

# Diplomado Ciencia de Datos, Módulo 3 Práctica 1

David Rincon Morales

Octubre 2022

## Contents

<b>1 Ejercicio 1</b>	<b>2</b>
1.1 Datos / EDA . . . . .	2
1.2 Análisis Factorial . . . . .	3
1.2.1 Análisis Estadístico - Esfericidad de Bartlett . . . . .	3
1.2.2 Análisis Estadístico - KMO Test . . . . .	3
1.3 Factores . . . . .	3
1.4 Interpretación de factores . . . . .	4
1.5 Conclusiones Ejercicio 1 . . . . .	6
<b>2 Ejercicio 2</b>	<b>6</b>
2.1 Datos / EDA . . . . .	6
2.1.1 Test Número Factores . . . . .	7
2.2 PCA . . . . .	8
2.3 t-SNE . . . . .	9
2.4 Conclusiones Ejercicio 2 . . . . .	10
<b>3 Ejercicio 3</b>	<b>11</b>
3.1 Datos / EDA . . . . .	11
3.1.1 Test Número Factores . . . . .	11
3.2 PCA . . . . .	12
3.3 t-SNE . . . . .	13
3.4 Conclusiones Ejercicio 3 . . . . .	14

# 1 Ejercicio 1

El objetivo de este primer ejercicio, era el poder realizar el análisis de factores para un grupo de datos, los cuales son las respuestas a una encuesta sobre una areolínea

## 1.1 Datos / EDA

En cuanto a los datos, teníamos 25976 registros con 15 columnas (id incluido), en lo cuales no tenemos registros faltantes ("missings").

Servicio wifi a bordo	0
Hora de salida/llegada conveniente	0
Facilidad de reserva en línea	0
Ubicación de la puerta	0
Alimentos y bebidas	0
Embarque en línea	0
Comodidad del asiento	0
Entretenimiento a bordo	0
Servicio a bordo	0
Servicio de sala de piernas	0
Manejo de equipaje	0
Servicio de facturación	0
Servicio a bordo.1	0
Limpieza	0
...	...

Una vez revisado esto, procedemos a eliminar la columna "Id", ya que no nos ayudará en nuestro análisis de factores. De igual manera, podemos darnos cuenta que hay dos variables que parecen ser la misma *Servicio a Bordo* y *Servicio a Bordo.1*; pero, al aplicar el método "equals" a ambas columnas, así como ver sus valores, podemos darnos cuenta que son diferentes variables, por lo tanto dejamos a ambas.

	Servicio a bordo	Servicio a bordo.1
0	5	5
1	4	4
2	4	2
3	1	1
4	2	2
...	...	...
25971	3	5
25972	4	5
25973	4	4
25974	3	5
25975	1	1

Al hacer una descripción de nuestras variables, podernos darnos cuenta que el 0.001 y el 0.999 percentiles no están muy lejanos de los valores mínimos y

máximos de nuestros datos, esto pasa porque nuestros datos están en un rango de 1-5.

## 1.2 Análisis Factorial

Una vez hecho un EDA básico, procedemos a realizar el Ánalisis Factorial. Para esto, primero estandarizamos nuestros datos sin importar que se encuentren dentro de la misma escala.

### 1.2.1 Análisis Estadístico - Esfericidad de Bartlett

Ahora, para poder ver si podemos realizar el análisis factorial necesitamos hacer un par de pruebas estadísticas a nuestros datos, una de ellas es la Esfericidad de Bartlett.

Con esta prueba, podemos ver si las variables tienen alguna relación entre sí, como en este caso tenemos un p-value < 0.05, podemos confirmar esta hipótesis

```
Esfericidad de Bartlett
Valor de Chi : 152172.39908232106
P - value : 0.0
```

### 1.2.2 Análisis Estadístico - KMO Test

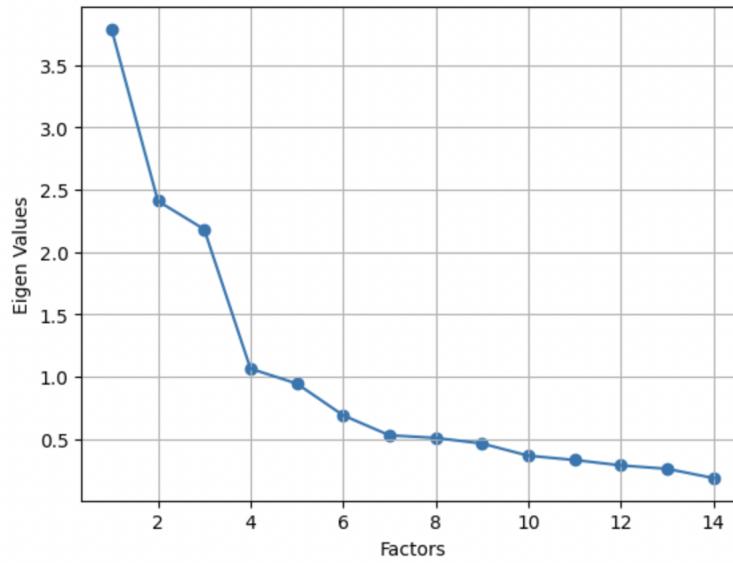
El siguiente estadístico que debemos revisar es la prueba de Kaiser-Meyer-Olkin (KMO), ya que nos dice si podemos realizar un análisis de factores en nuestros datos. Para ello, mientras el resultado sea > 0.5, podemos proceder con los siguientes procesos.

---

```
KMO Test Statistic 0.7824143507684171
```

## 1.3 Factores

Luego, para ver hasta cuántos factores podemos "dividir" nuestra información, procedemos a hacer una gráfica de "codo", la cual está basada en los eigenvalores; y , mientras que los eigenvalores sean > 1, podemos contarlos como número de factor.



En este caso, como el 5to factor ya tiene un eigenvalor  $< 1$ , vamos a tomar a 4 como número de factores a analizar.

#### 1.4 Interpretación de factores

- Carga de Factores: La carga de factores Ayudan a explicar cuánto una variable es explicada por un factor, en nuestro caso, podemos darnos cuenta que los dos primeros factores ayudan a explicar variable más relacionadas con el servicio dentro del avión; mientras que, los dos últimos factores ayudan a explicar variables más relacionadas con el aeropuerto.

	0	1	2	3
<b>Servicio wifi a bordo</b>	0.09	0.13	<b>0.62</b>	<b>0.45</b>
<b>Hora de salida/llegada conveniente</b>	-0.02	0.05	<b>0.59</b>	-0.00
<b>Facilidad de reserva en línea</b>	-0.03	0.03	<b>0.77</b>	<b>0.43</b>
<b>Ubicación de la puerta</b>	0.01	-0.05	<b>0.70</b>	-0.12
<b>Alimentos y bebidas</b>	<b>0.77</b>	0.00	0.02	0.04
<b>Embarque en línea</b>	0.28	0.11	0.13	<b>0.76</b>
<b>Comodidad del asiento</b>	<b>0.76</b>	0.07	-0.03	<b>0.22</b>
<b>Entretenimiento a bordo</b>	<b>0.78</b>	<b>0.46</b>	0.04	0.02
<b>Servicio a bordo</b>	0.08	<b>0.71</b>	0.01	0.05
<b>Servicio de sala de piernas</b>	0.06	<b>0.49</b>	0.05	0.09
<b>Manejo de equipaje</b>	0.04	<b>0.77</b>	0.04	-0.03
<b>Servicio de facturación</b>	0.10	0.29	-0.04	<b>0.15</b>
<b>Servicio a bordo.1</b>	0.05	<b>0.80</b>	0.04	-0.05
<b>Limpieza</b>	<b>0.85</b>	0.08	-0.01	0.10

- Diferencias: Este análisis ayuda a ver cuánta varianza explica cada factor, así como la varianza acumulada. En nuestro ejercicio, los 4 factores ayudan a explicar el 56% de la varianza

	0	1	2	3
<b>Variance</b>	2.614813	2.310657	1.846621	1.080291
<b>Proportional Var</b>	0.186772	0.165047	0.131902	0.077164
<b>Cumulative Var</b>	0.186772	0.351819	0.483721	0.560884

- Comunidades: Este nos ayuda a ver cuánta varianza es explicada por los factores dentro de nuestro análisis. Aquí, sólo en 3/14 variables tienen la mayoría de su varianza explicada por los 4 factores.

Unicity	
<b>Servicio wifi a bordo</b>	0.380922
<b>Hora de salida/llegada conveniente</b>	0.649000
<b>Facilidad de reserva en línea</b>	0.214819
<b>Ubicación de la puerta</b>	0.499015
<b>Alimentos y bebidas</b>	0.400940
<b>Embarque en línea</b>	0.313488
<b>Comodidad del asiento</b>	0.370441
<b>Entretenimiento a bordo</b>	0.180963
<b>Servicio a bordo</b>	0.493557
<b>Servicio de sala de piernas</b>	0.745325
<b>Manejo de equipaje</b>	0.406054
<b>Servicio de facturación</b>	0.881165
<b>Servicio a bordo.1</b>	0.354861
<b>Limpieza</b>	0.257067

## 1.5 Conclusiones Ejercicio 1

- Al hacer el análisis de factores, vimos que la cantidad "ideal" es que existan 4 factores en total.
- Con estos 4 factores, se explica solamente cerca del 56% de los datos.
- Sólo 3 de las variables de nuestro dataset tienen la mayoría de su varianza explicada con estos 4 factores.

## 2 Ejercicio 2

El objetivo del ejercicio 2 era poder comparar al menos 2 métodos de reducción de dimensiones distintos en un conjunto de datos de cancer.

### 2.1 Datos / EDA

En tema de datos, tenemos 32 columnas distintas, donde incluimos el id y el diagnóstico de las pruebas.

Para empezar con el EDA, vamos a tirar el id ya que no nos sirve para este proceso; luego revisamos si hay valores ausentes dentro de nuestros datos. A lo que no encontramos con eso (una muestra abajo)

	count
<b>radius_mean</b>	0.0
<b>texture_mean</b>	0.0
<b>perimeter_mean</b>	0.0
<b>area_mean</b>	0.0
<b>smoothness_mean</b>	0.0
<b>compactness_mean</b>	0.0
<b>concavity_mean</b>	0.0
<b>concave points_mean</b>	0.0
<b>symmetry_mean</b>	0.0
<b>fractal_dimension_mean</b>	0.0
<b>radius_se</b>	0.0
<b>texture_se</b>	0.0
<b>perimeter_se</b>	0.0

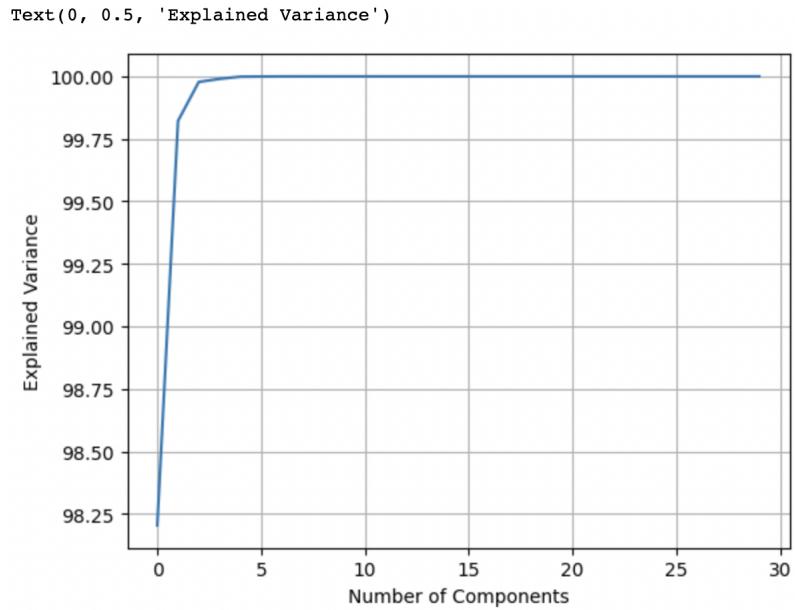
Luego, vemos la proporción de diagnósticos en nuestro data set, a lo cual, casi el 63% tiene diagnóstico "B" y el resto "M"

<b>B</b>	<b>0.627417</b>
<b>M</b>	<b>0.372583</b>

Al ver la descripción de nuestras variables continuas podemos ver que tampoco hay mucha diferencia entre los bajos y altos percentiles contra los valores mínimos y máximos de nuestro dataset.

### 2.1.1 Test Número Factores

Antes de hacer la reducción de dimensiones, hicimos un test para poder ver cómo se iban a comportar nuestros componentes y sea más fácil comprobar.

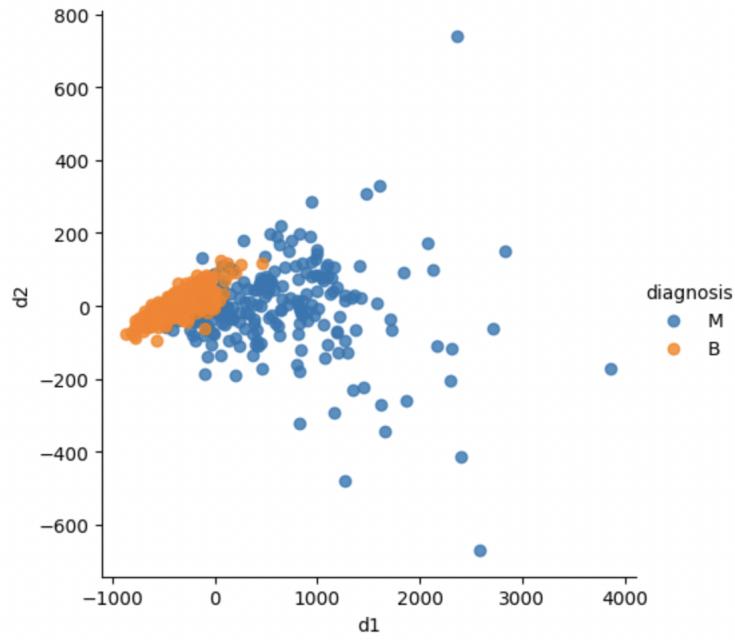


De esta manera, podemos ver que con relativamente pocos componentes se explica la mayor cantidad de nuestra varianza.

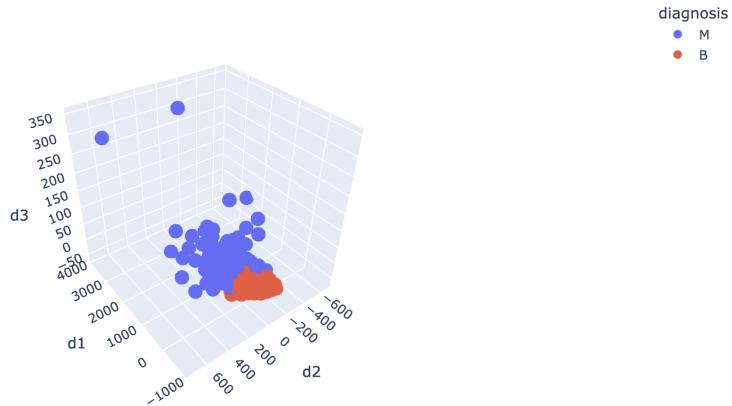
## 2.2 PCA

Para esta práctica, usamos los métodos de PCA y t-SNE. Antes pasamos a escalar nuestros valores para que no tengamos problemas con nuestros datos.

- 2 Componentes:



- 3 Componentes:

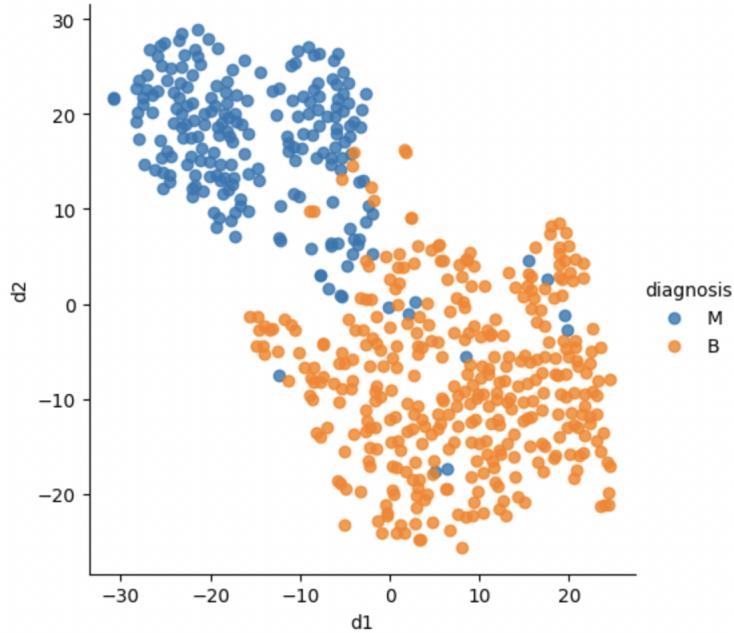


En este caso, desde los 2 componentes de PCA empezamos a ver la distinción más clara de nuestros datos.

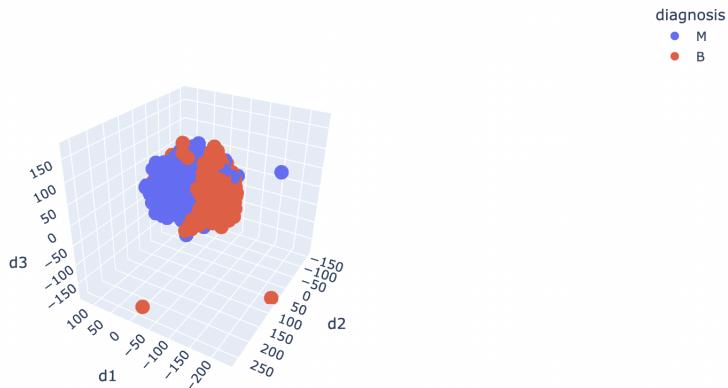
### 2.3 t-SNE

- 2 Componentes:

<seaborn.axisgrid.FacetGrid at 0x2838fcceb0>



- 3 Componentes:



Aunque con PCA ya empzábamos a ver nuestras variables bien definidas, con t-SNE podemos ver los diagnósticos mucho más definidos

## 2.4 Conclusiones Ejercicio 2

- Para el ejercicio 2, usamos PCA y t-SNE para reducir dimensiones.

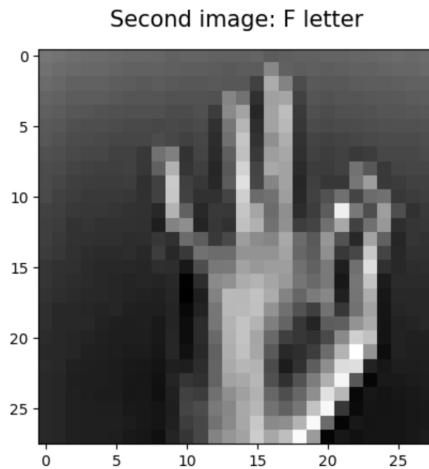
- Desde la primer implementación de PCA con 2 componentes, podemos apreciar la buena distinción de nuestros datos.
- La mejor implemntación para este proceso, podemos usar cualquier implementación de t-SNE (2 o 3 componentes), ya que se muestra de mejor manera la distribución 60-40 de los diagnósticos.

### 3 Ejercicio 3

El objetivo de la 3er parte, es muy similar al ejercicio 2 pero aplicado con imágenes.

#### 3.1 Datos / EDA

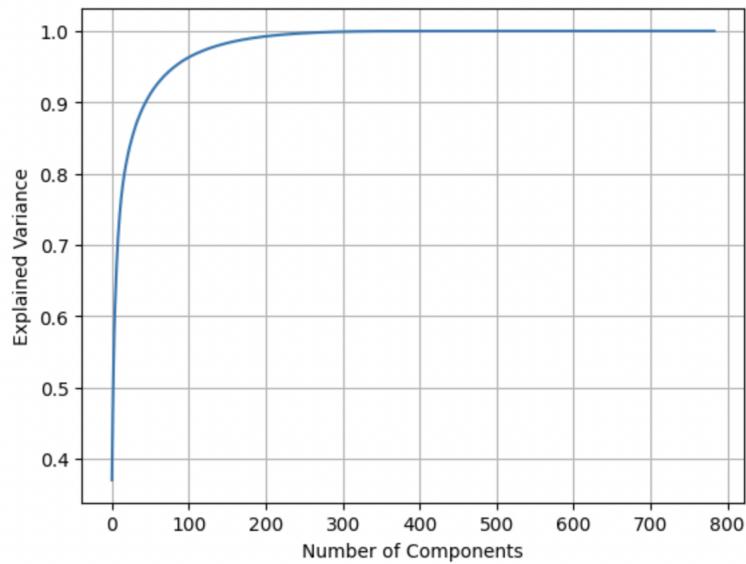
En estos datos, tenemos 7,172 registro y cada imagen representada con 784 pixeles (cada columna).



Como son imágenes, no hay mucho que podamos analizar. Luego de este breve EDA, escalamos nuestra información con StandarScaler.

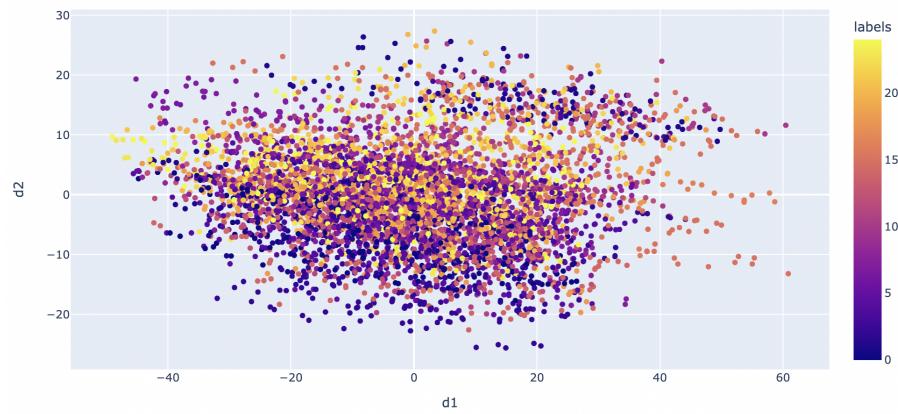
##### 3.1.1 Test Número Factores

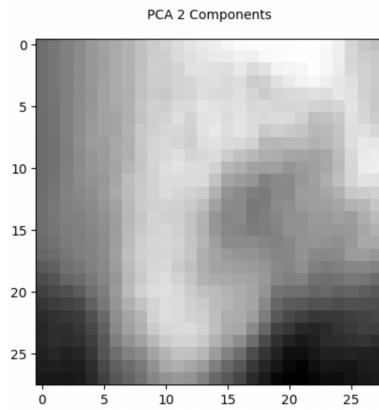
Igual que en el ejercicio 2, hicimos una prueba para poder ver cómo se van a comportar nuestros componentes dentro de nuestro análisis. A lo cual podemos apreciar que con sólo 2 o 3 componentes no vamos a tener una buena representación en las imágenes.



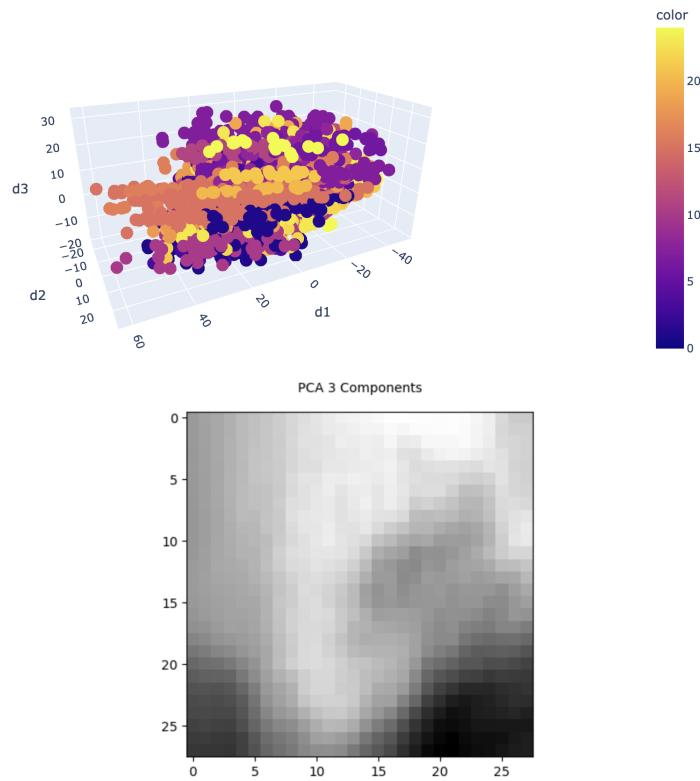
### 3.2 PCA

- 2 Componentes: El PCA de 2 componentes sólo logra explicar cerca del 46% de las varianzas.



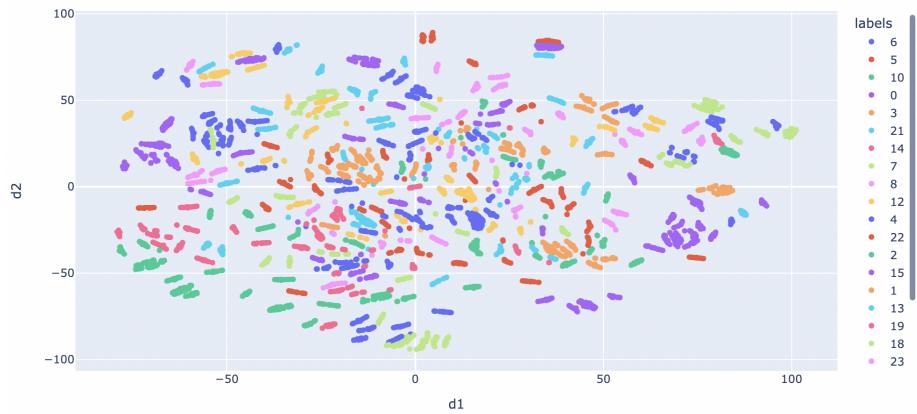


- 3 Componentes: Con los 3 componentes, se sigue explicando sólo cerca del 52%, poco más de los dos componentes.

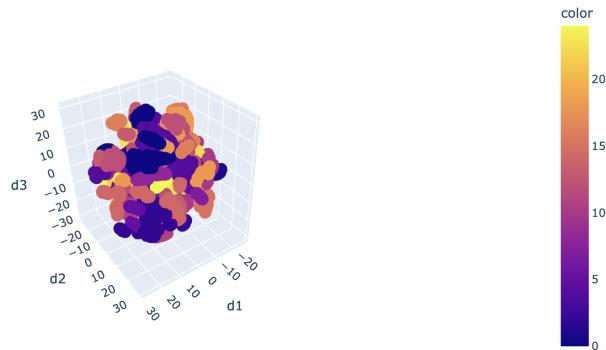


### 3.3 t-SNE

- 2 Componentes:



- 3 Componentes:



### 3.4 Conclusiones Ejercicio 3

- Para este ejercicio, nuestra varianza está muy poco explicada, por lo tanto las imágenes re-escaladas de PCA tienen muy poca definición.
- El PCA de 2 componentes sólo explica cerca del 46% de los datos, mientras que los 3 componentes explica sólo el 53% de los datos.
- Aunque no tengamos tanta definición de nuestros datos, la "mejor" implementación sería el t-SNE con 3 componentes.
- Preferiblemente, se deberían realizar por lo menos cerca de 25 componentes para tener una mejor definición.