Proyecto 1
B.I
Grupo 25

Tabima, Moreno, Nunes

Tabla de **contenidos**

01 Presentación del equipo

Grupo 25

02

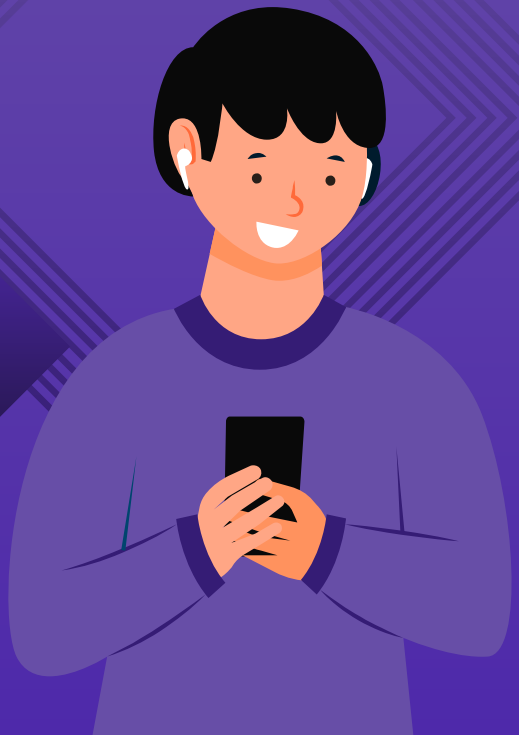
Presentación del problema

El problema:

Reddit es una de las redes sociales emergentes más usadas hoy en día, las personas aprovechan el anonimato de internet y el formato de foro para unirse en causas en común y mostrar apoyo.

En este caso manejaremos datos de un Reddit sobre intentos de suicidio, depresión y apoyo, en busca de detectar a tiempo publicaciones que puedan resultar en casos de suicidio. Lo anterior se logrará a través de generar modelos de machine learning que nos permitan detectar, en un post de reddit, si el usuaio podria ser considerado un caso de suicidio.

Oportunidad / Problema de Negocio	La oportunidad es poder prevenir a traves de prediccion de calidad el suicidio. El problema de negocio es que algunos usuarios de reddit se están suicidando y no se están detectando los casos de posible suicidio entre los usuarios de los Reddits con depresión a tiempo.	
Enfoque analítico	Se necesita analizar los textos para determinar adecuadamente cual es un relato de suicido que permita determinar si el relato es en realidad un post donde se pide ayuda por parte de la víctima y separarlo de un post en broma (shitpost)	
Organización y rol dentro de ella que se beneficia con la oportunidad definida	Se beneficia la comunidad de Reddit, en especial sus usuarios, que pueden recibir alertas y acompañamiento de parte de otras personas para ayudarlos en caso de necesitarlo, también ayuda a los moderadores del Thread sobre depresión para monitorear y tomar acciones ante sus usuarios	
Técnicas y algoritmos a utilizar	Se tiene planeado un árbol de decisión que permita determinar que casos un post de reddit sea un reddit determinante para un suicidio. También se planea usar KNN y Support Vector Machine, Baggin, AbaBoost, Bernoulli	Se van a procesar los datos tanto corrigiendo elementos de typos, asciis y números, como aplicando Stems y Lemmatización, para así poder generar tokens que puedan ser procesador por una vectorización y generar así un análisis completo del lenguaje.



03

Manejo de los datos

Eran muchos datos :(

Etapas de nuestro Procesamiento

Perfilamiento de los datos

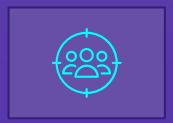
Pasamos los datos por Pandas Profilling report

Procesamiento Inicial

Removemos carácteres Ascii, lowercase, puntuación, remoción números

Normalización

Lemmatización, Steams y unificación de datos resultantes



Vectorización y declaración de variable Objetivo

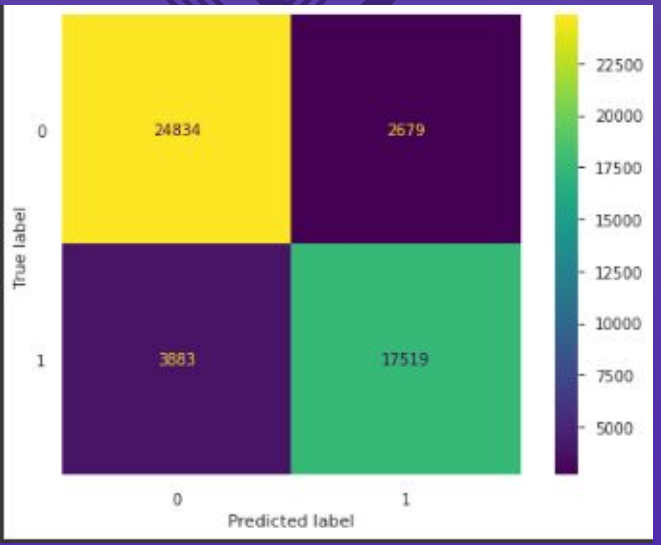
Se preparan los datos con el manejo de tokens Columnares para el manejo de los modelpos

	class	processed
0	suicide	want destroy myselff everyth start feel okay c...
1	non-suicide	kind get behind schedule learn next week testw...
2	suicide	im sur anymorefirst foremost im brazil judg me...
3	suicide	pleas giv reason liveth much dont reason liv l...
4	suicide	27f struggle find mean mov forward admit bite ...
...	...	...
195695	non-suicide	drop cool new cer idea lik would id cer
195696	non-suicide	unpopul opin cat deserv lov respect much dog k...
195697	non-suicide	hey guy yal doin
195698	non-suicide	uhm cov dog blanket light wont wak wok run wal
195699	suicide	_____god it end lif i tir couldnt want anyth mo...

04

Modelos, conclusiones

Árboles de decisión



- Matriz de elección del Modelo

Se aplicó Árboles de decisión teniendo como variable objetivo identificar la clase de un post en reddit, tomando como base las palabras claves con alegoría al suicidio (kill, pain, sadness, depression, etc) buscando así hallar cuales eran los posts que podían ser inferidos como alegorías al suicidio, lo encontrado es lo siguiente:

Este modelo es muy bueno eligiendo los positivos adecuados, teniendo una precisión del 86%, creemos que le aporta valor a análisis de posts
dado que es capaz de elegir entre un post a partir de indicadores
Claros del lenguaje natural que pueden ser de mucha utilidad
Para moderación o incluso ayuda por parte de admins
Del sub reddit de donde estamos analizando los datos.

A continuación las métricas del modelo.

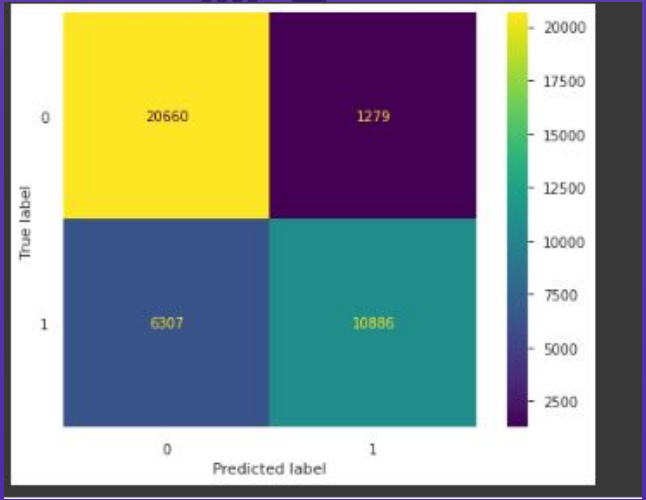
Métricas de árbol de decisión

	precision	recall	f1-score	support
0	0.86	0.90	0.88	27513
1	0.87	0.82	0.84	21402
accuracy			0.87	48915
macro avg	0.87	0.86	0.86	48915
weighted avg	0.87	0.87	0.87	48915

Random Forest

Se aplicó RF teniendo como variable objetivo identificar la clase de un post en reddit, porque queríamos saber si ser pueden desenglosar algunas palabras y volverlas un denotador congruente para un post sobre suicidio, depronto una combinación precisa de palabras puede ser un post de suicidio acertado y nosotros no tenerlo previsto.

Sin embargo este modelo resultó poco útil, dado que su precisión es menor a la requerida por nuestra selección de calidad, por esto decidimos dejarlo
De lado a la hora de escoger el mejor modelo.

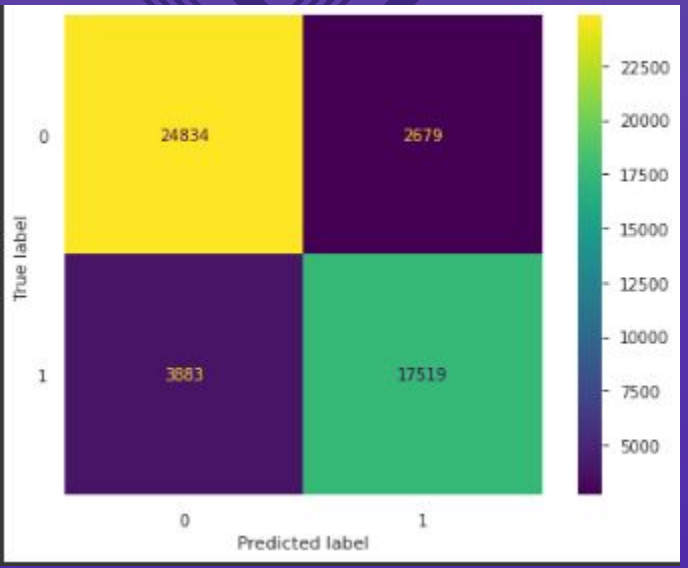


- Matriz de elección del Modelo

	precision	recall	f1-score	support
0	0.77	0.94	0.84	21939
1	0.89	0.63	0.74	17193
accuracy			0.81	39132
macro avg	0.83	0.79	0.79	39132
weighted avg	0.82	0.81	0.80	39132

K-nearest neighbors algorithm

Se aplicó el algoritmo KNN, el cual es muy útil para predecir valores discretos, en el caso del requerimiento de negocio, los nuevos comentarios que no son etiquetados,



- Matriz de elección del Modelo

Este modelo satisface los requerimientos del negocio y es capaz de predecir y etiquetar de manera correcta el 85 % de los datos.

Se realizó un modelo sin hiperparametros con f1 de 75 %

Se realizó un modelo con hiperparametros

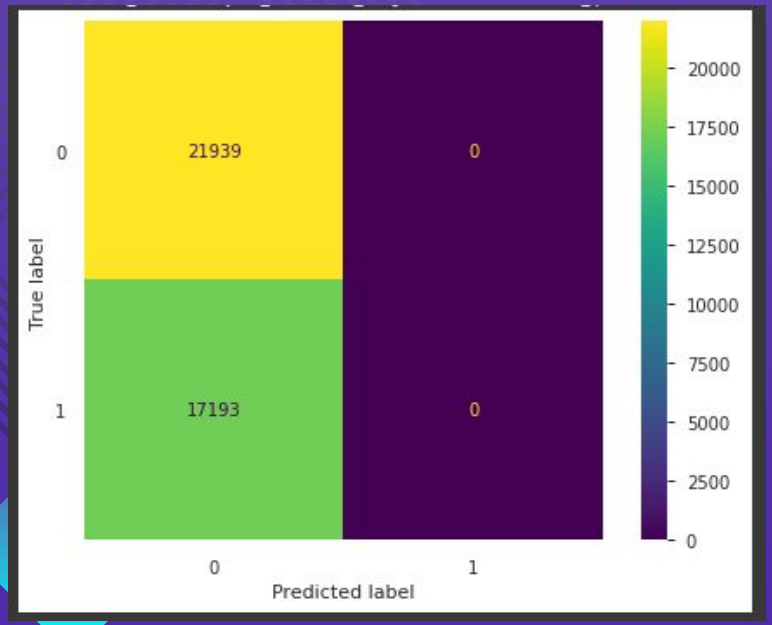
A continuación las métricas del modelo.

Metricas obtenidas con KNN

	precision	recall	f1-score	support
non-suicide	0.87	0.86	0.86	880
suicide	0.83	0.84	0.83	714
accuracy			0.85	1594
macro avg	0.85	0.85	0.85	1594
weighted avg	0.85	0.85	0.85	1594

Baggin Classifier

Para procesar los datos usamos también el modelo de Baggin Classifier, que es un metaclasificador enfocado en subsets aleatorios y genera predicciones, lo entrenamos para predecir diferentes casos de suicidio, para así presentar diferentes outcomes dependiendo de las palabras clave en los posts de Reddit.

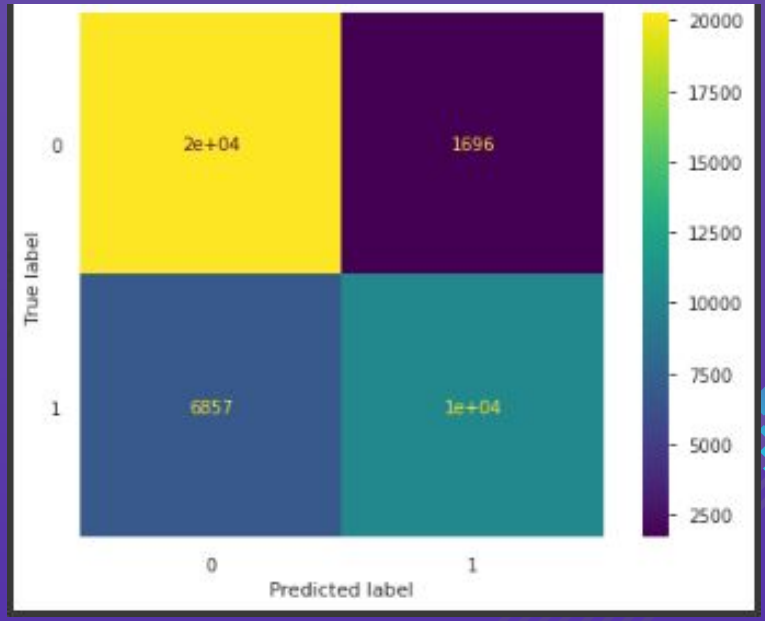


	precision	recall	f1-score	support
0	0.56	1.00	0.72	21939
1	0.00	0.00	0.00	17193
accuracy			0.56	39132
macro avg	0.28	0.50	0.36	39132
weighted avg	0.31	0.56	0.40	39132

Luego de procesar el modelo podemos ver que su precisión es de menos del 60% por lo que vamos a descartarlo a la hora de escoger un modelo, Bagging resulta apropiado para este contexto de análisis de texto, pero no en este caso donde los subsets pueden brindar información poco concluyente dado que se necesita la linealidad del texto para generar un procesamiendo eficiente y concluyente.

Bernoulli

Este modelo de machine learning es una distribución de probabilidad discreta, lo que significa que solo considera variables aleatorias discretas. Esto significa que no analiza variables continuas, esto es muy importante porque el modelo permite predecir y etiquetar probabilisticamente la posibilidad de que un usuario se suicide o no. El resultado es satisfactorio ya que se obtuvo cerca del 78% de accuracy.

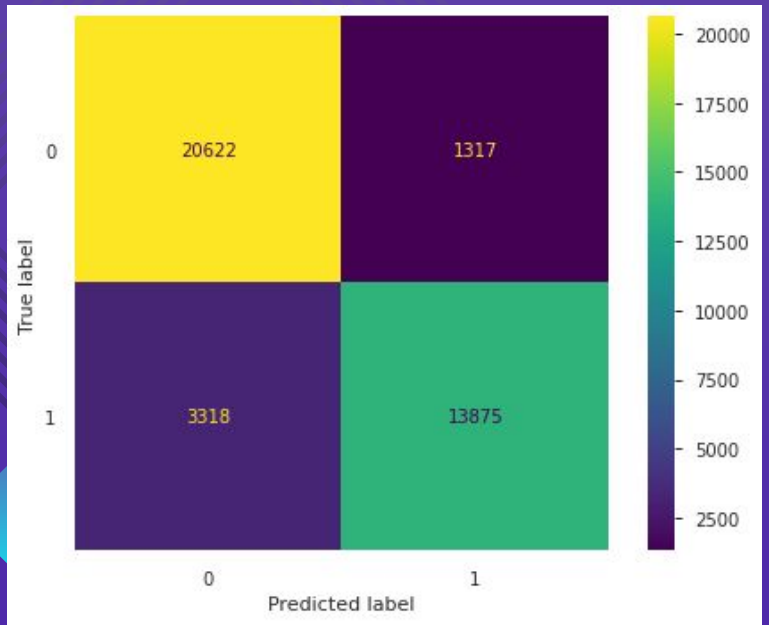


	precision	recall	f1-score	support
0	0.75	0.92	0.83	21939
1	0.86	0.60	0.71	17193
accuracy			0.78	39132
macro avg	0.80	0.76	0.77	39132
weighted avg	0.80	0.78	0.77	39132

Este modelo de machine learning pertenece al grupo de los modelos que utilizan deep learning probabilistico. La accuracy obtenida es del 78% y se uso en la variable objetivo las variables discretas suicido y non suicidio, teniendo resultados satisfactorios mas no optimos. Aunque este es un problemas con variables discretas encontramos que no tiene suficiente exactitud en este caso de negocio. Creemos que esto se puede deber a la interpretación necesaria para hacer PLN.

Adaboost:

Un modelo de machine learning que implemente AdaBoost es un metaestimador que ajusta inicialmente un clasificador en el conjunto de datos original, esto se usa para posteriomente volver mas eficientes otros estimadores y conseguir una sinergia entre modelos de machine learning.

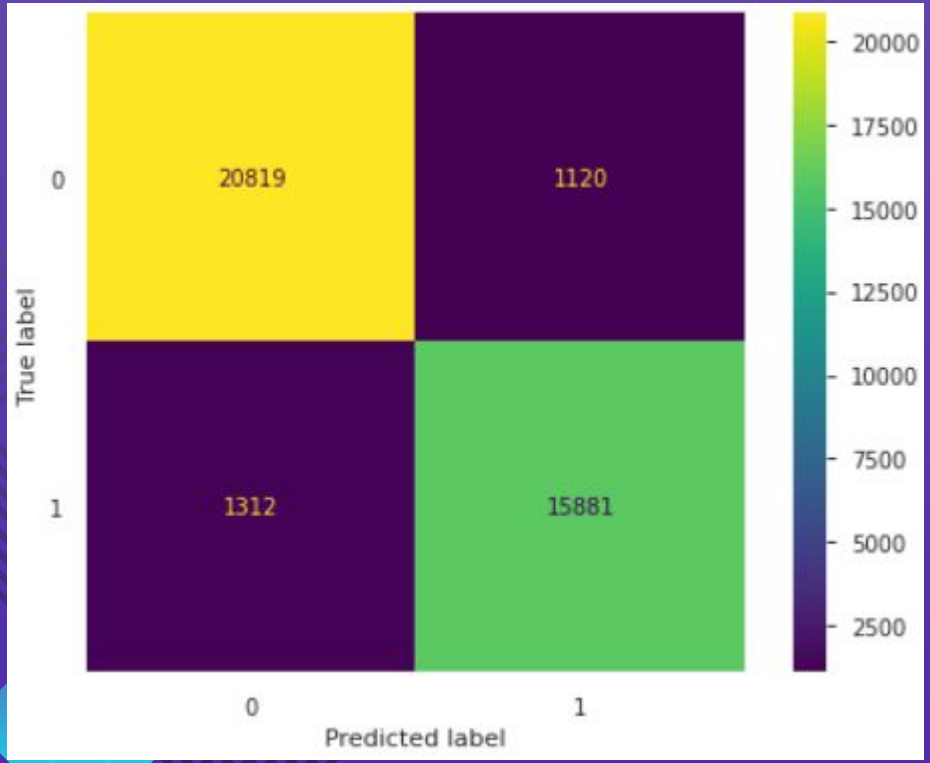


AdaBoost es un metaclasificador que corrige en los datos originales las ditancias entre las instancias con el fin de lograr modelos de machine learning mas optimizados. Nos arrojo una mejora considerable y un accuracy del 88 %. Podemos concluir que es un modelo de machine learning muy util para mejorar el rendimiento de modelos de clacificaion de aprendizaje no supervisado posteriores. La utilizacion de este metamodelo en casos donde el modelo se apoya en distancias como la distancia ecludiana causa una sinergia que mejora la predicción.

Indicadores adaboost

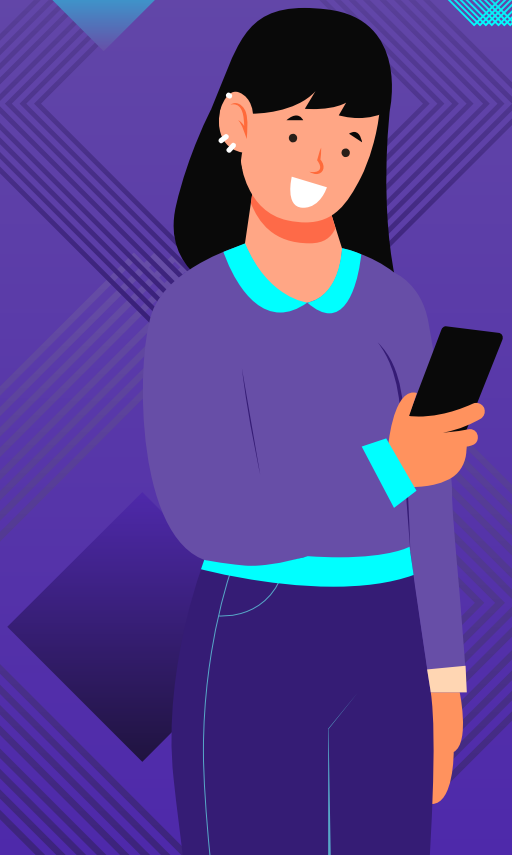
	precision	recall	f1-score	support
0	0.86	0.94	0.90	21939
1	0.91	0.81	0.86	17193
accuracy			0.88	39132
macro avg	0.89	0.87	0.88	39132
weighted avg	0.88	0.88	0.88	39132

Support Vector Machine



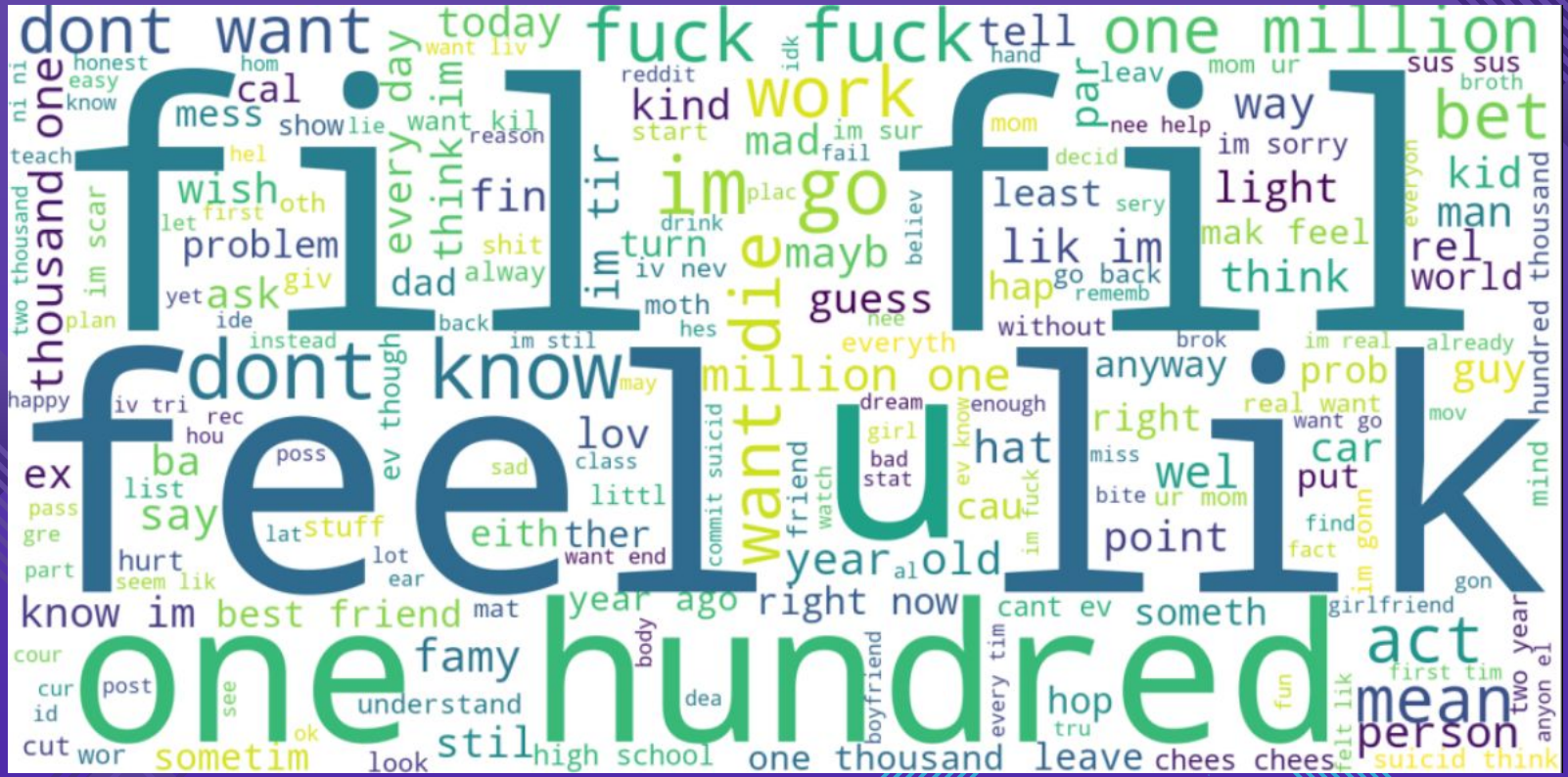
	precision	recall	f1-score	support
0	0.94	0.95	0.94	21939
1	0.93	0.92	0.93	17193
accuracy			0.94	39132
macro avg	0.94	0.94	0.94	39132
weighted avg	0.94	0.94	0.94	39132

Conclusiones finales

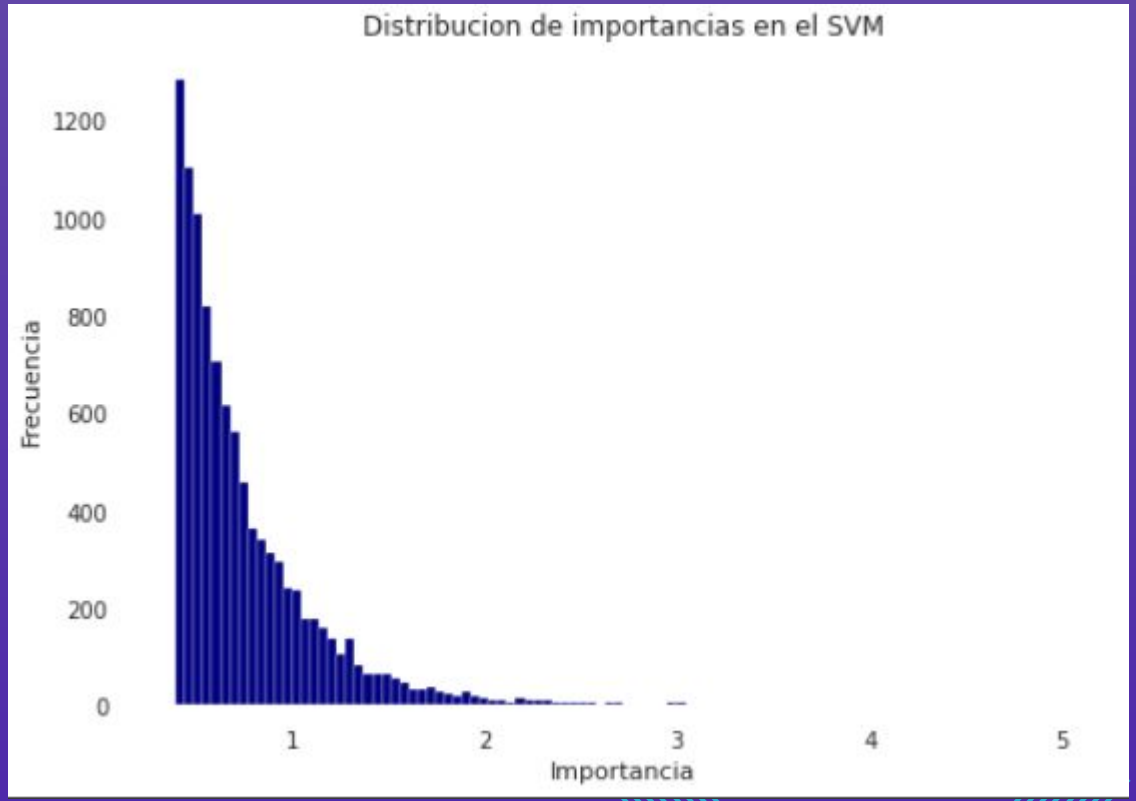


- El mejor modelo que encontramos para nuestro caso de negocio es Machine Vector Support.
- Los metaclasificadores mejoran notablemente el desempeño de los clasificadores usados posteriormente.
- Es de vital importancia el preprocesamiento de los datos para vectorizar y cuantificar los datos.

Importancia de tokens



Importancia de los tokens



¡Gracias!

¿Alguna duda?

(Por favor no que nos corchan)