Precisión en los modelos

Todos los modelos se evaluaron con 3 tipos de datos distintos, los primeros siendo los datos en crudo o base, sin modificación alguna, los segundos realizándose una limpieza de datos y el tercer conjunto de datos, siendo con la tokenización de los datos.

Datos en bruto
Códigos sin modificación

Datos limpios
Estos datos no tienen comentarios y se separan los caracteres especiales. Por ejemplo:

$$(a * b = c) \Rightarrow (\ a\ \ *\ \ b\ )\ \ =\ \ c$$

Datos tokenizados
En estos datos, se renombran los tokens de los datos ya limpios, como por ejemplo, las variables se renombran como variable, los operadores como operador, los comentarios como comentario, etc., para tener más estandarizado el código. Por ejemplo:

$$int\ a\ =\ 5\ \Rightarrow\ int\ variable\ operador\ número$$

Distribución de probabilidad

Datos en bruto
Se considera plagio a partir de 0.9, es decir a partir del 80.79% de precisión

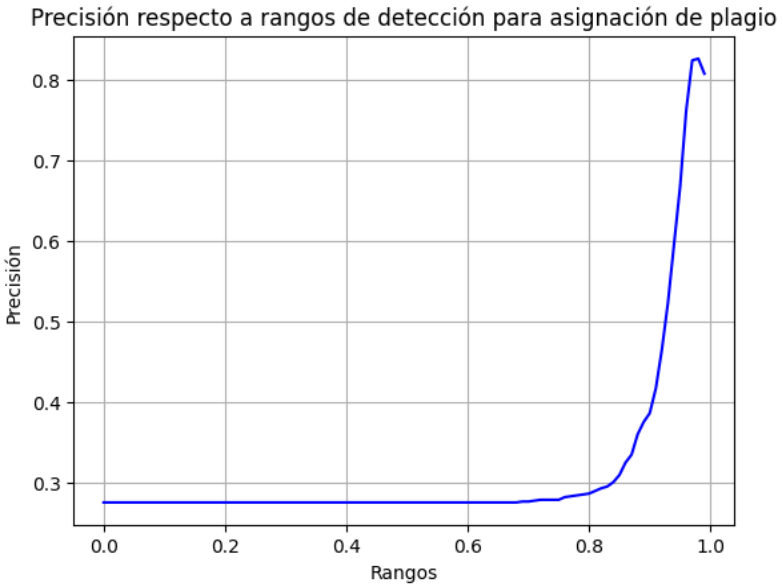
Gráfica 1. Distribución de probabilidad. Datos en bruto

El modelo identifica que es un plagio cuando la diferencia de cosenos es mayor a 0.95 (registros 72 de 911 = 7% de los casos)

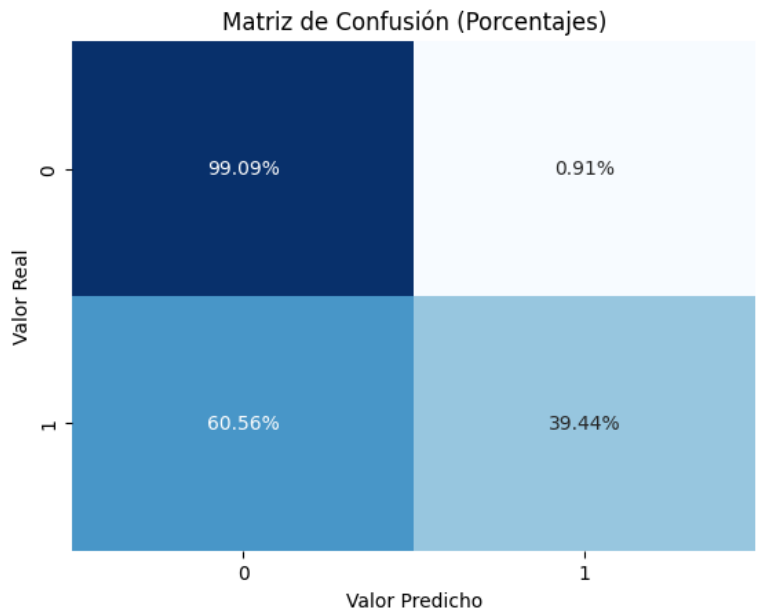
Gráfica 2. Distribución de probabilidad. Datos limpios

Datos limpios

Se considera plagio desde 0.98 es decir de 82.65% de precisión



Gráfica 3. Distribución de probabilidad. Datos tokenizados



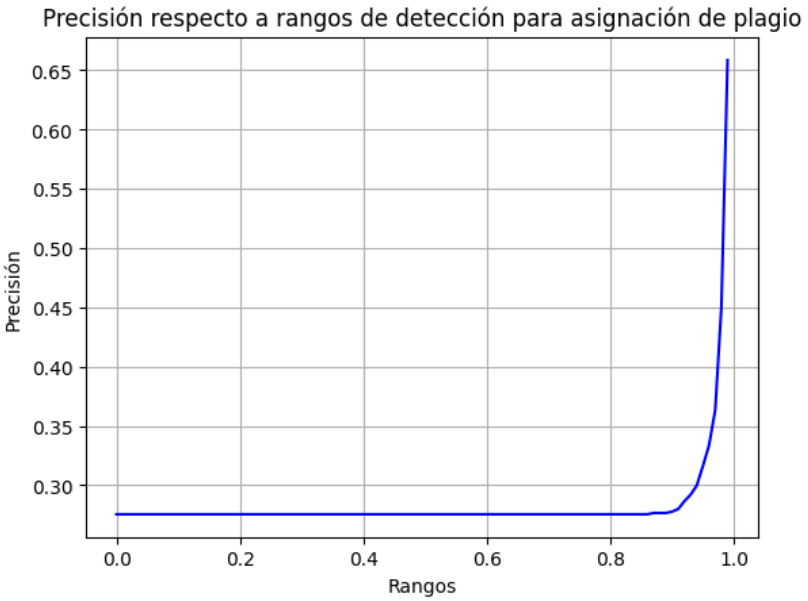
Gráfica 4. Precisión respecto a rangos. Datos limpios

El modelo identifica que es un plagio cuando la diferencia de cosenos es mayor a 0.99 (registros 78 de 911 = 8.5% de los casos)

Gráfica 5. Precisión respecto a rangos. Datos limpios

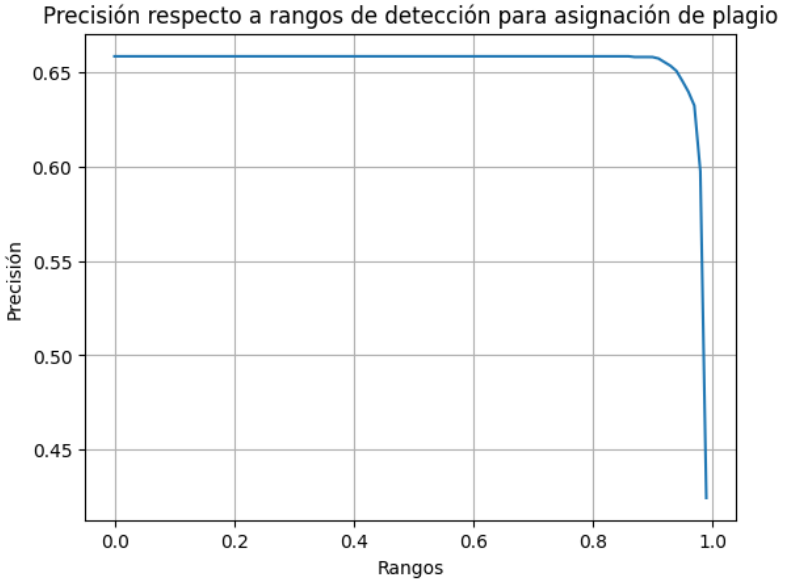
Datos tokenizados

Se considera plagio desde 0.99 es decir de 65.8% de precisión



Gráfica 6. Precisión respecto a rangos. Datos tokenizados

El modelo nunca se llega a una precisión por encima del 66%



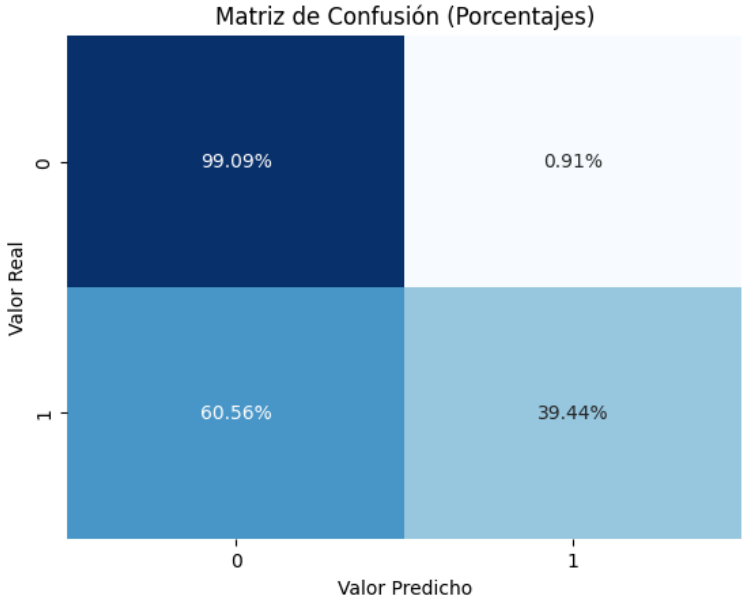
Gráfica 7. Precisión respecto a rangos. Datos tokenizados

Markov

Datos en bruto

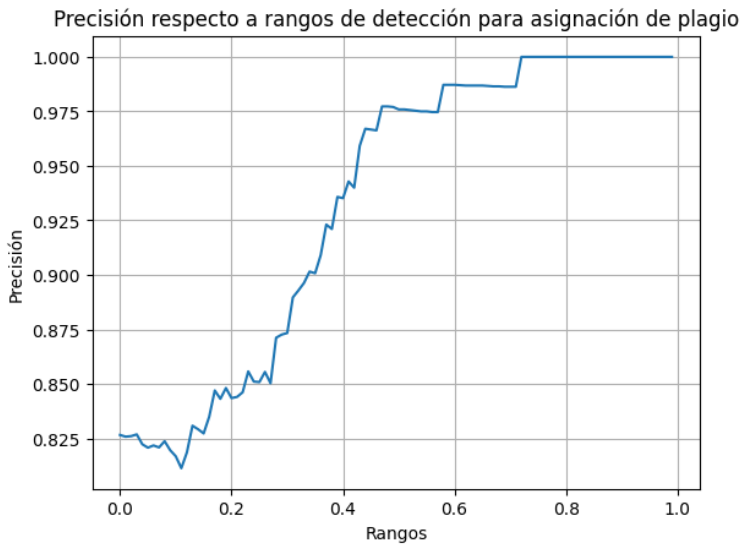
Se considera plagio a partir de 0.41, es decir a partir del 82.65% de precisión

Gráfica 8. Markov. Datos en bruto



Gráfica 9. Markov. Datos en bruto

El modelo identifica que es un plagio cuando la diferencia de cosenos es mayor a 0.58 (registros 78 de 911 = 8.5% de los casos)

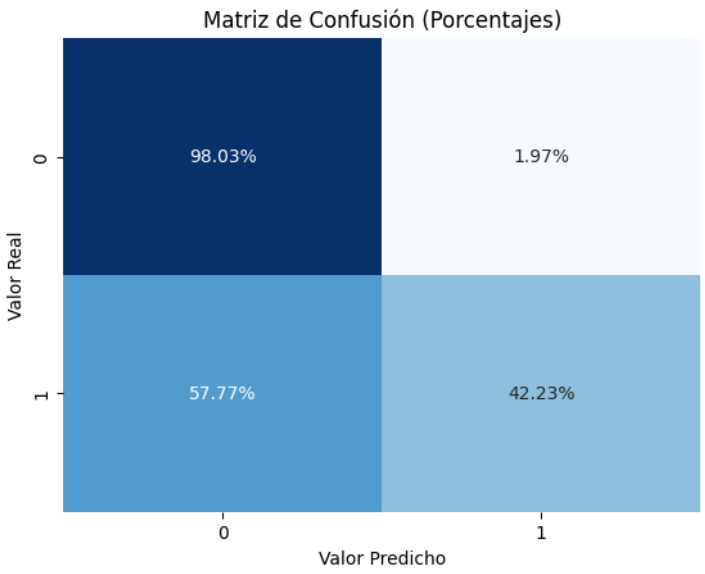


Gráfica 10. MARKOV. Datos en bruto

Datos limpios

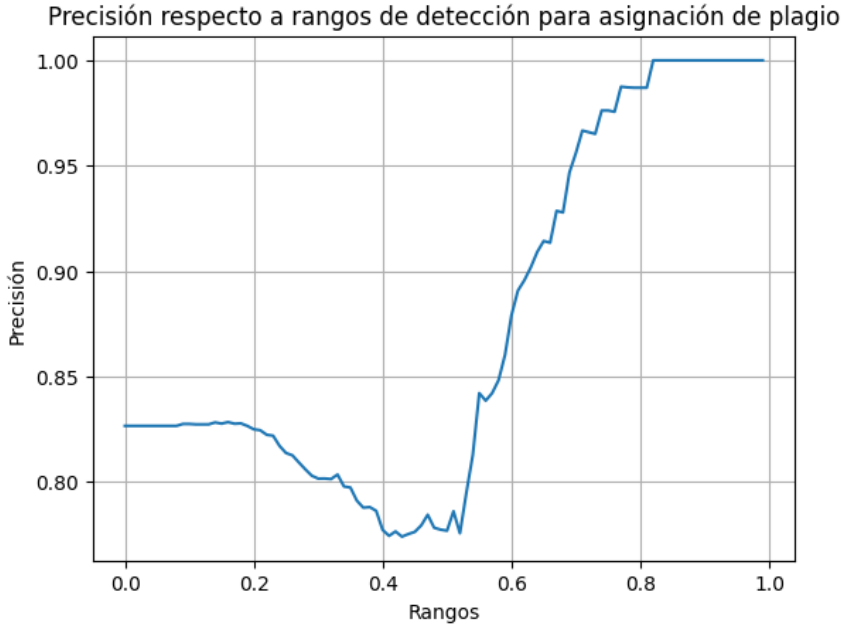
Se considera plagio a partir de 0.61, es decir a partir del 82.65% de precisión

Gráfica 11. MARKOV. Datos limpios



Gráfica 12. MARKOV. Datos limpios

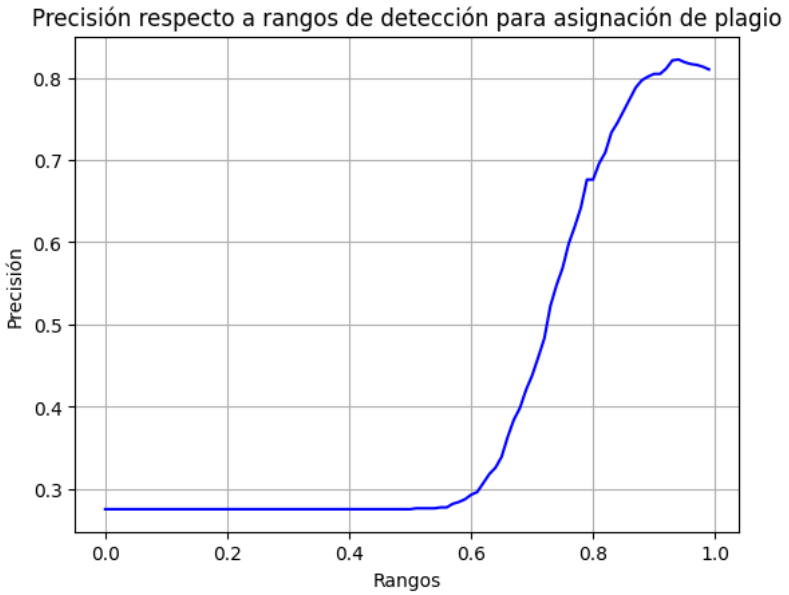
El modelo identifica que es un plagio cuando la diferencia de cosenos es mayor a 0.77 (registros 80 de 911 = 8.7% de los casos)



Gráfica 13. MARKOV. Datos limpios

Datos tokenizados

Se considera plagio a partir de 0.94, es decir a partir del 82.21% de precisión



Gráfica 14. MARKOV. Datos tokenizados

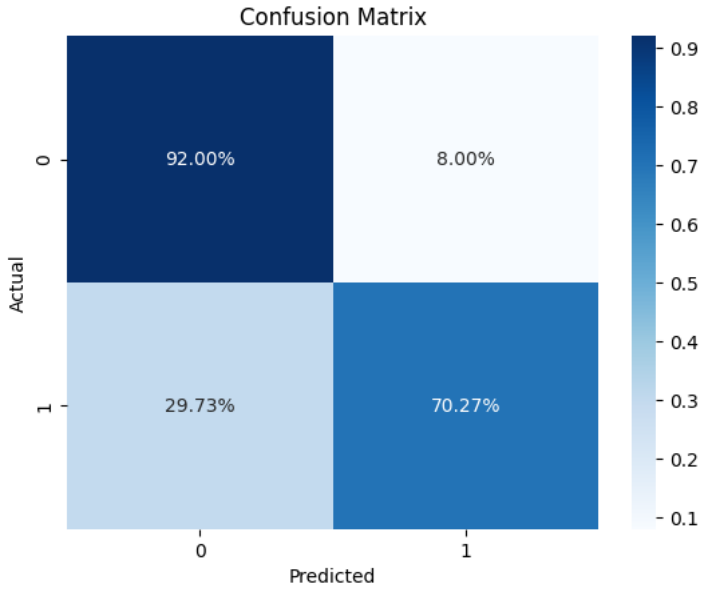
El modelo identifica que es un plagio cuando la diferencia de cosenos es mayor a 0.98 (registros 81 de 911 = 8.9% de los casos)

Gráfica 15. MARKOV. Datos tokenizados

Matrices de confusión XGBOOST

Datos en bruto

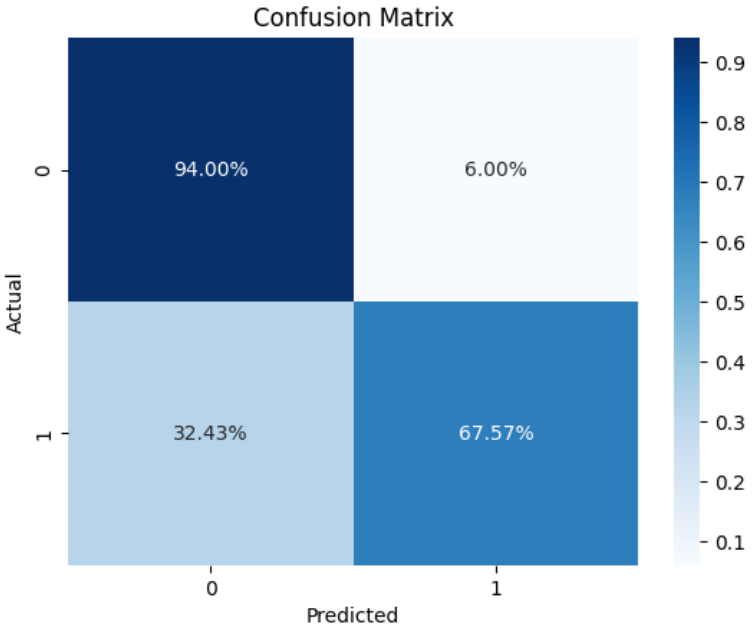
El modelo presenta 86.13% de acurracy



Gráfica 16. XGBOOST. Datos bruto

Datos limpios

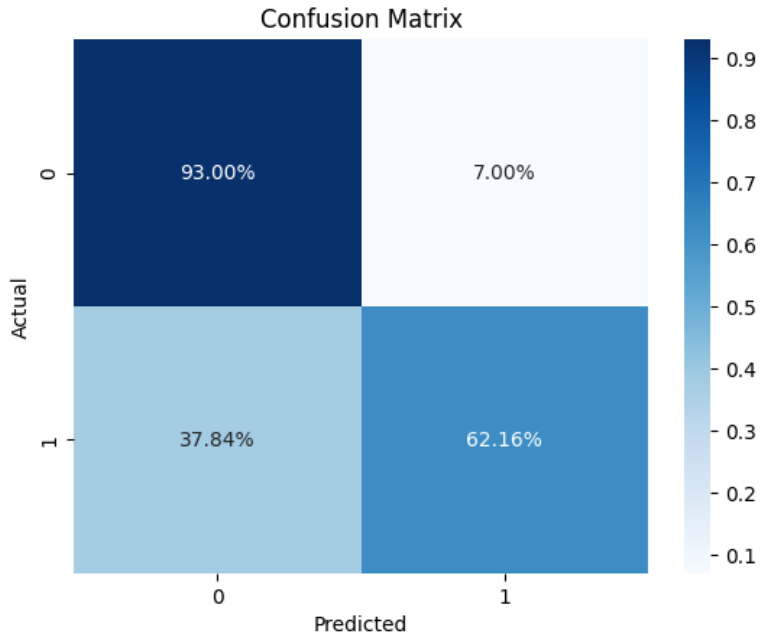
El modelo presenta 86.86% de acurracy



Gráfica 17. XGBOOST. Datos limpios

Datos tokenizados

El modelo presenta 84.67% de acurracy

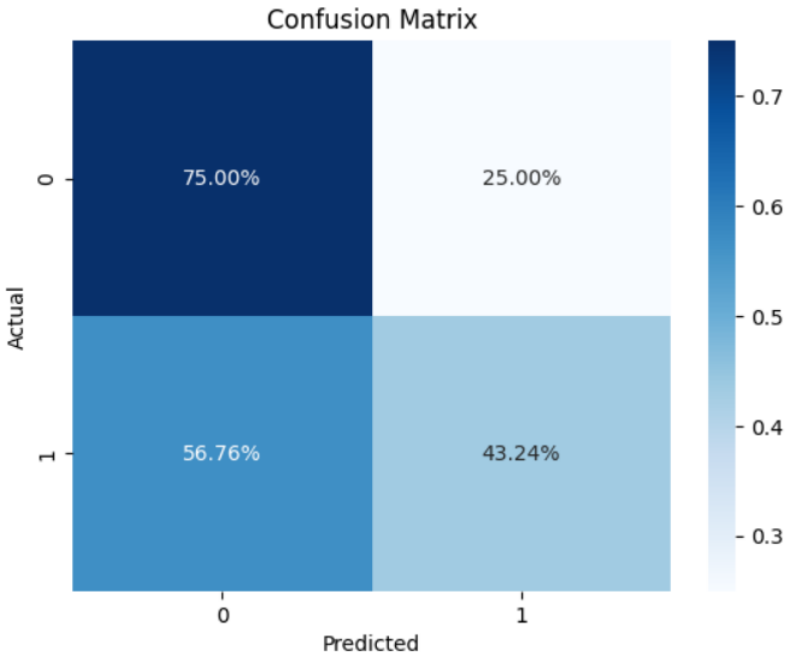


Gráfica 18. XGBOOST. Datos tokenizados

Neuronal Network

Datos en bruto

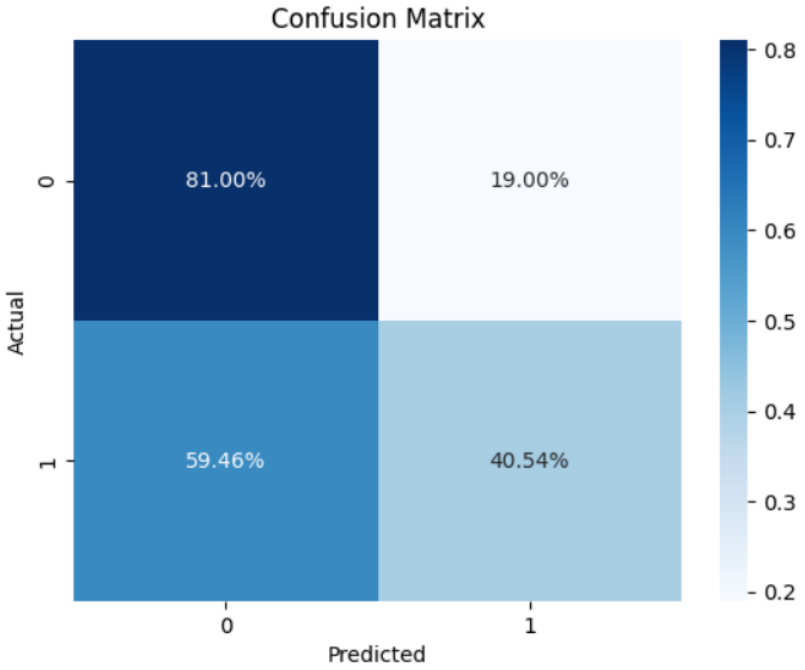
El modelo presenta 68.81% de acurracy



Gráfica 19. Neuronal network. Datos en bruto

Datos limpios

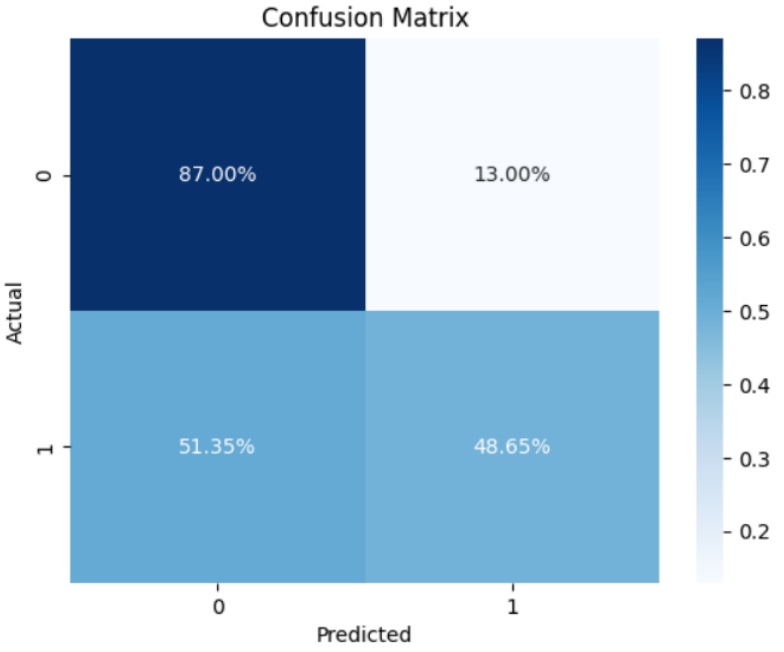
El modelo presenta 70.01% de acurracy



Gráfica 20. Neuronal network. Datos limpios

Datos tokenizados

El modelo presenta 70.01% de acurracy

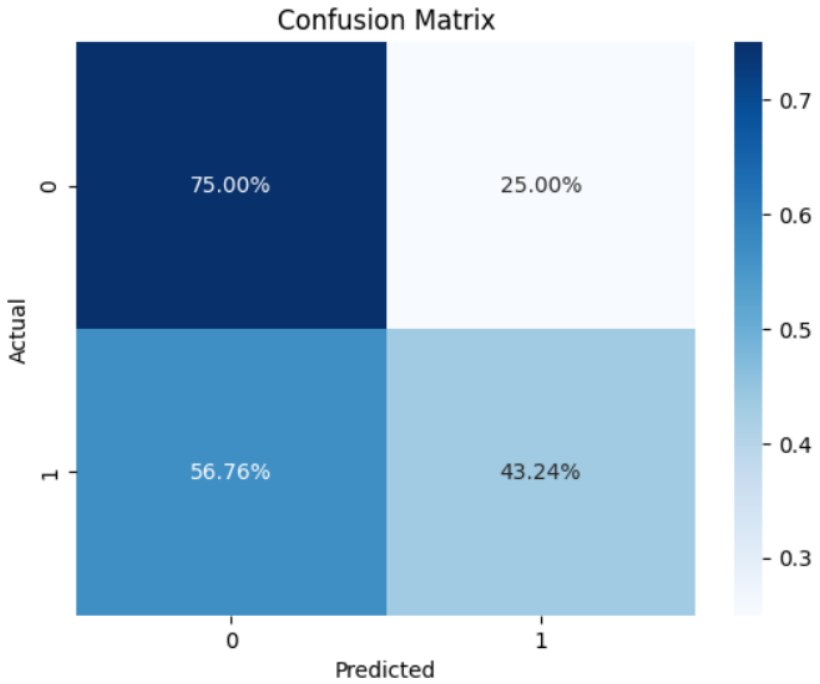


Gráfica 21. Neuronal network. Datos tokenizados

LSTM

Datos en bruto

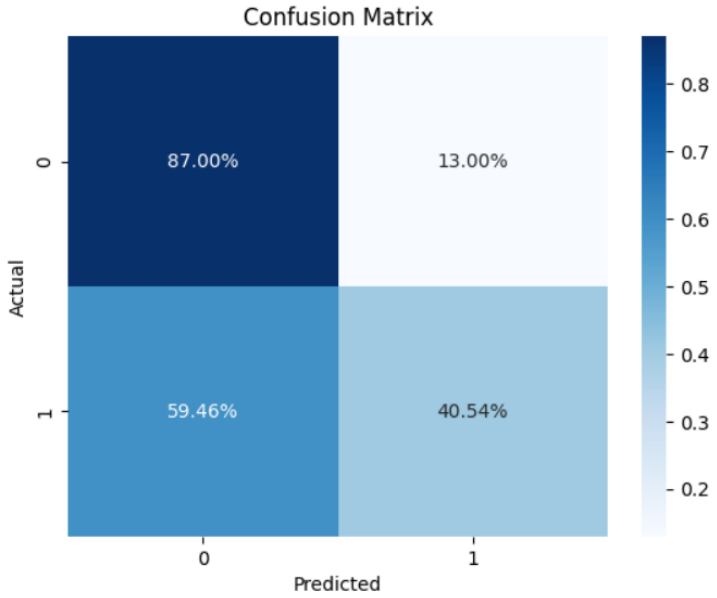
El modelo presenta 66.42% de acurracy



Gráfica 22. LSTM. Datos en bruto

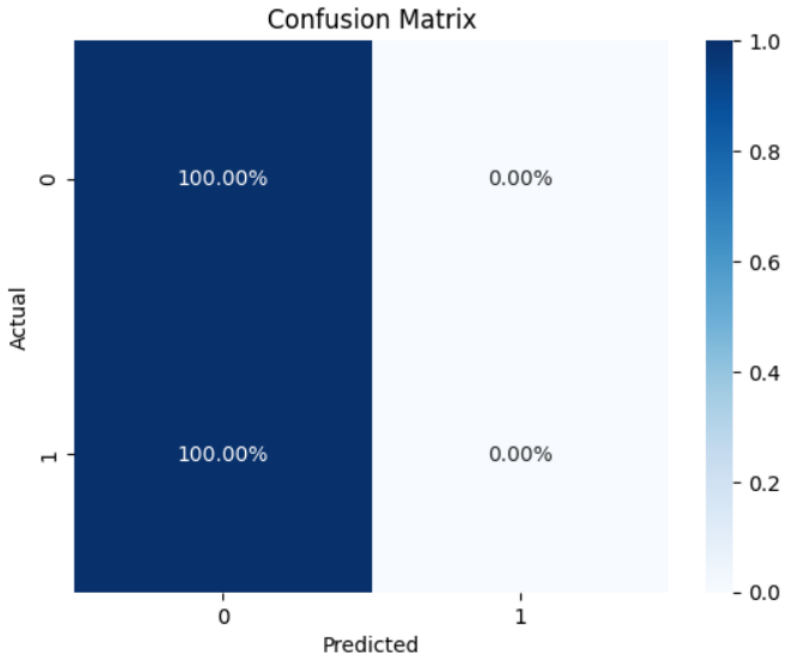
Datos limpios

El modelo presenta 74.45% de acurracy



Gráfica 23. LSTM. Datos limpios

Datos tokenizados



Gráfica 24. LSTM. Datos tokenizados