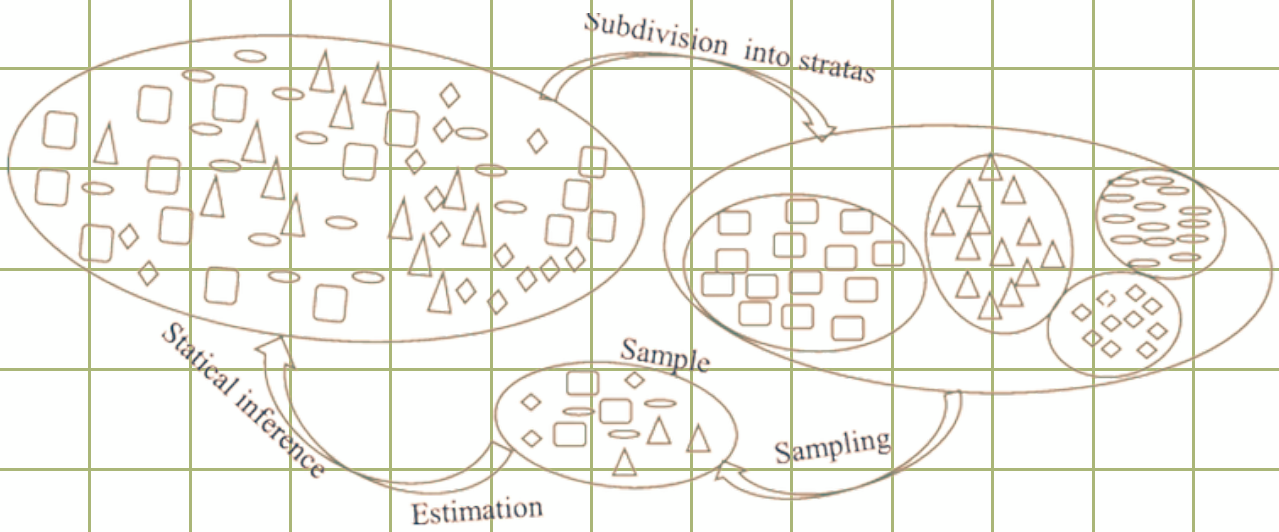
UNIVERSIDAD AUTÓNOMA DE NUEVO LEÓN
Facultad de Ciencias Físico Matemáticas

MAESTRÍA EN CIENCIAS DE DATOS

DATOS MASIVOS

PRACTICA



Cd Universitaria a 29 de Junior de 2023

Manejar Datos Masivos

Dataset

fue obtenido de kaggle.com a través de la API PushShift API, que corresponde a los comentarios generados en el grupo ADHD de la red social Rededit (https://www.reddit.com/r/ADHD/). El dataset contiene 3,356,541 registros lo que hace consumir muchos recursos.


```
#PASO 1. LEER LA BASE DE DATOS
PP = pd.read_csv('/content/drive/MyDrive/Datas/ADHD-comment.csv')
PP.head()
display(PP)
```

Se trabaja en preprocesamiento para eliminar [deleted] y datos duplicados.

	body	id	score	created_utc	created_datetime
	If I try to look this up right now I will get ...	c09y8qz	2.0	1.243790e+09	2009-05-31 17:08:19
	potassium is used as the thing that stops your...	c09yia6	2.0	1.243815e+09	2009-06-01 00:07:50
	I've love a link to anything about this. \n\n...	c0a81e6	3.0	1.244752e+09	2009-06-11 20:25:36
	I don't know anything specific, but I would *d...	c0aixrg	2.0	1.245813e+09	2009-06-24 03:04:51
	Despite continued controversy, powerful new ev...	c0fjlvp	1.0	1.257783e+09	2009-11-09 16:12:44
	...	...	...	...	...
	I completely agree, but it also makes me reall...	gwug5et	1.0	1.620085e+09	2021-05-03 23:44:19
	>They can't see how my brain shuts itself d...	gwug5t9	1.0	1.620085e+09	2021-05-03 23:44:24
	First, congrats, second, you got this!\n\nSo r...	gwug6h0	1.0	1.620085e+09	2021-05-03 23:44:33
	I'm completely the same. I absolutely hate wri...	gwug7do	1.0	1.620085e+09	2021-05-03 23:44:45

Eliminación de columnas

	body	created_datetime
	If I try to look this up right now I will get ...	2009-05-31 17:08:19
	potassium is used as the thing that stops your...	2009-06-01 00:07:50
	I've love a link to anything about this. \n\n...	2009-06-11 20:25:36
	I don't know anything specific, but I would *d...	2009-06-24 03:04:51
	Despite continued controversy, powerful new ev...	2009-11-09 16:12:44

Muestreo de Datos

Debido a la cantidad de información se utilizó el manejo de datos masivos con la técnica de obtener muestreos aleatorios simples que proporciona la librería Pandas

pandas.DataFrame.sample

Se tomaron dos muestras aleatorias de diferente tamaño para revisar si los resultados varian. La primer muestra fue de 12,000 registros y la segunda de 50,000 registros

1

```
dataset = dataset.sample(n=12000)
len(dataset)
```

```
12000
```

2

```
dataset = dataset.sample(n=50000)
len(dataset)
```

```
50000
```

MEDICACIÓN

Para el ejercicio de analizar los comentarios relacionados a la medicación, se consiguió una lista de medicamentos para ADHD de un dataset en https://archive.ics.uci.edu/datasets

El listado se trabajo para convertirlo en arreglo.

```
drugsdt = pd.read_excel('/content/drive/MyDrive/Datas/DRUGS+DATASET+ADHD.xlsx')
drugsdt.head()
display(drugsdt)
```

```
#creamos el arreglo que validara los medicamentos en los comentarios
arreglodrugs = []
existe = 0
cadena = ''
# Iterar sobre los índices del arreglo utilizando un bucle for
for i in range(len(drugsdataset)):
    elemento = str(drugsdataset[i])
    elementof = np.array(elemento.split())
    #arreglo = np.array(elementof)
    for j in range(len(elementof)):
      arreglo = []
      palabra = str(elementof[j])
      palabra= palabra.replace(" ","")
      palabra= palabra.replace("[","")
      palabra= palabra.replace("]","")
      palabra= palabra.replace("'","")
      #para no guardar formatos de duracion prolongada etc.
      if len(palabra)>=3:
        cadena = ', '.join(str(elemento) for elemento in arreglodrugs)
        existe = existe_palabra(cadena, palabra, 80)
        #existe es para no repetir el medicamento en el arreglo.
        if existe==1:
          arreglo = []
        if existe==0:
          arreglo.append(palabra)
          arreglodrugs += arreglo
print(arreglodrugs)
```

```
['Amphetamine', 'dextroamphetamine', 'Clonidine', 'Bupropion', 'Vyvanse', 'Atomoxetine', 'Dexmethylphenidate',
```

COMENTARIOS RESPECTO A MEDICACIÓN

Se utilizó la API fuzzywuzzy para crear una función que validara en los comentarios si hablan de alguna medicación en especificó estableciendo un umbral de 80 y con esto filtramos los comentarios por únicamente los que hablaban de medicación

```python
#funcion para encontrar palabras
from fuzzywuzzy import fuzz

def existe_palabra(texto, palabra_buscada, umbral):
    e = 0
    palabras = texto.split()
    for palabra in palabras:
        similitud = fuzz.ratio(palabra_buscada, palabra)
        if similitud >= umbral:
          e=1
    return e
```

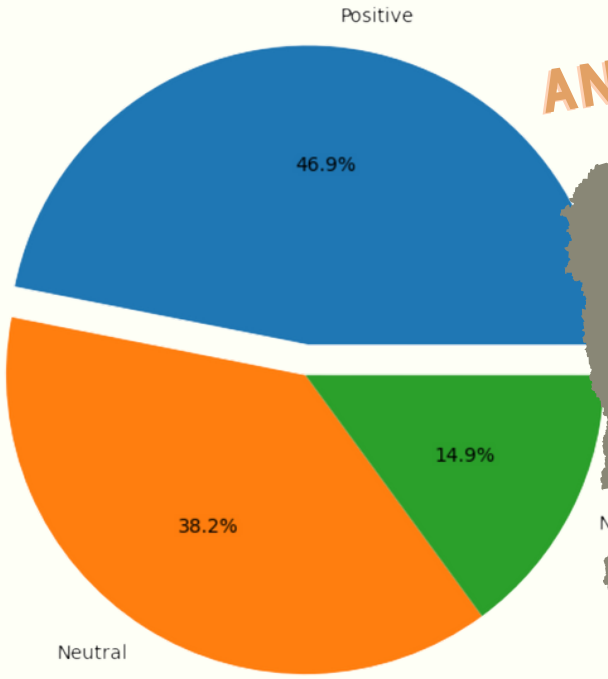
BUSQUEDA DE MEDICAMENTO EN COMENTARIOS

```python
existe = 0
j=0
#dataset.head()
for indice, registro in dataset.iterrows():
  j+=1
  cadena = registro['Cleaned Body']
  for i in range(len(arreglodrugs)):
    palabra = str(arreglodrugs[i])
    existe = existe_palabra(cadena, palabra, 80)
    if existe ==1:
      dataset.loc[indice, 'Tipo'] = 'Medicación'
      dataset.loc[indice, 'Medicación'] = palabra
      print(j)
      break
```

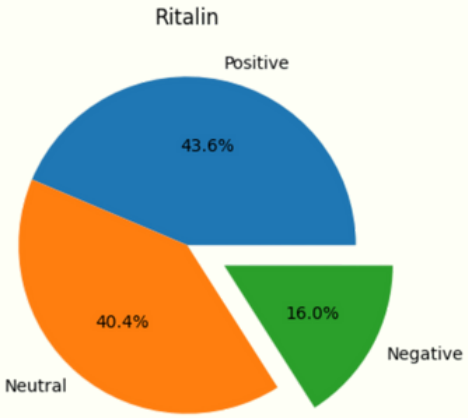
	body	created_datetime	Year	Month	Cleaned Body	Tipo	Medicación
489947	ooh well you see i have some janky shit going ...	2016-01-30 22:53:48	2016	1	ooh well you see i have some janky shit going...	Medicación	Adderall
1795892	i mean, yeah, it's definitely only in the last...	2019-02-06 23:45:14	2019	2	i mean yeah it s definitely only in the last o...	Medicación	Amphetamine
2717519	Happens to me all the time too, you're not alo...	2020-07-19 23:22:24	2020	7	Happens to me all the time too you re not alon...	Medicación	Vyvanse
2402254	A different med could help. I have only ever b...	2020-01-18 16:40:03	2020	1	A different med could help I have only ever be...	Medicación	Adderall
108954	I'm currently at 60mg and I haven't had any ap...	2013-04-25 20:10:53	2013	4	I m currently at mg and I haven t had any appe...	Medicación	Strattera

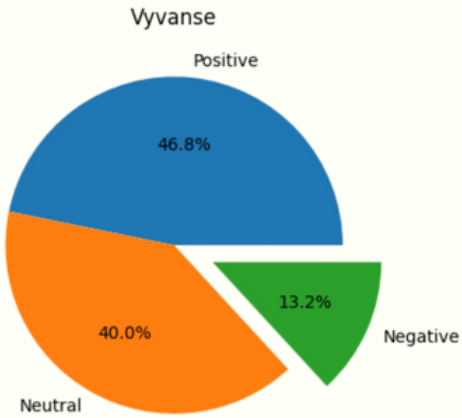
Con esto que se realizó se pudo filtrar unicamente los comentarios que hablaban sobre alguna medicación para el ADHD y se hizo el análisis de sentimientos

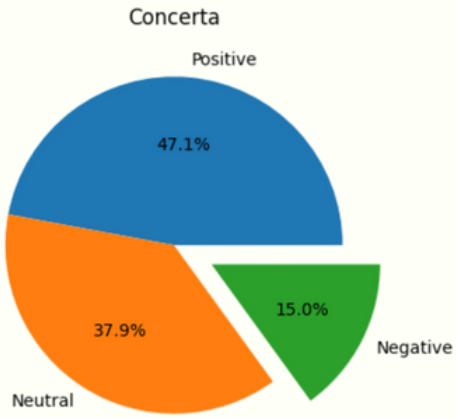
ANALISIS DE SENTIMIENTOS



En la muestra que se tomó por la cual se obtuvieron solo los comentarios que hablaban de medicación:
46.9% fueron positivos
38.9% fueron neutrales
14.9% fueron negativos







Se analizaron 3 de las meditaciones más comunes

- Ritalin
- Vyvanse
- Concerta

CONCLUSIÓN

Al tomar un muestreo aleatorio simple, los análisis fueron más rápidos, dando oportunidad de hacer búsquedas dentro de los comentarios de la meditaciones prescritas para el ADHD. Y con esto obtener únicamente comentarios hablando acerca.

Se pudo analizar 3 medicación dejando un poco más claro que Vyvanse dentro de la red social donde se obtuvieron los comentarios tiene más respuestas positivas de los otros dos.