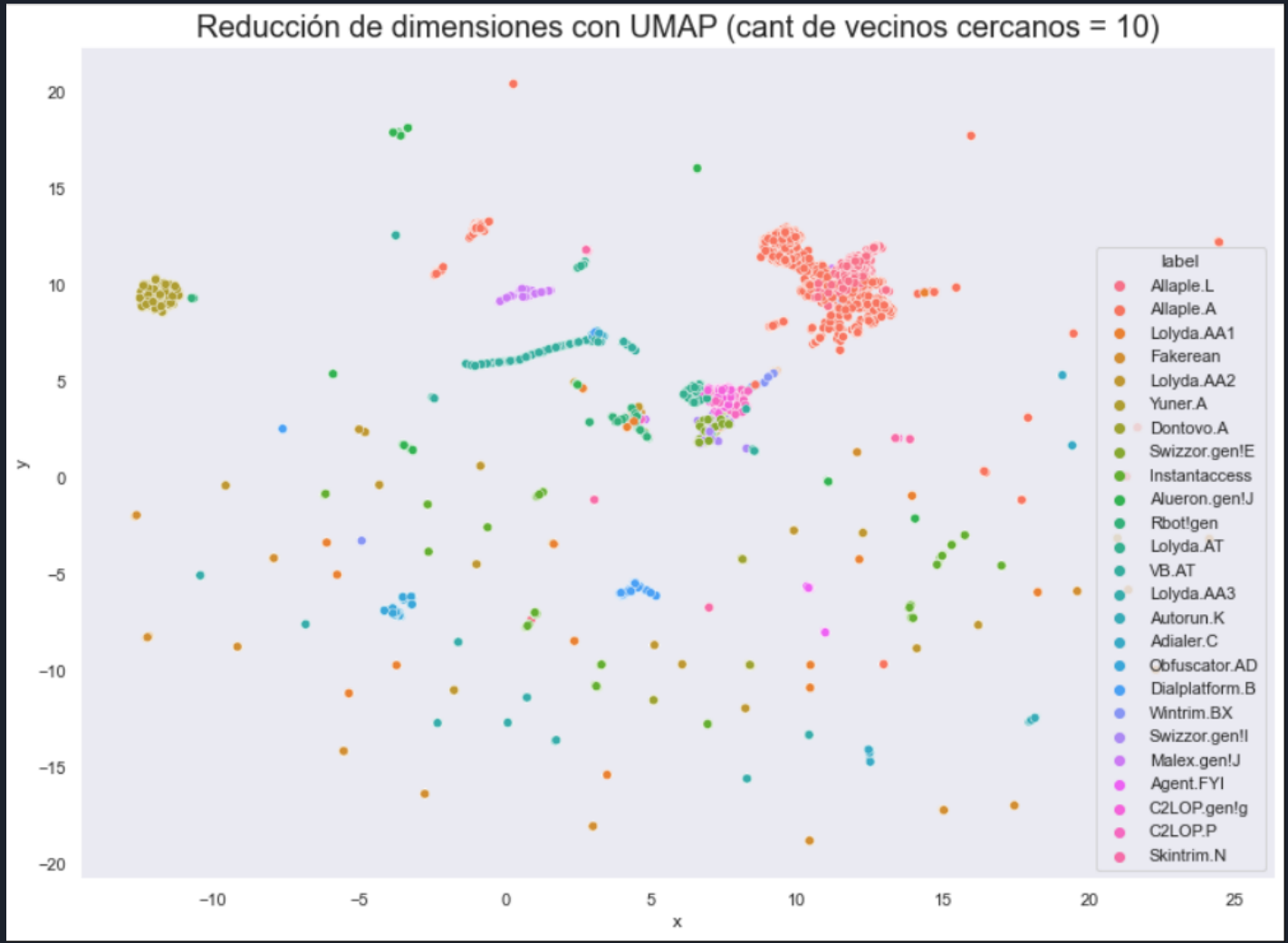
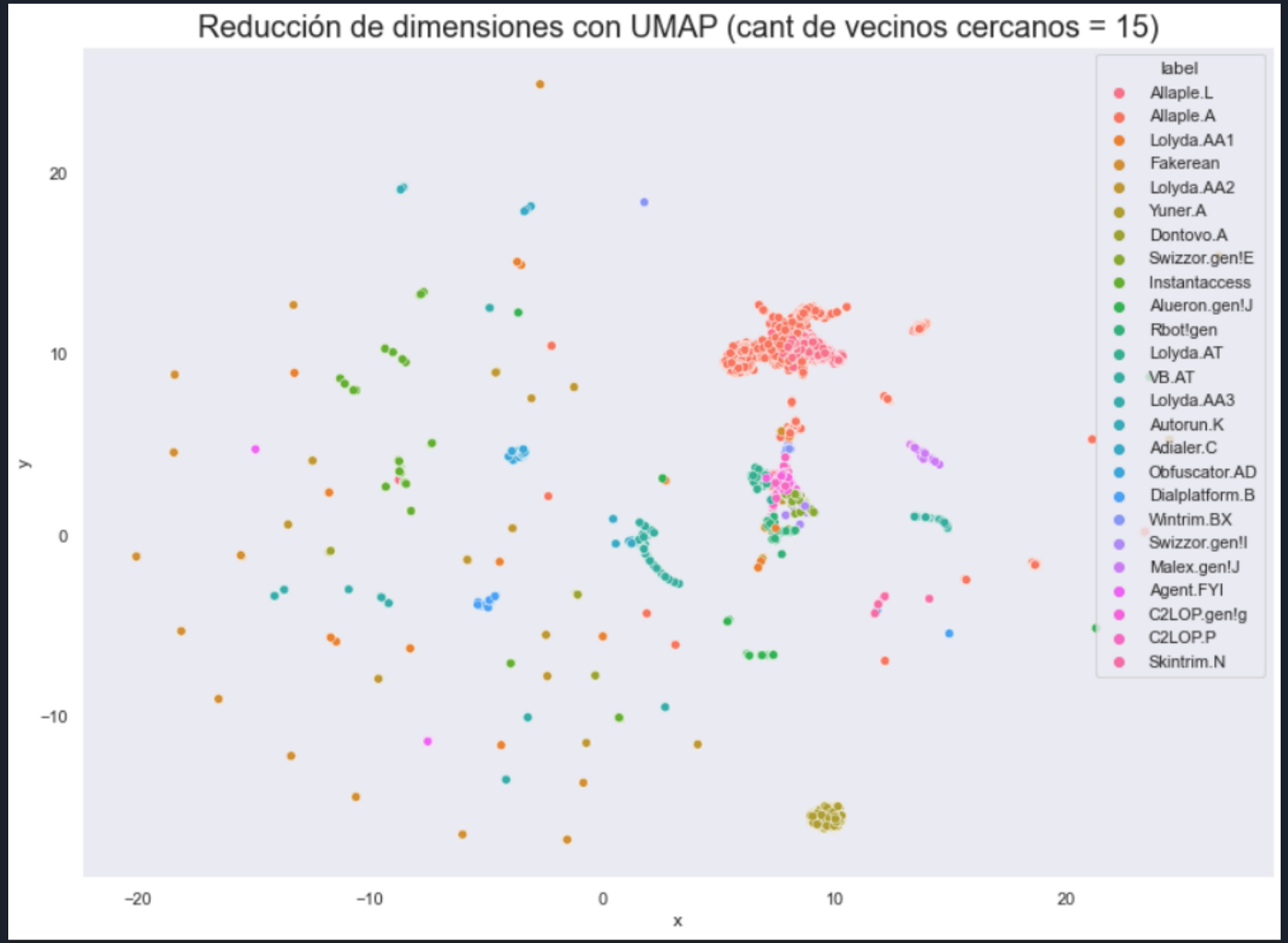
TP4

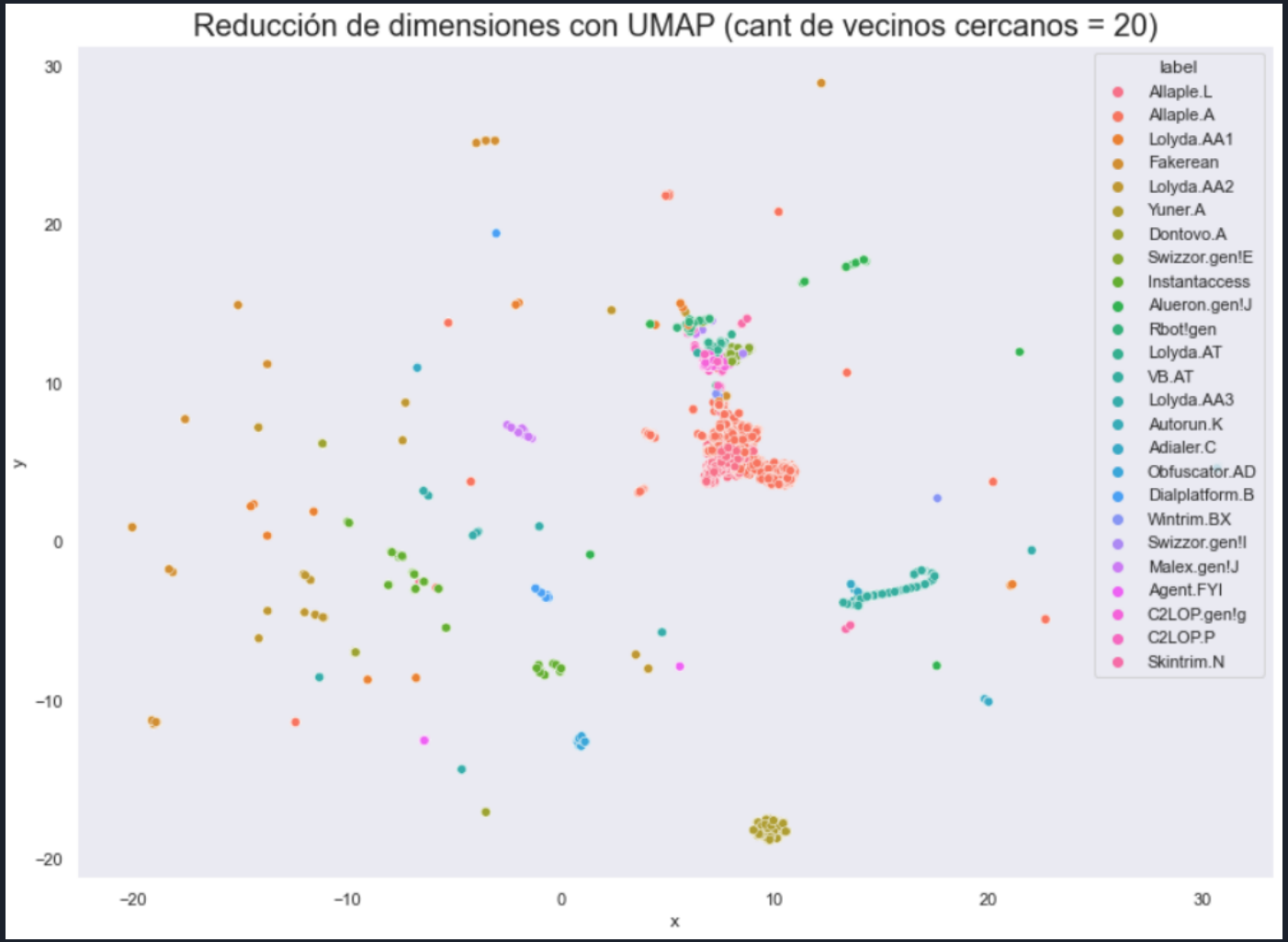
Gabriel Semorile

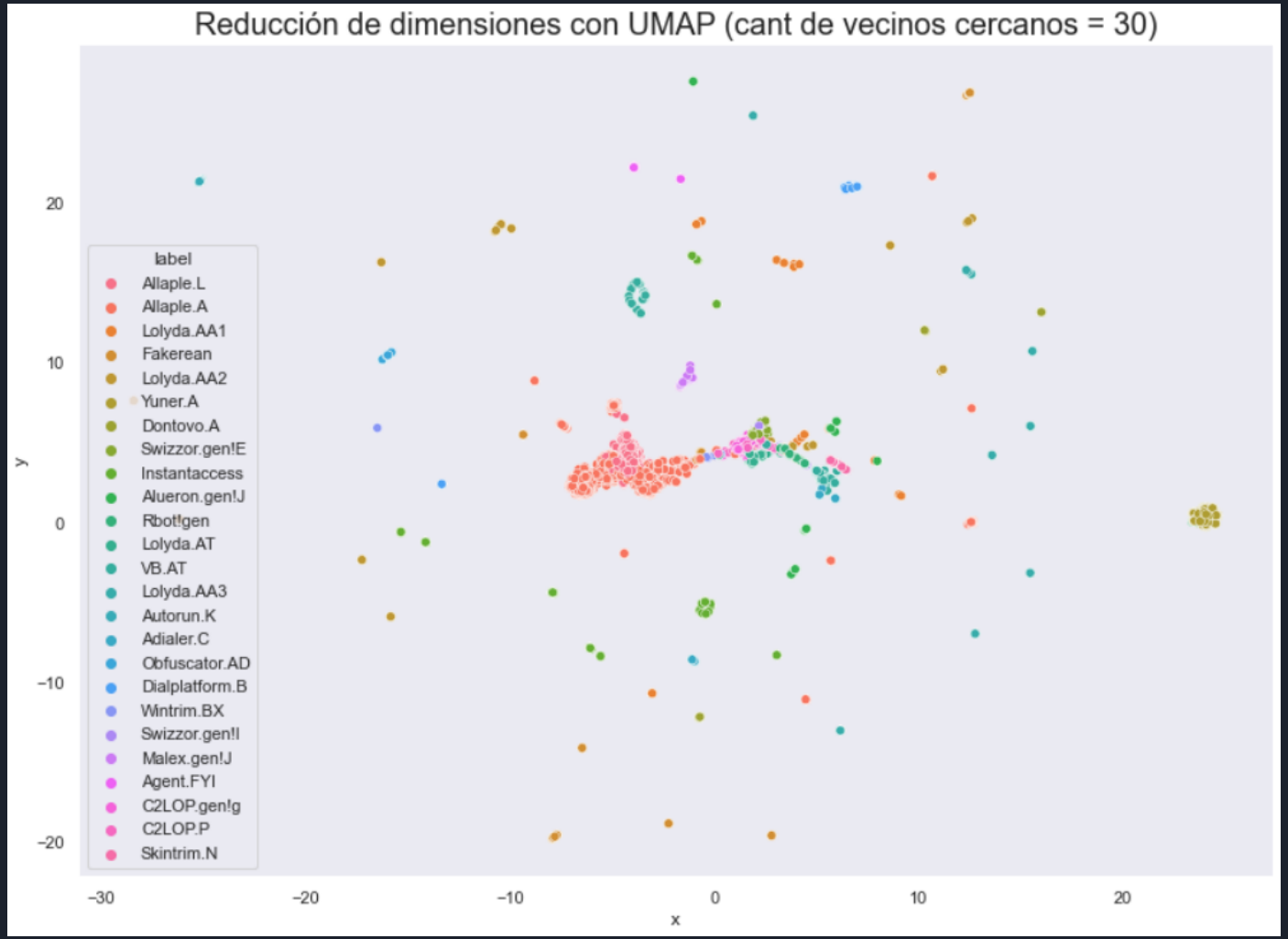
Preparación del Dataset

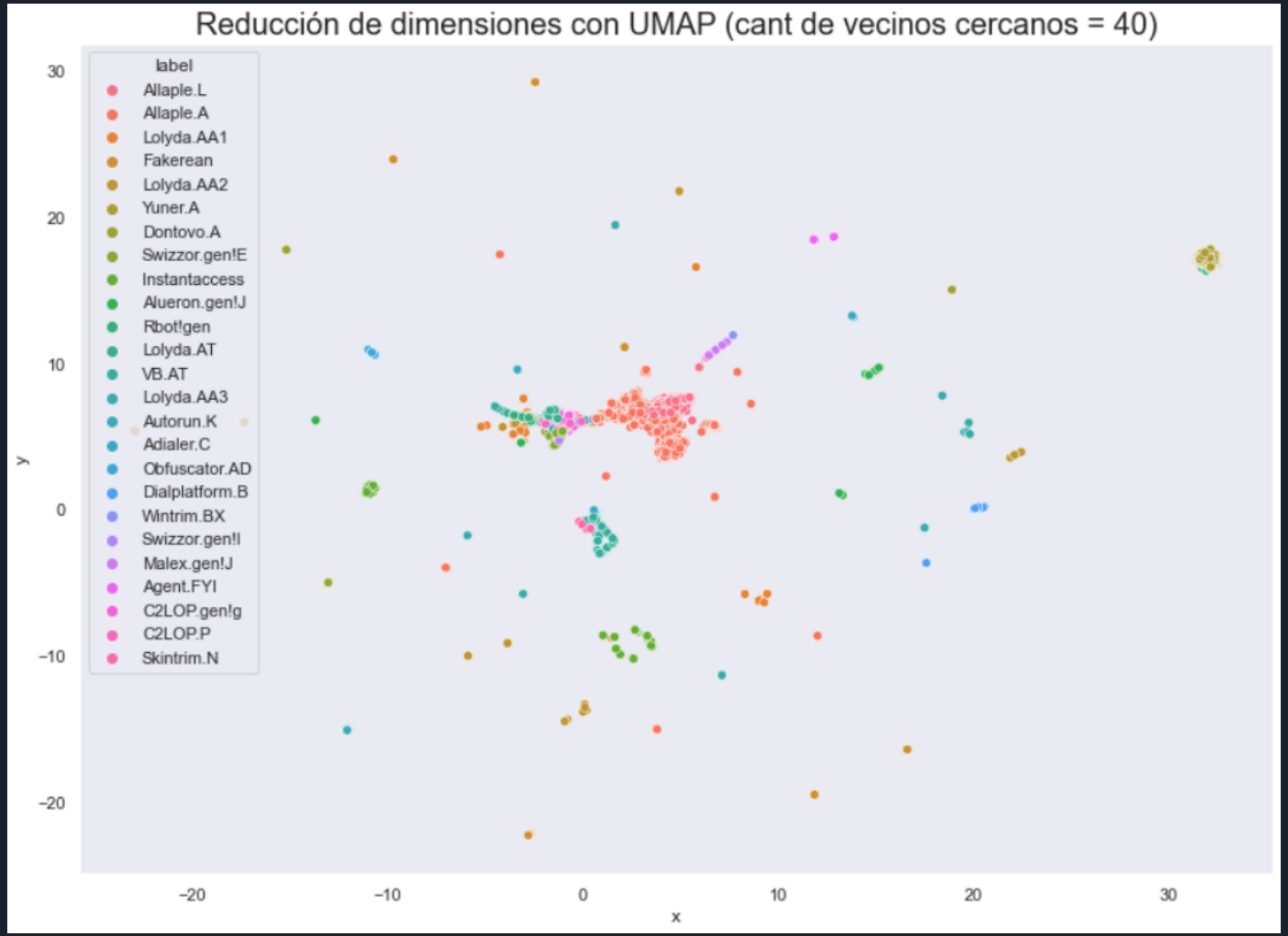
1) Separo de forma aleatoria las imagenes entre train, validation y test.
2) Redimensiono las imágenes.
3) Utilizo UMAP para la reducción de dimensiones previa al entrenamiento.
4) Primero reduzco a 2 dimensiones para graficar.

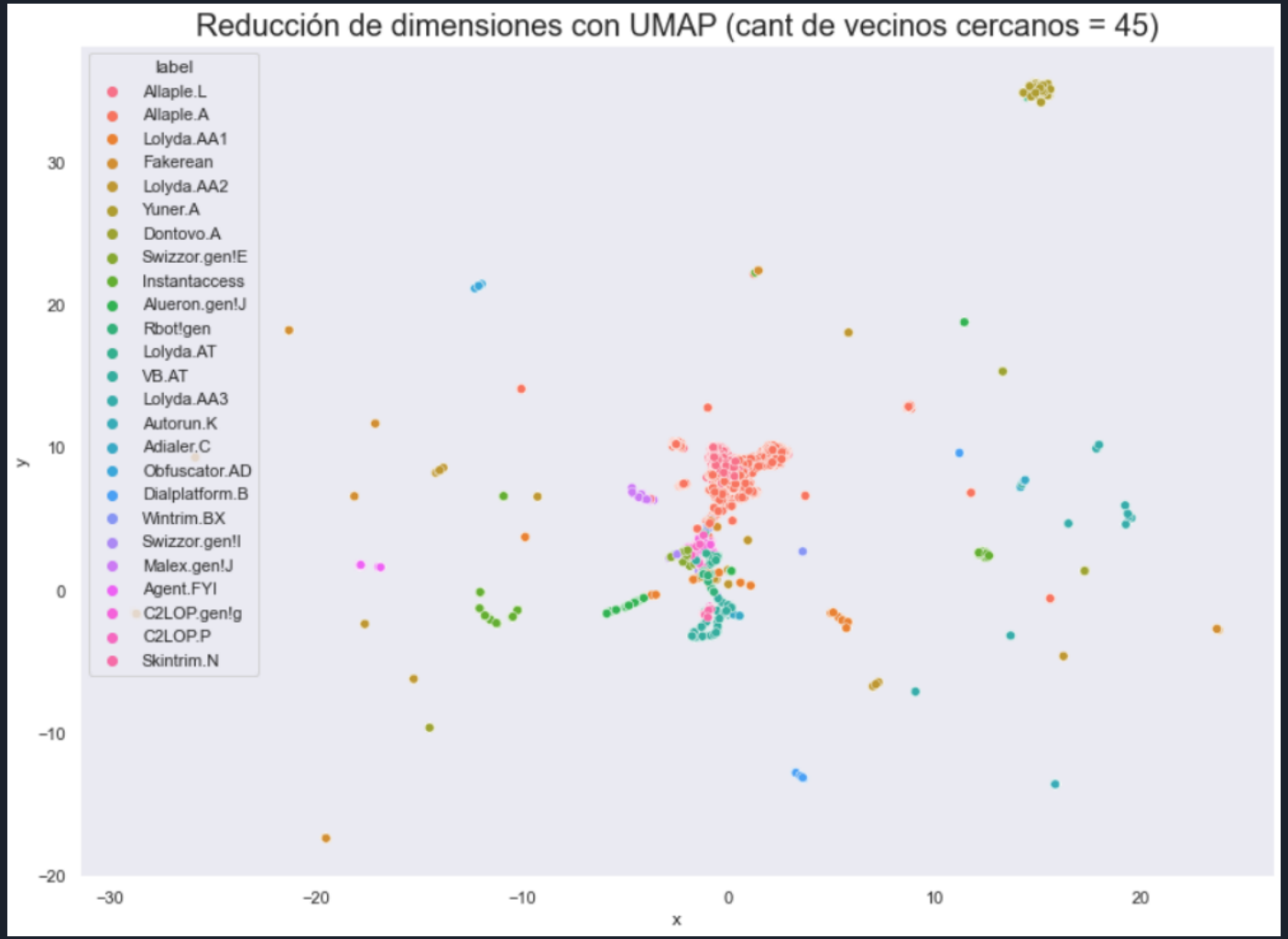

Reducción de dimensiones con UMAP (cant de vecinos cercanos = 10)
y
x
20
15
10
5
0
−5
−10
−15
−20
−10
−5
0
5
10
15
20
25
label
Allaple.L
Allaple.A
Lolyda.AA1
Fakerean
Lolyda.AA2
Yuner.A
Dontovo.A
Swizzor.gen!E
Instantaccess
Alueron.gen!J
Rbot!gen
Lolyda.AT
VB.AT
Lolyda.AA3
Autorun.K
Adialer.C
Obfuscator.AD
Dialplatform.B
Wintrim.BX
Swizzor.gen!I
Malex.gen!J
Agent.FYI
C2LOP.gen!g
C2LOP.P
Skintrim.N

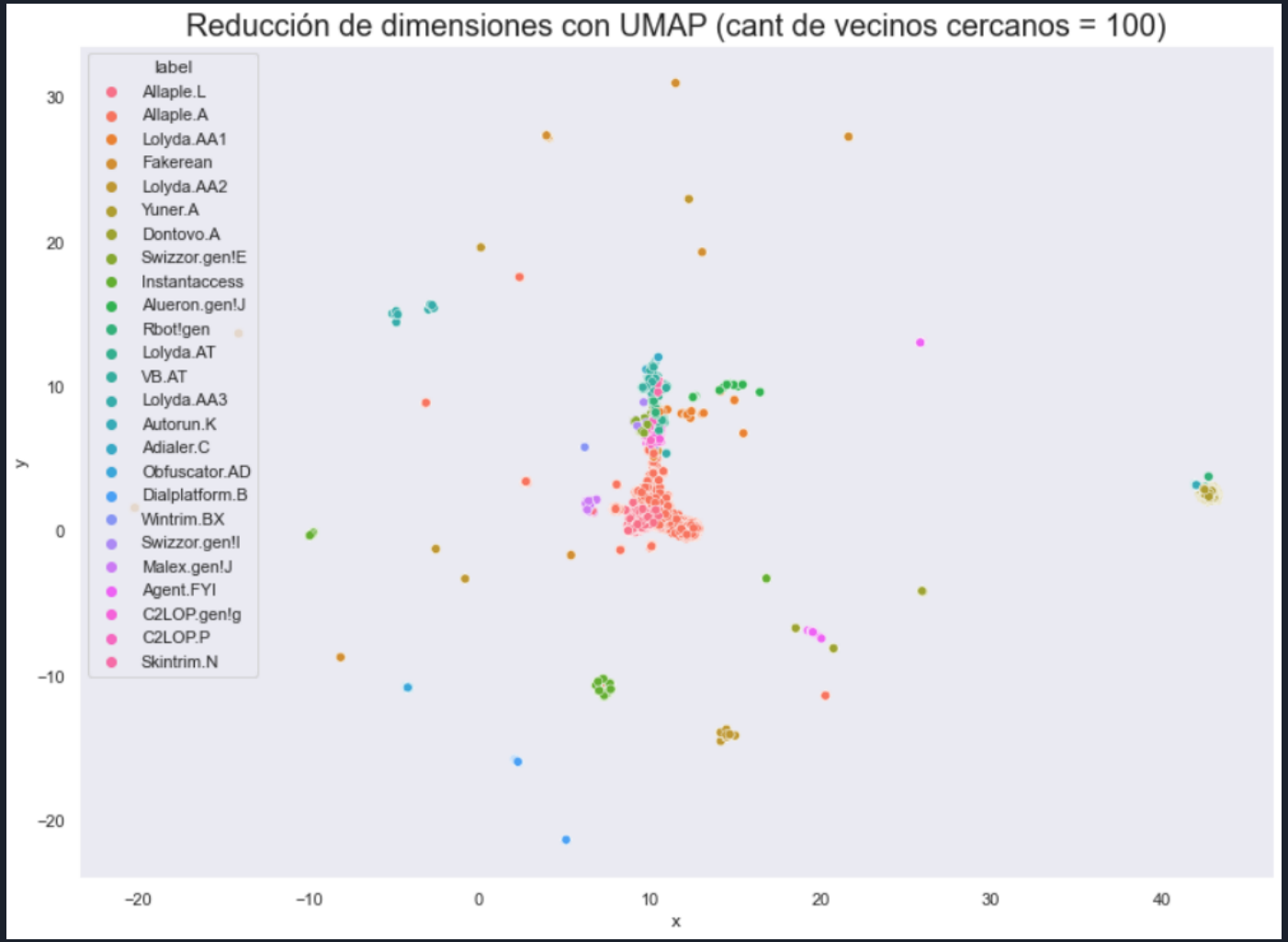

Reducción de dimensiones con UMAP (cant de vecinos cercanos = 15)
label
Allaple.L
Allaple.A
Lolyda.AA1
Fakerean
Lolyda.AA2
Yuner.A
Dontovo.A
Swizzor.gen!E
Instantaccess
Alueron.gen!J
Rbot!gen
Lolyda.AT
VB.AT
Lolyda.AA3
Autorun.K
Adialer.C
Obfuscator.AD
Dialplatform.B
Wintrim.BX
Swizzor.gen!I
Malex.gen!J
Agent.FYI
C2LOP.gen!g
C2LOP.P
Skintrim.N
20
10
0
−10
y
−20
−10
0
10
20
x


Reducción de dimensiones con UMAP (cant de vecinos cercanos = 20)
label
Allaple.L
Allaple.A
Lolyda.AA1
Fakerean
Lolyda.AA2
Yuner.A
Dontovo.A
Swizzor.gen!E
Instantaccess
Alueron.gen!J
Rbot!gen
Lolyda.AT
VB.AT
Lolyda.AA3
Autorun.K
Adialer.C
Obfuscator.AD
Dialplatform.B
Wintrim.BX
Swizzor.gen!I
Malex.gen!J
Agent.FYI
C2LOP.gen!g
C2LOP.P
Skintrim.N
30
20
10
0
−10
−20
y
−20
−10
0
10
20
30
x


Reducción de dimensiones con UMAP (cant de vecinos cercanos = 30)
label
Allaple.L
Allaple.A
Lolyda.AA1
Fakerean
Lolyda.AA2
Yuner.A
Dontovo.A
Swizzor.gen!E
Instantaccess
Alueron.gen!J
Rbot!gen
Lolyda.AT
VB.AT
Lolyda.AA3
Autorun.K
Adialer.C
Obfuscator.AD
Dialplatform.B
Wintrim.BX
Swizzor.gen!I
Malex.gen!J
Agent.FYI
C2LOP.gen!g
C2LOP.P
Skintrim.N
y
20
10
0
−10
−20
−30
−20
−10
0
10
20
x


Reducción de dimensiones con UMAP (cant de vecinos cercanos = 40)
label
Allaple.L
Allaple.A
Lolyda.AA1
Fakerean
Lolyda.AA2
Yuner.A
Dontovo.A
Swizzor.gen!E
Instantaccess
Alueron.gen!J
Rbot!gen
Lolyda.AT
VB.AT
Lolyda.AA3
Autorun.K
Adialer.C
Obfuscator.AD
Dialplatform.B
Wintrim.BX
Swizzor.gen!I
Malex.gen!J
Agent.FYI
C2LOP.gen!g
C2LOP.P
Skintrim.N
30
20
10
0
−10
−20
y
−20
−10
0
10
20
30
x


Reducción de dimensiones con UMAP (cant de vecinos cercanos = 45)
label
Allaple.L
Allaple.A
Lolyda.AA1
Fakerean
Lolyda.AA2
Yuner.A
Dontovo.A
Swizzor.gen!E
Instantaccess
Alueron.gen!J
Rbot!gen
Lolyda.AT
VB.AT
Lolyda.AA3
Autorun.K
Adialer.C
Obfuscator.AD
Dialplatform.B
Wintrim.BX
Swizzor.gen!I
Malex.gen!J
Agent.FYI
C2LOP.gen!g
C2LOP.P
Skintrim.N
y
30
20
10
0
−10
−20
−30
−20
−10
0
10
20
x


Reducción de dimensiones con UMAP (cant de vecinos cercanos = 100)
label
Allaple.L
Allaple.A
Lolyda.AA1
Fakerean
Lolyda.AA2
Yuner.A
Dontovo.A
Swizzor.gen!E
Instantaccess
Alueron.gen!J
Rbot!gen
Lolyda.AT
VB.AT
Lolyda.AA3
Autorun.K
Adialer.C
Obfuscator.AD
Dialplatform.B
Wintrim.BX
Swizzor.gen!I
Malex.gen!J
Agent.FYI
C2LOP.gen!g
C2LOP.P
Skintrim.N
y
30
20
10
0
−10
−20
−20
−10
0
10
20
30
40
x

Baseline con Kmeans

- Estandarizo la data para entrenar con StandardScaler
- Utilizo UMAP para reducir a 300 dimensiones.
- Entreno Kmeans con K = N (siendo N la cantidad de clases de virus).

Obtengo los siguientes Scores...

Datasets \ Scores	Purity Score	NMI
Validation	0.6323	0.7058
Test	0.6420	0.7005

Métricas a utilizar para Clustering

Purity:

Se basa en calcular para cada punto que proporción de los compañeros del cluster son del mismo tipo.

Normalized Mutual Info (NMI):

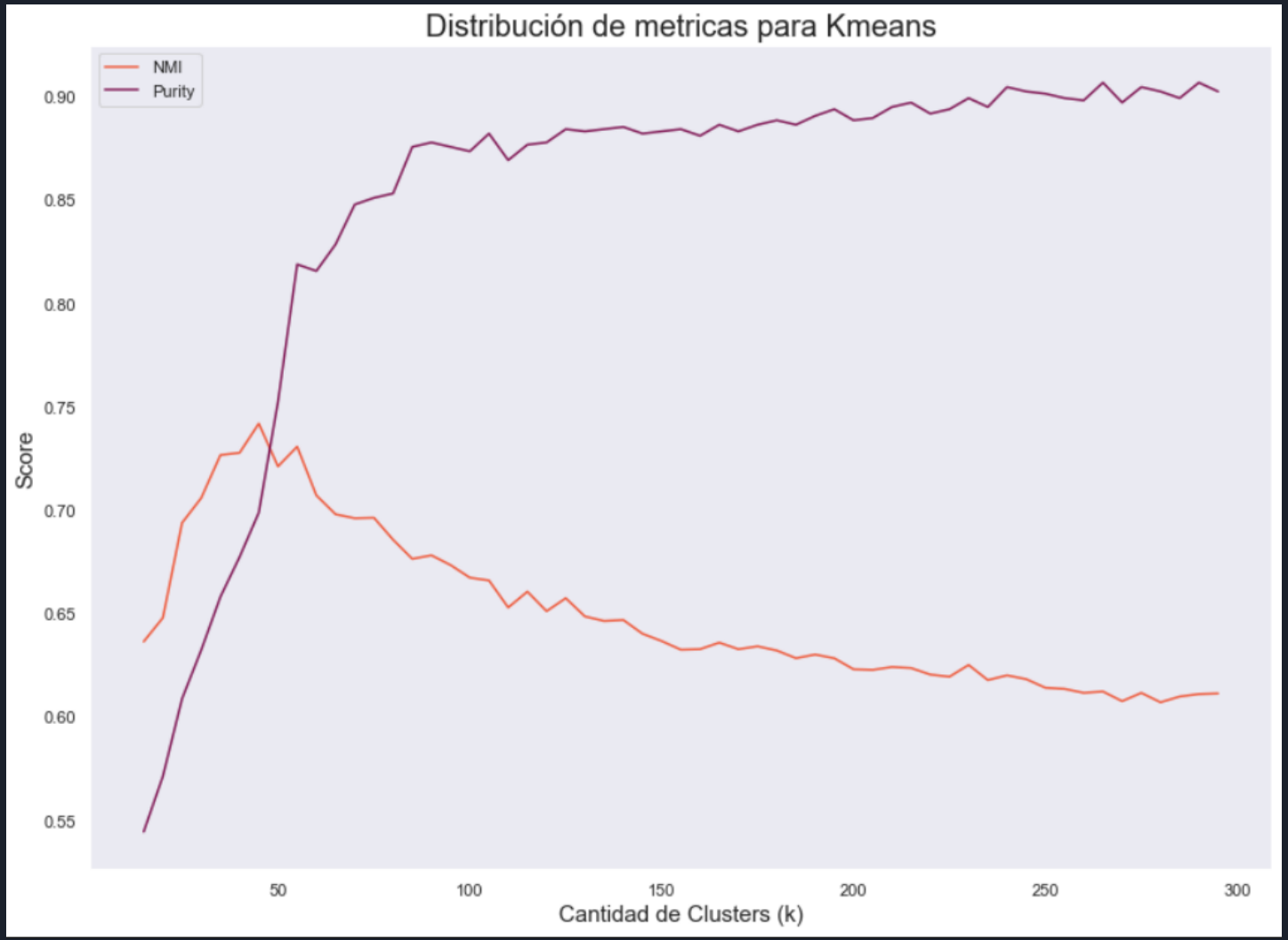
Se basa en normalizar Mutual Info, el cual es 0 si el clustering es aleatorio respecto a la clase de cada punto.

Características \ Scores	Purity Score	NMI
Ventajas	Muy Simple	Balance entre cantidad de clusters y calidad del clustering.
Desventajas	El score aumenta a medida que aumenta la cantidad de clusters.	Más Complejo

Kmeans con Búsqueda de Hiper-parámetros

- Hago la busqueda en torno a K (cantidad de clusters), por lo tanto uso NMI para conseguir el mejor modelo.
- Busco K entre 15 y 300 con un intervalo de 5.

El mejor NMI obtenido es: 0.7418


Distribución de metricas para Kmeans
NMI
Purity
Score
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0.55
50
100
150
200
250
300
Cantidad de Clusters (k)

HDBSCAN

- Hago un Grid Search entre min_cluster_size, min_samples y cluster_selection_epsilon

Los puntajes NMI obtenidos son:

Validation: 0.6705

Test: 0.6752

Muchas
Gracias!