

**Proyecto Integrador**  
**Maestría Ciencia de los Datos**  
**Detección de síntomas de depresión o tendencias suicidas en Twitter**

**Por:**

Jairo Andrés Ruiz Machado - jruizma2@eafit.edu.co

David Alzate Cardona - dalzatec1@eafit.edu.co

Laura Alzate Madrid - lalzatem@eafit.edu.co

**Universidad EAFIT**

**Medellín**

**2022**

## Contenido

<b>Introducción.....</b>	<b>3</b>
<b>Problema del negocio .....</b>	<b>4</b>
<b>Impacto de la solución .....</b>	<b>5</b>
<b>Análisis Exploratorio de Datos.....</b>	<b>6</b>
<b>Preparación de los datos.....</b>	<b>8</b>
<b>Análisis descriptivo .....</b>	<b>9</b>
<b>Selección de modelos .....</b>	<b>10</b>
<b>Análisis y Conclusiones de los modelos .....</b>	<b>15</b>
<b>Tecnología .....</b>	<b>16</b>
<b>Conclusiones .....</b>	<b>17</b>
<b>Referencias.....</b>	<b>18</b>

## **Introducción**

Los trastornos mentales son una problemática que ha existido en el mundo desde siempre, sin embargo, estos trastornos mentales se han visibilizado mucho más a lo largo del tiempo. Las personas que sufren de algún trastorno, en especial las que sufren de depresión, han encontrado una forma de desahogarse de manera pública o anónima en las redes sociales. De modo que hoy en día se puede encontrar una cantidad grande de publicaciones de personas con pensamientos depresivos. Se conoce también, que una de las redes sociales que más publicaciones depresivas tiene es Twitter, dado a que es una red social muy libre. Estas publicaciones han sido de gran ayuda para todas estas personas porque escribir es una forma de despejar su mente, pero lo ideal es que las personas empiecen los procesos de terapia que se brindan en el mundo.

## **Problema del negocio**

Las personas que sufren de depresión suelen esconder sus sentimientos de sus amigos y familia, y para poder liberarse han encontrado la manera de publicar sus sentimientos en redes sociales. Muchas de estas personas que sufren de depresión no asisten a terapia, y pasan desapercibidos frente a los demás. De modo que para efectos de la práctica y para prevenir un mayor índice de depresión se propone generar un modelo de análisis de textos, utilizando aprendizaje supervisado y no supervisado para hacer una clasificación sobre el texto de las publicaciones en redes sociales, en este caso en Twitter, con el fin de calcular la probabilidad de que una persona tenga depresión o incluso la probabilidad de que una persona pueda llegar a tener tendencias suicidas.

## **Impacto de la solución**

La solución por implementarse será un programa que sirva de frente al modelo de clasificación. Este programa recibirá como entrada el enlace del perfil de un usuario público de Twitter. Al recibir esto obtendrá mediante el api de Twitter los tweets y demás información del perfil y clasificará todos los tweets de esta persona.

Como salida el programa mostrara información de cuantos tweets leyó, la cantidad de tweets que clasifico como potencialmente indicadores de síntomas depresivos y en base a esta información y otros indicadores estadísticos generara una probabilidad de que esta persona sufra de depresión

## **Análisis Exploratorio de Datos**

### ***Data set: Depression: Twitter Dataset + Feature Extraction***

***Descripción:*** Este data set contiene datos ya extraídos de la API de Twitter, son tweets sin procesar y su contenido solo está en inglés. Esta tiene como contenido el índice del tweet, id post, fecha de creación del post, texto del tweet, id del usuario, número de seguidores, número de seguidos, me gustas de la publicación, cantidad de tweet que las personas han mandado en total y número de retweets. En este mismo orden, las variables serían:

#### ***Variables***

- index
- Post\_id
- Post\_created
- Post\_text
- User\_id
- followers
- friends
- favourites
- Statuses
- Retweet

#### ***Anonimización***

SI \_\_\_\_ | NO X

***¿Esta base de datos es pública?***

SI X | NO \_\_\_\_

### ***Twitter API:***

***Descripción:*** La API de Twitter nos proporciona los tweets necesarios. Para efectos de la práctica y del entrenamiento se usaron 20.000 datos.

***Variables:*** las variables que se tienen en cuenta a la hora de realizar el análisis son:

- Post\_text
- retweet

### ***Anonimización***

SI\_\_ | NO\_X

***¿Esta base de datos es pública?***

SI\_X | NO\_\_

## **Preparación de los datos**

Para la preparación de datos trabajamos con S3 de AWS, en este caso realizamos la ingesta de datos en nuestro S3 en archivo tipo parquet, para luego hacer una creación de una tabla en Glue con un Cluster. Posteriormente se eliminan los datos nulos y en este punto realizamos un mapeo de todas las columnas y cambiamos el tipo de datos de algunas columnas.

Luego, pasamos a la segunda capa del S3 y se realiza la eliminación de los outliers con ayuda de Spark (Proceso que se realiza en un notebook). Y con esto pasamos a la última capa del S3.



## **Análisis descriptivo**

Dentro del análisis exploratorio encontramos que los tweets con un gran número de Re-tweets, no son depresivos, ya que generalmente estos son usados para noticias, cosas de entretenimiento o hilos sobre cualquier tópico, mientras que los tweets que realmente muestran signos de depresión tienen pocos Re-tweets.

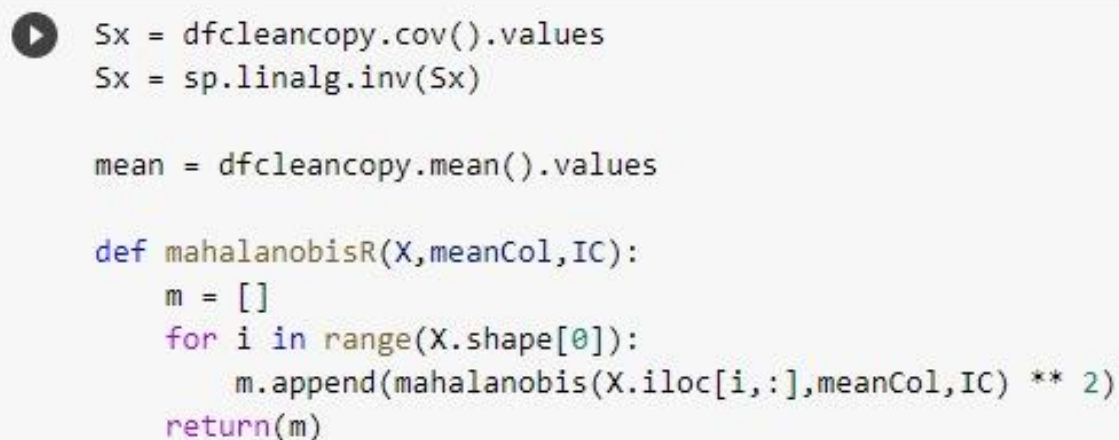
También encontramos relación en la hora donde se presentan más tweets depresivos, los cuales se hacen en las horas de la noche-madrugada, teniendo el pico a las 3 AM.

## Selección de modelos

Luego de un análisis profundo se vio la oportunidad de trabajar tanto con Modelos Supervisados como con Modelos no Supervisados.

Como generalidad para los modelos, utilizamos TF-IDF como reducción de dimensionalidad para poder entrenar los modelos.

Con el fin de mejorar los datos de entrenamiento, usamos la columna de Re-tweets para poder detectar los valores atípicos o outliers dado a que los Tweets con muchos Re-tweets rara vez son un problema de depresión personal. Para esto se hizo uso de la distancia de Mahalanobis respecto a la columna de Re-Tweets. (Ver *Imagen 1*)

A screenshot of a code editor showing Python code for calculating Mahalanobis distance. The code includes a play button icon at the start. It defines a function 'mahalanobisR' that takes a matrix 'X', a mean column 'meanCol', and a covariance matrix 'IC' as inputs. The function calculates the Mahalanobis distance for each row of 'X' and returns a list of these distances.

```
▶ Sx = dfcleancopy.cov().values
  Sx = sp.linalg.inv(Sx)

  mean = dfcleancopy.mean().values

  def mahalanobisR(X,meanCol,IC):
      m = []
      for i in range(X.shape[0]):
          m.append(mahalanobis(X.iloc[i,:],meanCol,IC) ** 2)
      return(m)
```

*Imagen 1*

Lo que se realizó para detectar outliers fue calcular el percentil 85, lo que significa que los valores que estén por encima de este percentil pertenecen a los outliers. Esto se puede ver representado en la *Imagen 2*.

```

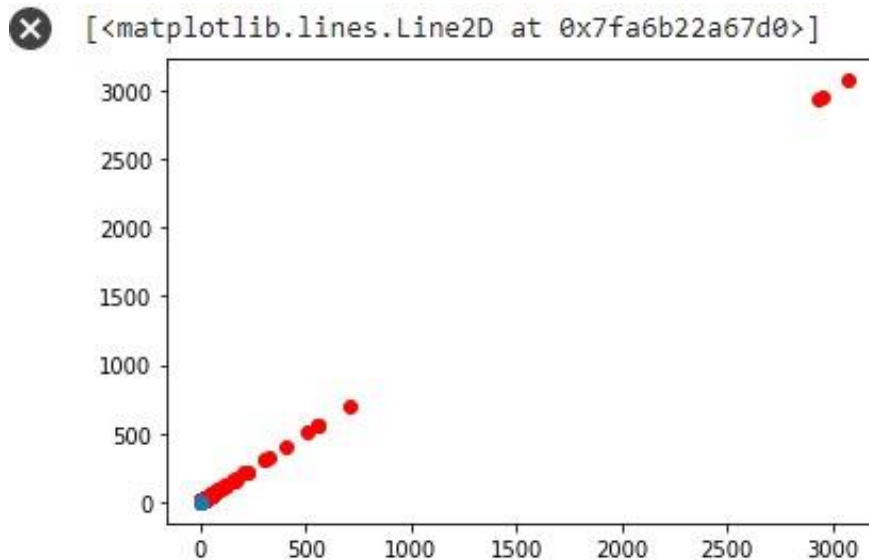
▶ mR = mahalanobisR(dfcleancopy,mean,Sx)
#percentil 85
per= np.percentile(mR,85)
print(per)
outliners=[]
for x in mR:
    if x>per:
        outliners.append(x)

plt.plot(mR,mR, 'o')
plt.plot(outliners,outliners, 'or')

plt.show()

```

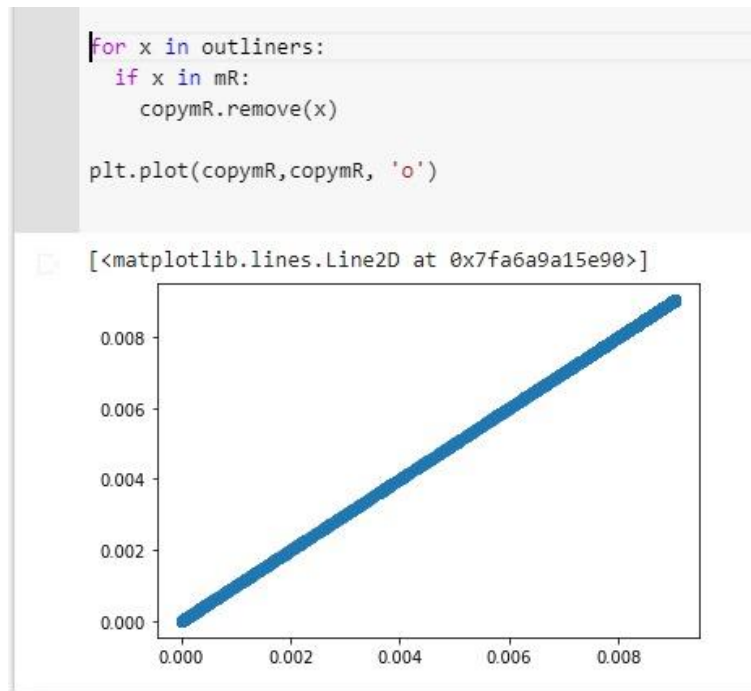
0.00904460310606404



*Imagen 2*

Como podemos observar, el 0.009... sería nuestro percentil 85, los valores en azul son los normales y los valores rojos son los atípicos.

En esta gráfica se ven en su gran mayoría puntos rojos debido a la escala, ya que los datos normales están por el rango del 0.009, mientras que estos valores “raros” llegan a 3000 de distancia respecto a la media. Al removerlos tenemos esto:



*Imagen 3*

## Modelos supervisados

- **Naive-Bayes**

**Entrenamiento:**

Para entrenar el modelo separamos los datos en 70 de prueba y 30 de testeo.

**Evaluación:**

Para evaluar el modelo usamos F1 score, este puntaje fue: 0.79

- **Random Forest:**

**Entrenamiento:**

Para entrenar el modelo separamos los datos en 70 de prueba y 30 de testeo.

**Evaluación:**

Para evaluar el modelo usamos F1 score, este puntaje fue: 0.53

## Modelos No Supervisados

- **LDA**

**Entrenamiento:**

Para el entrenamiento del modelo se utilizó la totalidad del dataset

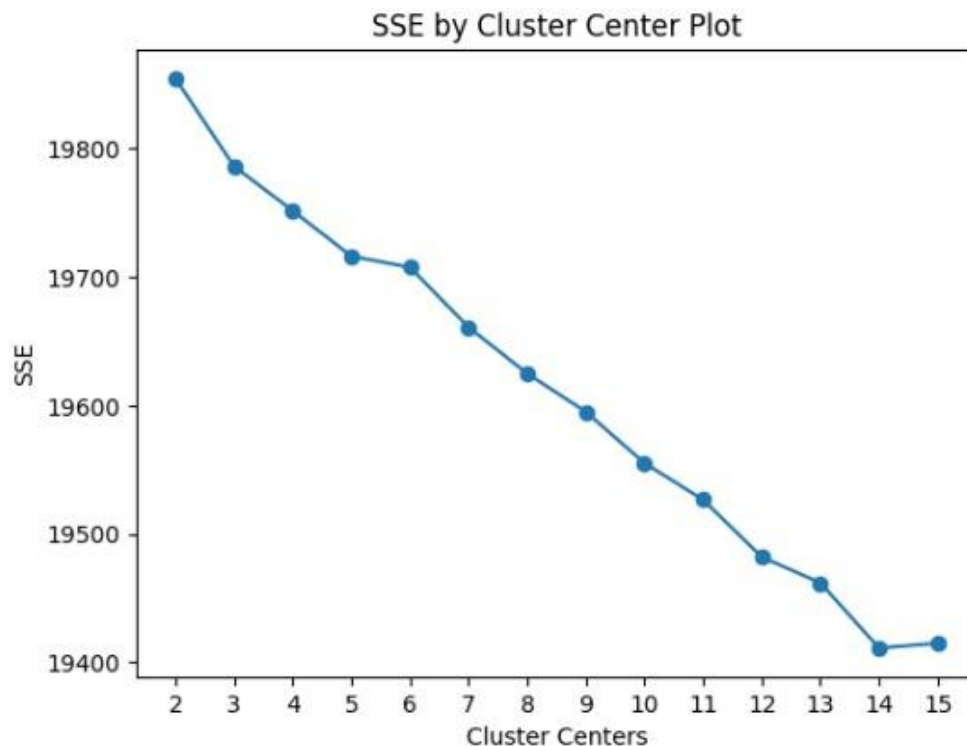
**Evaluación:**

La evaluación de este método se utilizó una métrica de coherencia para determinar el k tópicos optimo, el cual se determinó k=3

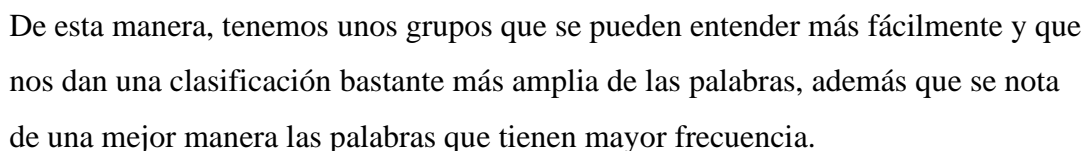
- **K-mean**

Como modelo no supervisado usamos k-means, esto con el fin de agrupar los tweets que tengan un tópico en común, lo primero que realizaremos es encontrar el k optimo.

Características e Ingeniería de Características



En un principio usamos  $k=6$  ya que vemos una gran caída entre 5 clústers y 6, pero este resultado nos daba grupos que no tenían mucho sentido ya que nos entregaban estas palabras:



## **Análisis y Conclusiones de los modelos**

K-means es un buen modelo y opción para clasificar los tweets ya que además de saber si tiene depresión una persona, los puede agrupar en otra categoría, aun así, como este no era nuestro foco principal, optamos por otro modelo.

Random forest nos otorgó una mala clasificación, teniendo un puntaje de x, lo cual concluimos que no es un buen modelo para clasificación de texto.

La mejor opción dentro de estos modelos es naive bayes, ya que nos otorga un buen resultado en la evaluación luego de procesar el texto, eliminar outliers y entrenarlo.

## **Tecnología**

Para el desarrollo del proyecto se hace uso de la plataforma de AWS Academy, en donde se utilizan la tecnología de Amazon S3 para almacenar nuestros archivos cvs modificados a parquet; se utiliza también AWS Glue para hacer la extracción, transformación y carga (ETL) de los datos. Por otro lado, el ambiente de procesamiento que se usó fue JupyterHub de AWS.

Nuestro modelo está siendo entrenado con datos tipo batch, pero la ingesta de datos se realiza de forma híbrida con batch y streaming gracias al acceso que se tiene a la API de Twitter.

El proyecto podrá detectar a los usuarios que sufren depresión y que comunican esto en la red social Twitter.

Un escenario en el que vemos la aplicabilidad e implementación de nuestro proyecto es una página web que permita a las personas ingresar el nombre de usuario de Twitter con el fin de detectar si ese usuario tiene depresión.



## **Conclusiones**

- Los resultados también indican que el uso de Twitter puede ser útil para monitorizar la salud mental de las personas y detectar posibles problemas de salud mental.
- Los tweets negativos no siempre indican depresión, y los tweets positivos no siempre indican ausencia de depresión.
- La detección de la depresión en Twitter puede ser útil para identificar a personas en riesgo de depresión, pero no es una prueba diagnóstica de depresión.
- La detección de la depresión en Twitter puede tener limitaciones, como la falta de contexto y la invisibilidad de la mayoría de los tweets positivos.
- Con el proyecto se logró llegar a varios resultados por métodos distintos lo cual permitió un mejor análisis y obtener mejores resultados.

## Referencias

- <https://www.kaggle.com/code/jbencina/clustering-documents-with-tfidf-and-kmeans/notebook>
- <https://spark.apache.org/docs/2.2.0/mllib-evaluation-metrics.html>
- <https://stackoverflow.com/questions/28786534/increase-resolution-with-word-cloud-and-remove-empty-border>
- [https://github.com/st1800eafit/st1800-2266/blob/main/hadoop\\_spark/spark/Data\\_processing\\_using\\_PySpark.ipynb](https://github.com/st1800eafit/st1800-2266/blob/main/hadoop_spark/spark/Data_processing_using_PySpark.ipynb)