Universidad de Chile
Facultad de Ciencias Físicas y Matemáticas
Departamento de Ingeniería Eléctrica

Tarea 4

Aprendizaje no supervisado

Integrantes: Vincko Fabres
Profesor: Pablo Estevez.
Auxiliar: Ignacio Reyes Jainaga
Ayudantes: Andrés González
Bastián Andreas
Daniel Baeza
Francisca Cona
Javier Molina
Óscar Pimentel
Pablo Montero
Roberto Cholaky

Fecha de entrega: Sábado 27 de Noviembre de 2021
Santiago, Chile

1. Parte teórica

1.1. ¿Qué significa que un algoritmo de aprendizaje de máquinas sea no supervisado?.

Esqto quiere decir que no se conocen las clases para la clasificación y las muestras se pueden agrupar mediante relaciones de similitud o proximidad en un conjunto finito.

1.2. El método PCA entrega una nueva base para describir un conjunto de muestras multidimensionales. Al usar PCA como método de reducción de dimensionalidad o método de visualización, los primeros vectores de la base se conservan, mientras que el resto son desechados. ¿Cuál es la justificación para esto? Explique considerando la relación entre los valores propios de la matriz de correlación y la varianza de los datos.

Los primeros vectores se conservan ya que PCA asume que las direcciones principales son las de mayor varianza, la que indica qué tan principal es una dirección, rankeando los vectores base, esto tiene trasfondo en minimizar los residuos en esa dirección.Los vectores propios de la matriz de correlación representan las direcciones principales a lo largo de las cuales las varianzas tienen sus valores extremos y los valores propios asociados definen los valores extremos de las varianzas.

1.3. Considere el método de Kernel PCA con kernel gaussiano, el cual mapea N muestras a un espacio distinto antes de aplicar PCA. ¿Cuántas dimensiones tiene dicho espacio para el caso del kernel gaussiano?

La dimensión de un kernel gaussiano es infinita.

1.4. Describa brevemente el algoritmo SOM y explique cómo se interpreta la visualización de la U-Matrix.

Se tiene el espacio original al que se le hace cuantización vectorial para poder mapear a través de una grilla 2-D la cual es topologicamente ordenada, a esta se le define la vecindad. Para esto se toman k prototipos, cuando se presenta un ejemplo se encuentra el prototipo mas cercano y luego se modifican los vectores de referencia del best match y la vecindad.

La visualización de Umatrix otorga una representacion de distancia entre vecinos, asignando un simil a curvas de nivel como tercer eje en colores, creando asi fronteras virtuales para una mayor interpretación.

2. Parte práctica

Se debe explorar una base de datos de la data del Observatorio de Complejidad Económica del MIT con métodos de aprendizaje no supervisado, la cual corresponde a las exportaciones de 97 productos, los métodos a utilizar son PCA, Kernel con PCA y SOM.

2.1.

Dado que cada país está descrito por 97 características, correspondientes a productos de exportación, lo primero es realizar las visualizaciones de PCA con 2 componentes principales comparando las diferencias de agrupar los productos en 15 categorías versus utilizar cada uno, es decir, 97 tipos.

Una vez realizadas ambas visualizaciones los resultados son los siguientes:

	Suma errores cuadraticos	Varianza PC	Suma PC
Sin agrupar	17815.8306	0.077540 0.042129	0.1196
Productos agrupados	1964.7359	0.154817 0.108092	0.2629

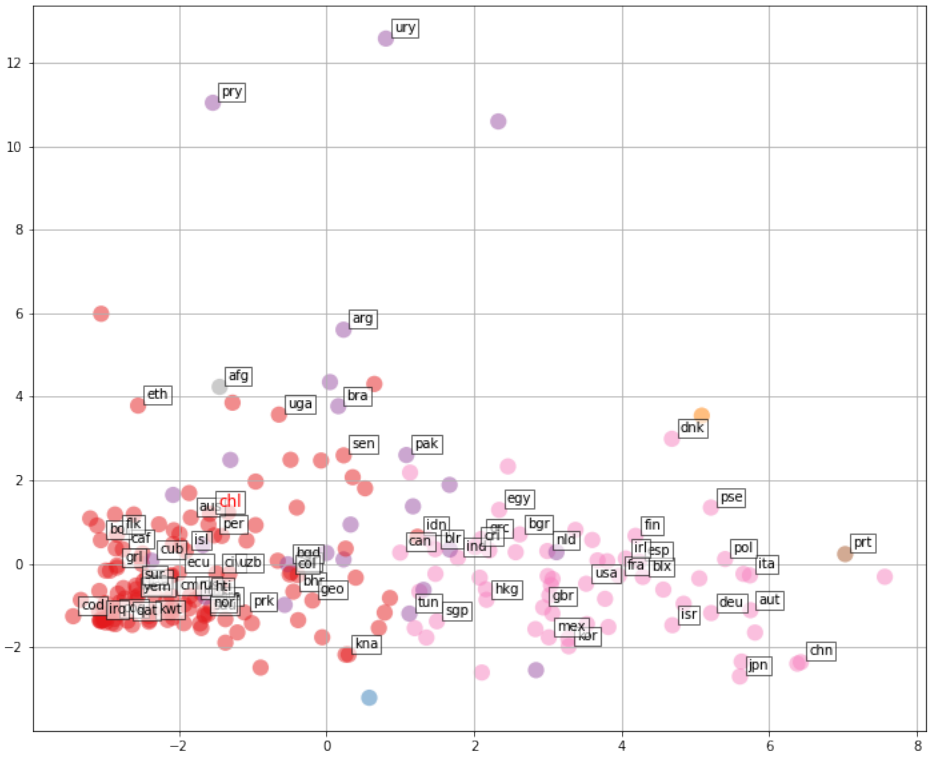

Figura 1: Visualización PCA 97 productos

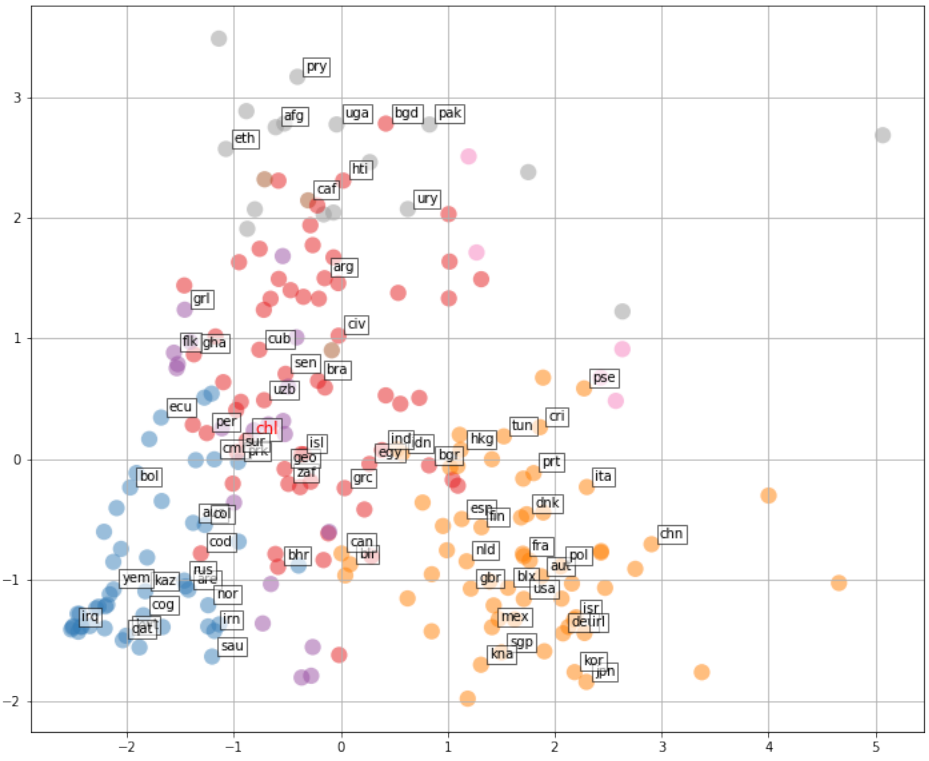

Figura 2: Visualización PCA 15 categorías

Al utilizar la información tal y como se presenta para cada país y utilizar PCA la visualización resulta dificil de analizar dado el solapamiento entre los paises, si el número de productos fuese aún mayor los clusters serían aún más aglomerados dificultando aún más el análisis, por otra parte el asignar categorías resulta en una ventaja a la hora de visualizar, ya que en este caso, países con exportaciones de productos similares estarán cercanos, mientras que países que no compartan categorías de productos en exportación estarás alejados, lo que en términos de anáisis de economías de exportación parecidas ayudará a su comprensión.

2.2.

A partir de ahora dadas las justificaciones anteriormente mencionadas sólo se trabaja con agrupación de productos por categoría.

Como es posible visualizar del resultado de Figura 2 los clusters visualizados presentan poca densidad, a simple vista es posible notar que los clusters comparten varios países según geografía,lo que se puede explicar por sus productos de exportación extraídos dada la localización, quedando algunos fuera del alcance de los clusters, es decir, outliers. Para verificar la hipótesis se revisan 3 clusters, con paises cercanos en cada uno revisando la data otorgada por el Observatorio de Complejidad Económica, resultando en el cluster azul paises con exporatacion de petroleo, el cluster anaranjado maquinarias y productos relaiconados con electricidad y para el cluster rojo productos de agricultura.

2.3.

Se debe comparar el resultado de PCA de Figura 2 versus Kernel PCA con kernel gaussiano, por lo cual se genera la siguiente visualización:

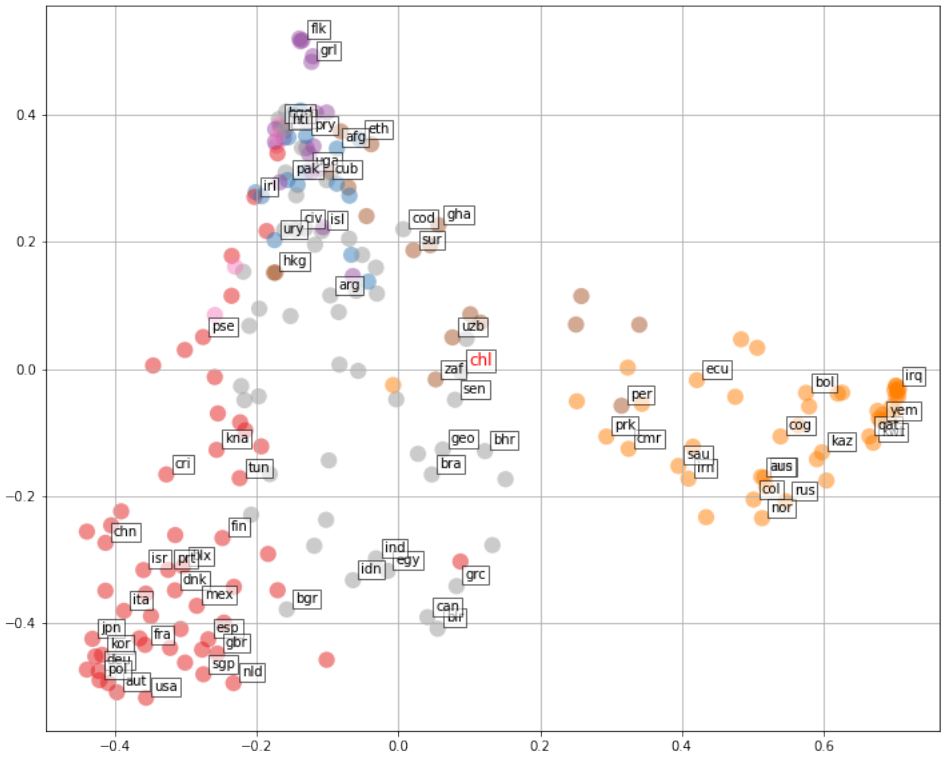

Figura 3: Visualización Kernel PCA

Al realizar una visualización simultanea, es posible apreciar la diferencia de clusters, donde el Kernel PCA genera una mayor separación de estos, generando 3 nubes separadas y un cluster bastante disperso, con poca densidad al medio. Por otra parte la nueva separación de países requiere una actualización en el análisis, ya que países que en la versión anterior estaban muy distanciados como irl y afg ahora se presentan en el mismo cluster.

2.4.

En PCA sin kernel se debe escoger el número de características que capturen al menos el 85 % de varianza de datos, es decir 11, ya que con 10 se captura el 83,2270 % y con 11 88,3317 %. Posterior a esto se debe construir un mapa auto organizado de Kohonen.

Para visualizar el SOM de usa U-Matrix, resultando lo siguiente:

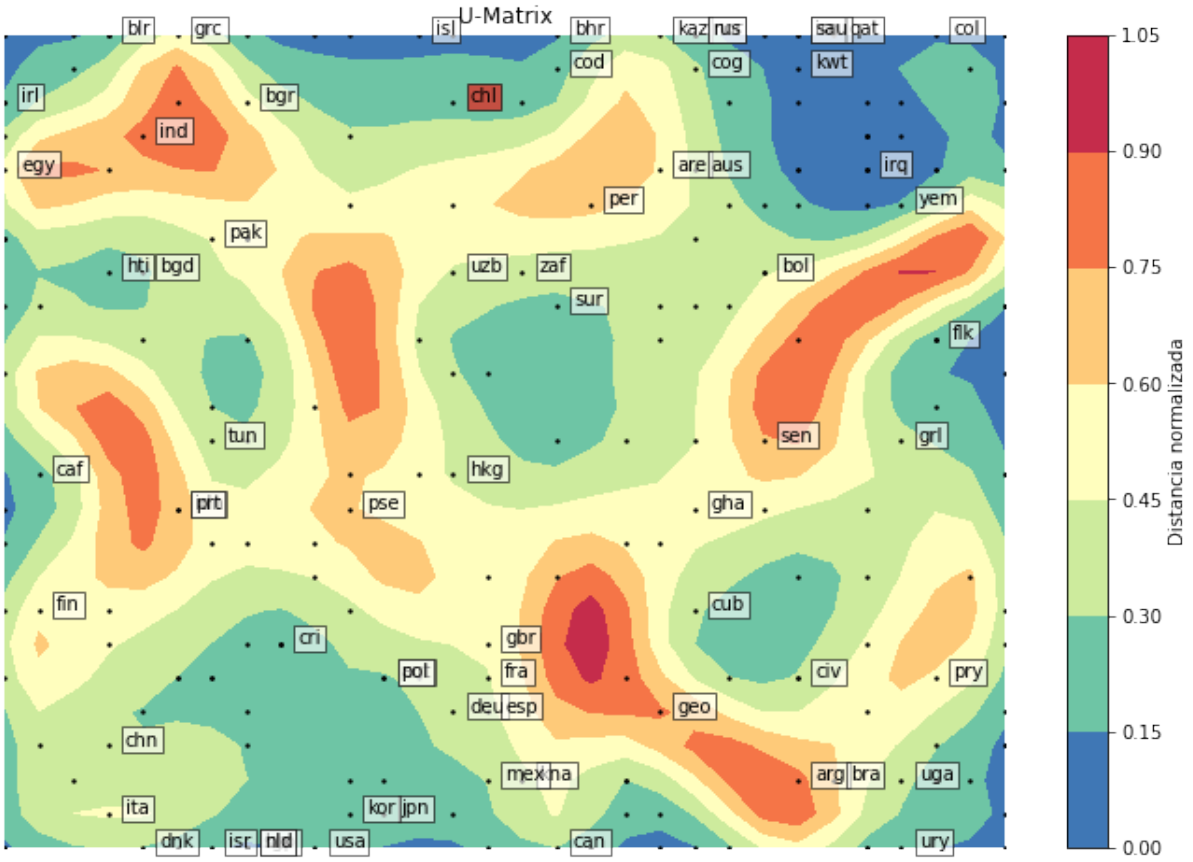

Figura 4: Visualización U-Matrix PCA

Donde es posible visualizar gracias a U-Matrix que existe una gran distancia o frontera para varios clusters, coloreados de rojo, como el caso de ind y egy, o arg que se encuentra solo. El cluster más denso es el correspondiente al color azul, donde se encuentra irq, kwt, sau, qat, entre otros. El espacio posee un grillado donde existen por decirlo así una ´cordillera de distancias´, ya que los colores pertenecientes a distancias más grandes están conectados (color amarillo) y en algunos casos llegando a presentar distancias enormes (color rojo).

2.5.

Esta vez se utiliza PCA con kernel gaussiano escogiendo el mismo número de características, es decir 11, para hacer un mapa SOM y una visualización usando U-Matrix. Resultando lo siguiente:

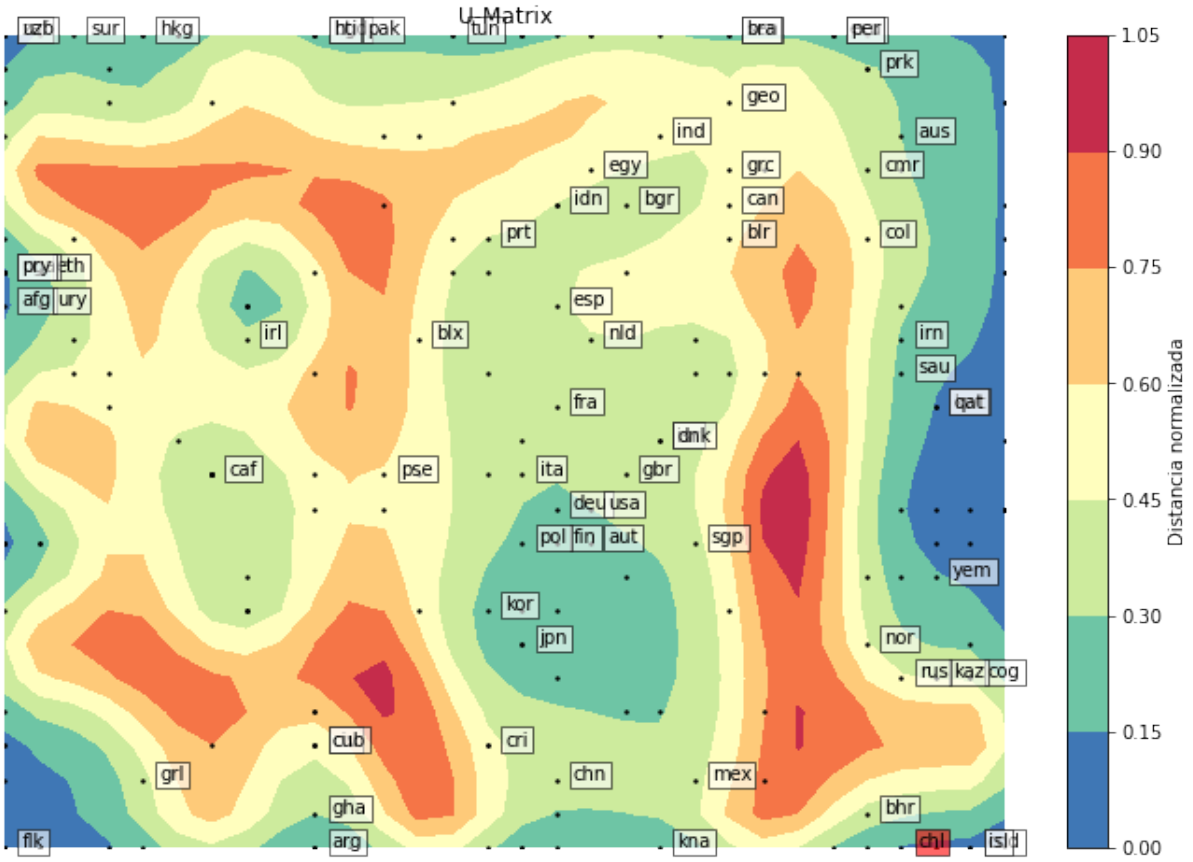

Figura 5: Visualización U-Matrix PCA Kernel gaussiano

La topología de la grilla varía bastante, esta vez las barreras de distancias se concentran en las esquinas derecha e izquierda, mientras que en la mitad se encuentran distancias menores, de igual forma existe un cumulo denso en la esquina extrema derecha perteneciente al cluster donde se encuentra yem.

2.6.

Dados todos los clusters es posible visualizar la semejanza de chl con Islandia, Baréin y la República Democrática del Congo, para los países latinoamericanos no es posible hayar clusters,no así con los países árabes o oriente medio, esto se debe a que existe una mayor correlación en las materias de exportación, donde el petroleo juega un papel importante en toda el área geográfica, en cambio latinoamerica extrae diversos recursos que varían según el área que es mucho más diverso.

2.7.

México se parece a San Cristobal y Nieves, en algunos clusters China y en otros, España. De acuerdo a ranking en PIB es el número 15, por lo que se puede hablar de desarrollo, México exporta

maquinarias en sumayor parte, mientras que Chile exporta recursos naturales tales como; cobre, pasta de madera y frutas. Siendo muy diferentes sus matrices productivas, la de Chile menos elaboración y diversa y la de México más centrada e industrializada.

Ambos están juntos en clusters y muy separados del resto, ambas matrices productivas se desempeñan en recursos marítimos como pescados y crustaceos.

3. Programación

Se pide lo siguiente:
Visualize la base de datos (agrupando los productos en las 15 categorías) utilizando la técnica t-SNE. Utilize la implementación de t-SNE disponible en sklearn. Ajuste la tasa de aprendizaje de tal manera que la proyección se vea razonable. Analize el resultado obtenido y compare con los experimentos anteriores. Evalúe el impacto de modificar el parámetro perplexity del algoritmo probando un valor pequeño, intermedio y grande.

Para esto se ejecuta el siguiente código:

```
from sklearn.manifold import TSNE
t_sne = TSNE(n_components=2, perplexity=20, learning_rate=100, random_state=419)
X = t_sne.fit_transform(world_data_scaled)

plt.scatter(X[:, 0], X[:, 1],c=pred_labels/clustering.n_clusters,cmap='Spectral', s=68)
plt.gca().set_aspect('equal', 'datalim')
plt.title('Visualización TSNE perplexity = 20', fontsize=15);
xscale = X[:, 0].max() - X[:, 0].min()
yscale = X[:, 1].max() - X[:, 1].min()
for i in range(N):
    if world_labels_short[i] in countries_subset:
        if world_labels_short[i] == "chl":
            ax.annotate(world_labels_short[i],
                        xy=(X[:, 0]+0.01*xscale, X[:, 1]+0.01*yscale), fontsize=12, color='r',
                        bbox={'facecolor':'white', 'alpha':100, 'pad':2})
        else:
            ax.annotate(world_labels_short[i],
                        xy=(X[:, 0]+0.01*xscale, X[:, 1]+0.01*yscale), fontsize=10,
                        bbox={'facecolor':'white', 'alpha':0.6, 'pad':2})
plt.grid()
```

Donde se varía el valor de perplexity en los valores [1,5,20],siendo este el parametro de vecindad y learning rate debe ser entre 100 y 1000, optando por el minimo, dando como resultado las siguientes visualizaciones:

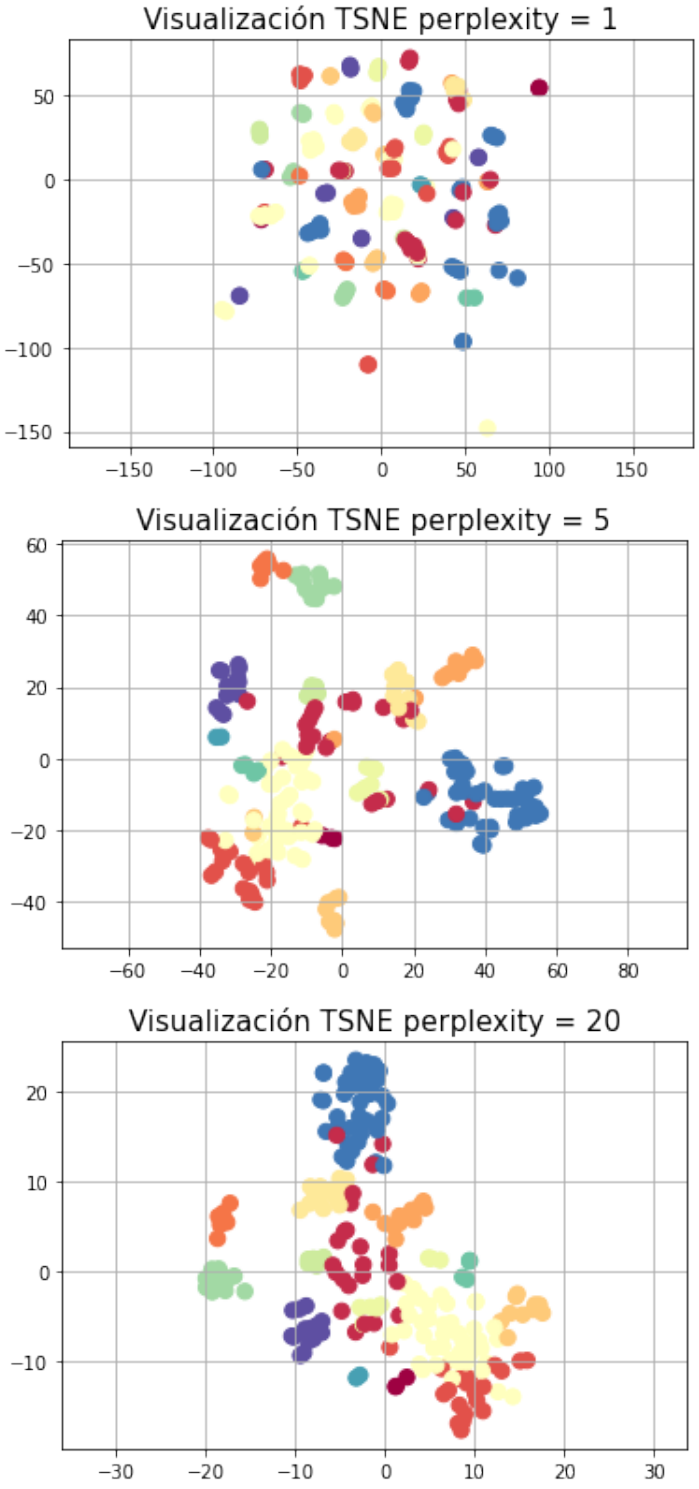

Figura 6: t-SNE variando perplexity

Al realizar t-Distributed Stochastic Neighbor Embedding, se puede visualizar otra forma de clusters, los cuales generan cúmulos, lo que se debe a que trabajan con la vecindad, preservando las

distancias locales, por lo que al escoger perplejidad baja, en este caso 1, se muestra la estructura global y al escoger un número grande en este caso aún preserva algo de la estructura del mapa local, aunque en el caso intermedio se puede notar una mayos distancia intercluster.