# *Analítica de textos en la salud mental*

Jesús David Barrios 201921887
Sergio Peñuela 201922873
Jhoan Diaz 201819861

# *Contexto*

- A partir de textos obtenidos de personas sufriendo de problemas de salud mental se quiere aplicar la analítica de textos para obtener modelos de machine learning que ayuden a detectar,a partir del mensaje escrito por la persona, si se va a llevar a cabo un intento de suicidio o no.

# *Descarga y entendimiento de los datos*

```
Número de filas:  195700
Número de columnas:  3
```

| | Unnamed: 0 | text | class |
|---|---|---|---|
| 21738 | 135566 | AAAA I'm literally the stupidest person in exi... | non-suicide |
| 26627 | 17383 | Is it just me or is the sound of rain One of t... | non-suicide |
| 135198 | 36707 | How close i amam tired of living, tired of bei... | suicide |
| 140848 | 186151 | Guess who kissed a girl? Not me.\n\n&#x200... | non-suicide |
| 79670 | 255320 | Should I delete my Reddit account? I mean shou... | non-suicide |

- En los datos encontrados hay 3 columnas. De estas 3 las columnas "text" y "class" son las que nos proporcionan la información necesaria para llevar a cabo el modelo.

# *Dimensiones de calidad*

- Completitud: no se encontraron valores nulos en los datos
- Unicidad: no hubo textos repetidos en los datos
- Consistencia y validez: la consistencia hace referencia a la integridad de datos entre fuentes y observaciones. Por su parte, la validez mide si los datos hacen sentido para el contexto específico. En este sentido, se considera que en términos generales se cumple con estas métricas.

# *Preparación de los datos*

- Se eliminó la columna "unnamed" pues no aportaba valor alguno a lo que se buscaba en el modelo.
- Se procedió a convertir los valores de la columna "text" a string para sus posteriores modificaciones.
- En la columan "class" se convirtió los valores a numéricos siendo 0 'non-suicide' y 1 'suicide'.

```
non-suicide    110165
suicide         85535
Name: class, dtype: int64
0    110165
1     85535
Name: class, dtype: int64
```

# *Procesamiento del texto*

| | text | class | tokens |
|---|---|---|---|
| 60891 | going to buy my first console ps what are some... | 0 | [going, buy, first, consol, gam, must, play, m... |
| 78088 | make it stop pleasei cannot do this anymore i ... | 1 | [mak, stop, please, anym, much, pain, everyday... |
| 169268 | i just wanted to share one of the embarrassing... | 0 | [want, shar, on, embarrassingawkward, tim, lif... |
| 159499 | can i roast your country if willing to take so... | 0 | [roast, country, wil, tak, crit, pleas, nam, c... |
| 82956 | when the imposter when the impogster sus jdksj... | 0 | [impost, impogst, sus, jdksjdjdhekehdidkdjj, s... |

- Todo carácter que no perteneciera al alfabeto inglés fue eliminado (e.g. @, #, %, !, etc.).
- Se pasaron todas las palabras a letras minúsculas.
- Se removieron las palabras que no contuvieran ninguna vocal o la letra "y", pues estas se consideraron como errores de escritura.
- Se removieron los "stop-words" pues estas palabras son utilizadas para la sintaxis del texto, pero no aportan significado a la frase que nos pueda dar indicios sobre el tipo de texto.
- Se eliminaron todos los tokens con una longitud menor o igual a 2
- Se cortaron las raíces de las palabras
- Se lematizaron los verbos

# *Modelamiento, validación y visualización*

Se hizo uso de los siguientes tres modelos:

- KNN
- Árbol de decisión
- Random Forest

# *Árbol de decisión*

- Se decidió implementar un modelo con árbol de decisión pues requiere de menor preprocesamiento de datos para obtener buenos resultados y tiene buenos tiempos de ejecución.
- Para poder utilizar el modelo óptimo se hizo uso de gridsearch para poder obtener los mejores hiperparámetros y utilizarlos en el modelo. Los mejores hiperparámetros obtenidos fueron los siguientes:

```
{'criterion': 'entropy', 'max_depth': 10, 'min_samples_split': 2}
```

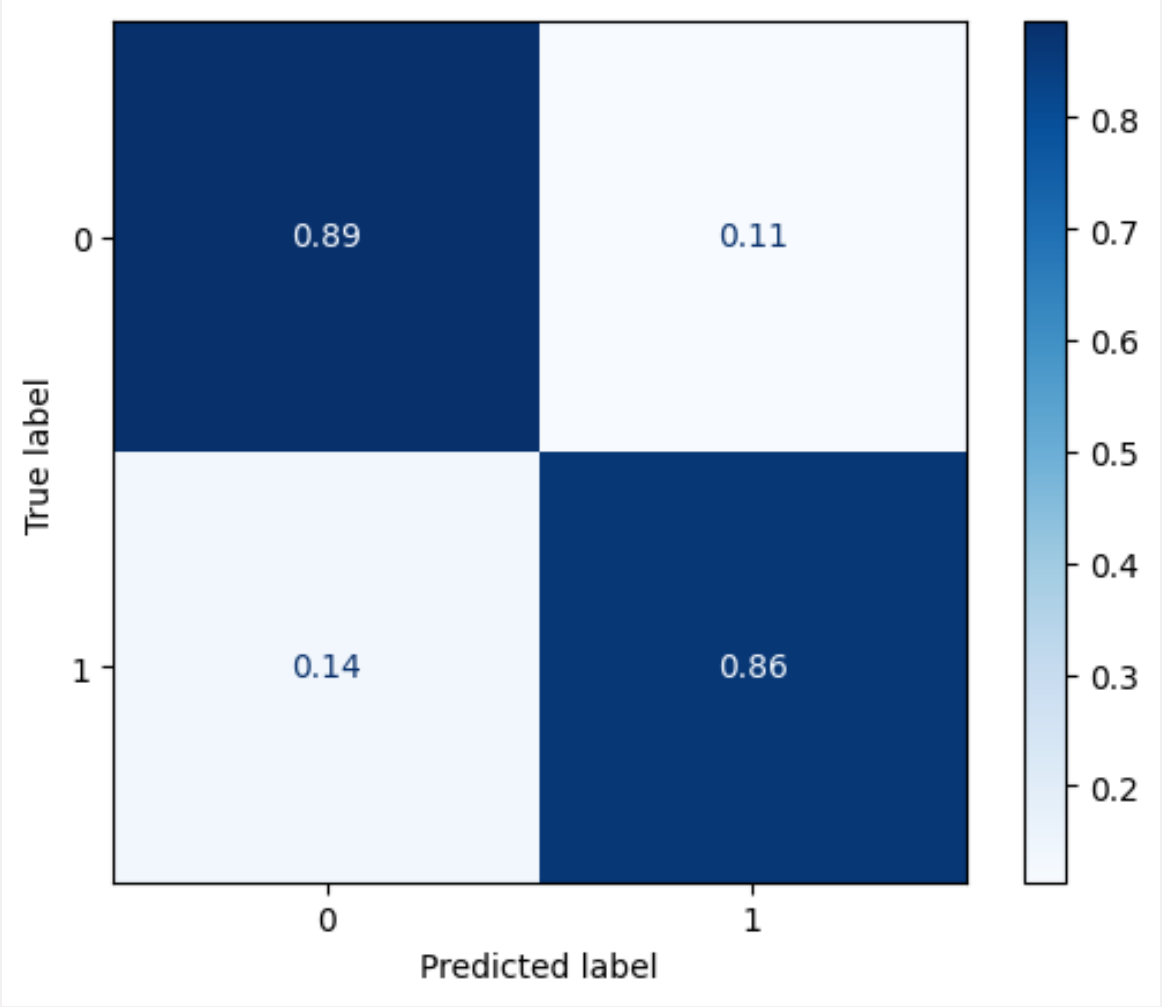
# *Resultados árbol de decisión*

- Como se puede ver se obtuvo un f1-score de 0.85 lo cual indica un buen nivel de precisión y de recall. Este modelo podría ser tenido en cuenta para identificar los intentos de suicidio a partir de textos.

```
Resultados:
Exactitud: 0.86
Recall: 0.7865829737151824
Precisión: 0.876541050974906
Puntuación F1: 0.8291291042924489
              precision    recall  f1-score   support

           0       0.85      0.91      0.88     16597
           1       0.88      0.79      0.83     12745

    accuracy                           0.86     29342
   macro avg       0.86      0.85      0.85     29342
weighted avg       0.86      0.86      0.86     29342
```

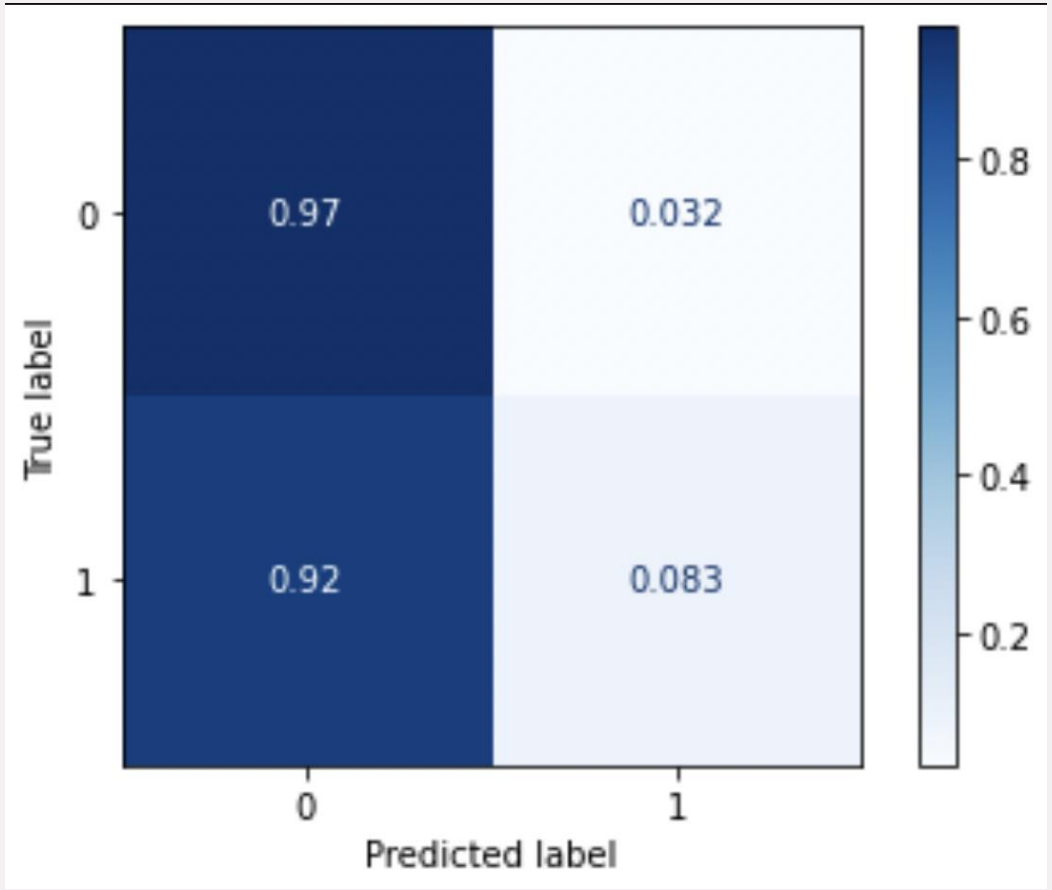
# *Random Forest*

```
Accuracy:  0.876869200685054
F1:  0.8769471295397717
Precision:  0.8770717948176725
Recall:  0.876869200685054
              precision    recall  f1-score   support

           0       0.89      0.89      0.89     22035
           1       0.85      0.86      0.86     17086

    accuracy                           0.88     39121
   macro avg       0.87      0.88      0.88     39121
weighted avg       0.88      0.88      0.88     39121
```



# *KNN*

```
Accuracy:  0.5816569106106695
F1:  0.47165582706008274
Precision:  0.6173809151323735
Recall:  0.5816569106106695
              precision    recall  f1-score   support

           0       0.58      0.97      0.72     22035
           1       0.67      0.08      0.15     17086

    accuracy                           0.58     39121
   macro avg       0.62      0.53      0.44     39121
weighted avg       0.62      0.58      0.47     39121
```



# *Conclusiones*

- A partir de los resultados obtenidos en los 3 modelos que se implementaron se recomienda utilizar el modelo de Random Forest pues es el que tiene un f1-score más alto por lo que es bueno identificando casos de suicido como casos de no suicidio, el árbol de decisión también se podría utilizar, pero tuvo un rendimiento menor de 3 puntos porcentuales. Además de eso se puede utilizar Random Forest para poder saber cuáles son las palabras que más influyen en la clasificación de suicidio. Por lo que este sería el modelo ideal para empresas que buscan identificar personas que puedan cometer un suicidio y evitar que lo hagan.