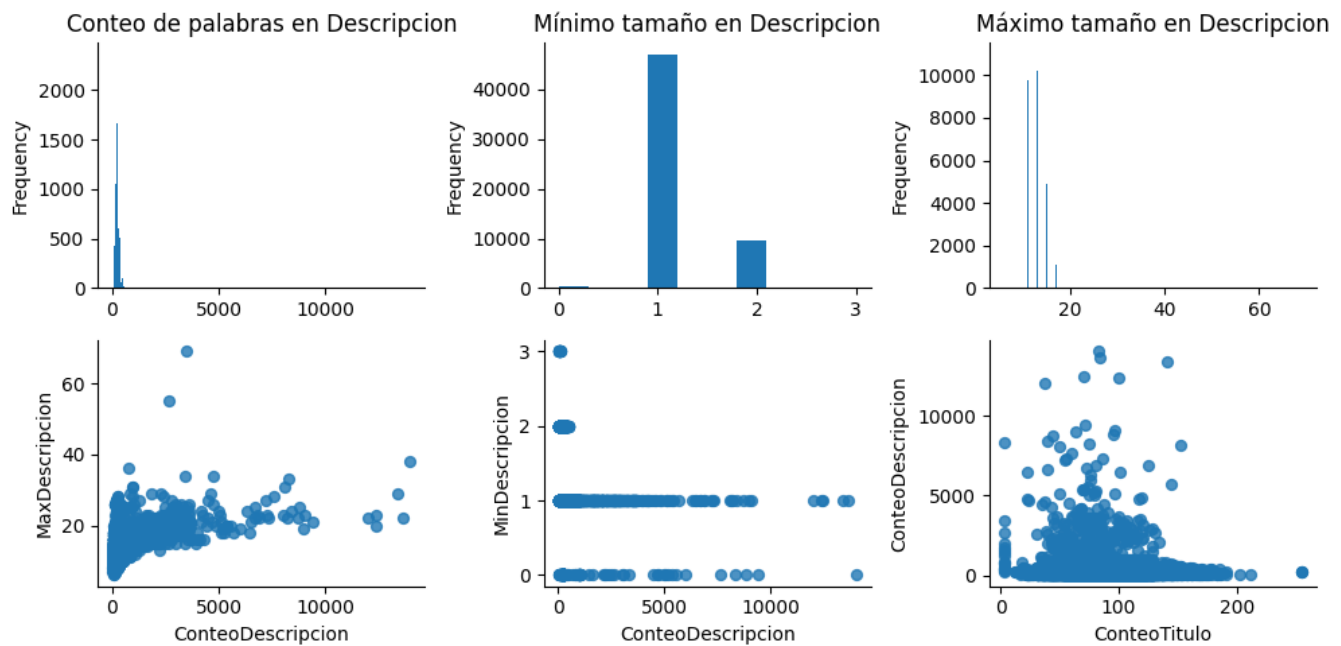


Figura 2. Estadística descriptiva para Descripción

Se realiza un preprocesamiento, compuesto de eliminar datos nulos, duplicados y vacíos posteriormente se aplicaron los siguientes pasos:

1. se aplicó **to\_lowercase**, que convierte todas las palabras a minúsculas para mantener la consistencia.
2. se implementó **remove\_non\_ascii**, que elimina caracteres especiales y no ASCII, asegurando que el texto contenga solo caracteres estándar
3. se aplicó **remove\_punctuation**, que elimina signos de puntuación y caracteres innecesarios que podrían interferir con el análisis.
4. Se aplicó **replace\_numbers**, que convierte los números en su forma textual usando la librería `num2words`.
5. Finalmente, se aplicó **remove\_stopwords**, que eliminó palabras irrelevantes como "el", "de", "y", que no aportan significado relevante al modelo.

Primero se **tokenizaron** los títulos y descripciones usando `word_tokenize()`, separando el texto en palabras individuales. Posteriormente se aplicaron las 5 fases descriptas anteriormente. Luego se reconstruyeron las frases para título y descripción. Sin embargo, tras probar el modelo con los datos preprocesados, se observó que las métricas de clasificación se vieron afectadas negativamente. La eliminación de stopwords y la normalización del texto redujeron la capacidad del modelo para identificar patrones importantes en los datos. Debido a esto, se decidió no aplicar el preprocesamiento y, en su lugar, concatenar el título y la descripción en una sola columna.

## Modelado y evaluación

### Modelo 1 – NAIVE BAYES – Juan Jose

El modelo de Naive Bayes es un clasificador probabilístico basado en el Teorema de Bayes, que asume independencia condicional entre las variables. En nuestro caso, este modelo fue utilizado para predecir la categoría de una determinada instancia a partir de las características seleccionadas. Sus métricas de desempeño muestran una precisión del 82.46 %, una recuperación del 97.75 % y una exactitud general del 86.61 %. Aunque tuvo un buen desempeño en la clasificación de la clase mayoritaria, presentó dificultades en la clase minoritaria, resultados que se pueden observar en la matriz de confusión generada por este modelo.

### Modelo 2 – Super Vector Machine – Nicolas

El modelo Support Vector Machine (SVM) fue elegido para la tarea de clasificación de noticias debido a su eficacia en problemas de clasificación binaria y su capacidad para manejar datos textuales de alta dimensión, como los presentes en la detección de noticias falsas. Utilizando un kernel lineal y transformando los textos mediante la técnica TF-IDF, el modelo logró separar de manera óptima las noticias reales de las falsas. Los resultados obtenidos muestran un rendimiento sólido con una precisión (accuracy) del 91.13%, una precisión específica para noticias falsas del 89.10%, una sensibilidad (recall) del 96.49% y un F1-Score de 92.64%, evidenciando un buen equilibrio entre la detección de noticias falsas y la minimización de falsos positivos. La matriz de confusión revela que 6,373 noticias falsas fueron correctamente identificadas, mientras que solo 232 pasaron desapercibidas y 780 noticias reales fueron clasificadas erróneamente como falsas. Estos resultados destacan la capacidad del modelo para cumplir con los objetivos de la organización al reducir la propagación de desinformación, mejorar la confianza de los usuarios y automatizar el proceso de moderación de contenido con un alto nivel de precisión y fiabilidad.

### Modelo 3 – Random Forest Classifier – Anderson Arevalo

Se desarrolla un modelo Random Forest Classifier, configurado con 50 árboles y una profundidad máxima de 2500. Antes de entrenarlo, se aplicó TF-IDF (Term Frequency - Inverse Document Frequency) para transformar los textos en representaciones numéricas. Este enfoque ayuda a identificar la importancia de las palabras en los títulos y descripciones de las noticias, permitiendo clasificar si son reales (0) o falsas (1). En cuanto a su rendimiento, el modelo obtuvo una exactitud (accuracy) del 94.09%, lo que indica un alto nivel de aciertos en la clasificación. La precisión para detectar noticias falsas (FAKE) fue del 92.62%, lo que significa que, de todas las noticias clasificadas como falsas, el 92.62% realmente lo eran. La sensibilidad (recall) para FAKE fue del 98.01%, lo que sugiere que el modelo identifica casi todas las noticias falsas correctamente. El F1-Score, que equilibra precisión y sensibilidad, fue de 95.24%, confirmando un buen desempeño general. Analizando la matriz de confusión, se observa que el modelo clasificó correctamente 3117 noticias reales y 5283 noticias falsas. Sin embargo, hubo 421 falsos positivos, es decir, noticias reales que fueron erróneamente clasificadas como falsas. También se encontraron 107 falsos negativos, lo que significa que algunas noticias falsas fueron clasificadas incorrectamente como reales. En términos de comportamiento, el modelo tiende a ser más riguroso al detectar noticias falsas, lo que se refleja en su alta sensibilidad. Sin embargo, su precisión es ligeramente menor, lo que indica que algunas noticias reales están siendo etiquetadas erróneamente como falsas.

## Resultados y conclusiones

- a) El modelo **Random Forest Classifier** demostró un excelente desempeño en la detección de noticias falsas, alcanzando una precisión del **94.07%**, un recall del **98.20%** y un F1-Score de **95.24%**, lo que refleja un equilibrio sólido entre identificar correctamente las noticias falsas y minimizar los errores al clasificar contenido legítimo. Con solo **97 falsos negativos** y una baja tasa de **432 falsos positivos**, el modelo asegura que la mayoría de las noticias falsas sean detectadas, alineándose con los objetivos de la organización de reducir la desinformación y fortalecer la confianza en la plataforma. Este alto rendimiento permite automatizar la moderación de contenido, mejorando la eficiencia operativa y reduciendo la carga manual, al tiempo que mantiene la libertad de expresión al evitar una censura excesiva. Se recomienda un monitoreo continuo y reentrenamiento periódico para mantener el rendimiento frente a nuevas tácticas de desinformación, junto con estrategias de educación al usuario para aumentar la transparencia y confianza en el sistema.
- b) El análisis de las palabras clave extraídas por el modelo **Random Forest Classifier** revela patrones distintivos entre noticias falsas y reales, donde términos sensacionalistas como “**conspiración**”, “**urgente**”, y “**revelado**” son indicadores frecuentes de noticias **FAKE**, mientras que palabras como “**oficial**”, “**reporte**”, y “**evidencia**” se asocian a contenido **REAL**. Esta información es crucial para la organización, ya que permite optimizar la detección de noticias falsas, implementar filtros preventivos basados en estas señales léxicas y educar a los usuarios sobre patrones comunes de desinformación. Además, este enfoque fortalece la confianza del usuario al asegurar que el contenido publicado sea veraz y fiable, contribuyendo directamente a los objetivos estratégicos de la organización.

## Trabajo en equipo

Juan José: Realizó el Canvas y su modelo correspondiente de aprendizaje automático, el cual fue Naive Bayes. Usó ChatGPT para encontrar algunos ejemplos de modelo para realizar y corregir algunos errores en el código.

Anderson: Realizó el entendimiento y preparación de datos y su modelo correspondiente de aprendizaje automático. Usó ChatGPT para mejorar la redacción en la explicación del Notebook y resultados.

Nicolás: Realizó el análisis de resultados y su modelo correspondiente de aprendizaje automático. Usó chatGPT para encontrar la utilidad del modelo para los objetivos del negocio y corregir algunos errores en el código.

Todos realizamos el trabajo que se nos asignó, por lo cual el porcentaje de contribución de cada uno es de 33.3%

Se realizó una reunión de planeación el 3 de febrero, en la cual se define cronograma forma de trabajo y las secciones que cada uno desarrolla. Las reuniones de ideación y de seguimiento no se realizan dado que se resuelven en la clase y por medio del grupo de WhatsApp.

Como oportunidades de mejora se plantea asignar las reuniones de seguimiento semanales virtuales para validar el avance correspondiente. También se propone mejorar la comunicación en el grupo de WhatsApp con el fin de generar mayor interacción y poder compartir mejor los avances del proyecto