Informe Proyecto Analítica de textos 20251 - Etapa 1

ISIS-3301 – Inteligencia de Negocios

Universidad de los Andes

Juan Felipe Puig Pardo - 202221336

Mateo Parra Ochoa - 202213933

Juan Nicolás García - 201717860

1. Introducción

   El siguiente informe tiene como fin documentar el proceso de creación de un modelo basado en aprendizaje automático para detectar noticias falsas. Se considera sumamente relevante para atacar la desinformación y mitigar su impacto en la sociedad, proporcionando una herramienta que permita identificar y clasificar contenido engañoso de manera automatizada.

   Para ello, se implementaron distintas técnicas de procesamiento de lenguaje natural (NLP) y se evaluaron modelos de clasificación, incluyendo Regresión Logística, Random Forest y Máquinas de Soporte Vectorial (SVM).

2. **CanvasMl 1.2.1**

3. Impacto y enfoque analítico

La solución propuesta tiene un impacto significativo en la detección automática de noticias falsas, ayudando a reducir la desinformación en plataformas digitales y medios de comunicación. El enfoque analítico es predictivo, ya que el modelo clasifica nuevas noticias como falsas o verdaderas. Se emplea aprendizaje supervisado, dado que el modelo se entrena con un conjunto de datos etiquetado. La tarea de aprendizaje es clasificación binaria, utilizando técnicas como la vectorización de texto con TF-IDF y modelos como Regresión Logística y Random Forest. Estos algoritmos fueron seleccionados por su capacidad de interpretar datos textuales y ofrecer buen rendimiento en términos de precisión y escalabilidad para la toma de decisiones automatizada.

4. Entendimiento de datos

En la etapa de entendimiento de los datos, se carga el dataset y se realiza un análisis exploratorio para comprender su estructura. Se revisa la cantidad de datos, los tipos de variables y se analiza la distribución de etiquetas de noticias falsas y verdaderas. Además, se estudia la longitud de los títulos y descripciones, y se generan gráficos y estadísticas para identificar patrones en los datos antes del modelado.

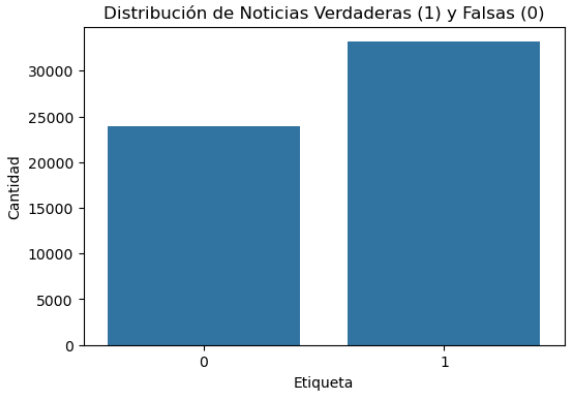

Figura 1: Distribución de noticias falsas y verdaderas

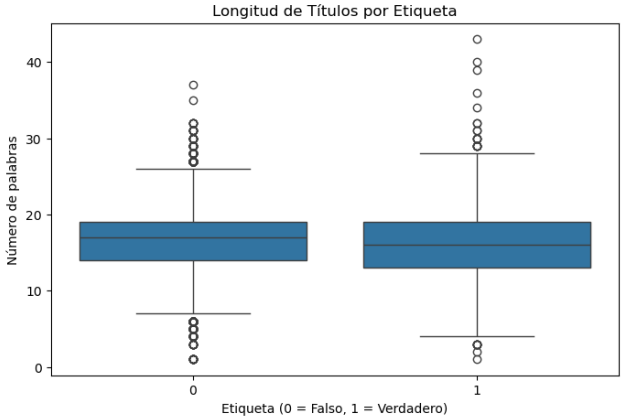

Figura 2: Longitud de títulos por etiqueta

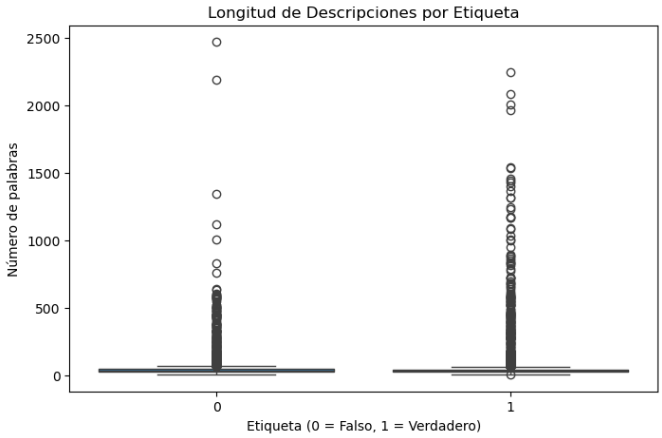

Figura 3 Longitud de Descripciones por Etiqueta

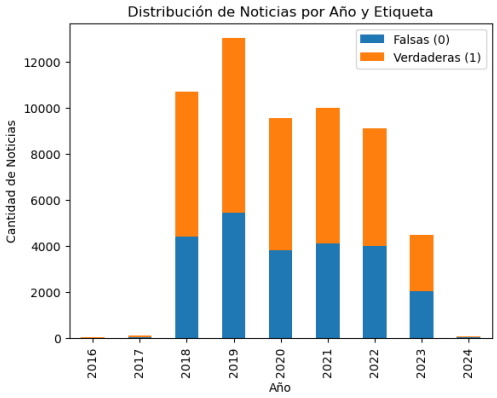

Figura 4: Distribución de noticias por año y etiqueta

5. Preparación de datos

En la etapa de preparación de datos, se limpia y normaliza el texto eliminando stopwords (palabras vacías) y aplicando técnicas de preprocesamiento como lematización o stemming. También se manejan posibles valores nulos y se estandariza el formato del texto, eliminando caracteres especiales y convirtiéndolo a minúsculas. Este proceso mejora la calidad de los datos y optimiza su representación para el modelado.

6. Modelos

Regresión Logística:
Se eligió porque es un modelo rápido, interpretable y eficiente para tareas de clasificación binaria como la detección de noticias falsas. Funciona bien cuando las clases son linealmente separables y permite obtener probabilidades de predicción, lo que ayuda a interpretar los resultados.
Random Forest:
Se seleccionó por su capacidad de manejar relaciones no lineales y capturar patrones complejos en los datos. Es un modelo robusto contra sobreajuste, ya que combina múltiples árboles de decisión para mejorar la precisión. Es especialmente útil cuando las características textuales transformadas con TF-IDF presentan interacciones difíciles de modelar con métodos lineales.
SVM:
Fue escogido debido a su capacidad para trabajar con datos que presentan muchas características (dimensiones), algo típico en el análisis de texto. Con el uso de kernels adecuados, SVM puede capturar relaciones no lineales en el texto representado con TF-IDF, resultando especialmente útil en la clasificación de noticias falsas. Además, su estrategia de maximizar el margen entre clases suele ofrecer un buen desempeño y

una notable capacidad de generalización, incluso en escenarios con datos relativamente limitados o desbalanceados.

7. Resultados

Resultados Regresion Logistica

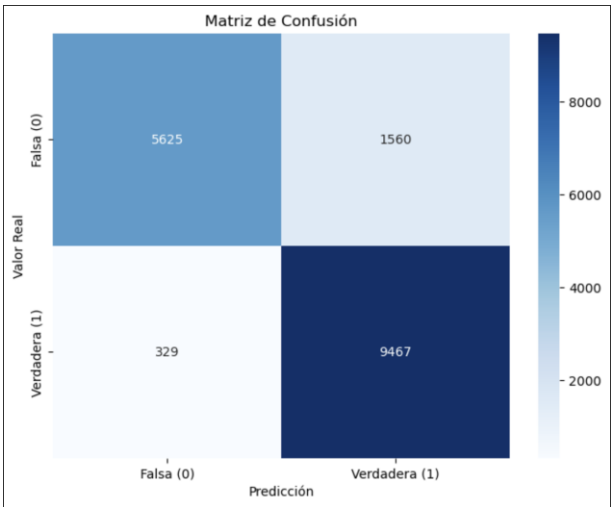

Figura 5: Matriz de confusión Regresión logistica

```
Precisión del modelo: 0.8887580236735174

Reporte de clasificación:
              precision    recall  f1-score   support

    Falsa (0)       0.94      0.78      0.86      7185
Verdadera (1)       0.86      0.97      0.91      9796

     accuracy                           0.89     16981
    macro avg       0.90      0.87      0.88     16981
 weighted avg       0.90      0.89      0.89     16981
```

Ejecución del modelo Regresión

Random Forest:

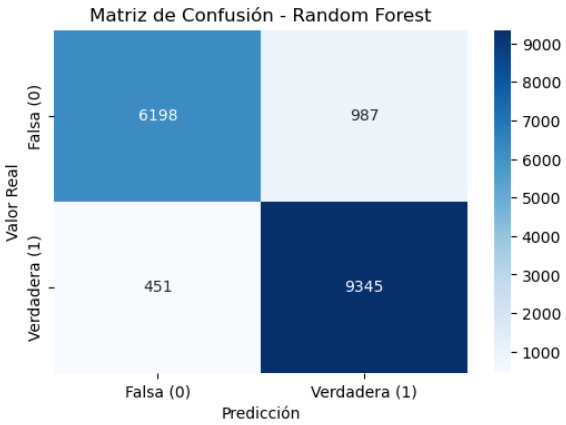

Figura 6: Matriz de confusión RF

```
Precisión del modelo Random Forest: 0.9153171191331488

Reporte de clasificación:
              precision    recall  f1-score   support

    Falsa (0)       0.93      0.86      0.90      7185
Verdadera (1)       0.90      0.95      0.93      9796

     accuracy                           0.92     16981
    macro avg       0.92      0.91      0.91     16981
 weighted avg       0.92      0.92      0.91     16981
```

Ejecución del modelo RF

Principales Características del Modelo Random Forest

Precisión general: 91.53%

Detección de noticias falsas (Clase 0):

Precisión: 93%

Recall: 86%

F1-Score: 90%

Detección de noticias verdaderas (Clase 1):

Precisión: 90%

Recall: 95%

F1-Score: 93%

Accuracy total: 92%

El modelo muestra un desempeño sólido con un accuracy del 92%, clasificando correctamente la mayoría de las noticias. Sin embargo, se detectan mejor las noticias verdaderas (recall del 95%) que las noticias falsas (recall del 86%), lo que indica que puede que en ejecución aparezcan falsos negativos. A pesar de esto, el balance entre precisión y recall es adecuado, por lo que se puede decir que el modelo es confiable para la clasificación de noticias. Existe un gran espacio de mejora relacionado a la optimización en la detección de noticias falsas.

Resultados SVM

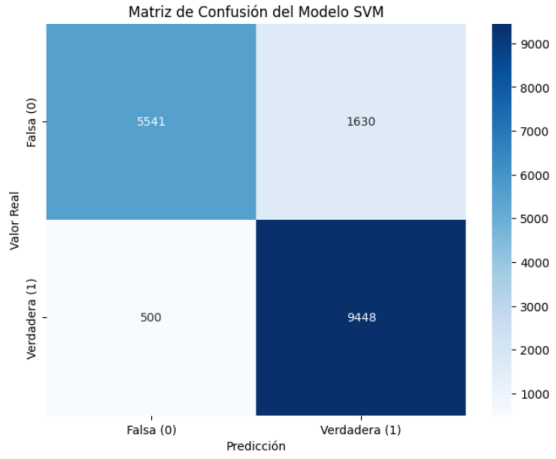

Figura 7: Matriz de confusión SVM

La matriz nos sugiere que el modelo tiende a desempeñarse ligeramente mejor en la identificación de noticias verdaderas que en la detección de las falsas.

```
Precisión del modelo SVM: 0.8755768444418482

Reporte de clasificación:
              precision    recall  f1-score   support

    Falsa (0)       0.92      0.77      0.84      7171
Verdadera (1)       0.85      0.95      0.90      9948

     accuracy                           0.88     17119
    macro avg       0.89      0.86      0.87     17119
 weighted avg       0.88      0.88      0.87     17119
```

Se resalta del modelo una precisión global de 88%, lo cual refleja un buen rendimiento general. Para la clase Falsa (0), obtiene una precisión de 92%, mientras que en la clase Verdadera (1) alcanza un 85%. Además, el F1-score promedio es de 0.87, que indica un equilibrio adecuado entre precisión y recall en ambas clases.

8. Estrategia

Podemos decir que, para la organización, estos modelos pueden ser clave en la detección automatizada de desinformación, ayudando a mejorar la credibilidad del contenido digital en medios de comunicación y plataformas en línea. Pues los resultados de los modelos indican que Random Forest es el modelo con mejor equilibrio entre precisión y recall, lo que lo hace adecuado para minimizar tanto falsos positivos como falsos negativos. Sin embargo, si la prioridad es detectar la mayor cantidad de noticias falsas posibles, SVM y Regresión Logística pueden

9. Descripción de tareas realizadas

Juan Felipe Puig Pardo – líder de proyecto

Fue responsable de la organización del proyecto, estableció fechas de reuniones, pre-entregas y asegurando una distribución equitativa de las tareas.

Aporto en el desarrollo del proyecto en las etapas:

- Entendimiento de los datos
- Preparación de los datos
- Implementación de modelo de regresión logística

Mateo Parra Ochoa – líder de datos

Se encargó de gestionar los datos utilizados en el proyecto y de asignar tareas relacionadas con ellos.

Aporto en el desarrollo del proyecto en las etapas:

- Entendimiento de los datos
- Preparación de los datos
- Implementación de modelo de Random Forest

Juan Nicolás García - líder de analítica

Se encargó de gestionar las tareas de analítica del grupo, asegurando que los entregables cumplieran con los estándares de análisis.

Aporto en el desarrollo del proyecto en las etapas:

- Entendimiento de los datos
- Preparación de los datos
- Implementación de modelo de SVM