Tarea 3 EL4106 SVM y Random Forests

Joaquín Zepeda

November 27, 2021

1 Parte Teórica

1. ¿Qué significa que un algoritmo de aprendizaje de máquinas sea no supervisado?
Esto se refiere a que el conjunto de datos no tiene etiquetas o targets definidos, es decir no se conocen o no se desean utilizar las clases de los datos. En otras palabras, solo se tienen los datos de entrada y no se poseen los datos de salida.

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.

La justificación para esto es que cada componente principal corresponde a un vector propio los cuales se ordenan decreciente. Gracias a las propiedades de los vectores propios son combinaciones lineales de los vectores originales de los datos, pro otro lado se elige el vector propio que tenga asociado el valor propio más alto, con el fin de que el vector propio sea la dirección en donde los datos se encuentren más dispersos y así este pueda representar de mejor manera los datos al reducir la dimensión.

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?

4. Describa brevemente el algoritmo SOM y explique cómo se interpreta la visualización de la U-Matrix.
El algoritmo SOM corresponde a una grilla de puntos o neuronas la cual busca similitudes o correlaciones en los datos de entrada con el fin de ir modelando en cada iteración una estructura interna, es decir se auto organiza con los datos de entrada. Finalmente se lleva al espacio de la grilla los puntos con el fin de posicionarlos en el punto con el cual tenga mayor similitud.

2 Parte Práctica

2.1 Análisis de Componentes principales (PCA)

1. Como se puede observar en la tabla 1, los datos agrupados presentan una mayor varianza explicada por sus componentes principales, esto indica que los datos agrupados capturan un mayor porcentaje de la información original del conjunto. Luego, considerando las varianzas y observando las figuras 1 y 2, es posible observar que la visualización que revela más información es la de los datos agrupados, pues se distinguen mejor los datos y los clusters a los cuales corresponden, además contienen más información (por lo dicho anteriormente sobre las varianzas).
La mayor dificultad al utilizar la información de los 97 productos para describir a cada país corresponde a que al reducir la dimensionalidad de los datos no es posible capturar suficiente información, en efecto como se ve en la tabla 1, se logra capturar menos del 20 % de la información de cada pais. Esto es algo común cuando se trabaja con dimensionalidades muy altas, pues bajar un conjunto de datos de dimensión 97 a dimensión 2 produce que se pierda mucha información.

	Varianza explicada por el primer componente	Varianza explicada por el segundo componente
No agrupados	0.07754024	0.04212973
Agrupados	0.15481744	0.10809297

Table 1: Varianza explicada por los primeros componentes principales con los datos agrupados y no agrupados

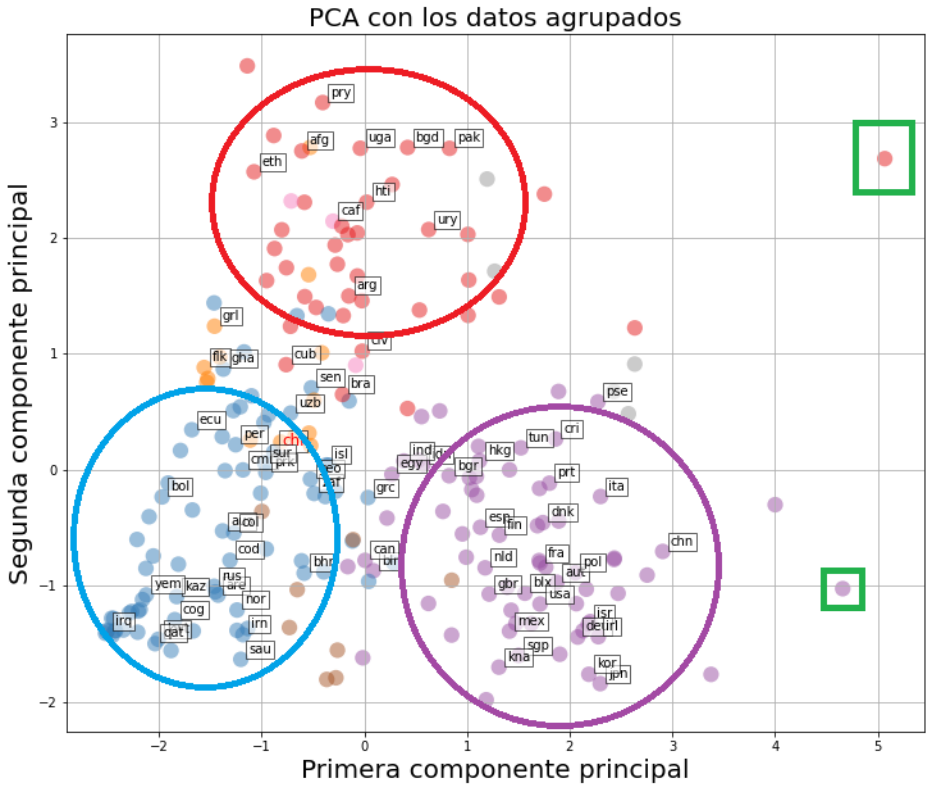

Figure 1: PCA usando los datos agrupados en 15 categorías.

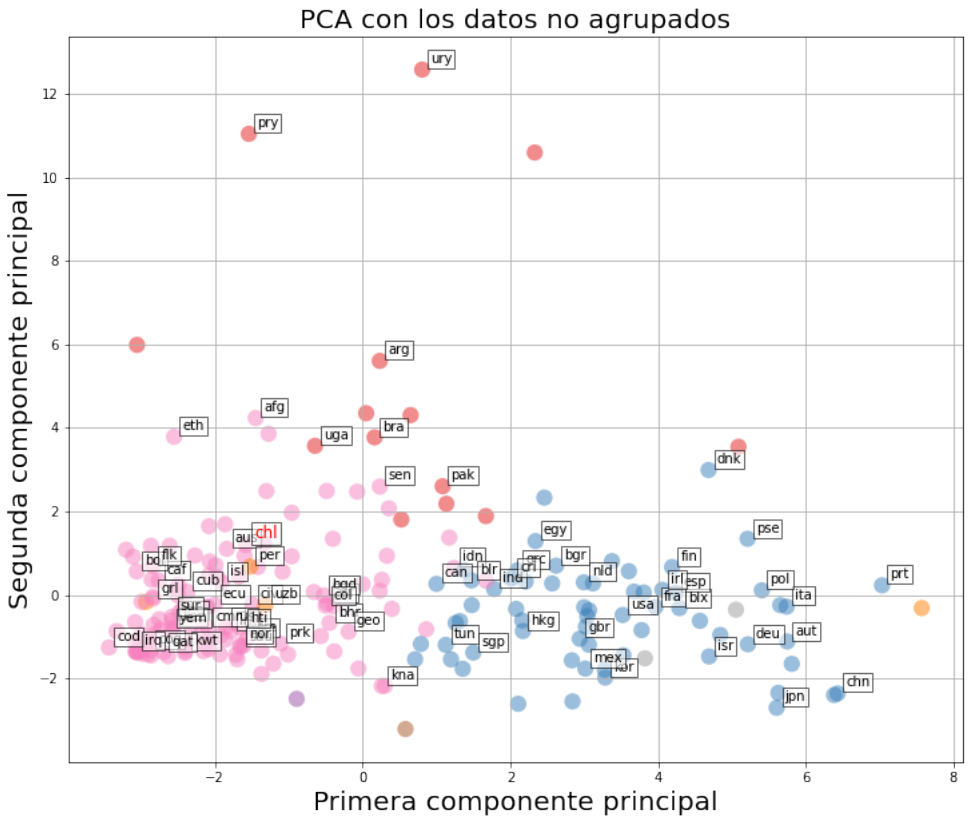

Figure 2: PCA usando los 97 tipos de productos.

2.1.1 Análisis PCA usando los productos agrupados en 15 categorías

Como se puede observar en la figura 1, si es posible observar outliers, estos corresponderían a los datos encerrados en recuadros verdes, son outliers pues están distanciados de su respectivo cluster o región de datos (Cluster rojo y morado).

Al observar la figura 1, fue posible distinguir 3 grandes grupos o cluster y 1 grupo más pequeño entre estos 3. A continuación se describen las características de estos:

1. El grupo marcado por el circulo azul presenta países del medio oriente tales como Rusia, Qatar, irán, Afganistan, Pakistan.
   Hipótesis: La característica común que tienen estos países es la exportación de petroleo.

2. El grupo marcado por el circulo morado presenta países europeos tales como Francia, España, Polonia y además presenta países como México, estados unidos, japón, entre otros.
   Hipótesis: La característica común que tienen estos países puede ser la exportación de autos.

3. El grupo marcado por el circulo rojo presenta países como Central African Republic, Uganda, Haiti, Argentina, entre otros.
   Hipótesis: La característica común que tienen estos países es la exportación de Textiles.

4. El grupo más pequeño, que no esta remarcado pero es representado por los puntos de color amarillo de la figura 1, presenta a Chile, Perú,Falkland Islands (Malvinas), Greenland.
   Hipótesis: La característica común que tienen estos países es la exportación de Productos del mar, frutos y cobre.

Luego se observan las exportaciones de los países con los datos provistos por el observatorio de Complejidad Económica a través de su página https://oec.world/ nos permite concluir lo siguiente:

1. El grupo marcado por el circulo azul efectivamente se caracteriza por exportar petroleo en sus distintas formas, por lo que se confirma esta hipótesis.

2. El grupo marcado por el circulo morado se confirma en forma parcial la hipótesis, pues se caracterizan por exportar no solo autos, si no partes de autos, tecnología y medicamentos.

3. El grupo marcado por el circulo rojo se caracteriza por la exportación de metales, en especifico la exportación de oro, por lo que la hipótesis era incorrecta.

4. El grupo más pequeño, que no esta remarcado pero es representado por los puntos de color amarillo de la figura 1, se caracteriza principalmente por la exportación de metales como el cobre y oro, y en un menor porcentaje por la exportación de frutas y productos marinos, a pesar de esto como igualmente se presentan en las exportaciones se confirma la hipótesis.

2.1.2 Kernel PCA

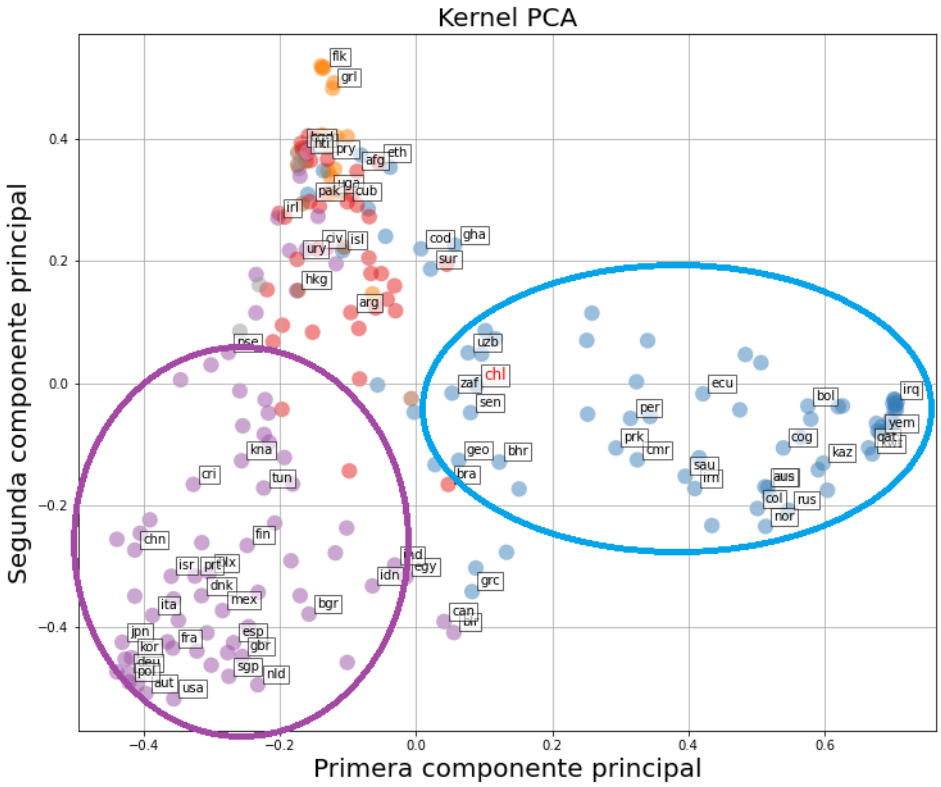

Figure 3: Kernel PCA usando filtro gaussiano.

AL comparar los resultados al utilizar PCA versus Kernel PCA con kernel gaussiano es posible distinguir cambios en la forma en la que se distribuyen los puntos, pero ambos graficos presentan información similar aunque se muevan de posición los paises y en como los países se agrupan, en la figura del kernel PCA 3, se pueden distinguir 3 grupos o clusters de datos. Por un lado se distinguen los paises encerrados en un circulo morado el cual se mantuvo después de realizar el kernel PCA (representa el cluster morado del PCA normal), por otro lado los países encerrados en un circulo azul cambiaron, si bien los países como Rusia, Iraq, Qatar ya estaban presentes en este cluster, se agregaron varios países que antes no estaban en este grupo, acá se considera Chile, South Africa, entre otros. Por otro lado se agruparon 2 tipos de cluster, correspondientes a el cluster de color rojo y amarillo. El resultado del Kernel PCA permite distinguir de mejor manera los diferentes grupos de los datos.

2.2 Mapa auto-organizativo de Kohonen (SOM)

Se eligen 11 componentes para la realización del PCA sin kernel, con esto es posible capturar un 88.3% de la varianza de los datos.

En la figura 4 es posible observar que aparecen clusters nuevos en comparación al PCA y kernel PCA realizado anteriormente, por un lado Brasil, Paraguay y Uganda forman un cluster entre ellos. Por otro lado se observa que se mantiene el cluster de países del medio oriente compuestos por Irac, Qatar pero Rusia se separa en mayor medida de este cluster, cosa que no se apreciaba en los anteriores visualizaciones pero que es razonable pues la economía de Rusia es similar a la de una potencia económica. Con respecto a Chile, este se ve cercano a Islandia pero lejano a Findlandia. Se observa que los Paises Europeos como Francia, italia, Polonia, entre otros estan presentes en un mismo valle en conjunto con Estados unidos y alejandose de ellos se muestran Japon y Corea del Sur, los cuales se presentan más cercanos a Alemania que a los otros paises mencionados. Una diferencia que se observa es que España se muestra lejano a los paises europeos mencionados, cosa que cambia con respecto a lo visto en los PCA anteriores.

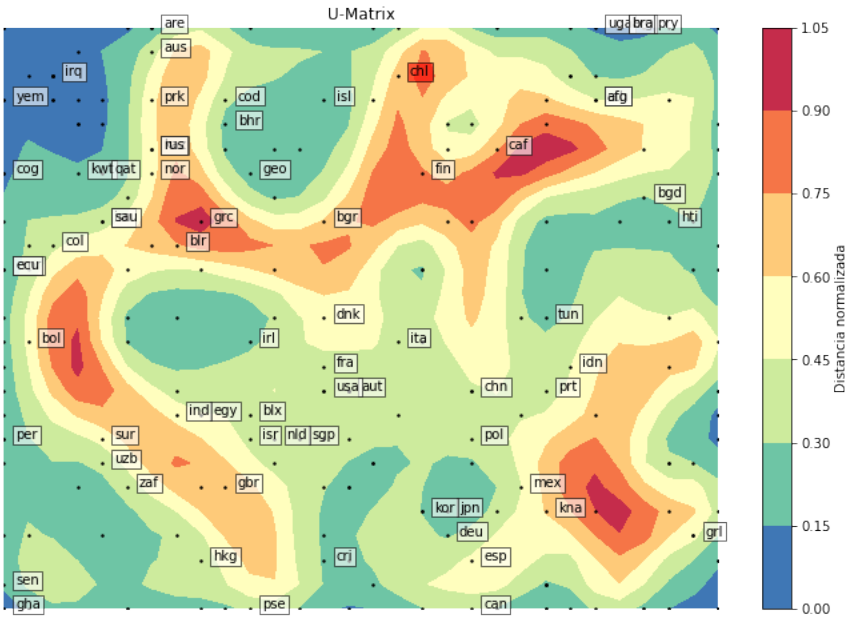

Figure 4: Visualización SOM sobre PCA.

En la figura 5 se puede observar la visualización del mapa de Kohonen sobre el Kernel PCA. En esta figura se observan cambios en las distancias del mapa (que se muestran a través de los colores), por otro lado la distribución de los países cambia, pero los clusters se mantienen relativamente iguales, pero se muestran desplazamientos entre ellos. Chile sigue estando cercano a Islandia, pero ahora se encuentra a una distancia bastante más cercana en comparación con la figura del SOM PCA anterior. Al usar el kernel se observa una disminución en las zonas naranjas y rojas, es decir, una disminución en las zonas con gran distancia (las cuales se pueden pensar como barreras en el mismo mapa) lo cual nos indica que ahora los países están más cerca que en el mapa realizado con el PCA sin kernel.

2.3 Análisis de países

Según lo visto en las visualizaciones anteriores Chile presenta similitud con Islandia, Perú, Sudáfrica, Senegal, entre otros.

Con respecto a si existe un cluster para países latinoamericanos, esto es parcialmente correcto,en la visualización del kernel PCA es posible observar un cluster que cuenta con una gran cantidad de países latinoamericanos, pero hay excepciones como México y Argentina, los cuales corresponden a otros clusters.

Si existe un cluster para países del medio oriente aunque existen algunos paises que son la excepción a esto, en la visualización del kernel PCA es posible observar un cluster que cuenta con una gran cantidad de países del medio oriente, logrando así agruparlos, a pesar de esto este cluster también cuenta

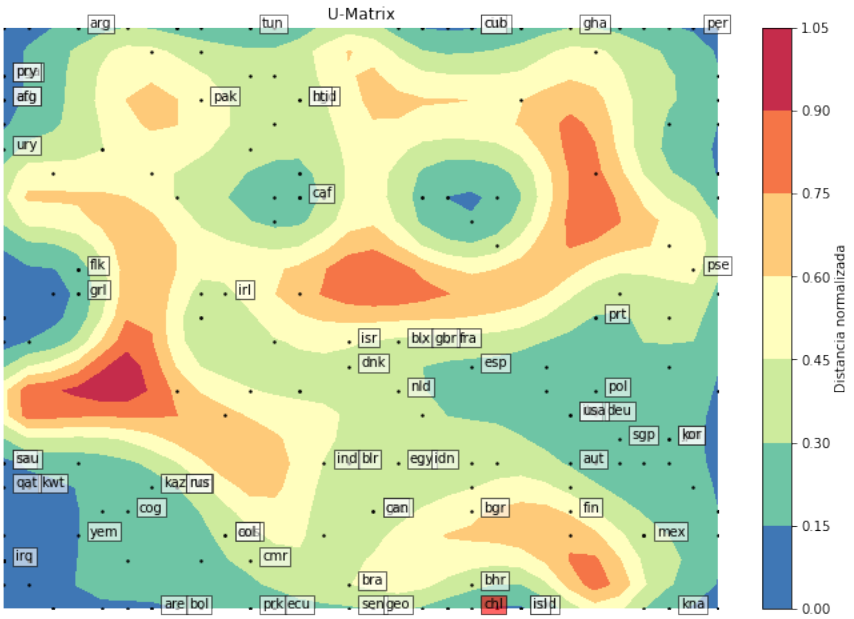

Figure 5: Visualización SOM sobre Kernel PCA.

con Países de otras zonas.

Algunos clusters tienen coherencia con la ubicación geográfica debido a la presencia de recursos naturales por esas zonas, como lo es en la zona de Perú y Chile con el Cobre, por otro lado en Países del medio oriente ocurren cosas similares con respecto al Petroleo, con los Países africanos con le Oro, a pesar de esto hay clusters en donde se exportan elementos industriales y estos van dependiendo de cada País y no con su ubicación geográfica.
Con respecto a México, en las visualizaciones se ve cercano a los Países europeos como España, Francia, Polonia y a Países desarrollados como Japón y Estados Unidos. Según el ranking de PIB 2020, México se encuentra en la posición 63, estando abajo de Chile (puesto 61), por lo que no se considera un país desarrollado, a pesar de esto debido al tipo de exportaciones que realiza México es la razón por la cual pertenece a ese grupo, pues México exporta en gran cantidad Autos, partes de autos y tecnología, por otro lado Chile exporta Cobre, productos del Mar y Frutos, por lo que en cuanto al tipo de elementos que exportan Chile y México difieren bastante.

En la visualización de la figura 1, es posible observar que Groenlandia y las Islas Malvinas se encuentran bastante cerca, esto se debe a que exportan productos del mar, tales como pescados congelados, moluscos, crustáceos, entre ortos.

3 Programación

Al analizar este modelo, es posible observar que se distinguen de mejor manera los clusters vistos anteriormente, al comparar esta figura con las visualizaciones anteriores, es posible concluir que al menos para este ejemplo el uso de TSNE permite obtener una mejor visualización de la distribución de los países (esto usando como referencia el modelo que hacia más sentido, el cual era usando un valor de perplexity medio).

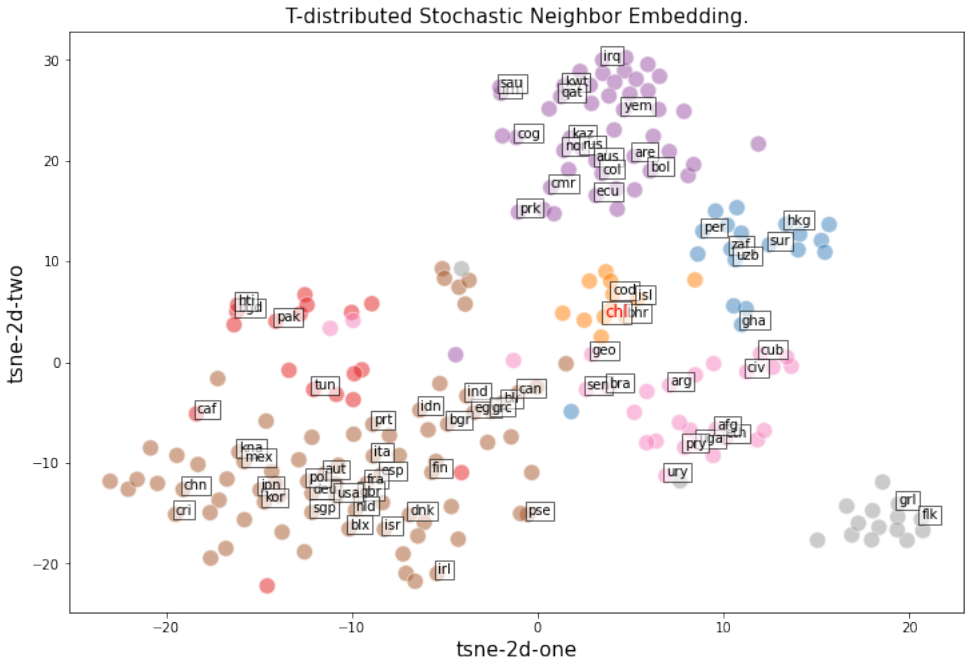

Figure 6: TSNE

3.1 Cambiando el valor del perplexity

Se cambio el valor de perplexity y se observo los diferentes resultados en las figuras 7,8 y 9 para un perplexity bajo (igual a 3), un valor medio de perplexity (de 30) y un valor alto de perplexity (de 50), esto se realizó considerando que esta variable toma valores desde 3 a 50. Este valor es el que configura que tan concentrados estan los datos, como se puede observar en la figura con un bajo perplexity 7, los datos se ven muy concentrados lo cual no permite distinguir de buena manera los clusters, por otro lado, como se puede observar en la figura con un alto perplexity 9, los datos se ven muy dispersos lo cual no permite distinguir de buena manera los clusters. Finalmente con un valor medio de perplexity es posible encontrar

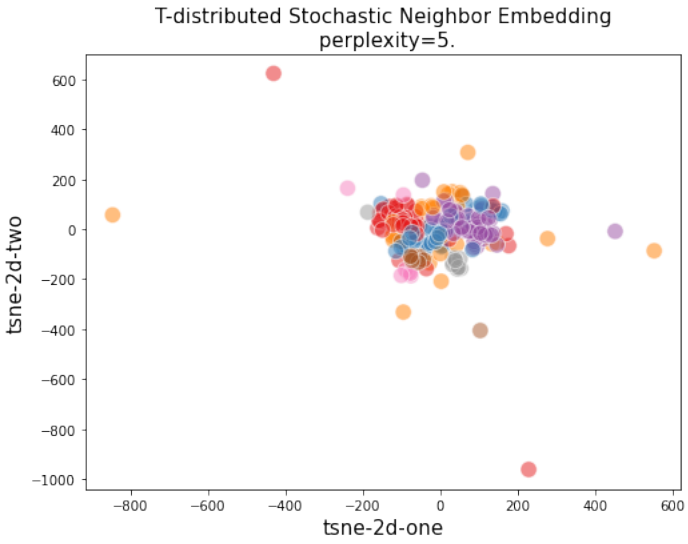

Figure 7: TSNE utilizando perplexity = 3

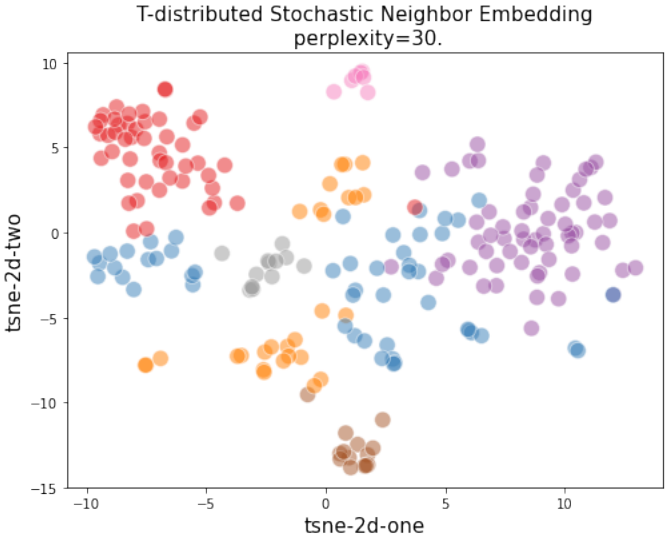

Figure 8: TSNE utilizando perplexity = 30

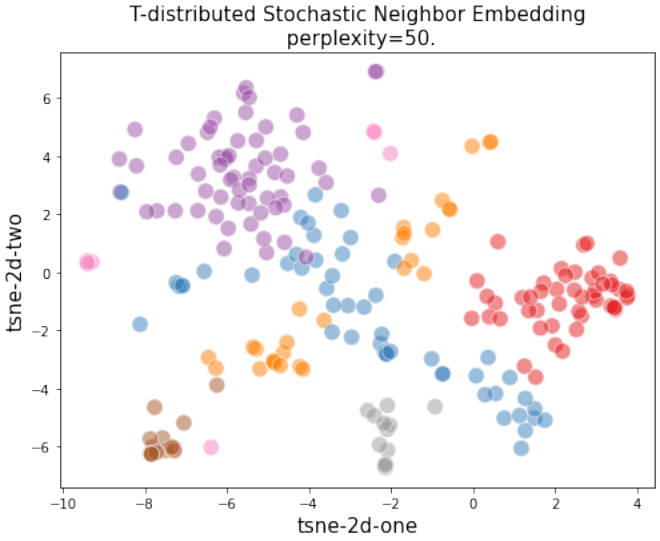

Figure 9: TSNE utilizando perplexity = 50

```
from sklearn.manifold import TSNE
import pandas as pd
import seaborn as sns

#The learning rate for t-SNE is usually in the range [10.0, 1000.0]
#perplexity: Consider selecting a value between 5 and 50.

#Inicializamos el modelo usando una tasa de aprendizaje de 500.
#Para los otros gr ficos se cambian los parametros de la siguiente linea.
tsne = TSNE(learning_rate=500, perplexity=20, n_iter=600)
#Entrenamos con los datos escalados
tsne_digits = tsne.fit_transform(world_data_scaled)


#Creamos un dataframe para guardar los resultados del TSNE
#y manejarlos de forma m s f cil
results= pd.DataFrame()
results['label']=world_labels_short
results['Tsne-2d-one']=tsne_digits[:,0]
results['Tsne-2d-two']=tsne_digits[:,1]
plt.figure(figsize=(12,8))

#graficamos usando un scatterplot con la libreria seaborn
sns.scatterplot(
    x='Tsne-2d-one', y='Tsne-2d-two',
    c=pred_labels/clustering.n_clusters,
    data=results,
    palette=sns.color_palette("hls", 10),
    alpha=0.5,
    s=150,
    cmap='Set1'
)
plt.title('T-distributed Stochastic Neighbor Embedding.',size=15)
plt.xlabel('tsne-2d-one',size=15)
plt.ylabel('tsne-2d-two',size=15)

#Agregamos las etiquetas
for i, txt in enumerate(results['label']):
    if world_labels_short[i] in countries_subset:
        if world_labels_short[i] == "chl":
            plt.text(results['Tsne-2d-one'][i]+0.01,
                     results['Tsne-2d-two'][i]+0.01,
                     s=txt,c='r',
                     bbox={'facecolor':'white', 'alpha':0.6, 'pad':2},
                     fontsize=12)
        else:
            plt.text(results['Tsne-2d-one'][i]+0.01,
                     results['Tsne-2d-two'][i]+ 0.01,
                     s=txt,
                     bbox={'facecolor':'white', 'alpha':0.6, 'pad':2},
                     fontsize=10)
```