Proyecto analítica de textos - El Culto de Dijkstra

- **Comprensión del negocio y enfoque analítico**.

Oportunidad/problema Negocio	Poder identificar usuarios que tengan tendencias suicidas por medio de sus comentarios en reddit para poder ofrecerles asistencia en salud mental	
Enfoque analítico (Descripción del requerimiento desde el punto de vista de aprendizaje de máquina)	Haciendo uso de algoritmos de clasificación y análisis de sentimientos utilizando tokens obtenidos a partir de comentarios (por medio de NLP) para poder clasificar a los usuarios	
Organización y rol dentro de ella que se beneficia con la oportunidad definida	- Empresa de seguros de vida para determinar la probabilidad de que un cliente tenga tendencias suicidas. - Centro de atención psicológica que puede brindarles apoyo a aquellas personas que puedan tener tendencias suicidas antes de algún accidente.	
Técnicas y algoritmos a utilizar	TF-IDF	Random forest RoBERTa SVM

- **Entendimiento y limpieza de los datos**

Para hacer un entendimiento y limpieza de los datos se utilizó principalmente la herramienta de Tableau Prep Builder. Esta herramienta se utilizó para remover todos los registros que contuvieran nulos, remover caracteres especiales, remover números y remover puntuación de la columna de texto. Como paso adicional se removieron los emojis, ya que se consideró que estos no aportaban un gran significado a la determinación de la columna objetivo.

Por otro lado, para realizar la tokenización requerida para aplicar los modelos de clasificación/ análisis de sentimientos, se utilizó TF-IDF. Se escogió este método de tokenización, puesto que este permite tener en cuenta la rareza de las palabras en cuenta a la hora de hacer una predicción. Adicionalmente, el método BoW requiere de una gran cantidad de memoria teniendo en cuenta la cantidad de datos que se está trabajando. Al aplicar este método se obtuvo un vocabulario compuesto por 120865 palabras.

- **Modelos aplicados**

1. **Random Forest (Gabriel Beltrán).**
   Teniendo en cuenta la tarea que se quiere desempeñar para este conjunto de datos, que es clasificación, uno de los modelos que mejores resultados da es el de Random Forest. Al aplicar este modelo con los tokens que se obtuvieron por TF-IDF

se obtuvieron valores de F1 de 99.99% en los datos de entrenamiento y 89.50% en los de prueba.

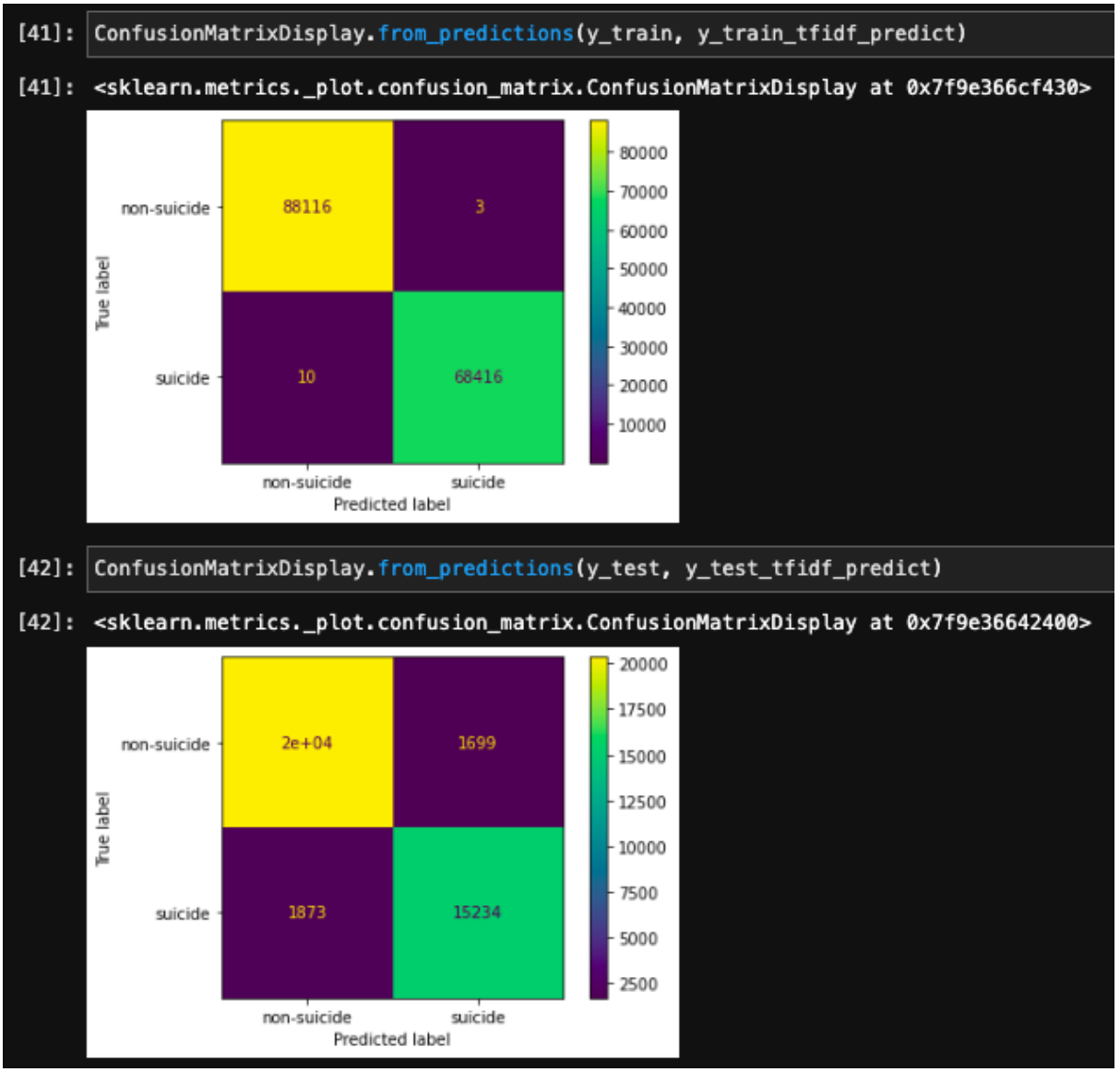


Imagen 1: Matrices de confusión modelo Random Forest

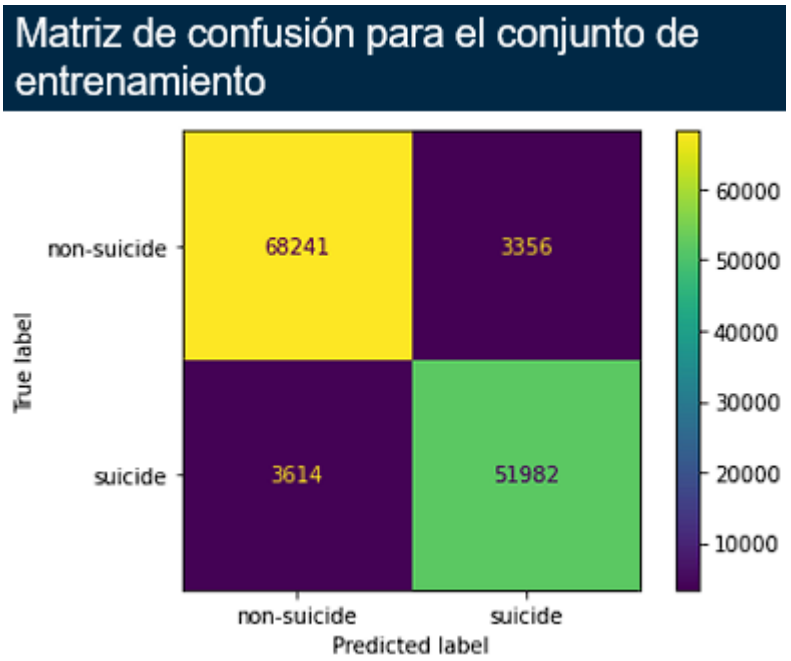
Adicionalmente, se puede ver que hubo un total de 3572 registros que fueron erróneamente clasificados y al hacer un análisis de estos se puede ver que los Falsos Positivos que el modelo clasificó se debe mayormente a que en los comentarios utilizaban palabras que pueden ser consideradas sensibles, pero el contexto en el que lo utilizaban es distinto.

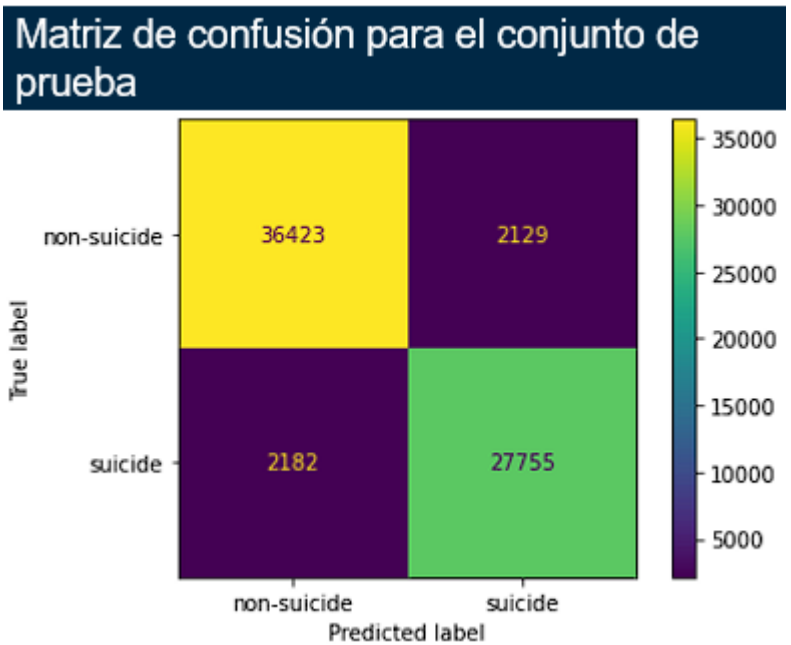
Para la búsqueda de los hiperparametros del modelo RandomForest se realizó un GridSearchCV con varios parametros. Sin embargo, al ver los resultados del modelo obtenido por el GridSearch se observó que eran menos precisos que los hiperparametros por defecto. Por lo tanto, se concluye que los mejores hiperparametros para este modelo son 100 estimadores con una profundidad media de 1790.

Finalmente, los resultados de este modelo serian de gran ayuda para los negocios propuestos anteriormente. Tomando los resultados de las métricas de F1 se podría

asegurar que el modelo puede ayudar a clasificar la mayoría de los casos en los que una persona tenga tendencias suicidas.

2. **Logistic Regression (Nicolás Fajardo).**
   Para el desarrollo de este modelo se parte de un proceso de tokenización con TF-IDF, para alimentar estos datos a un modelo de regresión logística.
   En general se obtuvieron resultados bastante buenos, un valor de f1 de alrededor de 93% para ambos conjuntos de datos.
   En cuanto a la matriz de confusión, se puede observar que el conjunto de entrenamiento presenta una cantidad considerable de falsos negativos, bastante mayor que el modelo de Random Forest, pero un porcentaje más bajo de estos en el conjunto de prueba, un 3%.





3. Análisis Sentimental con VADER y RoBERTa(Daniel Zambrano)

El enfoque que se usará en última instancia para procesar el texto estructurado que ya tenemos, es el análisis de sentimientos en las palabras o estructuras ya identificadas.

En primera instancia se hizo un análisis en base a una herramienta basada en léxico que usa diccionarios para clasificar las estructuras del texto como positivas, negativas o neutrales.

Se utilizará **VADER** *(Valence Aware Dictionary and sEntiment Reasoner)* que es es la herramienta de análisis de sentimientos basada en reglas y léxico escogida, que está específicamente entrenada con los sentimientos expresados en las redes sociales y funciona bien en textos de otros dominios. Esta herramienta nos permite trabajar con un enfoque en donde se usa una bolsa o diccionario gigante de palabras ya clasificadas como buenas, neutrales o malas para hacer la tarea de clasificar los tokens del texto estructurado en alguna de estas categorías y al final computar una fórmula para calcular los puntajes del texto en general en cada categoría.

En segunda instancia se hará un análisis de sentimientos en base a un modelo ya entrenado usando una tarea de *masked language modeling* (MLM) que es simplemente hacer que el modelo pueda predecir qué palabra debería llenar los espacios en blanco de una oración, recibe de entrada una máscara de texto como "la película estuvo una [MASCARA]", y retorna las posibles palabras que podrían llenar tal máscara.

El modelo a usar se llama **RoBERTa** (Robustly Optimized BERT Pre Training Approach) y es un modelo de transformadores pre-entrenado en un gran cuerpo de datos en inglés de manera auto-supervisada. **RoBERTa** Está destinado principalmente a ajustarse en una tarea específica de modelado de lenguaje, que es nuestro caso es de sentimientos.

Haremos que también clasifique estructuras del texto en positivas, negativas o neutras pero,con la particularidad de que en este modelo ya se consideran las relaciones entre palabras y frases, así como también el sarcasmo y más comportamientos exhibidos por las personas cuando escriben algún texto y, en principio se espera que el resultado de este análisis sea mas preciso y/o confiable que el de **VADER.**

Sección de trabajo en equipo:

Gabriel Beltrán - 201921903 - Líder de datos - Líder de proyecto
Daniel Zambrano - 201914912 - Líder de Analítica
Nicolás Fajardo - 201914912 - Líder de Negocios

- **Reunión de lanzamiento y planeación:**
  - Se define el tema a trabajar
  - Se recluta un miembro más para el equipo
  - Se definen roles que cada uno iba a desempeñar

- **Reunión de ideación:**
  - Se idean negocios que se benefician con la solución al problema
  - Se establece la tarea que se iba a desempeñar
  - Se definen los algoritmos que cada integrante va a desarrollar

- **Reuniones de seguimiento:**
    - Se realizan principalmente por chat para revisar el avance que va teniendo cada uno con respecto a las tareas que tiene asignadas
- **Reunión de finalización:**
    - Se discuten los resultados
    - Se consolidan los entregables en un repositorio
    - Se finalizan los entregables