

Trabajo Práctico N° 1

Minería de datos

Tecnicatura Universitaria en Inteligencia Artificial

FCEIA - UNR

2do año

Integrantes

Fernández, Florencia

Salvañá, Leandro

1. Introducción.....	2
2. Análisis Exploratorio (EDA).....	3
2.1. Importar el dataset y visualizar el dataframe.....	3
2.2. Conocer las columnas.....	3
2.3. Medidas estadísticas y de localización.....	4
2.4. Visualización de gráficos y matriz de correlación.....	6
2.5. Estandarización.....	24
3. PCA.....	24
4. Isomap.....	30
5. T-SNE.....	39
6. K-Means.....	53
7. Clustering jerárquico.....	62
Conclusiones:.....	69

1. Introducción

Este informe se enfoca en aplicar técnicas de análisis de datos y aprendizaje no supervisado para resolver un problema práctico relacionado con la agricultura y los cultivos. Las actividades incluyen análisis de datos, estandarización, reducción de dimensionalidad con PCA, uso de Isomap y t-SNE, clustering con K-Means, y clustering jerárquico. Estas técnicas se aplican con el objetivo de recomendar cultivos de manera efectiva según las características del suelo y la zona.

La agricultura de precisión (AP) es una estrategia de gestión agrícola basada en observar, medir y responder a la variabilidad temporal y espacial para mejorar la sostenibilidad de la producción agrícola. El objetivo de la investigación en agricultura de precisión es definir un sistema de apoyo a las decisiones.(DSS) para la gestión de toda la finca con el objetivo de optimizar el rendimiento de los insumos y al mismo tiempo preservar los recursos.

Ayuda a los agricultores a tomar decisiones informadas sobre la estrategia agrícola. Aquí, se presenta un conjunto de datos cuyo análisis permitiría a los usuarios tener disponible información ordenada sobre los factores que influyen en los cultivos, como así también construir un modelo para agrupar los cultivos según patrones detectados en función de varios parámetros.

Nota: los resultados del código aplicado se analizan en este informe y también se presentan las tablas y gráficos pertinentes correspondientes a cada sección del trabajo. Se agregan aquí para que las explicaciones dadas sobre los resultados obtenidos estén acompañadas visualmente y sea más sencillo de entender. Además, correr el código insume mucho tiempo ya que la cantidad de gráficos es extensa.

El código utilizado se anexa en un archivo .py y las secciones se encuentran claramente distinguidas en forma de comentarios. Dichas secciones coinciden con las del informe en cada punto del trabajo. Además, hay comentarios en el código que indican cuándo se debe referir al informe para obtener el análisis correspondiente de lo aplicado.

Dicho mensaje es el siguiente:

```
*****  
*****  
"""Observaciones pertinentes redactadas en el informe en la sección 'se  
inserta la sección que corresponda en el informe'"""  
*****  
*****
```

Si desea correr por su cuenta el código, deberá instalar previamente las librerías necesarias en su entorno de trabajo. Las librerías utilizadas están detalladas al inicio del archivo .py.

2. Análisis Exploratorio (EDA)

Se comenzará haciendo un análisis exploratorio del dataset para conocer sus características principales y evaluar si es necesario hacer algún manejo de datos faltantes, outliers, codificar variables categóricas, o algún otro proceso antes de comenzar.

Para esto se hallarán:

1. Medidas de centralidad
2. Cuartiles y percentiles
3. Medidas de dispersión
4. Valores atípicos
5. Matriz de correlación
6. Métricas de distancia y similaridad: distancia de mahalanobis

En esta primera etapa se obtendrá información sobre el dataset y se realizarán gráficos que permitan observar los datos sin realizar modificaciones previas sobre el set de datos, es decir, se analizará el dataset tal como proviene de la fuente.

2.1. Importar el dataset y visualizar el dataframe

Descripción de las columnas

- N: proporción del contenido de nitrógeno en el suelo (mg/kg)
- P: proporción del contenido de fósforo en el suelo (mg/kg)
- K: proporción del contenido de potasio en el suelo (mg/kg)
- temperature: La temperatura en grados Celsius.
- humidity: Porcentaje de humedad registrado.
- ph: Valor de ph medido.
- rainfall: La cantidad de lluvia registrada en mm.
- label: etiqueta con el nombre del alimento.

El dataset posee 2200 filas y 8 columnas correspondientes a variables que describen los datos

El dataset no posee filas duplicadas. Es decir, no existe información redundante.

2.2. Conocer las columnas

En este paso se imprime una lista de las columnas con sus respectivos tipos y cantidad de datos.

Column	Non-Null Count	Dtype
0 N	2200	non-null int64
1 P	2200	non-null int64
2 K	2200	non-null int64
3 temperature	2200	non-null float64

```
4 humidity    2200 non-null float64
5 ph          2200 non-null float64
6 rainfall    2200 non-null float64
7 label       2200 non-null object
```

Se observa que no hay presencia de valores faltantes y que el tipo de dato especificado para cada columna se corresponde con lo que representan. Es decir, las columns N, P y K se corresponden con el tipo de dato numérico discreto, temperature, humidity, ph y rainfall con un valor numérico con decimales y label con texto.

Por lo divisoado, no será requerido completar valores faltantes en el dataset ni realizar un cambio en el tipo de dato de las columnas del mismo.

Valores únicos en la columna 'label':

- rice
- maize
- chickpea
- kidneybeans
- pigeonpeas
- mothbeans
- mungbean
- blackgram
- lentil
- pomegranate
- banana
- mango
- grapes
- watermelon
- muskmelon
- apple
- orange
- papaya
- coconut
- cotton
- jute
- coffe

Existe la misma cantidad de registros para cada alimento. Se analizó si había errores en la entrada de datos por ejemplo si había porcentajes fuera del rango de [0, 100] % o si había valores de pH fuera del rango [0,14]. No se observa presencia de valores fuera de rango.

2.3. Medidas estadísticas y de localización

En este paso se estudian, para cada columna, medidas de localización como mínimo, máximo, cuartiles, y de centralidad como la mediana y la media.

Estas mediciones proporcionan una visión resumida de la distribución de los datos. Esto puede aportar valiosa información para el análisis e interpretación de los datos de suelo y su relación con los alimentos.

Observaciones:

Columna N: proporción del contenido de nitrógeno en el suelo (mg/kg)

Los datos valores mostrados muestran una amplia variabilidad en los datos, con una media de alrededor de 50.55. La presencia de un valor máximo de 140, tan alejado del tercer cuartil puede ser indicador de valores atípicos. Además, hay una considerable dispersión en los datos, como indica la desviación estándar relativamente alta. También, una mediana menor a la media, hace pensar que el gráfico de caja mostrará el lado derecho de la caja más amplio que el izquierdo.

Columna P: proporción del contenido de fósforo en el suelo (mg/kg)

Los datos de la proporción de fósforo también muestran variabilidad, con una media de aproximadamente 53.36. La presencia de un valor mínimo de 5 sugiere que no hay muestras de granos que no hayan tenido proporción de fósforo en el suelo, es decir, no se llega a 0. La desviación estándar indica una dispersión considerable en los datos. El valor máximo está muy alejado del tercer cuartil, lo que es un indicio de presencia de valores atípicos.

Columna K: proporción del contenido de potasio en el suelo (mg/kg)

La desviación estándar es alta en comparación con la media, lo que indica una amplia variabilidad en los datos. La presencia de un valor mínimo de 5 sugiere que no hay valores extremadamente bajos. La discrepancia entre la media y la mediana, junto con la alta desviación estándar, indica que la distribución de los datos es altamente asimétrica y está sesgada hacia los valores más bajos. Además los valores máximo y mínimo y su diferencia con el tercer y el primer cuartil respectivamente, indican una considerable variabilidad y dispersión en los datos, con la presencia de valores atípicos o extremos.

Columna temperature: Temperatura en grados Celsius

La temperatura media y la mediana son muy similares, dando indicio de una distribución simétrica. La desviación estándar es relativamente baja en comparación con la media, lo que sugiere que los datos no se encuentran muy dispersos, sino que están relativamente cerca de la media.

Columna humidity: Porcentaje de humedad registrado

El porcentaje de humedad tiene una media de aproximadamente 71.48. La desviación estándar es considerable, lo que indica una amplia variabilidad en los datos. La presencia de un valor mínimo de 14.26 sugiere que no hay valores extremadamente bajos. Los valores del primer y tercer cuartil indican que se registran mayormente valores elevados de humedad en el suelo.

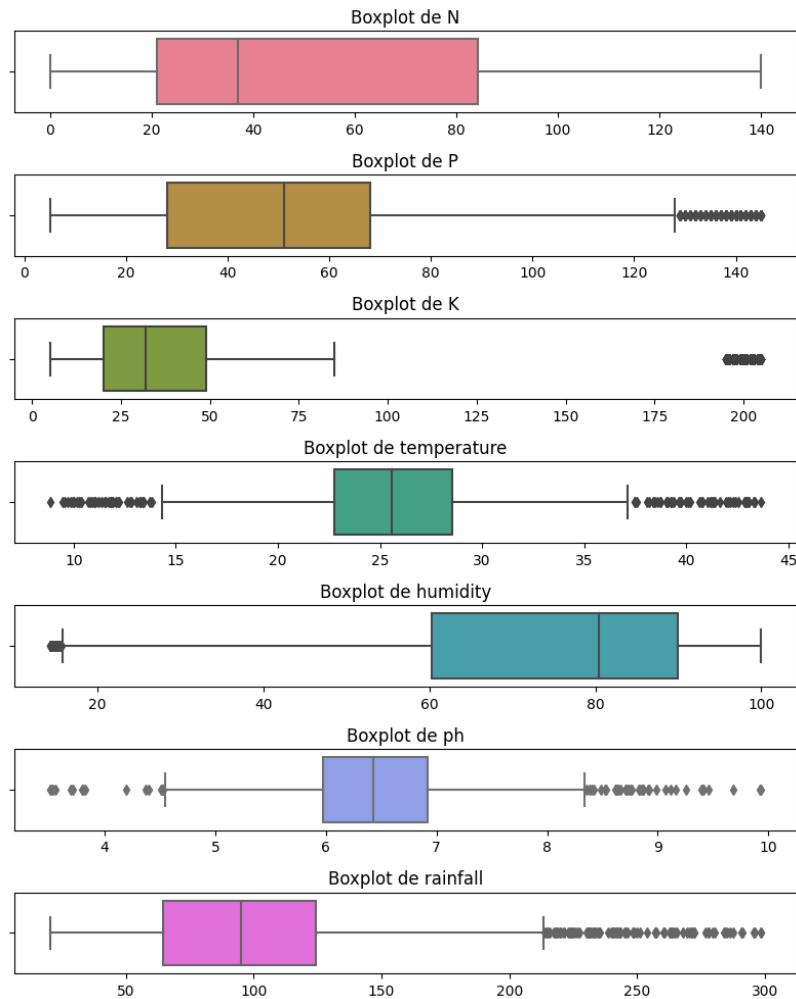
Columna ph: Valor de ph medido

El pH medio y la mediana medidas son casi iguales, indicando que la distribución es simétrica. La desviación estándar es relativamente baja en comparación con la media, lo que sugiere que los datos están relativamente cerca de la media. Los valores mínimos y máximos no se encuentran extremadamente alejados de los cuartiles q1 y q3.

Columna rainfall: Cantidad de lluvia registrada en mm

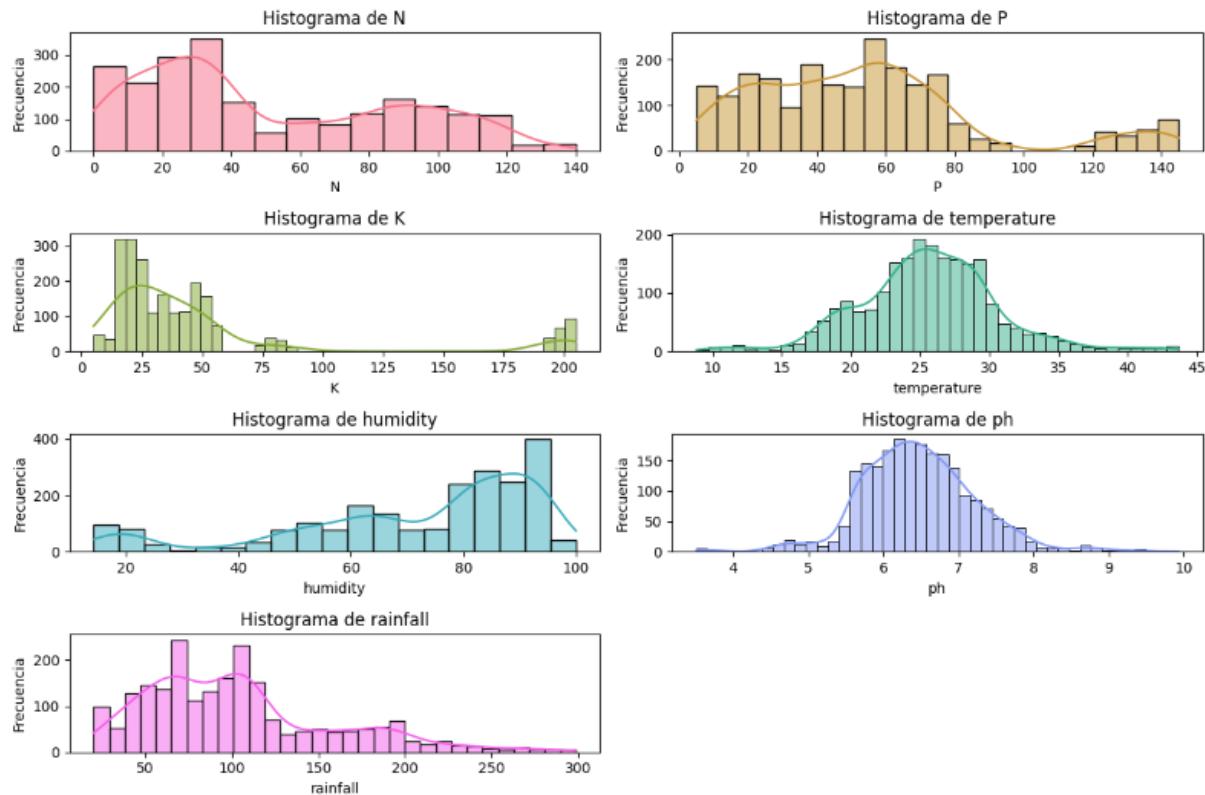
Los datos de cantidad de lluvia están sesgados hacia la izquierda, ya que la mediana es mayor que la media. La amplia variabilidad en los datos, indicada por la alta desviación estándar, sugiere que la cantidad de lluvia puede variar significativamente.

2.4. Visualización de gráficos y matriz de correlación



En la mayoría de las columnas, los gráficos de caja muestran presencia de outliers. No solo detecta valores atípicos, sino que, para las columnas que los tienen se observa una gran cantidad de los mismos, estando en general relativamente cercanos entre ellos.

Teniendo lo anterior en cuenta, se realizará otro gráfico que permita conocer las frecuencias de los valores para cada variable y comprender mejor la distribución de los datos.



Se observa que las distribuciones para los nutrientes K, P y N son similares, junto con la de rainfall.

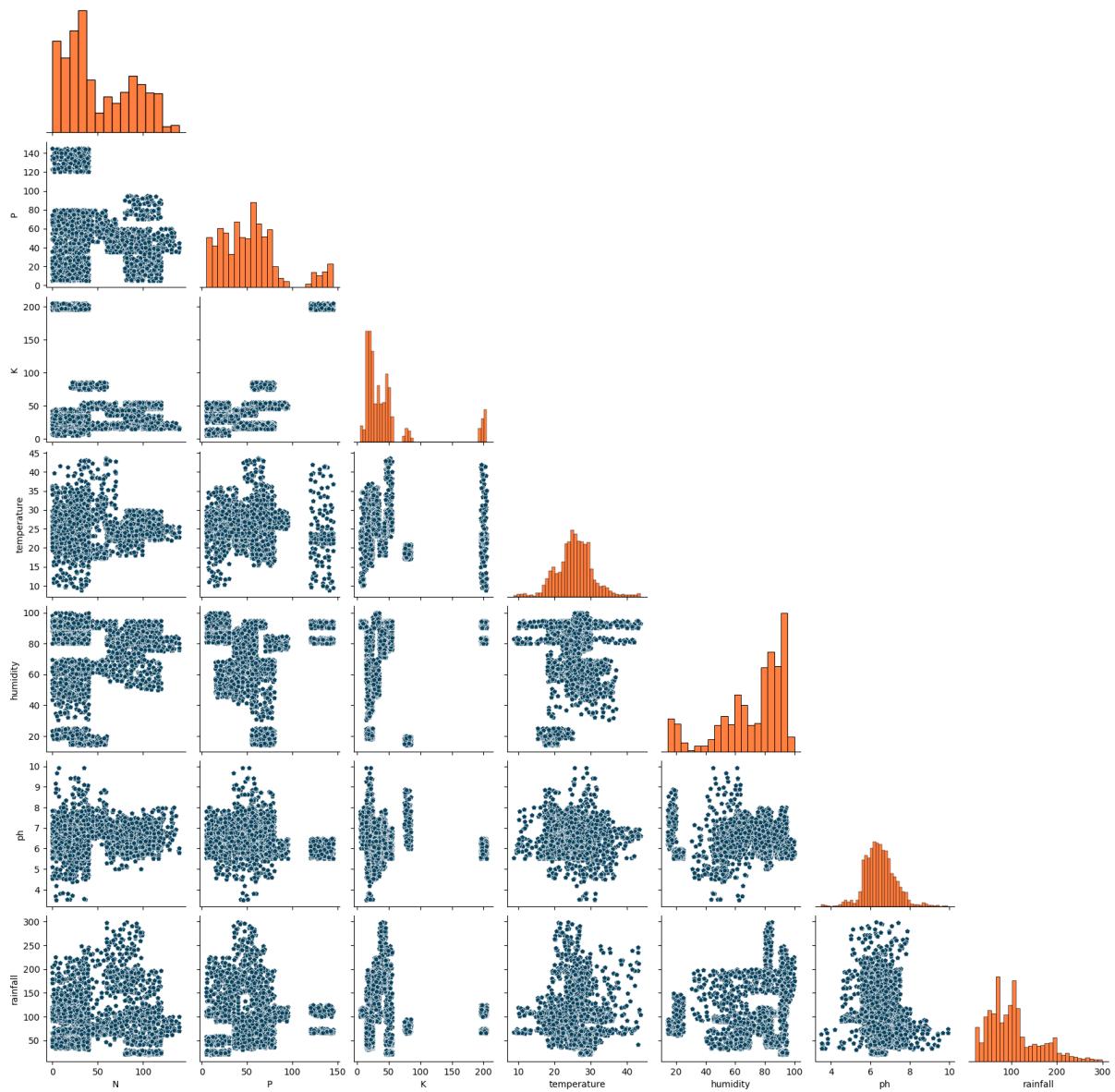
Se destaca además que para dichas variables existe gran amplitud en los valores de los datos y que, como la distribución a simple vista se nota sesgada a la izquierda, los valores que se encuentran en el extremo derecho pueden corresponderse con los atípicos definidos en los boxplot. Otro detalle importante es que la frecuencia de los mencionados valores no es despreciable. Es importante determinar si son valiosos para el análisis.

Además, en vista de las similitudes en distribución que muestran ciertas variables, se procederá a crear la matriz de correlación para las mismas a fin de observar el nivel de correlación lineal entre ellas.



Se observa que no se detectan correlaciones lineales entre las variables. Solo se encuentra una muy alta, en relación a las demás entre los nutrientes fósforo (P) y potasio (K). Dicha correlación es positiva, indicando que a medida que el valor de una aumenta, el de la otra también lo hace.

Se observará la dispersión entre las columnas para conocer cómo se representa ése valor de correlación gráficamente.



Se observa que en la gran mayoría de gráficos, los valores se distribuyen en una especie de bloques cuadrados algunos de los cuales se encuentran cerca, formando una concentración de puntos, mientras otros se observan visiblemente lejos.

A continuación, como los diagramas de caja han mostrado presencia de outliers y, en los gráficos anteriores se encuentran nubes de puntos visiblemente alejados del resto de datos, se procederá a usar la detección de outliers con la distancia de Mahalanobis.

La elección de éste método, se basa en que se han hallado variables con cierta correlación, por lo que se intentará identificar outliers teniendo esto en cuenta.

La distancia de Mahalanobis (MD) es la distancia entre dos puntos en un espacio multivariado . En un espacio euclíadiano regular , las variables (por ejemplo, x, y, z) se representan mediante ejes trazados en ángulo recto entre sí; La distancia entre dos puntos cualesquiera se puede medir con una regla. Para variables no correlacionadas, la distancia

euclíadiana es igual a la MD. Sin embargo, si dos o más variables están correlacionadas , los ejes ya no están en ángulo recto y las mediciones se vuelven imposibles con una regla. Además, si tiene más de tres variables, no puede trazarlas en un espacio 3D regular. El MD resuelve este problema de medición, ya que mide distancias entre puntos, incluso puntos correlacionados para múltiples variables.

El uso más común de la distancia de Mahalanobis es encontrar valores atípicos multivariados , lo que indica combinaciones inusuales de dos o más variables. Por ejemplo, es bastante común encontrar una mujer de 6 pies de altura que pese 185 libras, pero es raro encontrar una mujer de 4 pies de altura que pese tanto.

Cantidad de datos antes de eliminar outliers: 2200

Cantidad de datos después de eliminar outliers: 2173

Se quiere conocer qué valores fueron eliminados para comprobar que no corresponden todos a la misma clase.

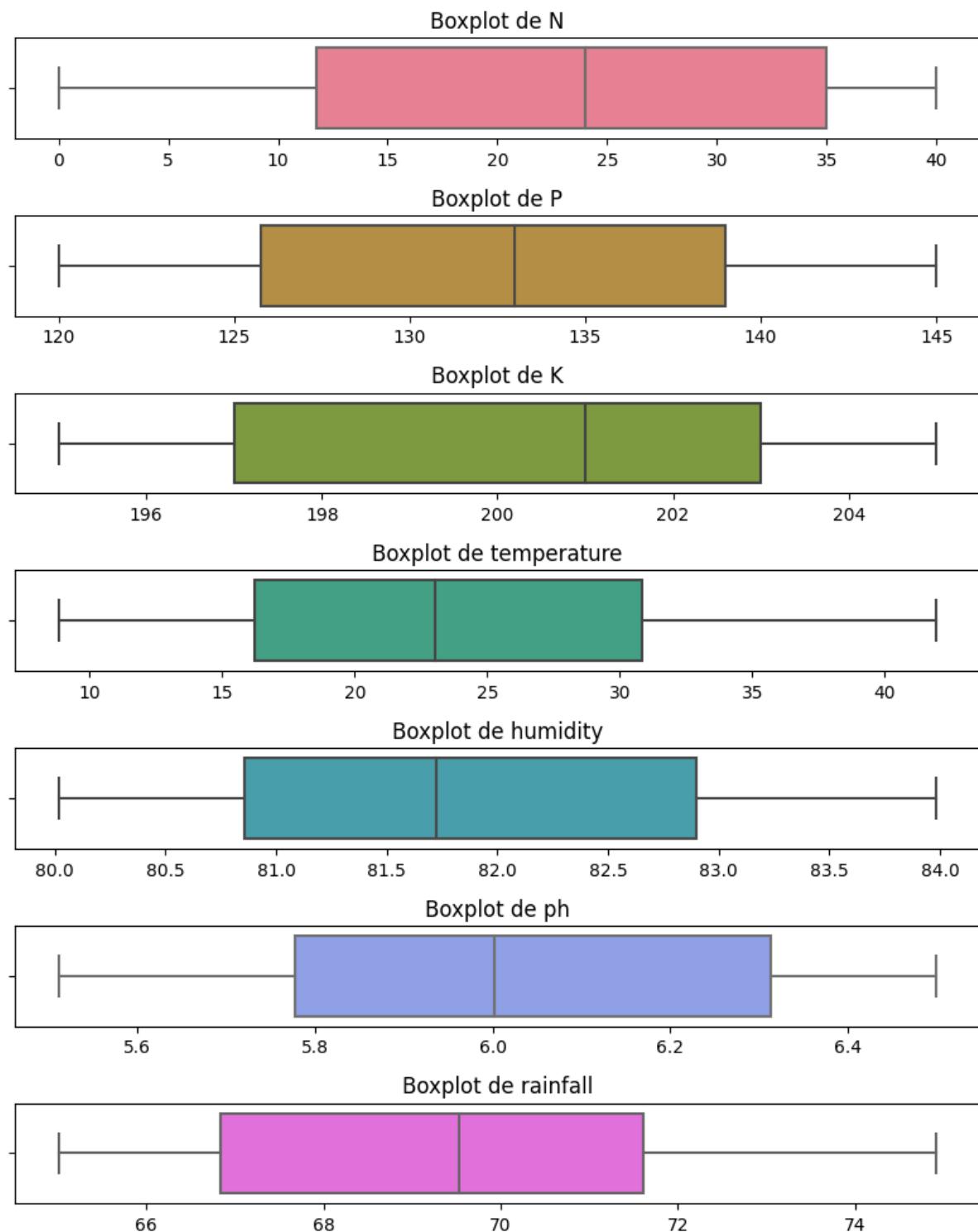
Si esto fuera así, se estaría eliminando gran parte de los datos que corresponden a un único grupo. El tener menos datos puede llevar a perder representatividad y aumentar el sesgo.

Queremos conocer si aquellos registros que se consideran outliers corresponden a una única clase ya que de ser así, tendríamos que considerarlos para el análisis.

	N	P	K	temperature	humidity	ph	rainfall	label
1208	6	123	203	12.756798	81.624974	6.130310	68.778448	grapes
1211	27	145	205	9.467980	82.293355	5.800243	68.027652	grapes
1213	16	139	203	17.828037	80.960934	6.275641	65.847488	grapes
1214	32	141	204	8.825675	82.897537	5.536646	67.235765	grapes
1216	31	144	202	11.021054	80.555572	5.870601	68.239632	grapes
1217	3	136	205	17.586294	80.848066	6.334771	71.406545	grapes
1225	24	140	205	12.087022	83.593987	5.932029	68.668134	grapes
1227	5	126	197	12.800004	81.208784	6.417501	67.104394	grapes
1231	7	126	203	16.762017	82.003356	5.662140	73.287128	grapes
1234	20	142	196	10.898759	80.016394	6.207601	68.694204	grapes
1239	20	122	204	11.797647	80.863254	6.487370	65.069625	grapes
1240	40	126	201	11.363009	80.031000	6.116983	71.182894	grapes
1250	32	120	204	10.380048	83.445181	6.138959	67.391738	grapes
1254	21	134	202	10.723025	80.021306	6.425420	65.298211	grapes
1263	37	135	205	11.827682	80.282719	5.510925	74.102251	grapes
1270	6	140	205	17.665584	82.929034	6.313086	69.887126	grapes
1288	37	144	197	11.189043	80.808431	6.415556	66.342349	grapes
1291	14	121	203	9.724458	83.747858	6.158889	74.464111	grapes
1293	32	138	197	9.535586	80.731127	5.908724	69.441152	grapes
1294	11	124	204	13.429886	80.066340	6.361141	71.400430	grapes
1295	23	138	200	9.851243	80.226317	5.965379	68.428024	grapes
1299	35	134	204	9.049929	82.551390	5.841138	66.008176	grapes

Considerando que casi la totalidad de los datos eliminados en la consideración de outliers son grapes se procederá a conocer si efectivamente corresponden a valores outliers dentro de esta clase específica o si los datos eliminados integran el conjunto de observaciones a analizar.

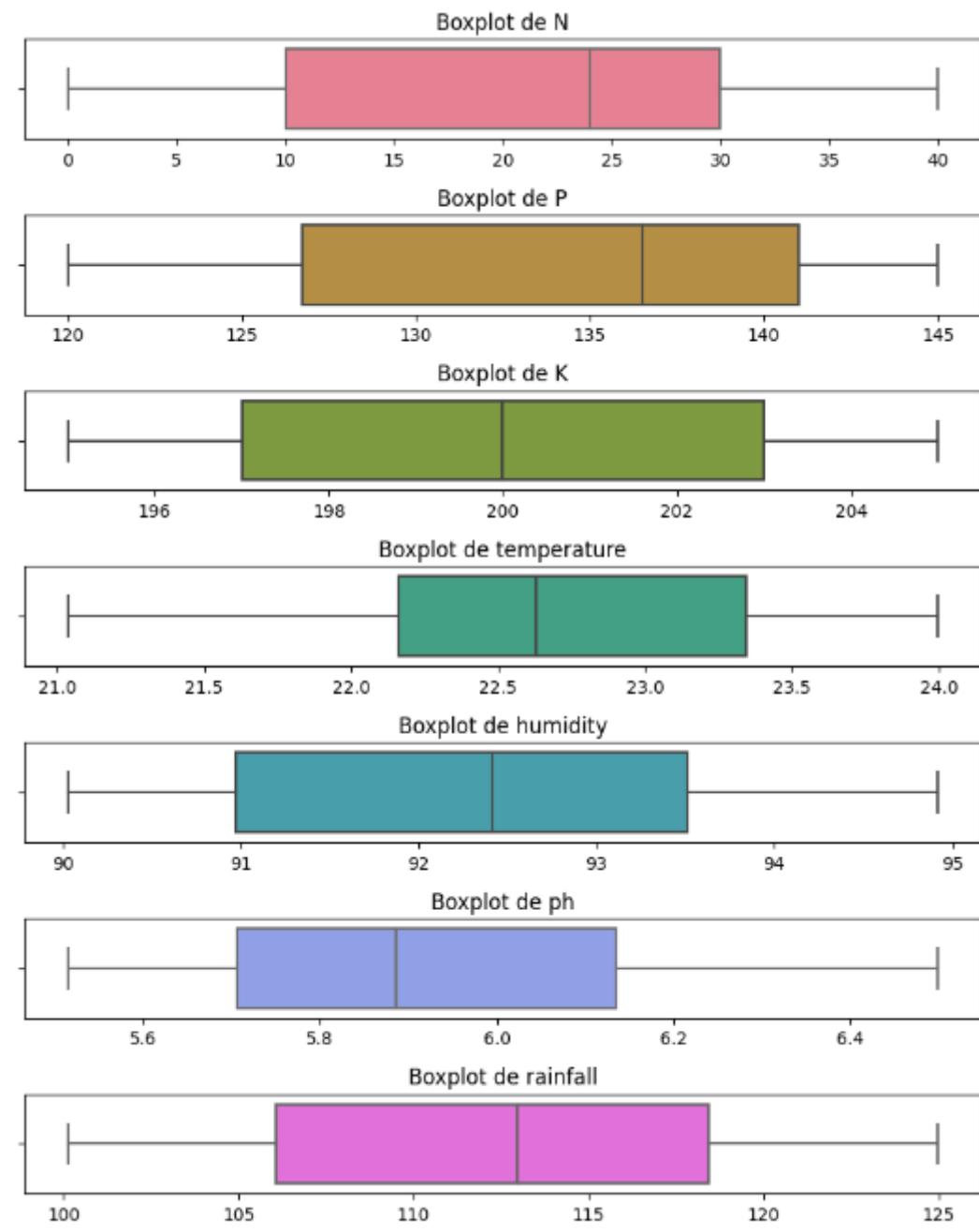
Para ello se realizará un boxplot únicamente para la clase grapes.



Cantidad total de filas con el label grape en el dataset original: 100

Cantidad de filas con el label grapes eliminadas como outliers: 25

Se repiten las observaciones para la clase apple, para la que se habían eliminado dos registros

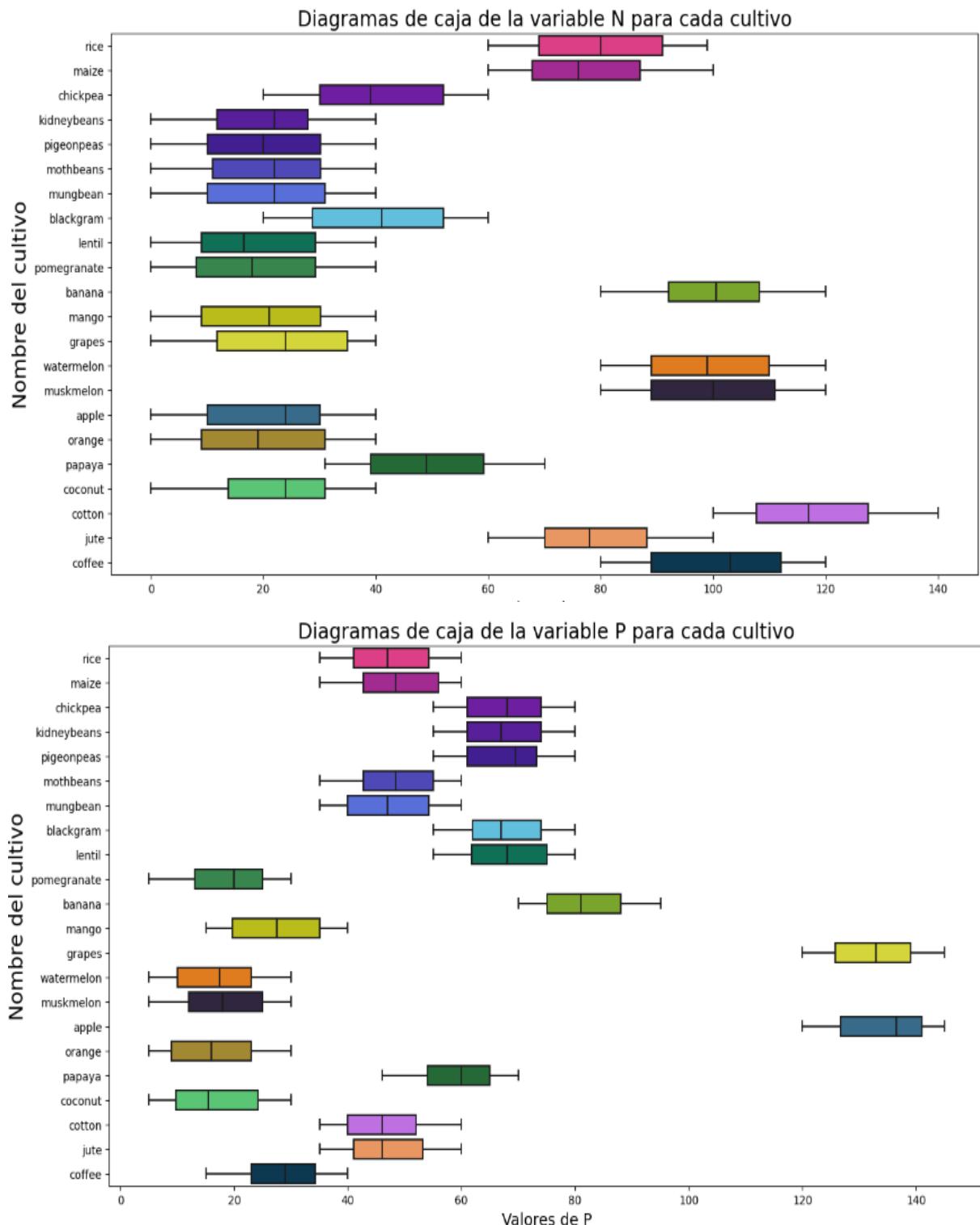


Cantidad total de filas con el label grape en el dataset original: 100

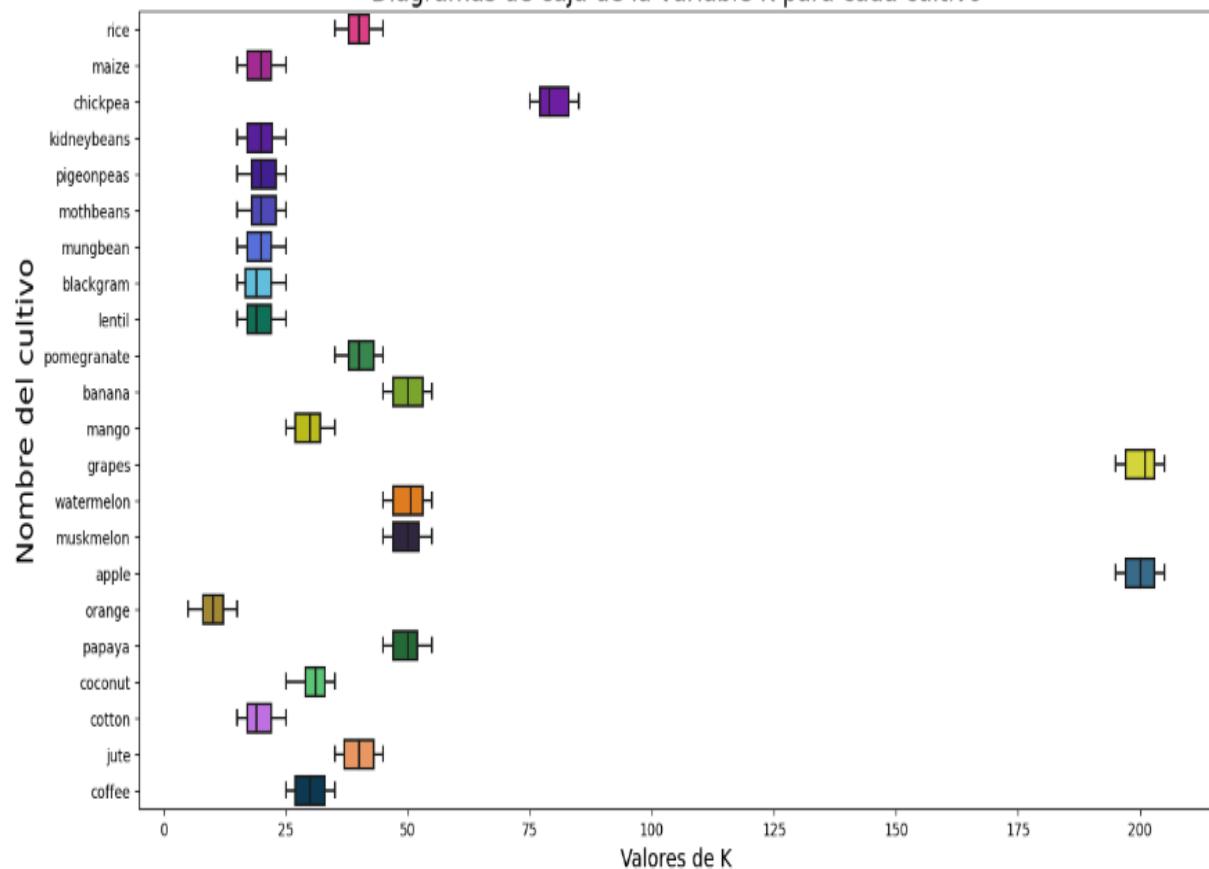
Cantidad de filas con el label apple eliminadas como outliers: 2

En resumen, no se eliminarán datos ya que los mismos representan mayormente a un cultivo y eliminarlos podría quitar representatividad.

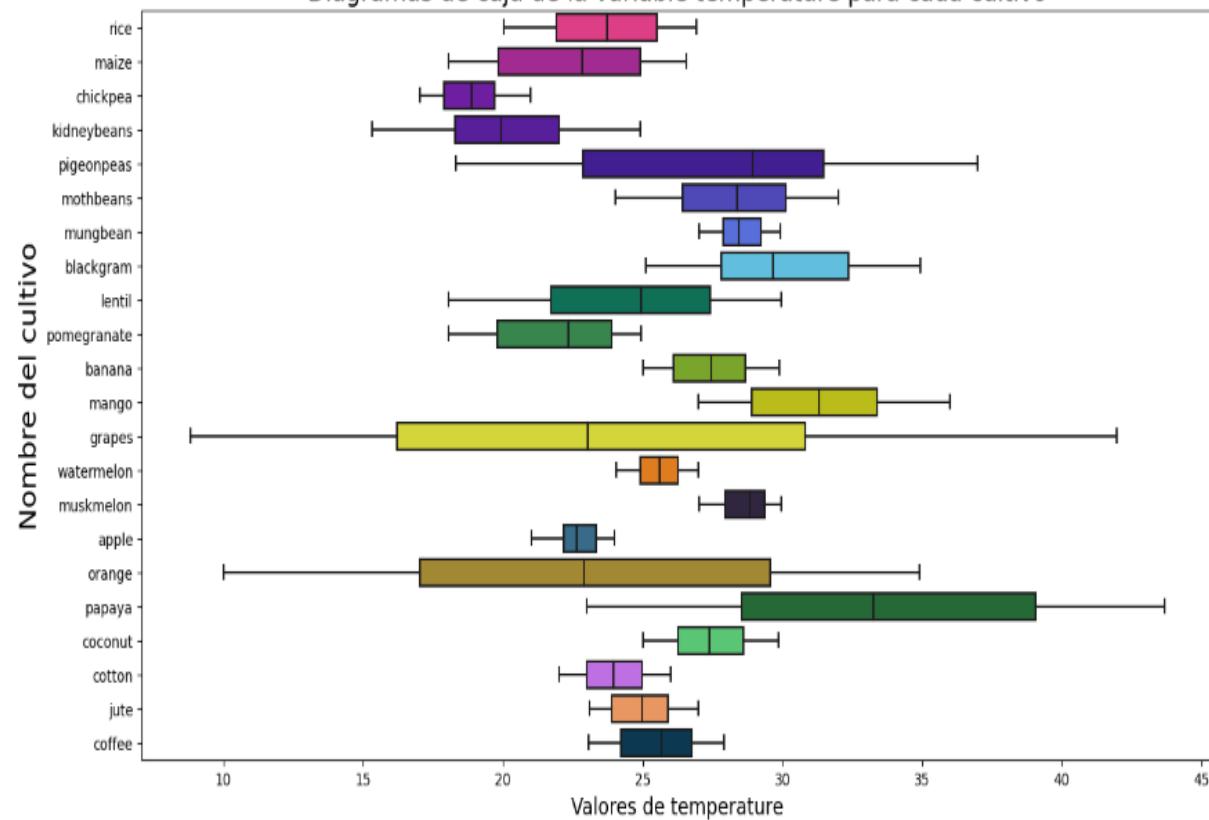
Como lo que se tiene en el dataset proporcionado es, en definitiva, información sobre distintos tipos de cultivos, se procederá a explorar las distintas variables por cada cultivo representado en los datos.



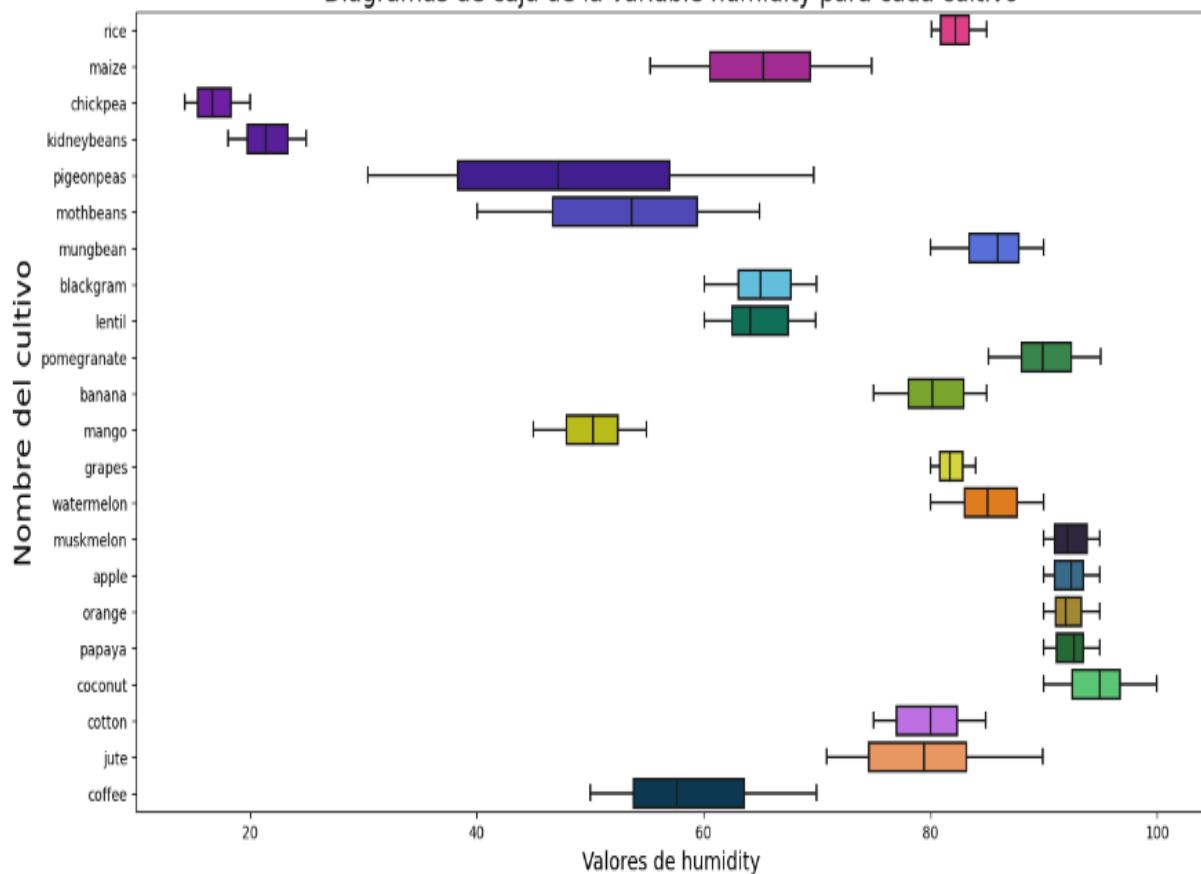
Diagramas de caja de la variable K para cada cultivo



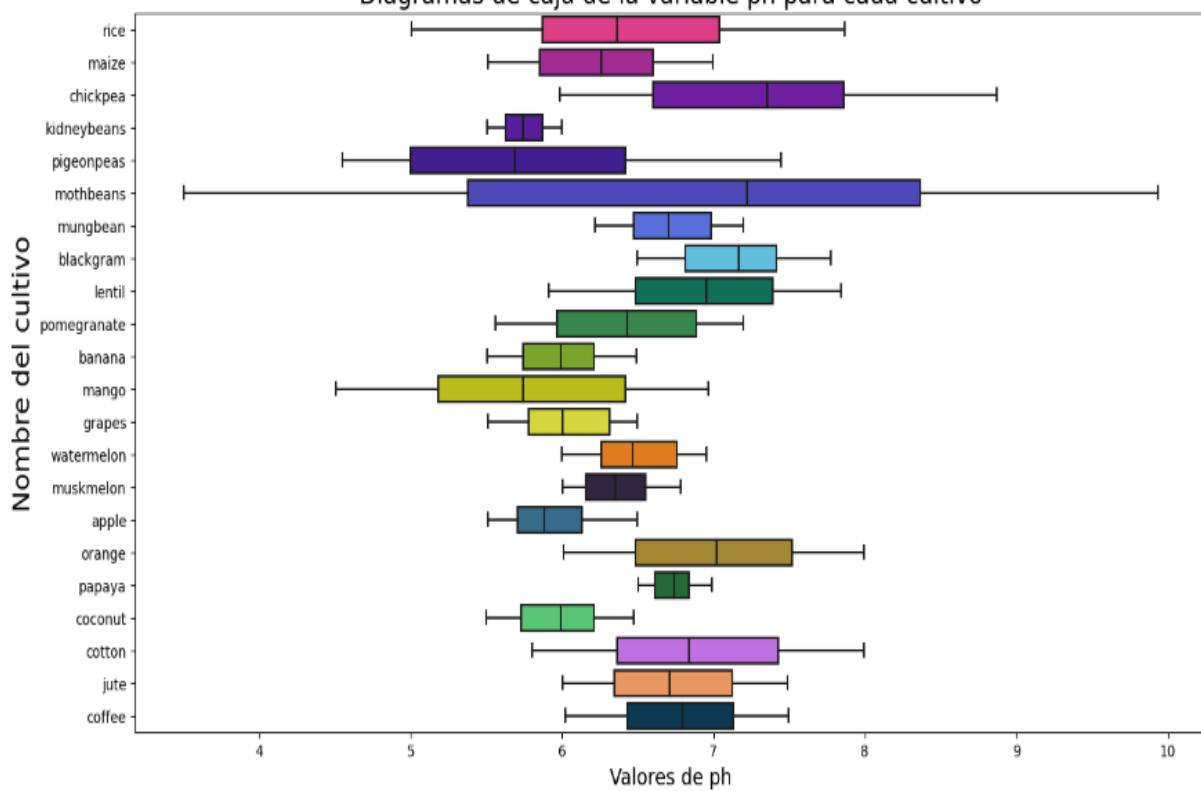
Diagramas de caja de la variable temperature para cada cultivo

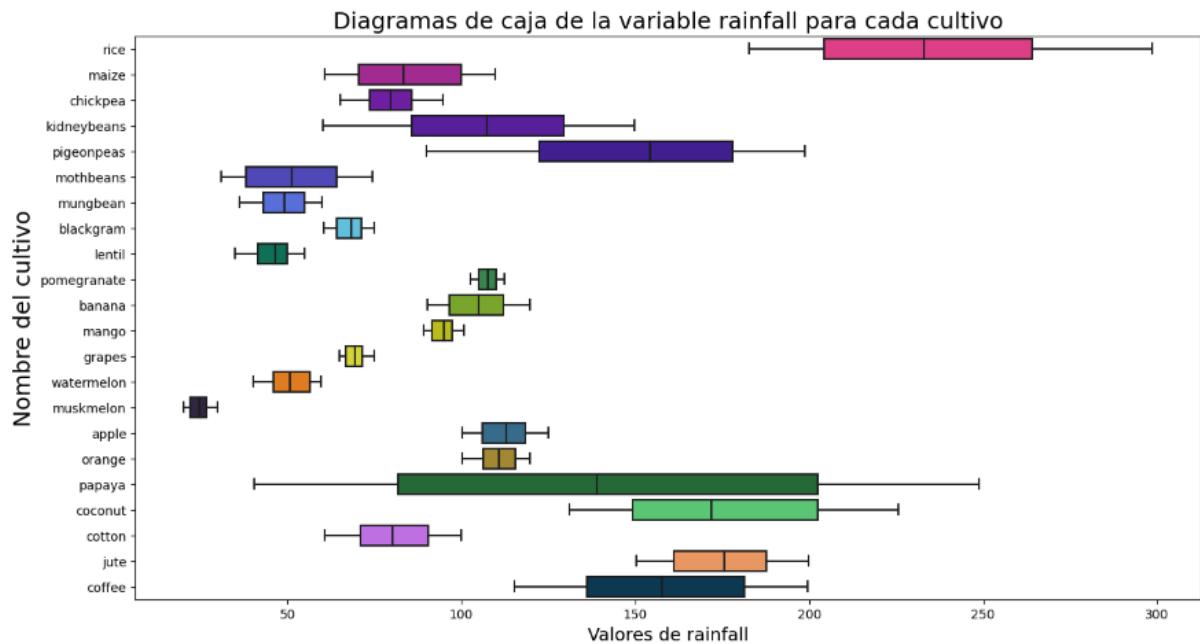


Diagramas de caja de la variable humidity para cada cultivo



Diagramas de caja de la variable ph para cada cultivo



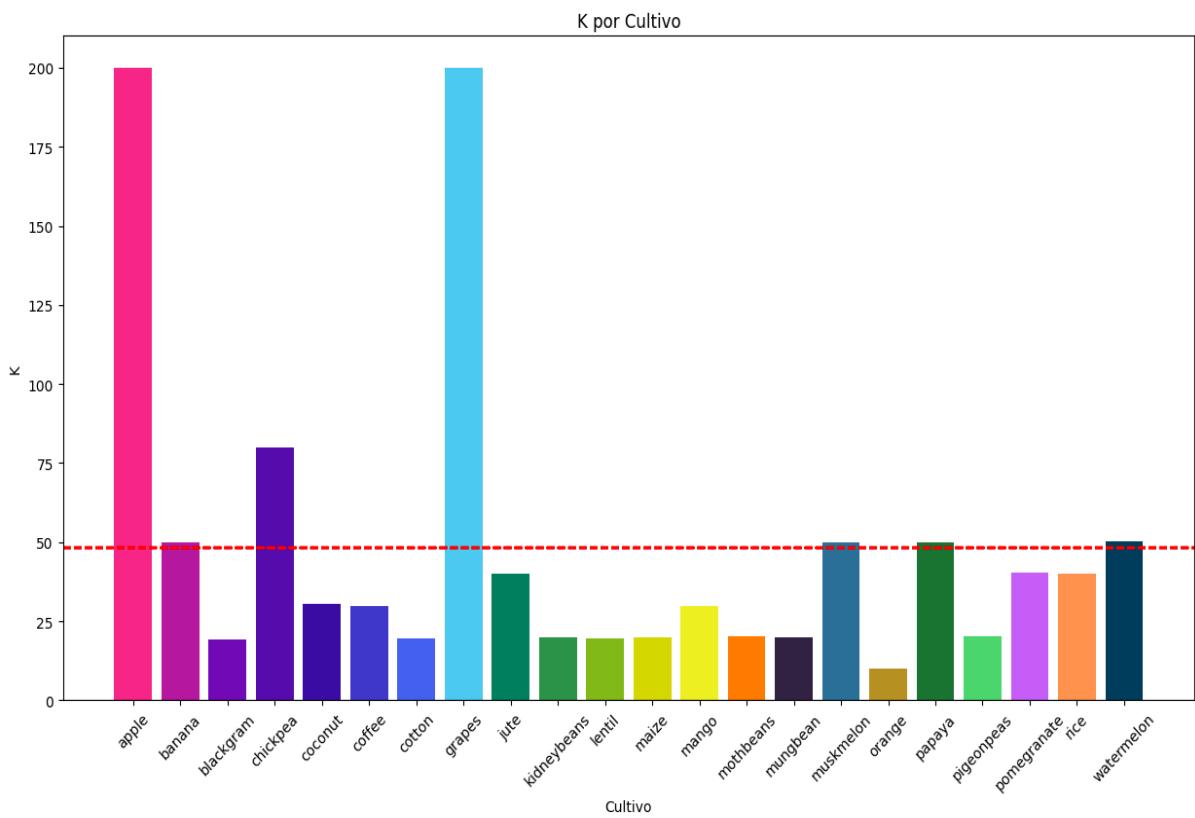
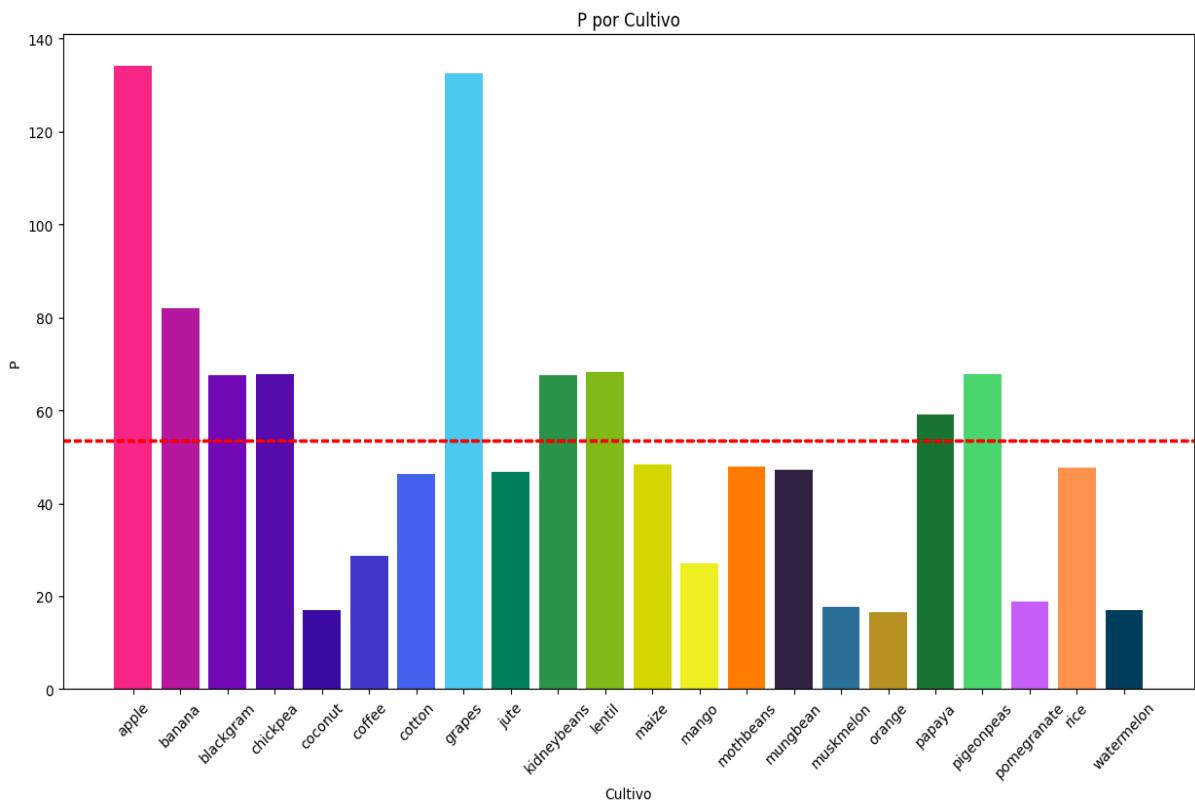


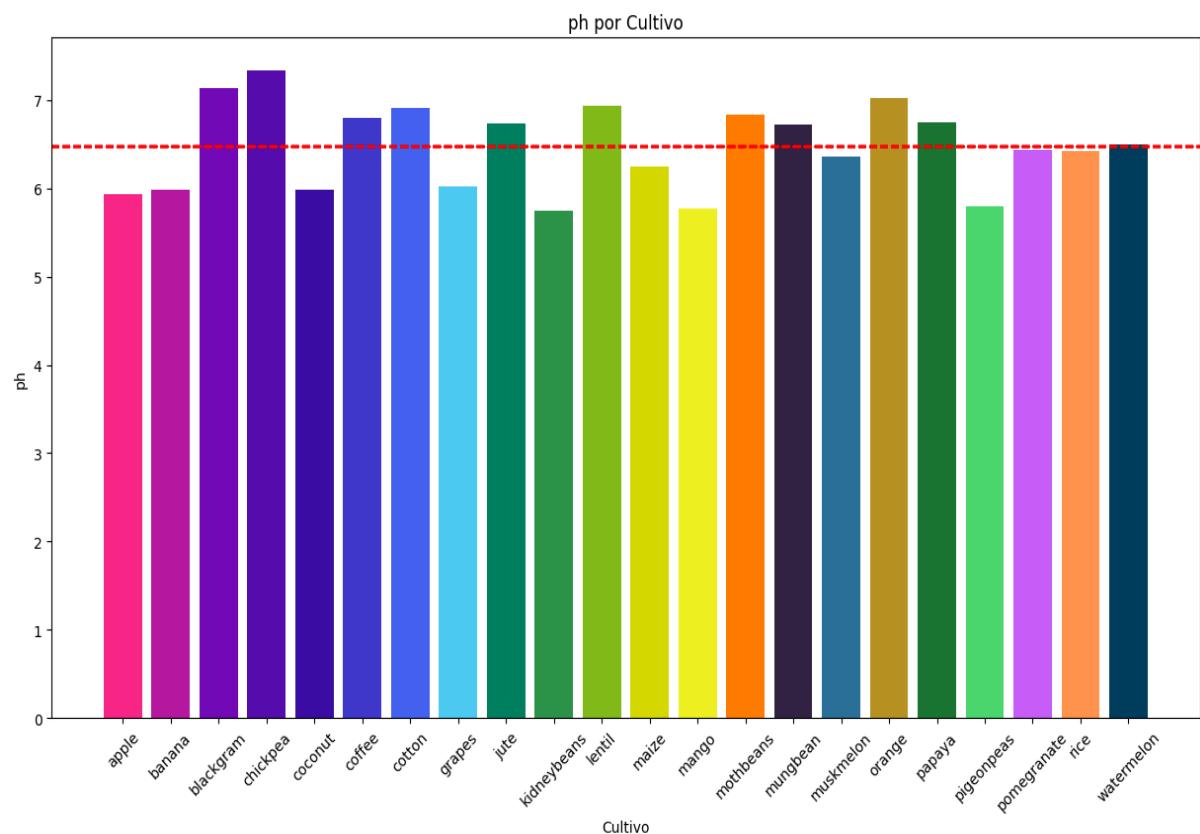
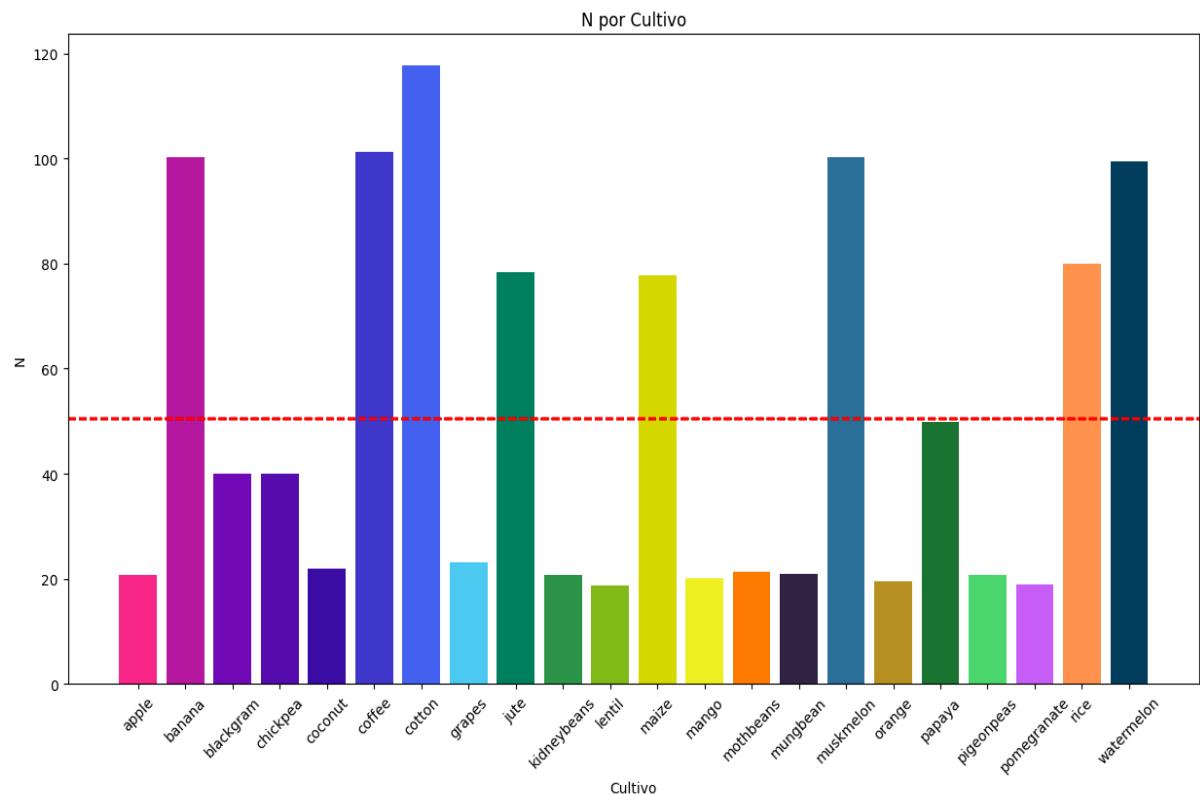
Creación del dataframe con las medidas estadísticas por cultivo: `estadisticas_por_cultivo_df` contiene la media, mediana y desvío estándar para cada columna por cultivo.

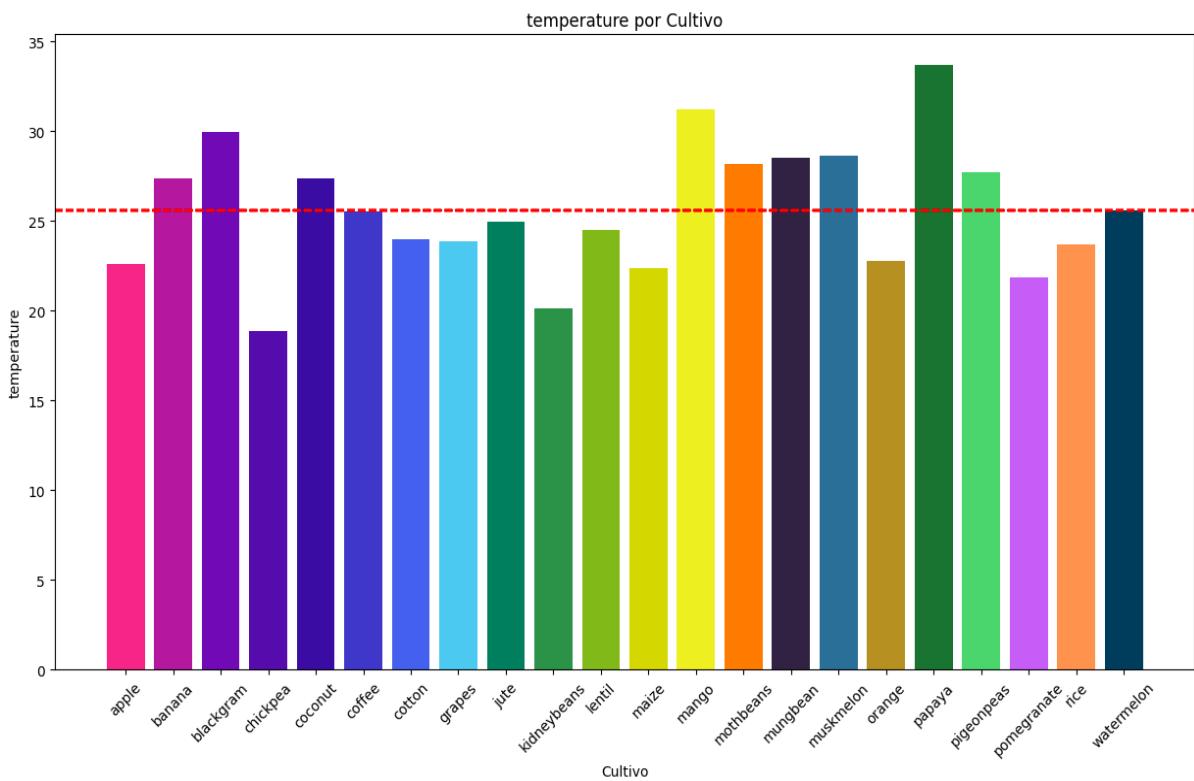
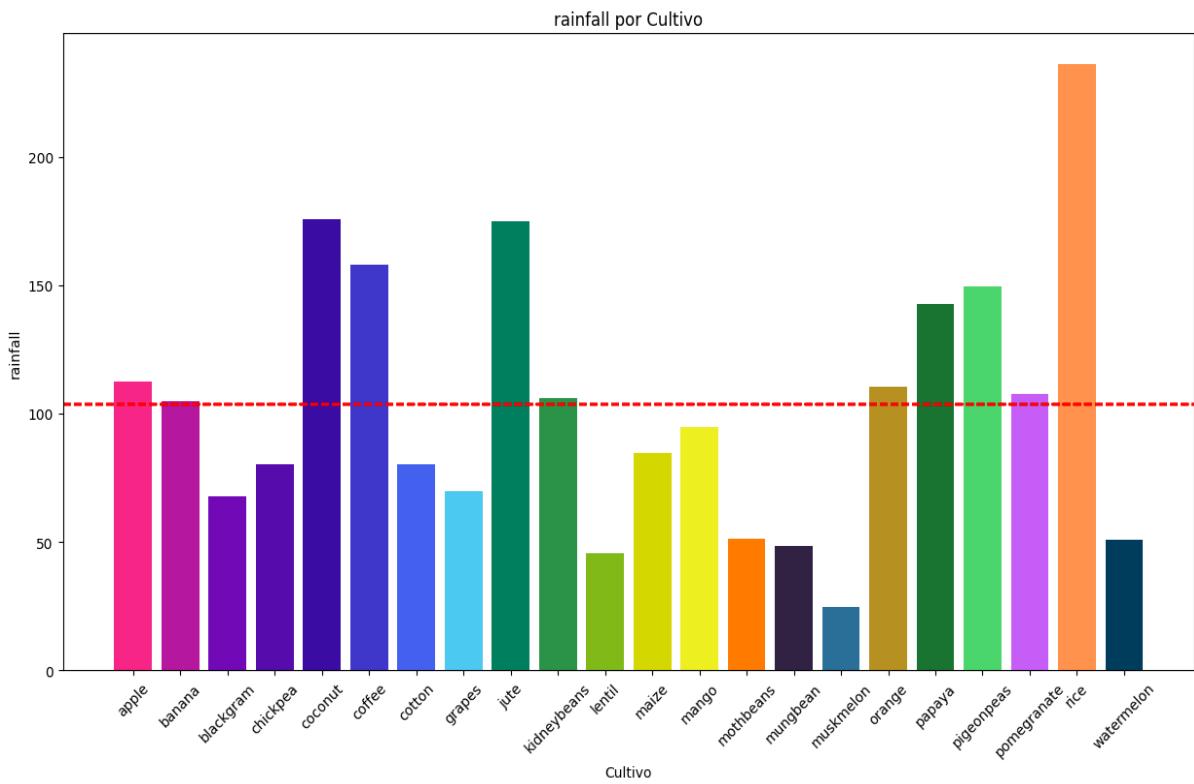
estadistica	label	K	desvio_estandar	media	mediana	desvio_estandar	N \
0	apple	3.320871	199.89	200.0	199.89	11.863784	
1	banana	3.382591	50.05	50.0	50.0	11.107241	
2	blackgram	3.188109	19.24	19.0	19.0	12.664258	
3	chickpea	3.261901	79.92	79.0	79.0	12.150649	
4	coconut	2.998636	30.59	31.0	31.0	11.761931	
5	coffee	3.246817	29.94	30.0	30.0	12.345203	
6	cotton	3.169680	19.56	19.0	19.0	11.628817	
7	grapes	3.265662	200.11	201.0	201.0	12.466829	
8	jute	3.313563	39.99	40.0	40.0	10.968274	
9	kidneybeans	3.102215	28.05	28.0	28.0	10.834266	
10	lentil	2.968164	19.41	19.0	19.0	12.196915	
11	maize	2.941500	19.79	20.0	20.0	11.949490	
12	mango	3.096691	29.92	30.0	30.0	12.329037	
13	mothbeans	3.047950	20.23	20.0	20.0	11.343418	
14	mungbean	3.148368	19.87	20.0	20.0	11.510641	
15	muskmelon	3.218256	50.08	50.0	50.0	12.176215	
16	orange	3.056687	10.01	10.0	10.0	11.941930	
17	papaya	3.097474	50.04	50.0	50.0	12.219607	
18	pigeonpeas	2.815165	20.29	20.0	20.0	11.849950	
19	pomegranate	3.032800	40.21	40.0	40.0	12.617652	
20	rice	2.946167	39.87	40.0	40.0	11.917981	
21	watermelon	3.264687	50.22	50.5	50.5	12.565127	

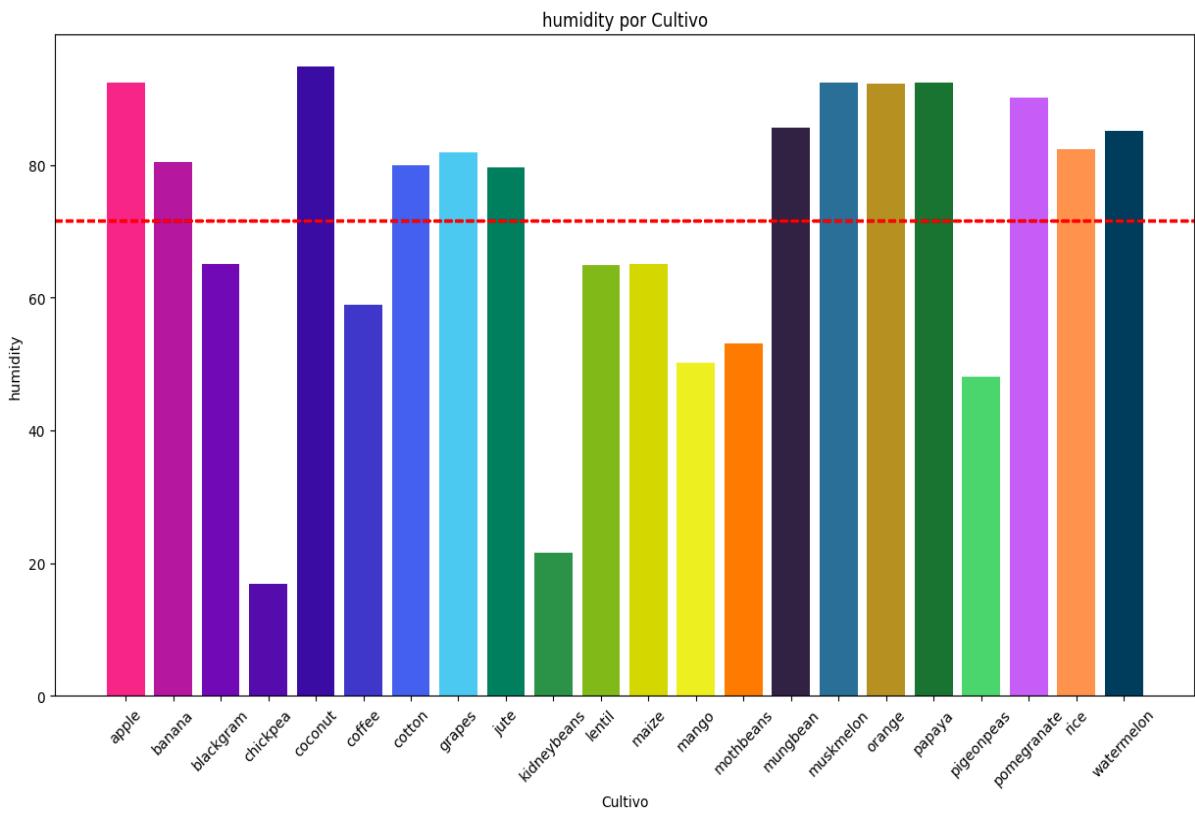
Continua para todas las variables evaluadas.

Se utiliza el dataframe creado y se generan gráficos de barra para visualizar la media por cultivo. Se traza una línea horizontal para la media de todos los datos.









En general, excepto para las columnas ph y temperature son pocos los cultivos que superan el valor de la media ampliamente, mostrando características más alejadas del resto.

- Rainfall: coconut, coffee, jute, papaya, pigeonpeas y rice son los que presentan valores de lluvia que superan por más amplitud la media. Kidneybeans, mungbean, mothbeans, muskmelos, watermelos son los cultivos que menor lluvia registran. Legumbres y frutas con mucho contenido acuoso.
- Humidity: chickpea, kidneybeans y pigeonpeas son los cultivos que menos contenido de humedad registran. Todas legumbres.
- P y K: Apple y grapes son los cultivos que presentan mayores valores para dichos nutrientes. Coconut y orange son de los que menores valores presentan en ambos. Todas frutas.
- N: Apple y grapes son los cultivos que presentan menores valores para dicho nutriente. Coton, coffee, banana, muskmelon y watermelon son los que presentan los valores más elevados para N.

Los cultivos que superan la media de N son: banana coffee cotton jute maize muskmelon rice watermelon

Los cultivos que no superan la media de N son: apple blackgram chickpea coconut grapes kidneybeans lentil mango mothbeans mungbean orange papaya pigeonpeas pomegranate

Los cultivos que superan la media de P son: apple banana blackgram chickpea grapes kidneybeans lentil papaya pigeonpeas

Los cultivos que no superan la media de P son: coconut coffee cotton jute maize mango mothbeans mungbean muskmelon orange pomegranate rice watermelon

Los cultivos que superan la media de K son: apple banana chickpea grapes muskmelon papaya watermelon

Los cultivos que no superan la media de K son: blackgram coconut coffee cotton jute kidneybeans lentil maize mango mothbeans mungbean orange pigeonpeas pomegranate rice

Los cultivos que superan la media de temperature son: banana blackgram coconut mango mothbeans mungbean muskmelon papaya pigeonpeas

Los cultivos que no superan la media de temperature son: apple chickpea coffee cotton grapes jute kidneybeans lentil maize orange pomegranate rice watermelon

Los cultivos que superan la media de humidity son: apple banana coconut cotton grapes jute mungbean muskmelon orange papaya pomegranate rice watermelon

Los cultivos que no superan la media de humidity son: blackgram chickpea coffee kidneybeans lentil maize mango mothbeans pigeonpeas

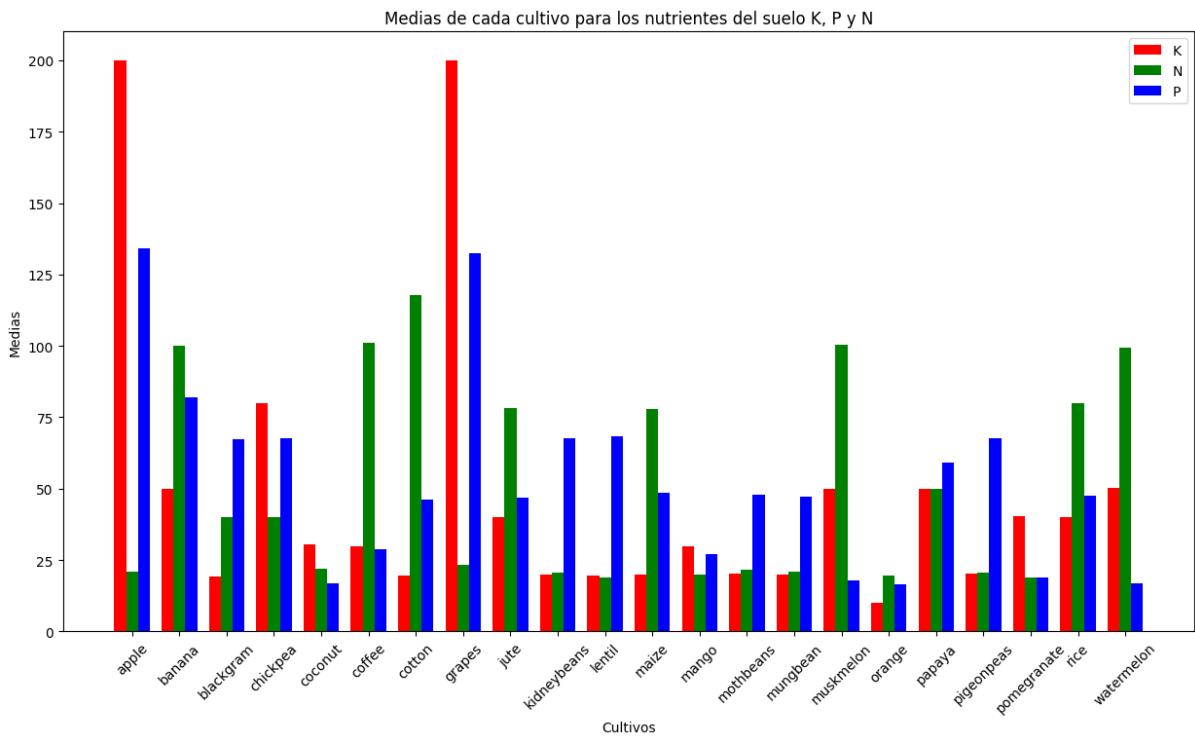
Los cultivos que superan la media de ph son: blackgram chickpea coffee cotton jute lentil mothbeans mungbean orange papaya watermelon

Los cultivos que no superan la media de ph son: apple banana coconut grapes kidneybeans maize mango muskmelon pigeonpeas pomegranate rice

Los cultivos que superan la media de rainfall son: apple banana coconut coffee jute kidneybeans orange papaya pigeonpeas pomegranate rice

Los cultivos que no superan la media de rainfall son: blackgram chickpea cotton grapes lentil maize mango mothbeans mungbean muskmelon watermelon

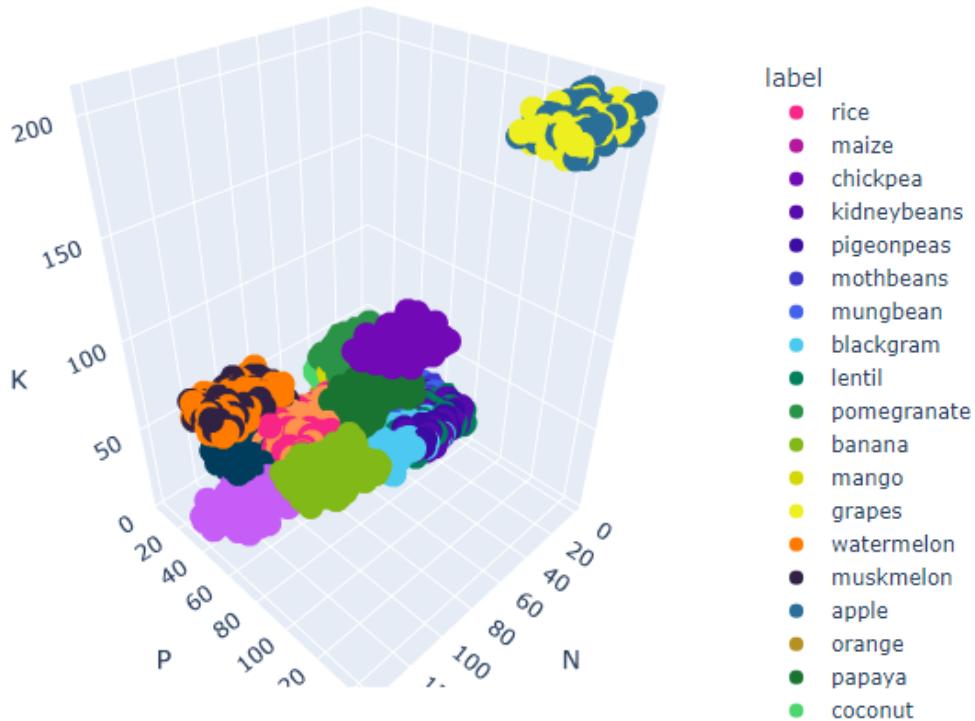
Se hará un gráfico de barras solo para los nutrientes puesto que se observa una distribución similar en ellos.



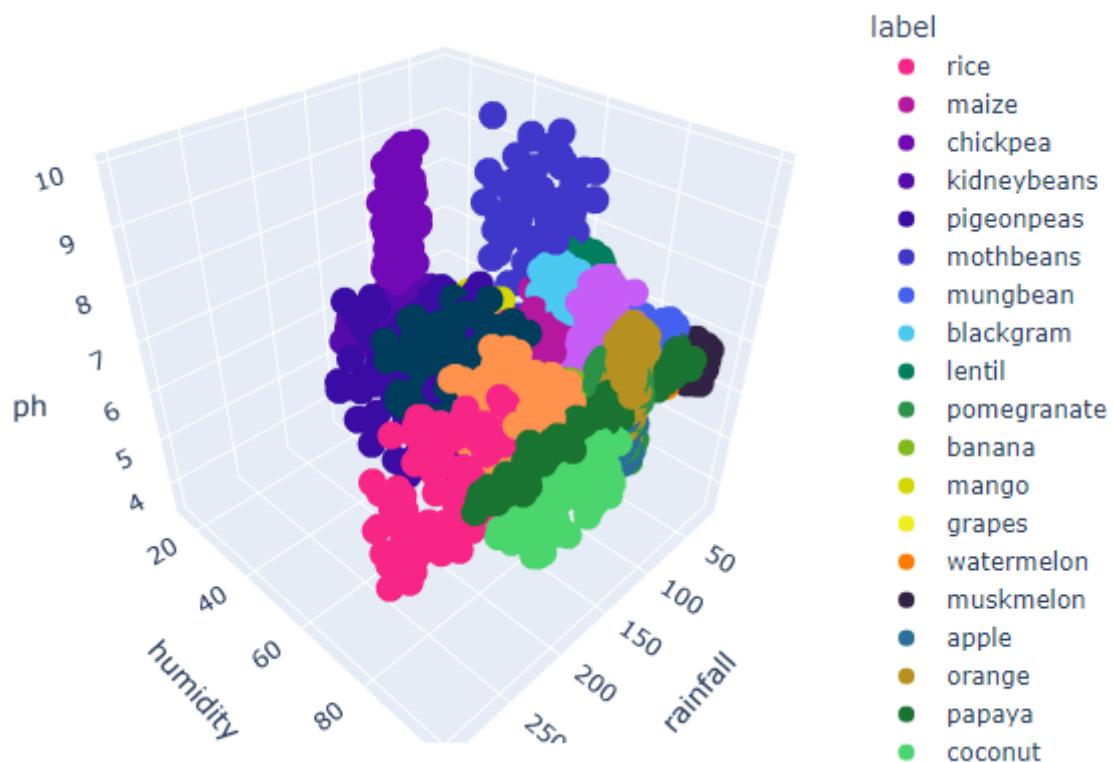
Se observa que los cultivos que mayores valores tienen de nutrientes K, P, y N son: apple, grapes, banana, whatermelon, muskmelos, cotn y coffee. En su mayoría frutas.

A continuación se analizará la distribución de los cultivos en gráficos tridimensionales usando distintas variables.

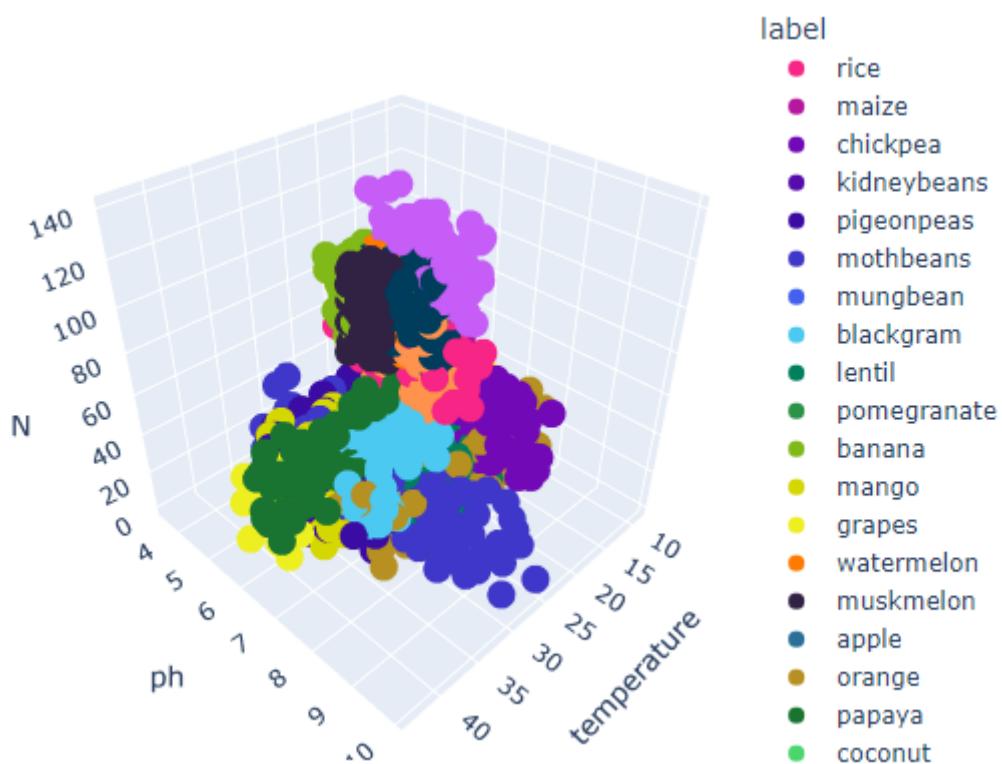
Columnas graficadas: N, P y K



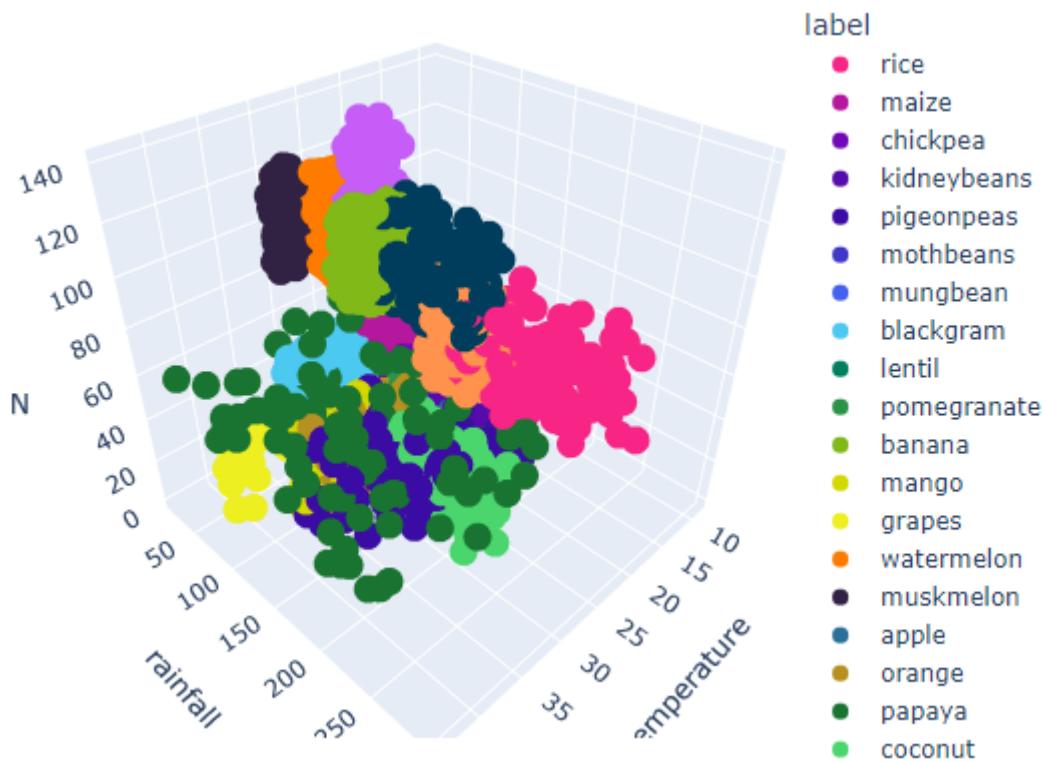
Columnas graficadas: rainfall, humidity, ph.



Columnas graficadas: temperature, ph, N.



Columnas graficadas: temperature, rainfall, N.



En los primeros dos gráficos es donde se aprecia una separación más clara entre los cultivos a simple vista. En el primero, apple y grapes son los que se alejan claramente de los demás en el espacio de dimensiones. En el segundo, chickpea y kidneybeans se encuentran un poco más distanciados del resto.

2.5. Estandarización

El proceso de PCA identifica aquellas direcciones en las cuales la varianza es mayor. Como la varianza de una variable se mide en su misma escala elevada al cuadrado, si antes de calcular las componentes no se estandarizan todas las variables para que tengan media 0 y desviación estándar 1, aquellas variables cuya escala sea mayor dominarán al resto.

Por esto se eligió la estandarización Z-Score. Este método escala cada característica restando la media de la característica y dividiendo por su desviación estándar.

3. PCA

Realizar PCA y determinar el número de componentes principales considerando alguno de los 3 criterios dados en la práctica. Graficar la varianza acumulada y las componentes de PCA en un gráfico 2D o 3D con sus respectivas clases.

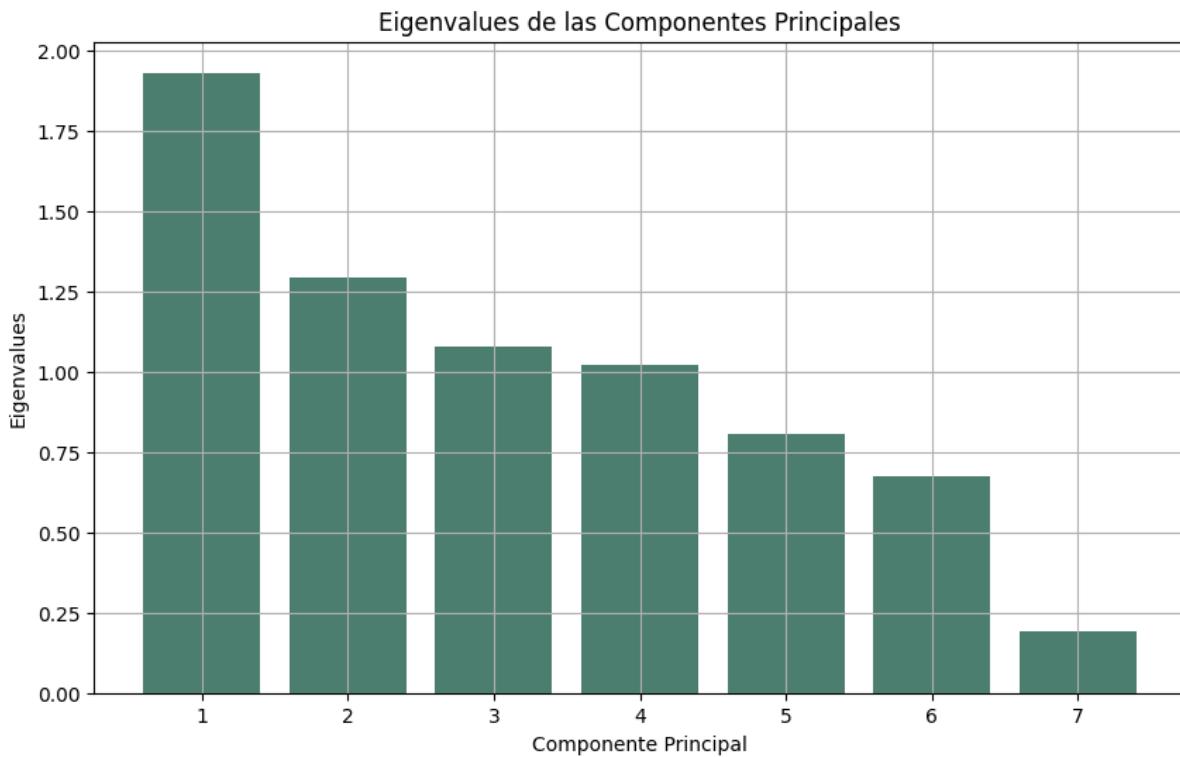
	PC1	PC2	PC3	PC4	PC5	PC6	PC7	label
0	-0.582869	-0.844586	1.373343	-1.614129	0.308224	-0.095997	-0.025239	rice
1	-0.474635	-0.784895	1.252178	-1.792762	1.107745	-0.532255	-0.280543	rice
2	-0.634068	-0.694522	1.179332	-1.818106	2.523263	-0.538551	-0.105967	rice
3	-1.047920	-1.087658	1.393351	-0.982401	1.448781	-0.656929	0.275272	rice
4	-0.873258	-0.658673	1.455685	-2.335012	1.959633	-0.318025	0.052740	rice
...
2195	-1.260921	-0.618363	0.711297	-1.014970	0.133230	-1.122259	0.307025	coffee
2196	-1.355583	-0.154043	0.701649	-0.197683	-0.898658	-0.771476	0.684957	coffee
2197	-1.158384	-0.640475	1.045840	-1.301841	-0.491765	-0.885129	0.203463	coffee
2198	-1.219188	0.052390	0.180855	-0.990499	-0.601388	-1.308967	0.509656	coffee
2199	-1.373004	0.055818	0.500654	-1.219230	-0.346351	-0.572730	0.577993	coffee

2200 rows × 8 columns

En el anterior dataframe vemos la proyección de cada registro en cada una de las nuevas variables.

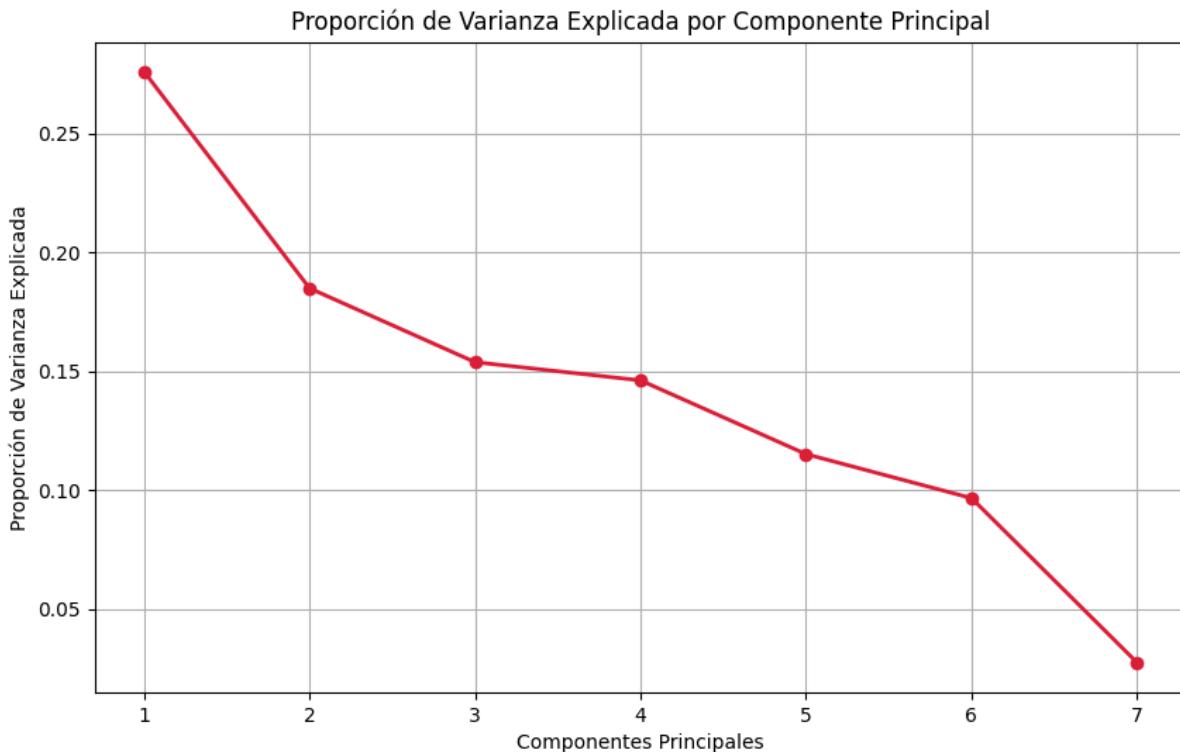
A continuación, se observarán en forma de tabla y gráficamente las representaciones de los eigenvalues para cada componente.

	Eigenvalues	Proporción de varianza explicada	Proporción acumulada de varianza explicada
0	1.931218	0.275888	0.275888
1	1.293910	0.184844	0.460733
2	1.076509	0.153787	0.614520
3	1.022891	0.146127	0.760647
4	0.805928	0.115133	0.875780
5	0.676562	0.096652	0.972431
6	0.192981	0.027569	1.000000

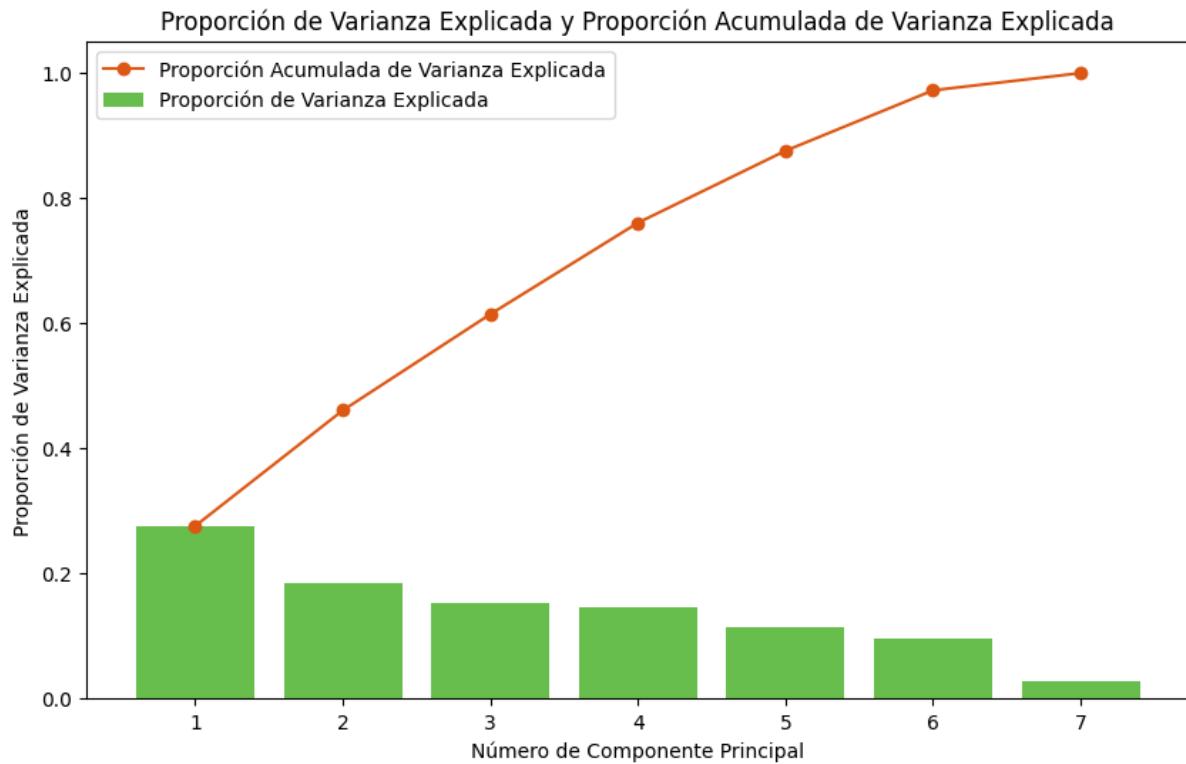


Vemos que las primeras 4 componentes tienen eigenvalues mayores a uno, indicando una mayor contribución a la explicación de la varianza en los datos.

A continuación, se visualizará gráficamente la varianza explicada por cada componente y la acumulada.

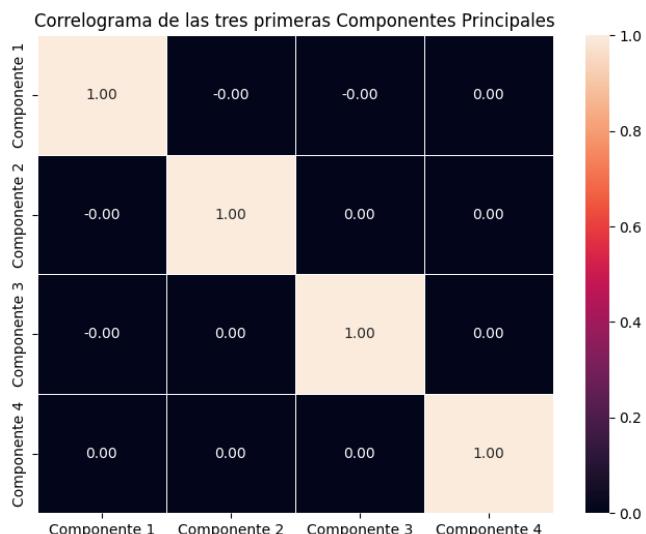


Al observar el gráfico, puede verse que a partir de la quinta componente la proporción de varianza explicada decae considerablemente en relación a las primeras cuatro componentes.



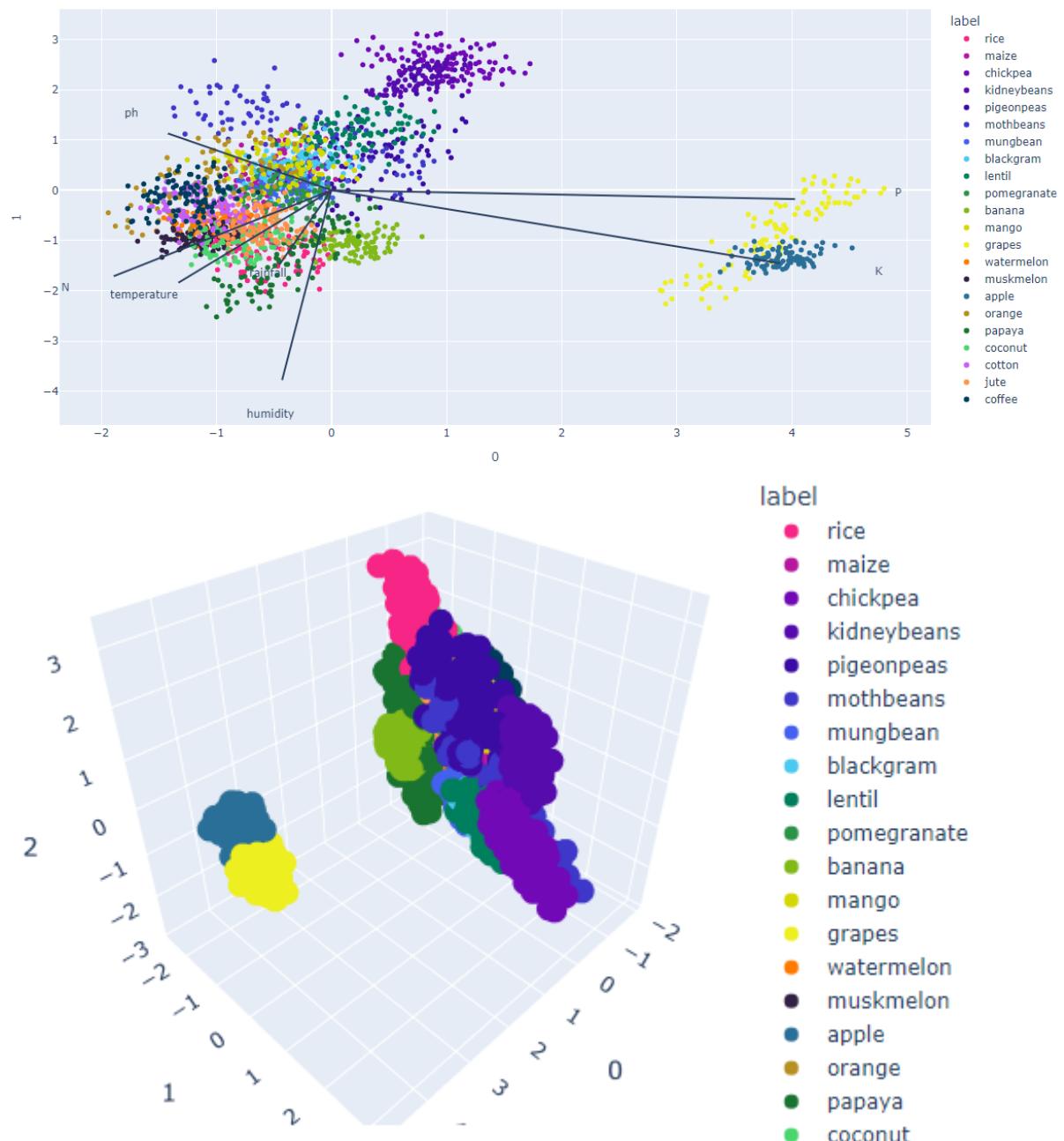
Las primeras cuatro componentes son las más influyentes en cuanto a proporción de varianza explicada, logrando entre las cuatro una proporción de varianza explicada de aproximadamente 0.75 ($\sim 75\%$).

Se tuvo en cuenta el criterio de la varianza explicada (~75% -80%), la regla de Kaiser (eigenvalues > 1) y el gráfico Scree para tomar las primeras cuatro variables. Dichas variables acumulan más del 76% y tienen eigenvalores superiores a 1. Además, la gráfica Proporción de variabilidad explicada - Componentes principales muestra el cambio de la variabilidad explicada en la componente 4.



Como es de esperar, la correlación entre las componentes principales es nula, es decir que las mismas son ortogonales.

Dado que solo es posible graficar en dos dimensiones o tres, se seleccionarán las primeras 2 y 3 componentes(ya que son las que tienen las proporciones más elevadas de todas) para realizar los mismos y observar visualmente los resultados de la distribución de los datos en el nuevo espacio de dimensiones.



Se observa que la distribución de los datos en el nuevo espacio de dimensiones es muy similar a la que se observó anteriormente para el espacio de dimensiones original utilizando las variables correspondientes a los nutrientes K, P y N. Esto puede ser un indicio de que las variables originales que más contribuyeron a las componentes principales en el nuevo espacio de dimensión son, en efecto, K, P y N.

Se observa claramente que los cultivos grapes y apple se encuentran significativamente alejados de los demás. Esto, también había sido destacado en el análisis exploratorio, por lo que se entiende que se respeta la distribución original.

En el contexto del Análisis de Componentes Principales (PCA), los componentes principales son combinaciones lineales de las variables originales. Cada componente principal es una nueva "dimensión" en el espacio de características que es una combinación ponderada de las variables originales.

Se quiere entender a qué variables originales hacen referencia las componentes principales.

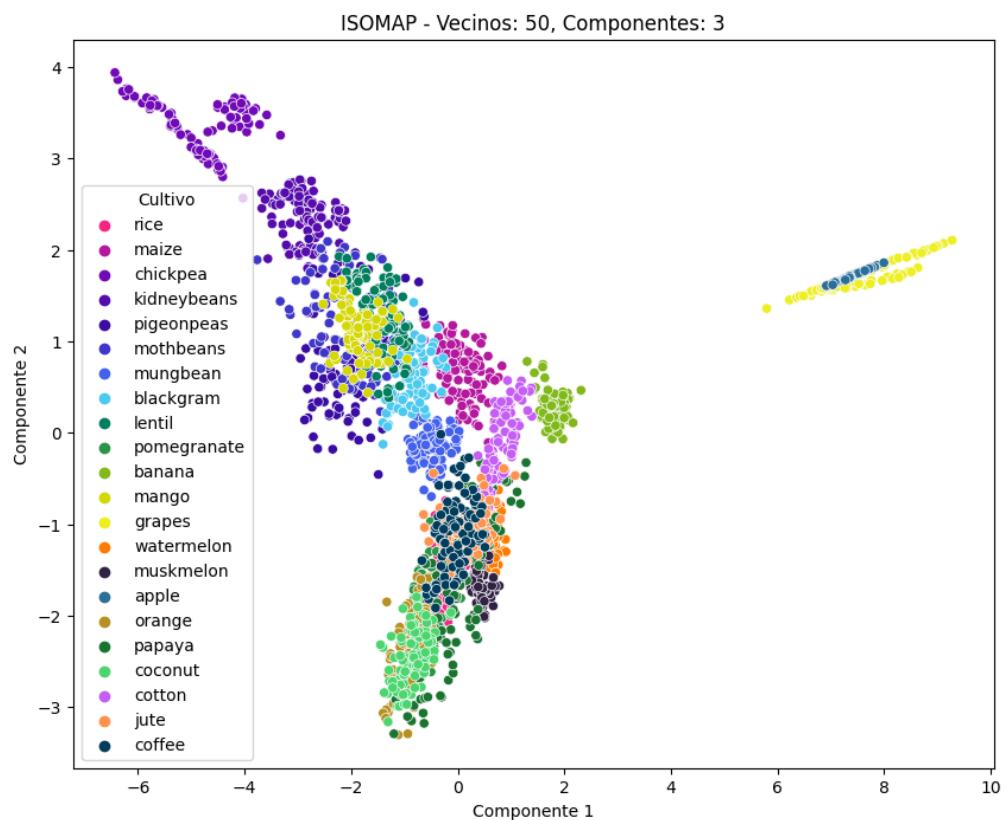
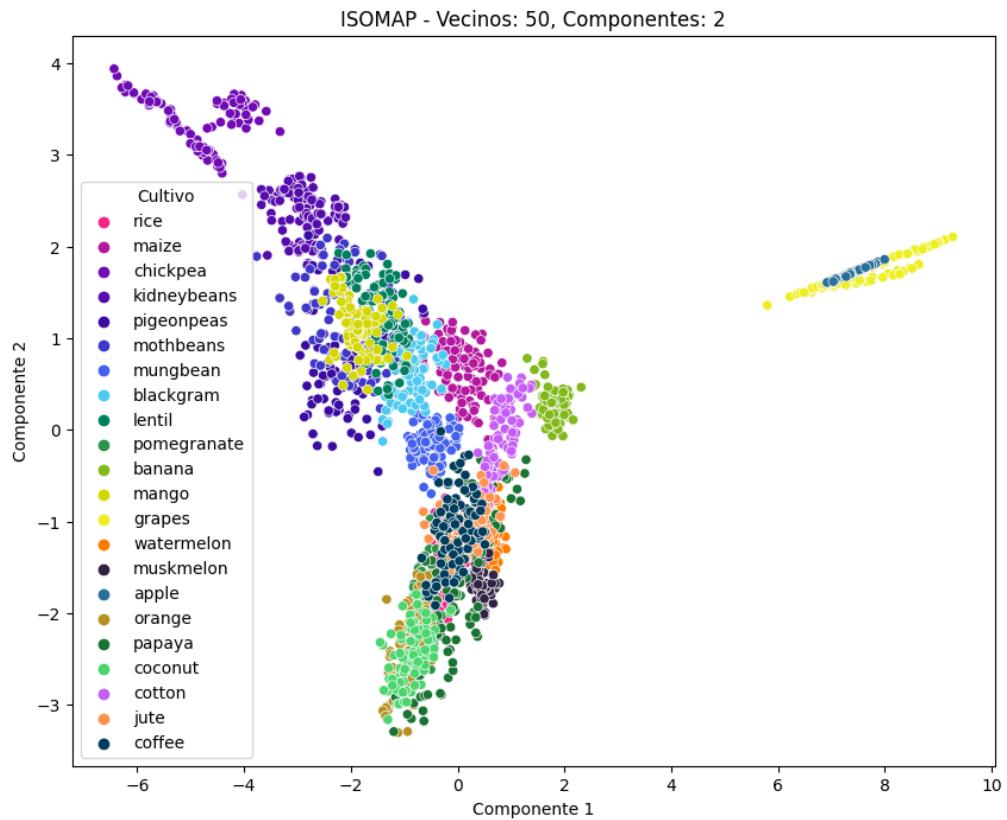
	N	P	K	temperature	humidity	ph	rainfall
0	-0.302191	0.643787	0.622607	-0.212428	-0.068483	-0.226943	-0.072532
1	-0.334107	-0.034358	-0.283829	-0.359487	-0.737917	0.220657	-0.290158
2	-0.112045	-0.109939	-0.163173	-0.248228	-0.213599	-0.548520	0.735267
3	-0.541651	-0.046293	-0.154867	0.690826	-0.067171	-0.395700	-0.205318
4	-0.507785	0.082331	0.033425	0.154865	0.128871	0.651881	0.518382
5	-0.482904	-0.376847	-0.028967	-0.500418	0.547871	-0.125712	-0.239930
6	-0.008473	-0.649104	0.692268	0.111282	-0.289624	0.040028	0.038577

Se observa que para la primera componente, que es la que más contribuye a la proporción de varianza explicada, las variables N, P y K son las que tienen un aporte más significativo.

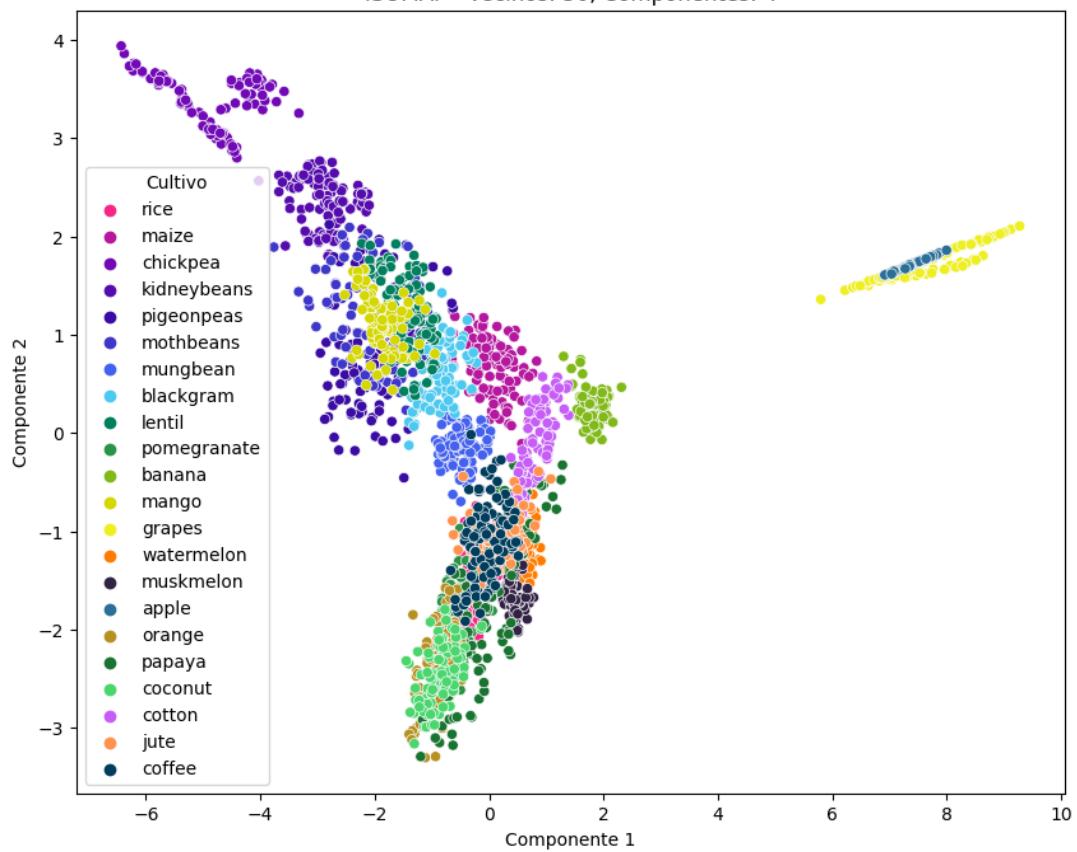
Se recuerda entonces que el gráfico 3D del dataset original usando las variables K, P y N se observa muy similar al 3D representando los datos en función de las 3 primeras componentes halladas mediante PCA.

4. Isomap

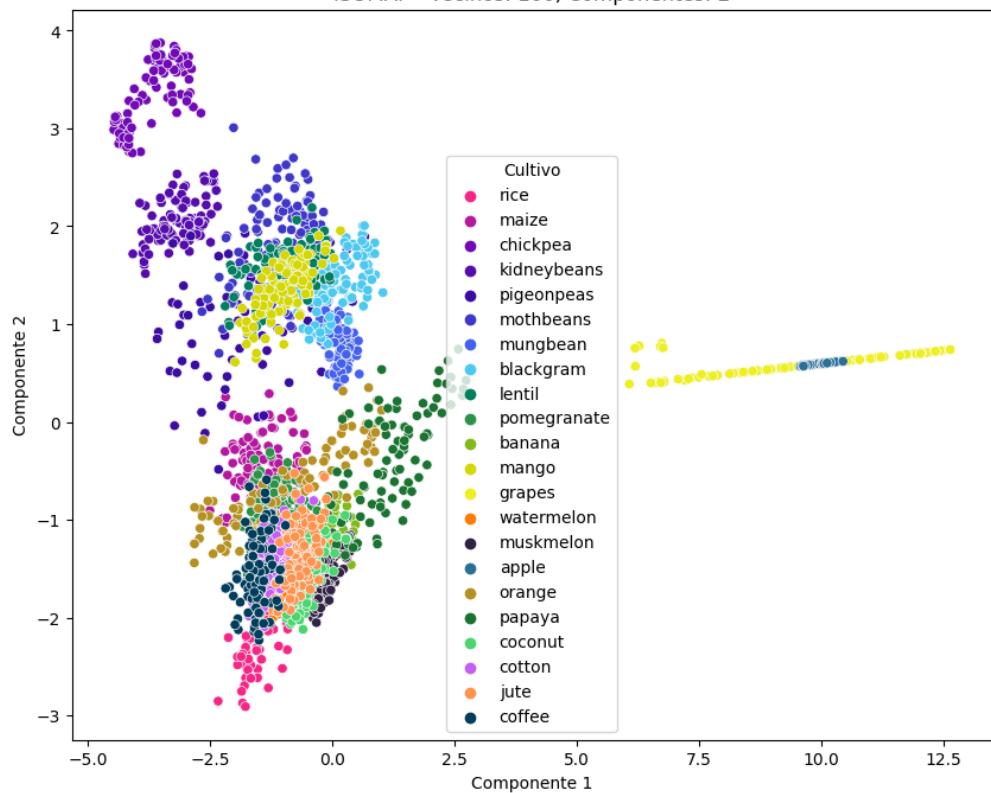
Aplicar Isomap y analizar los resultados obtenidos variando el número de vecinos y componentes. Realizar un gráfico en 2D utilizando dos componentes.



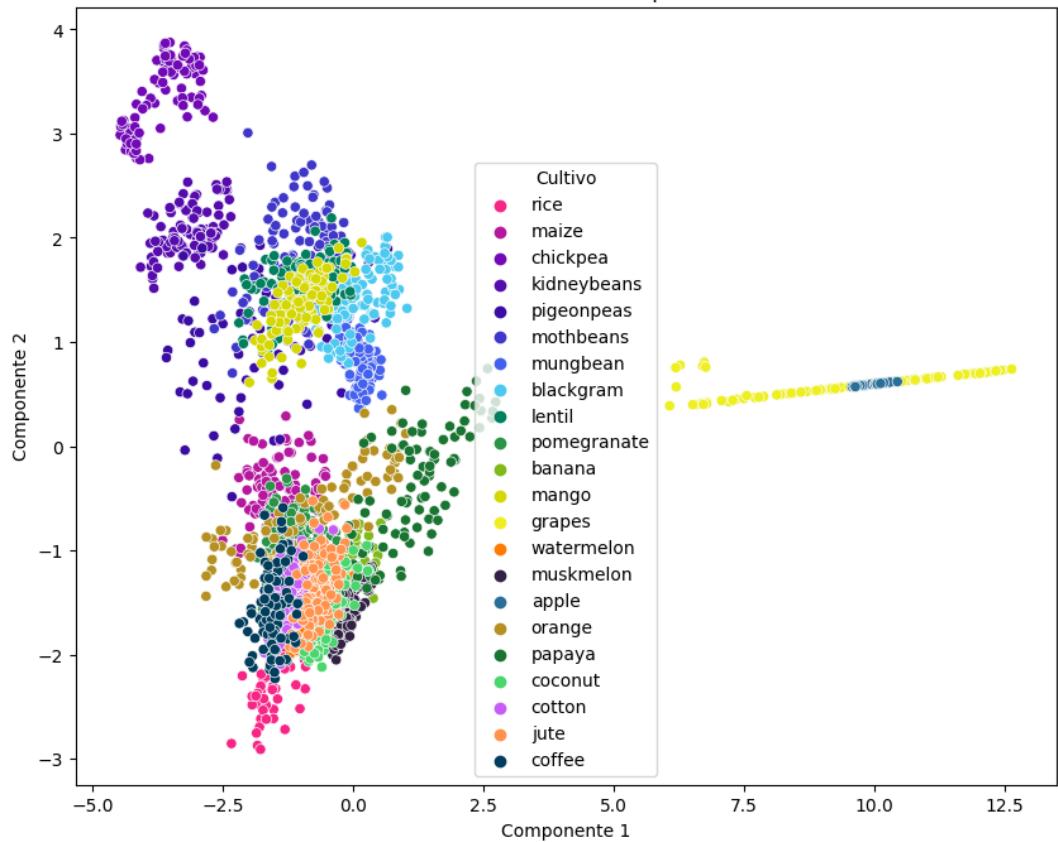
ISOMAP - Vecinos: 50, Componentes: 4



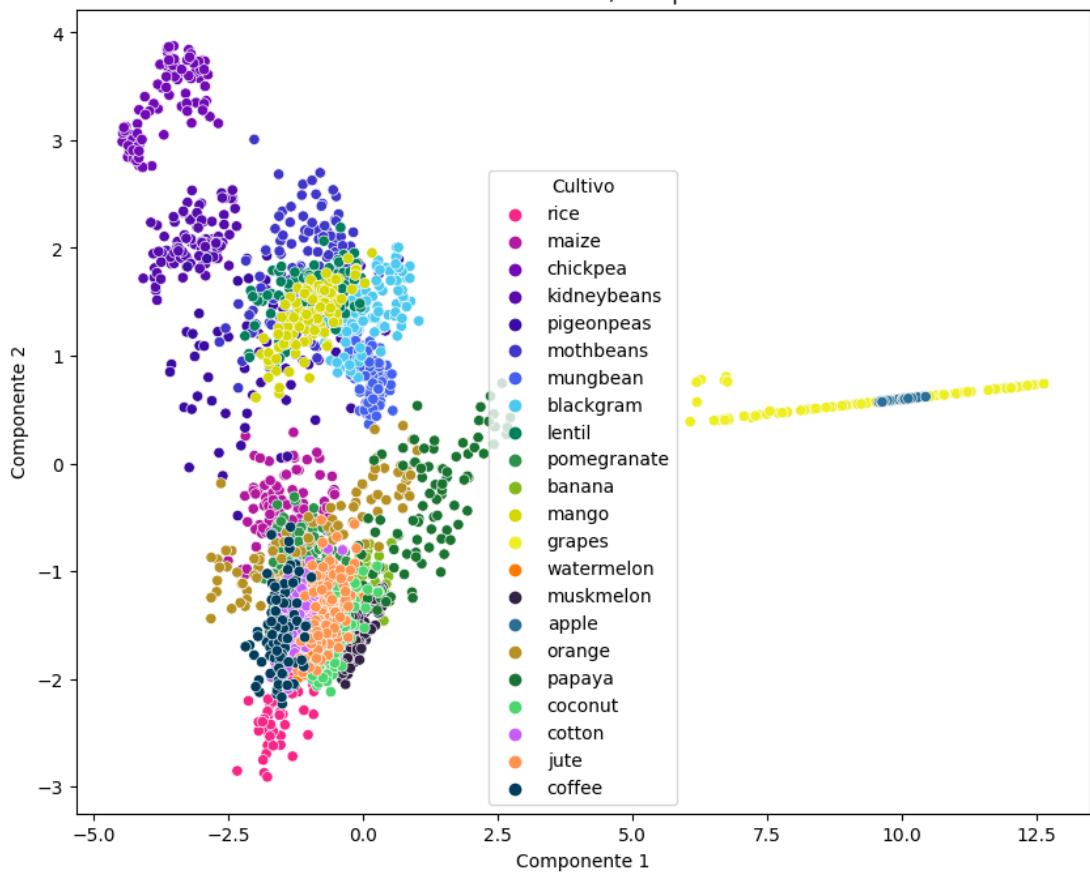
ISOMAP - Vecinos: 100, Componentes: 2



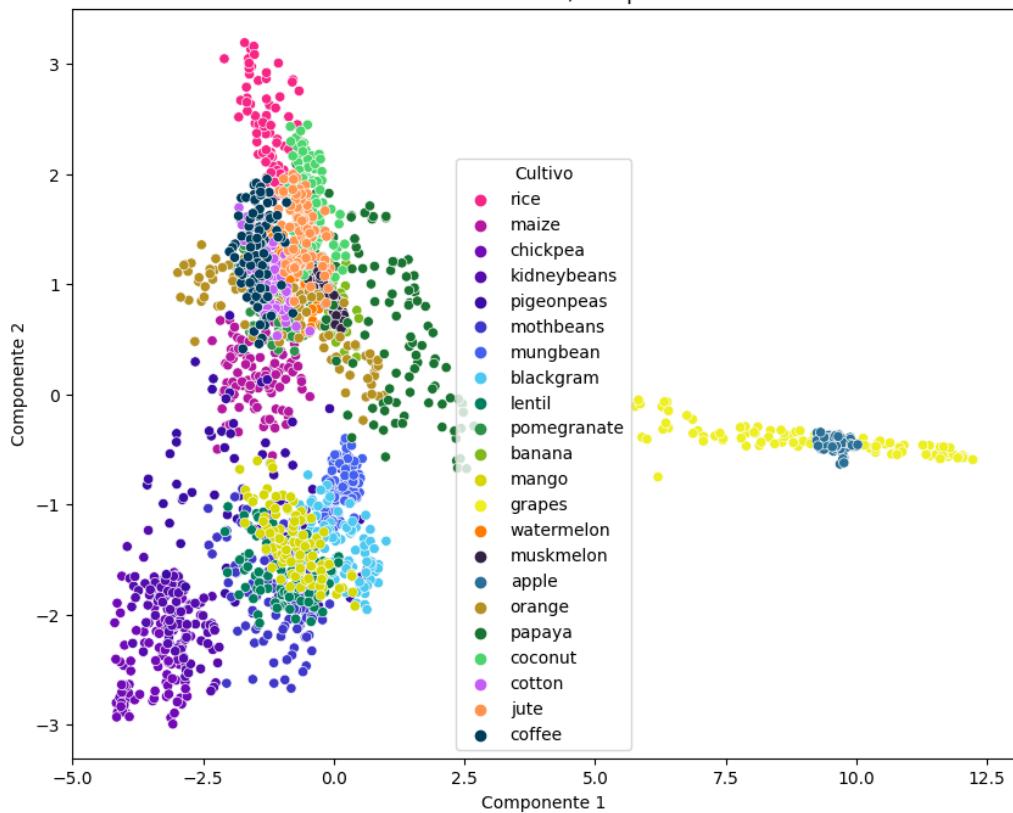
ISOMAP - Vecinos: 100, Componentes: 3



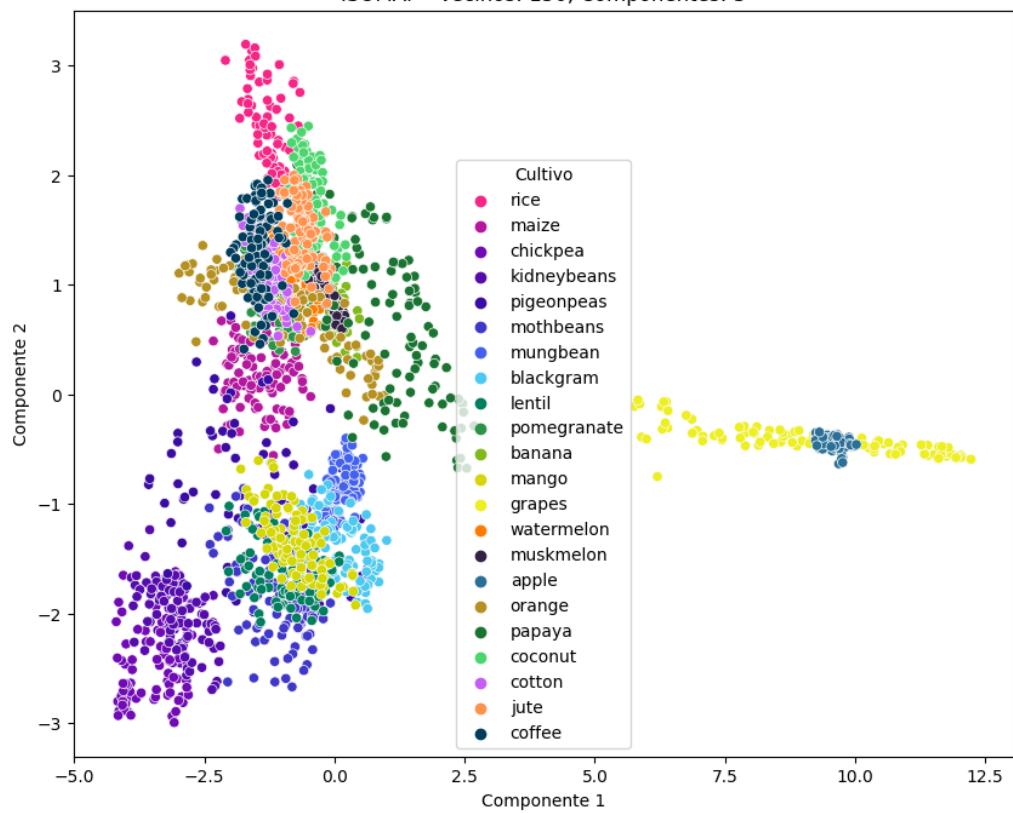
ISOMAP - Vecinos: 100, Componentes: 4



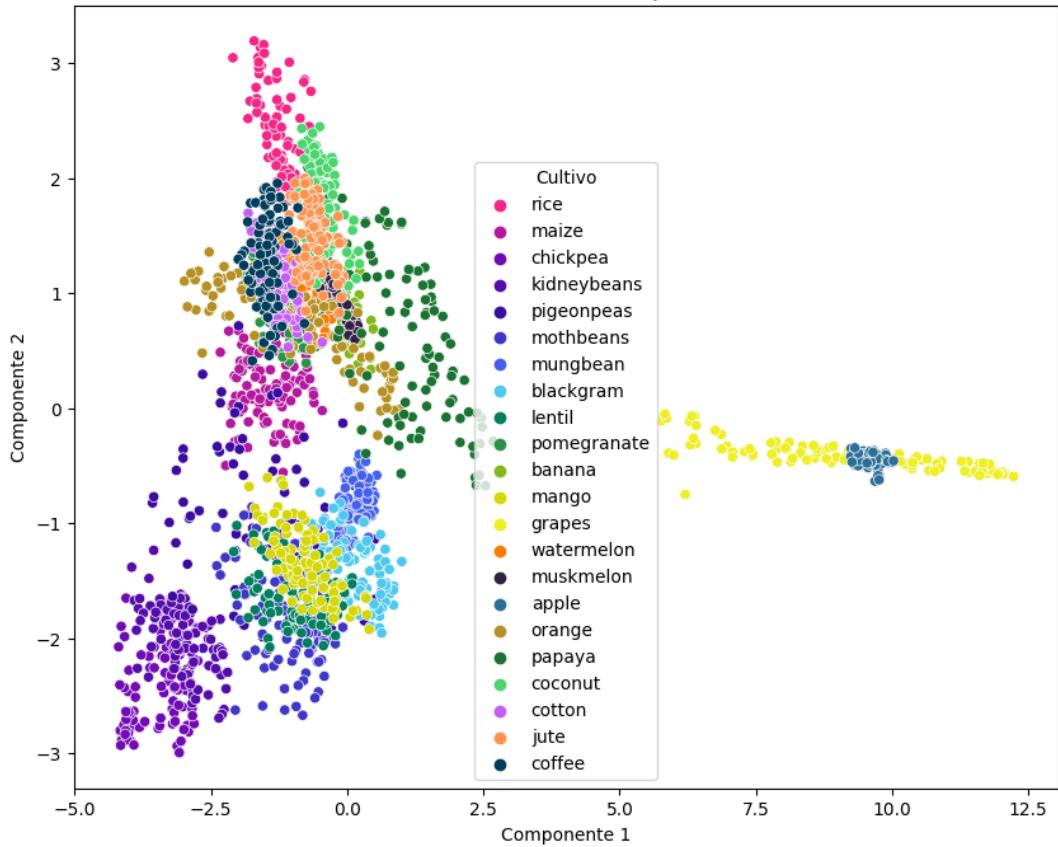
ISOMAP - Vecinos: 150, Componentes: 2



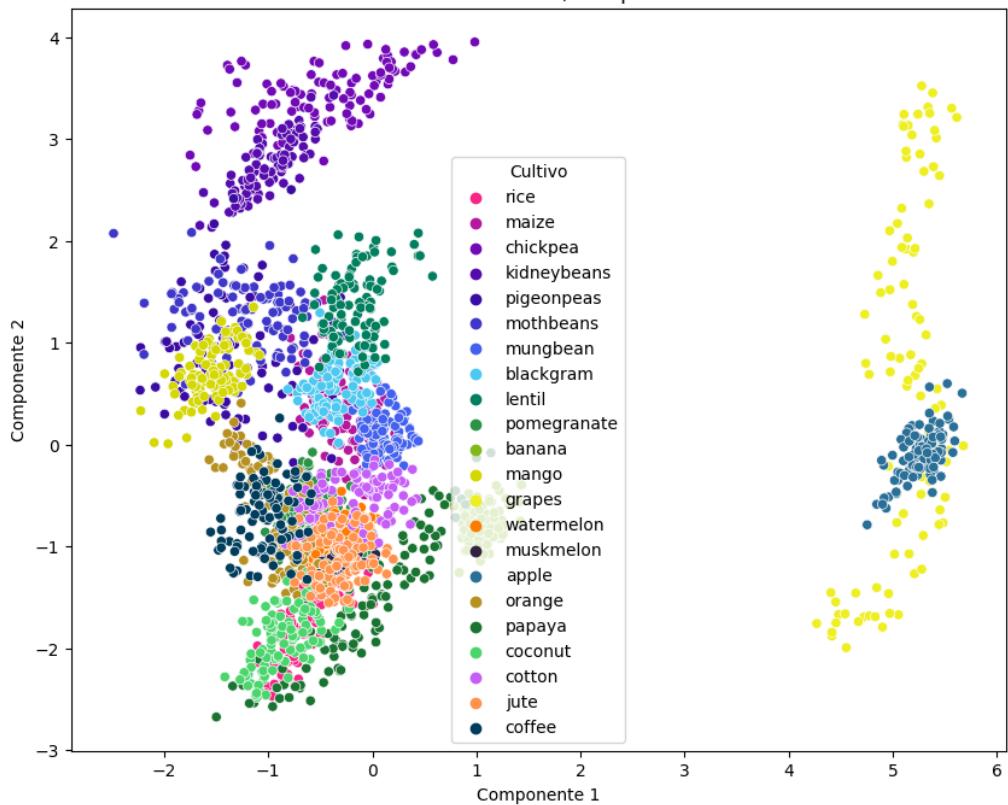
ISOMAP - Vecinos: 150, Componentes: 3



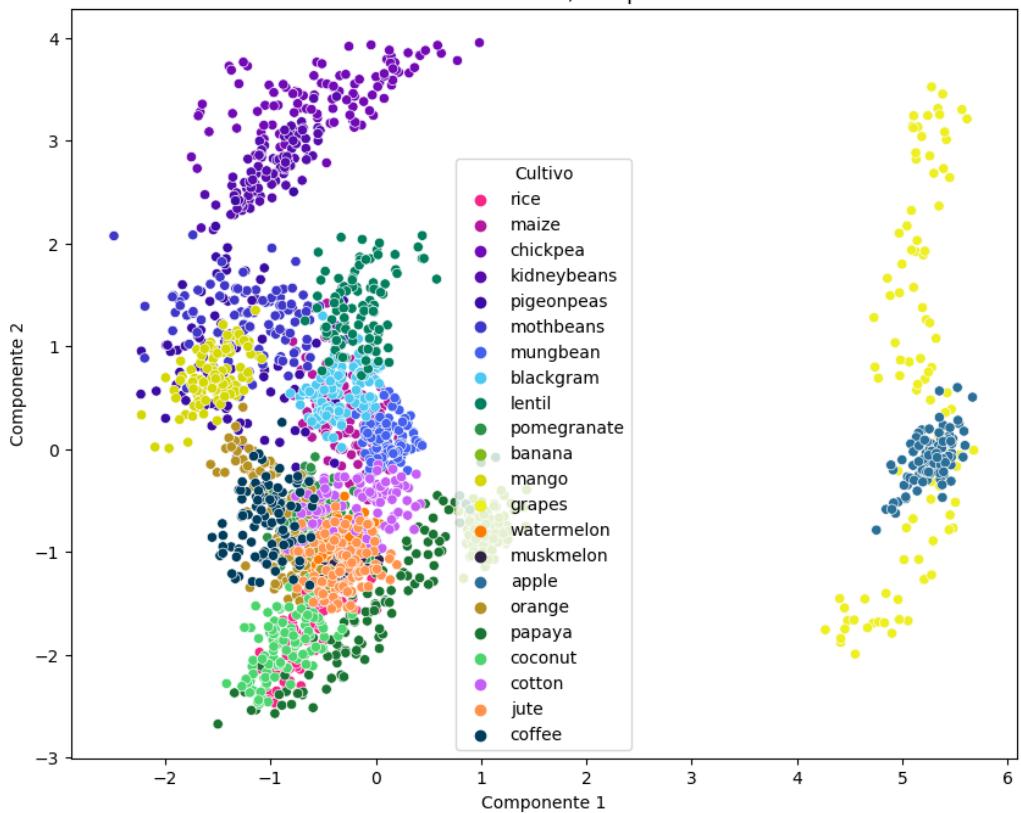
ISOMAP - Vecinos: 150, Componentes: 4



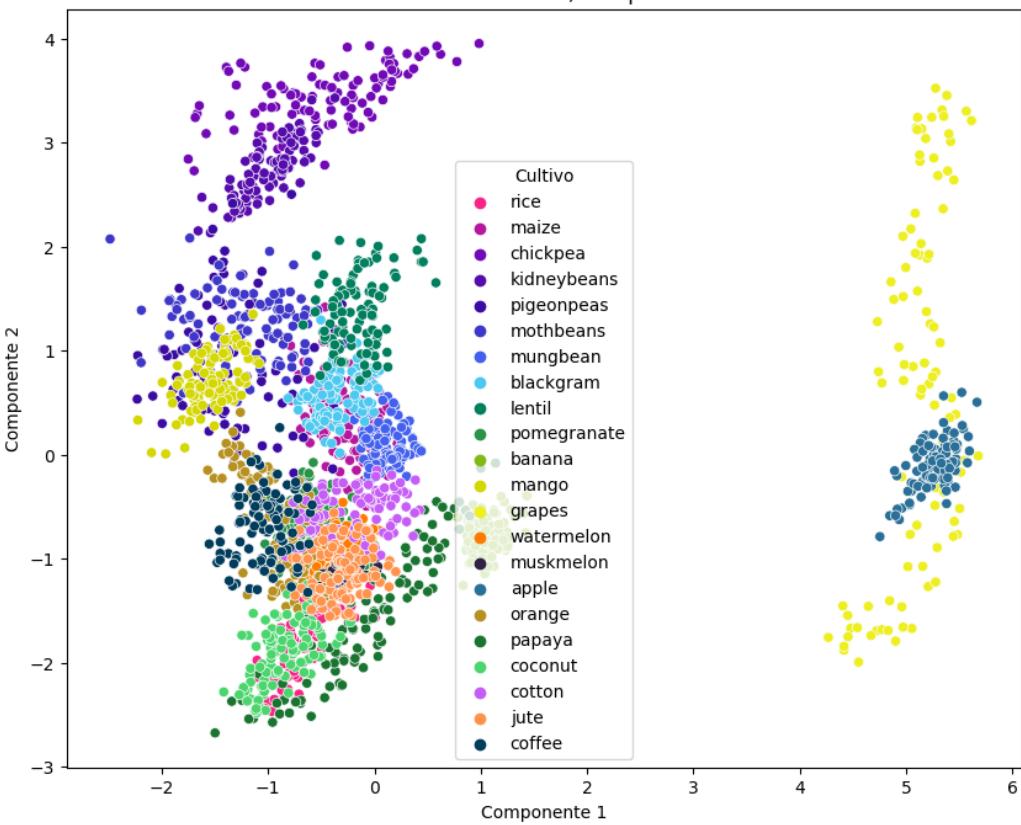
ISOMAP - Vecinos: 200, Componentes: 2

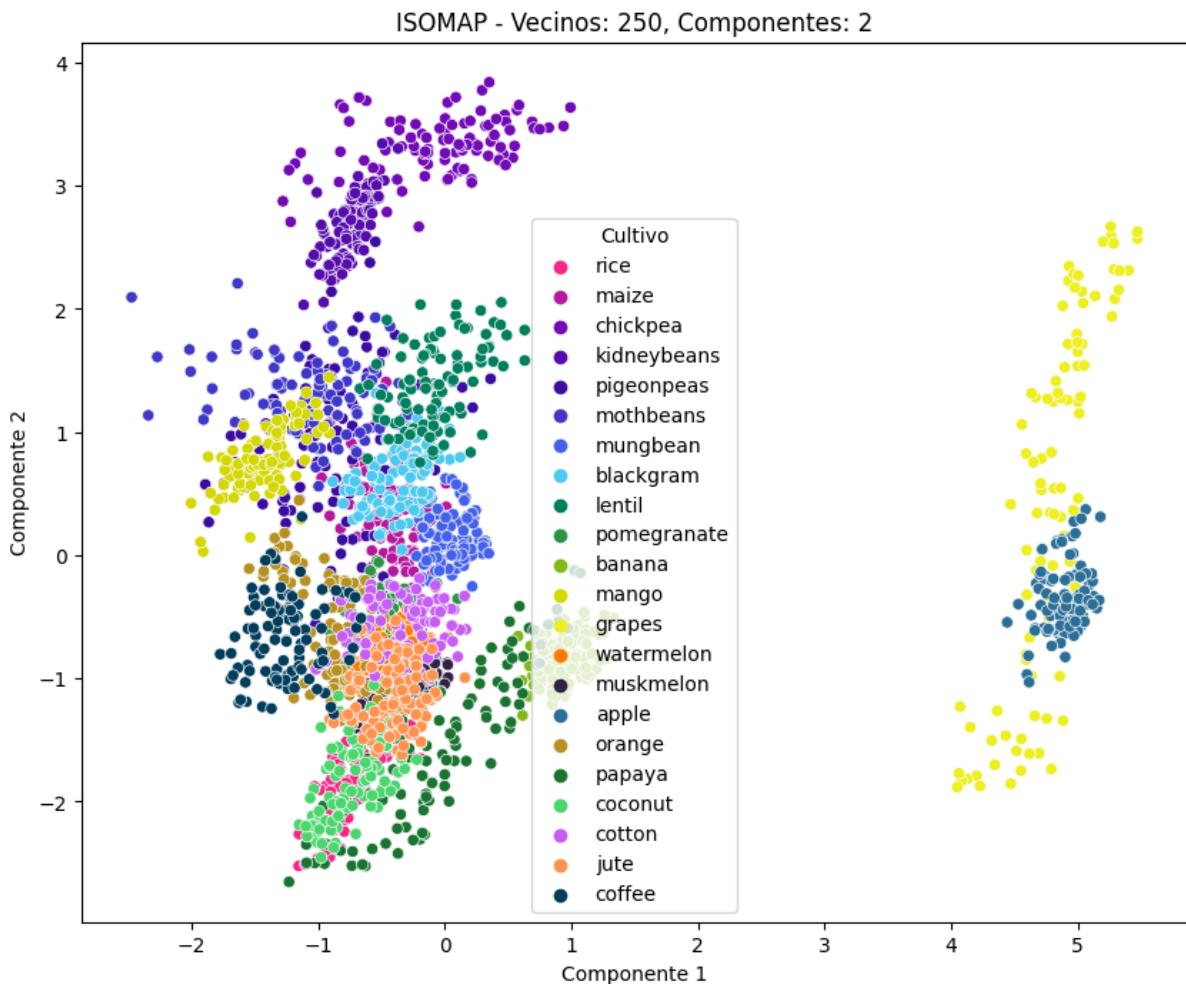


ISOMAP - Vecinos: 200, Componentes: 3



ISOMAP - Vecinos: 200, Componentes: 4





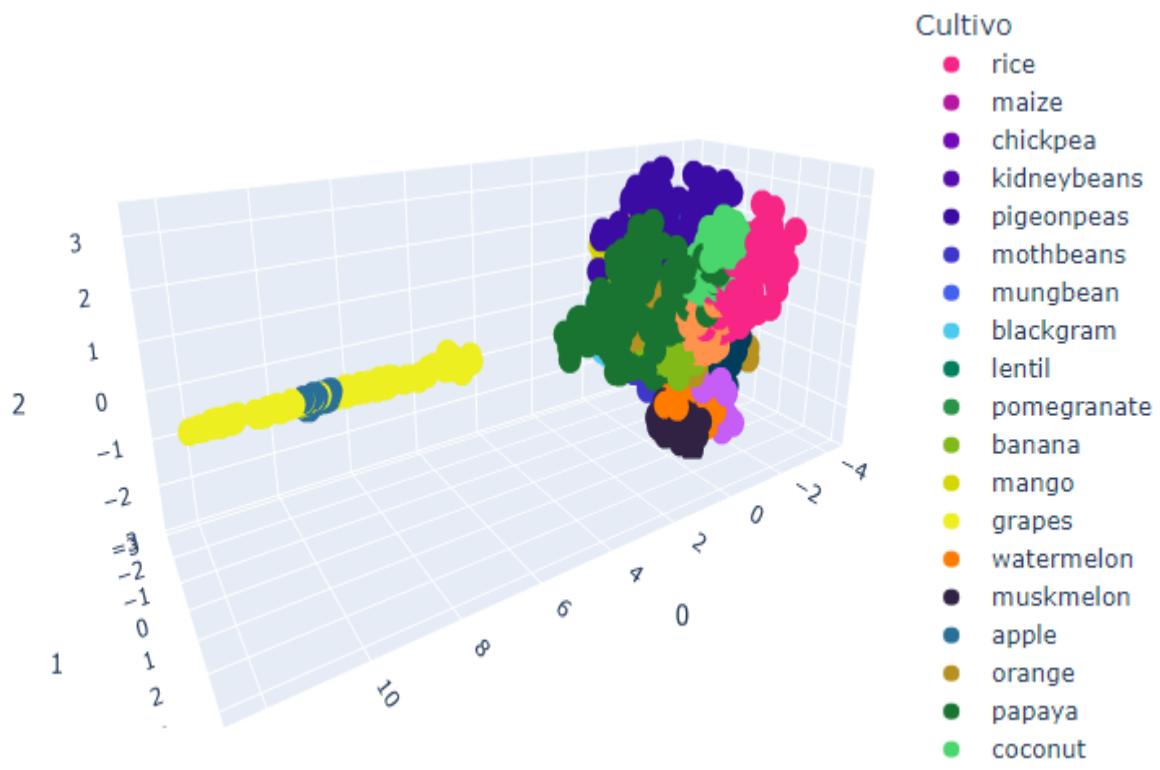
Al realizar los gráficos se observa en todos ellos que hay 2 cultivos que se encuentran sustancialmente alejados de los demás. Otros detalles que se presentan:

- A medida que aumenta el número de vecinos los grupos forman asociaciones más en forma de nube y menos lineales y también se visibiliza mejor su separación.
- Ésa diferenciación se observa clara al pasar de 150 a 200 vecinos independientemente de la cantidad de componentes elegida.

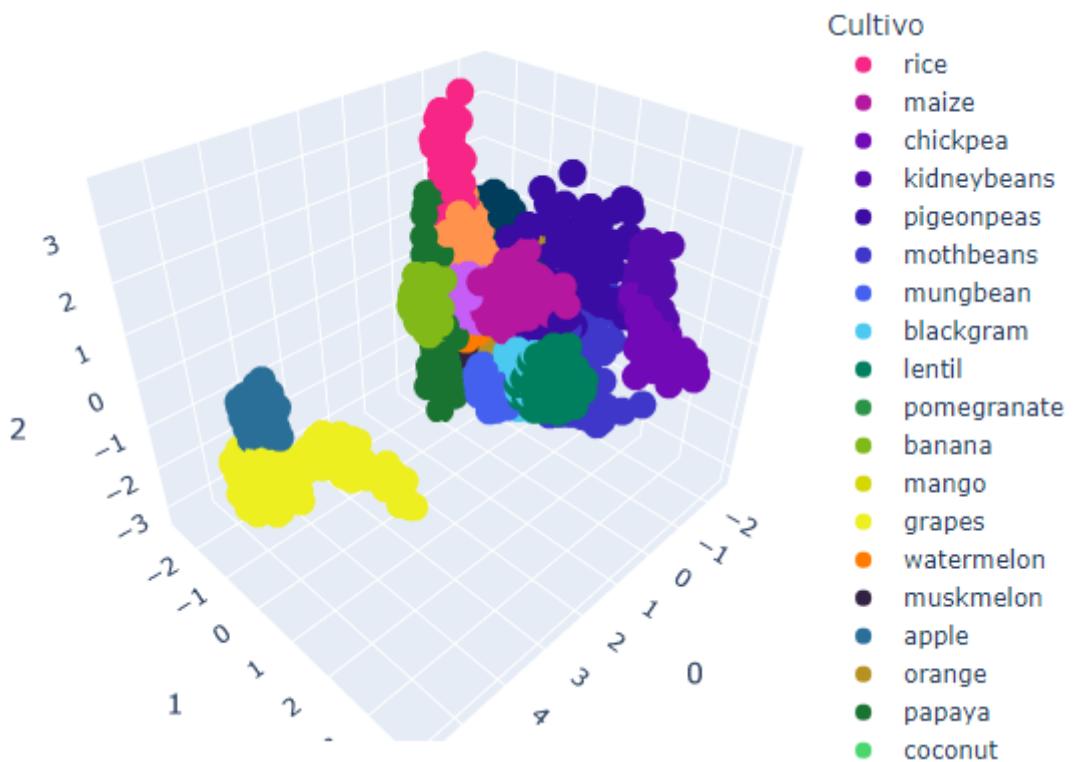
- Al pasar la cantidad de 200 vecinos ya no se observan mayores cambios en la distribución de los puntos. Continúan distribuidos de la misma forma y se compactan ligeramente.
- Al pasar de 150 vecinos a 200, el grupo que estaba sustancialmente alejado de los demás dejó de observarse de forma horizontal en el gráfico 2D y pasó a posicionarse de manera vertical. A su vez, la posición del resto de los puntos se invierte horizontalmente. Es decir, los puntos que se encontraban arriba en el gráfico 2D, pasaron a estar abajo y viceversa.

A continuación, se visualizarán en gráficos tridimensionales los resultados de ISOMAP para la cantidad de vecinos detectada como punto de inflexión en las observaciones anteriores (200), como así también la cantidad utilizada anterior a ella y la posterior. De esta manera se quiere tener un acercamiento más detallado del cambio producido.

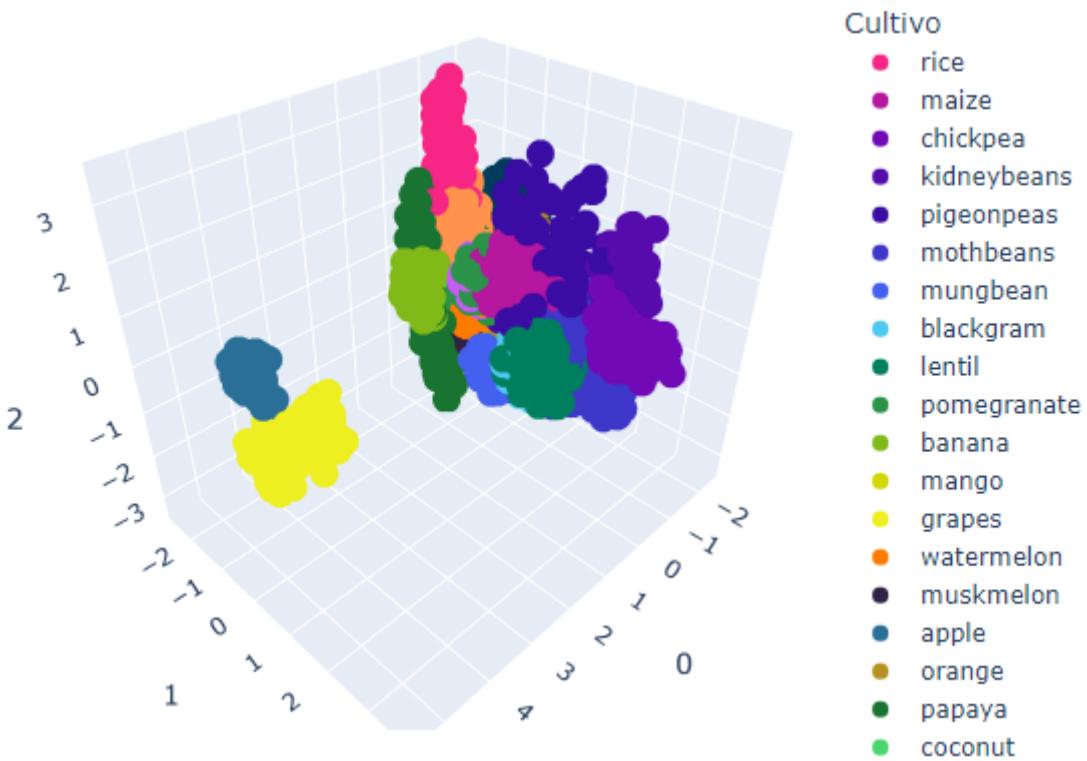
- Resultados para 150 vecinos y 3 componentes:



- Resultados para 200 vecinos y 3 componentes:



- Resultados para 300 vecinos y 3 componentes:



En los gráficos 3D se produce una mejora al momento de apreciar los grupos de datos.

- Apple y grapes están sustancialmente alejados de los demás cultivos.
- Al apreciar el gráfico desde la perspectiva de las variables 1 y 2 se observa una separación más clara de los cultivos chickpea, kidneybeans y pidgeonpeas del resto de los cultivos
- El resto de los cultivos tienen separaciones menos notables y se encuentran más compactados en el espacio.

Todas las observaciones realizadas hasta el momento sugieren varias reflexiones:

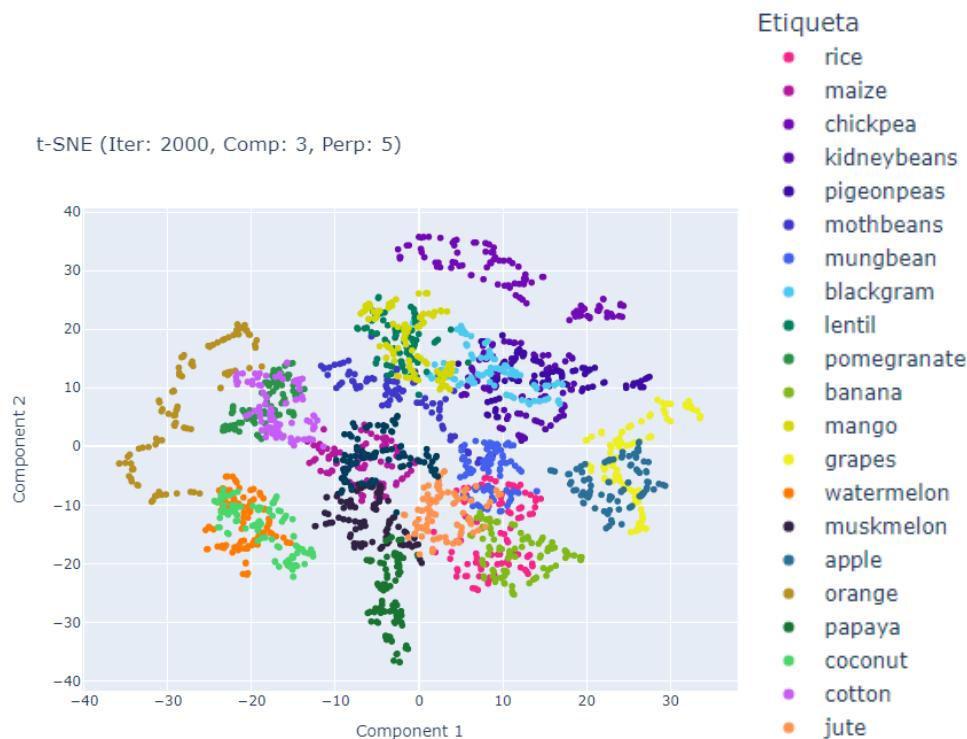
- Sensibilidad a la cantidad de vecinos: El número de vecinos en el algoritmo ISOMAP tiene un impacto significativo en la forma en que se agrupan los datos en el espacio reducido. Esto sugiere que la elección de este hiperparámetro es crucial y debe ajustarse según la naturaleza de los datos. En este caso, el dataset posee 2200 registros y parece que 200 vecinos es la cantidad acorde para realizar la reducción de dimensionalidad.
- Número de cultivos y su distribución: Dado que hay 22 tipos diferentes de cultivos, el hecho de que la mayoría de ellos estén cercanos en el espacio reducido puede indicar que comparten similitudes en términos de las características que se están considerando. Sin embargo, los dos cultivos que están sustancialmente alejados podrían ser muy diferentes del resto en términos de esas mismas características.
- Estabilización: El hecho de que a partir de 200 vecinos no se observen cambios significativos en la distribución de los puntos sugiere que el patrón subyacente de similitud y distancia entre los cultivos podría haberse estabilizado a una estructura específica.

- Inversión en la posición de los puntos: El cambio en la orientación y posición de los puntos entre 150 y 200 vecinos puede indicar una reorganización en la estructura de similitud entre los cultivos. Esto puede estar relacionado con cómo se están considerando las relaciones de vecindad en el espacio de alta dimensión.

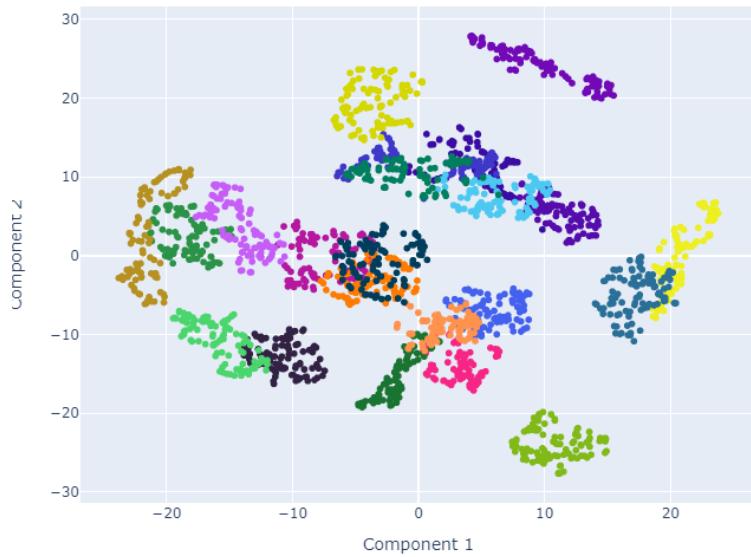
5. T-SNE

Aplicar t-SNE y analizar los resultados obtenidos variando el número de iteraciones, componentes y perplejidad. Realizar un gráfico en 2D de utilizando dos componentes.

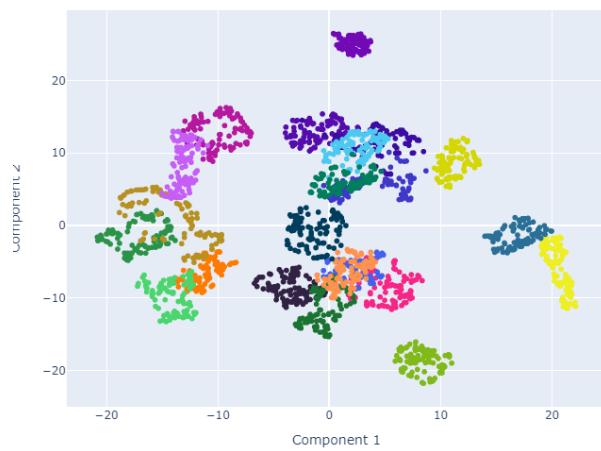
- 2000 iteraciones, 3 componentes y distintas perplejidades:



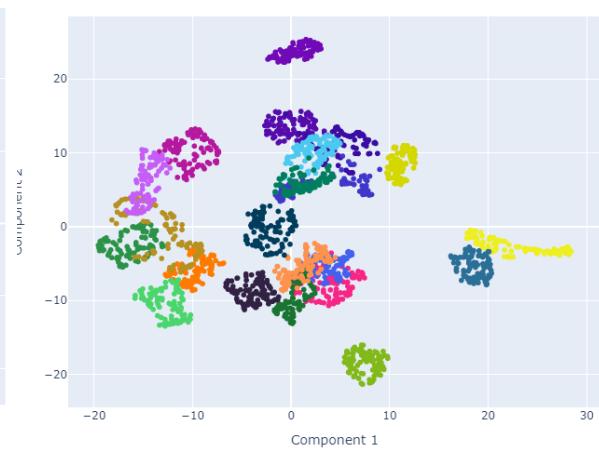
t-SNE (Iter: 2000, Comp: 3, Perp: 15)



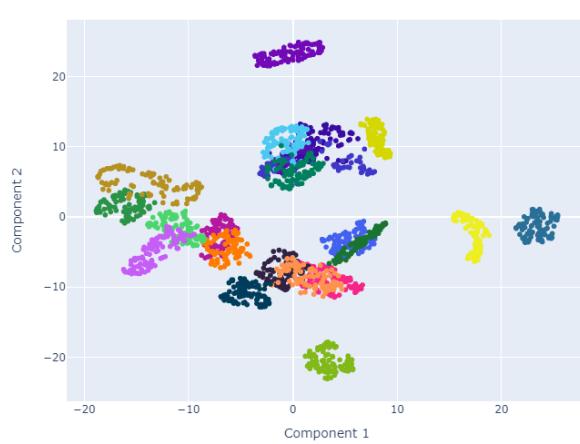
t-SNE (Iter: 2000, Comp: 3, Perp: 25)



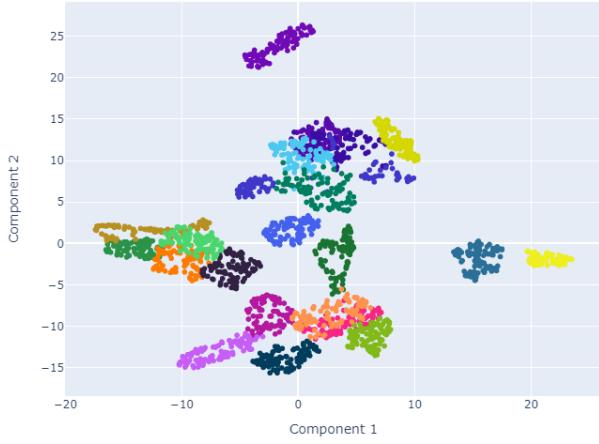
t-SNE (Iter: 2000, Comp: 3, Perp: 30)



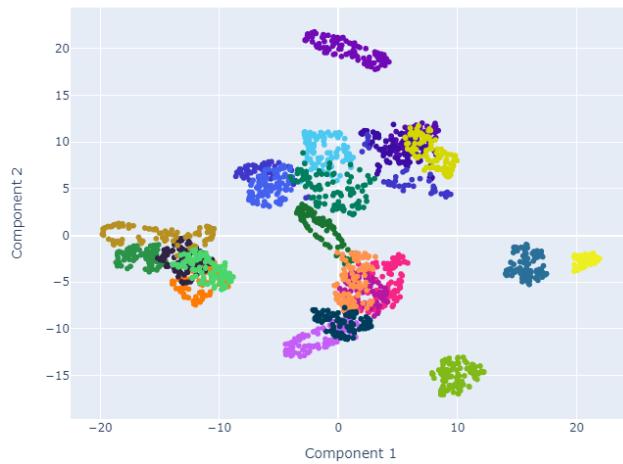
t-SNE (Iter: 2000, Comp: 3, Perp: 35)



t-SNE (Iter: 2000, Comp: 3, Perp: 40)

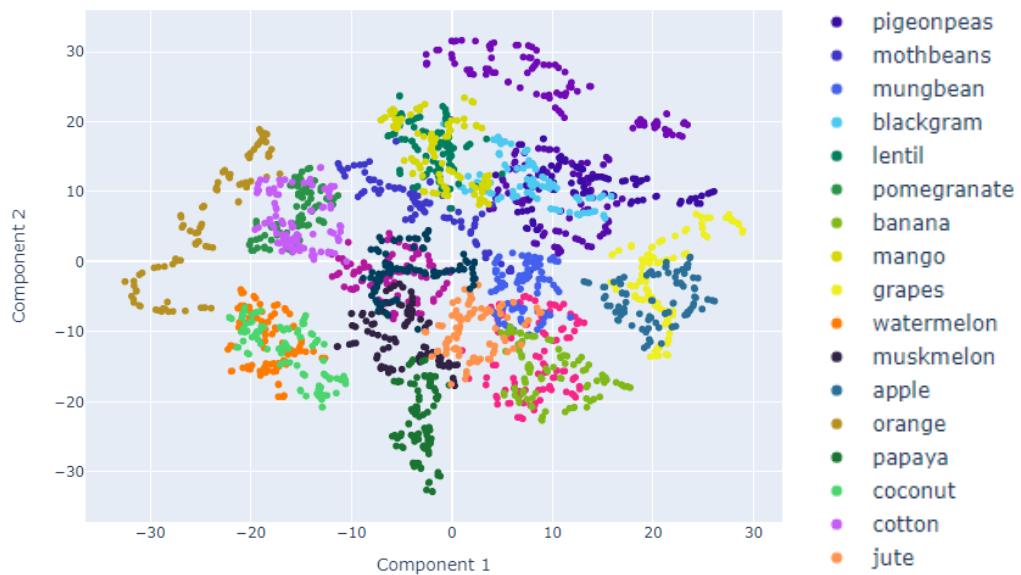


t-SNE (Iter: 2000, Comp: 3, Perp: 45)

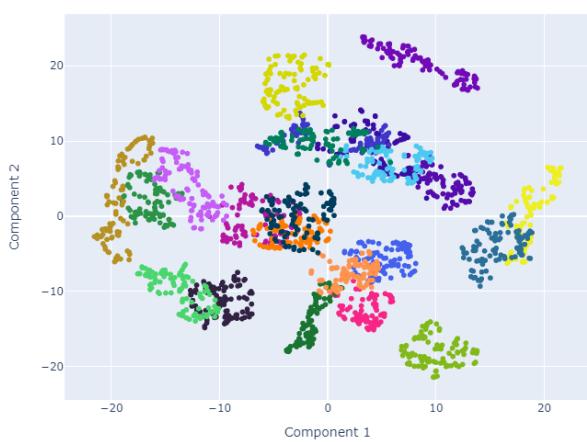


- 1000 iteraciones, 3 componentes y distintas perplejidades:

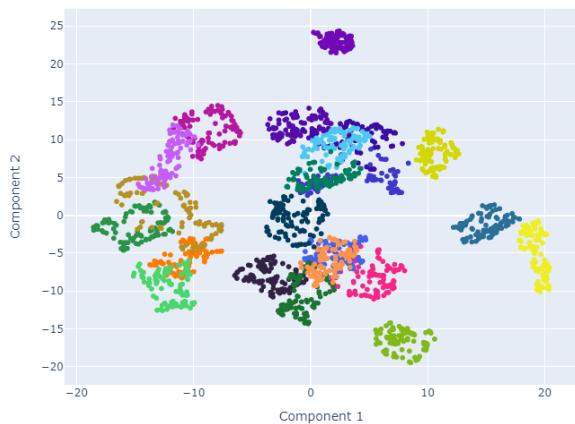
t-SNE (Iter: 1000, Comp: 3, Perp: 5)



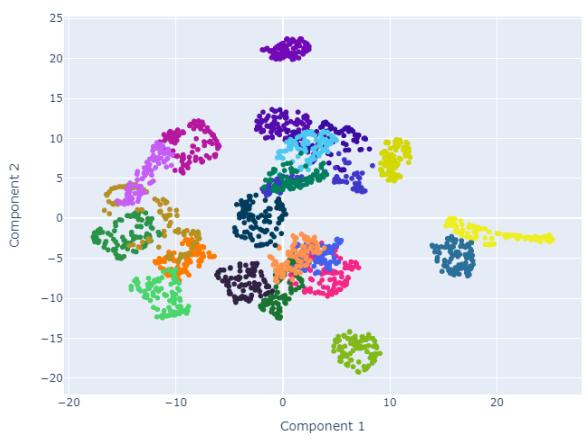
t-SNE (Iter: 1000, Comp: 3, Perp: 15)



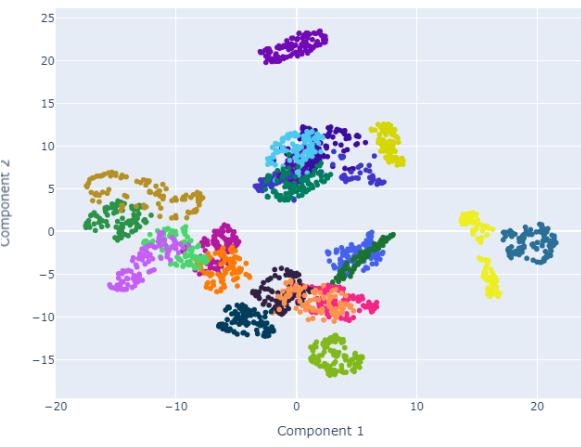
t-SNE (Iter: 1000, Comp: 3, Perp: 25)



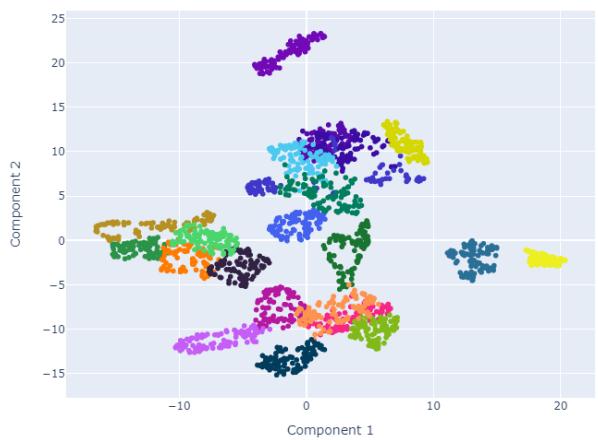
t-SNE (Iter: 1000, Comp: 3, Perp: 30)



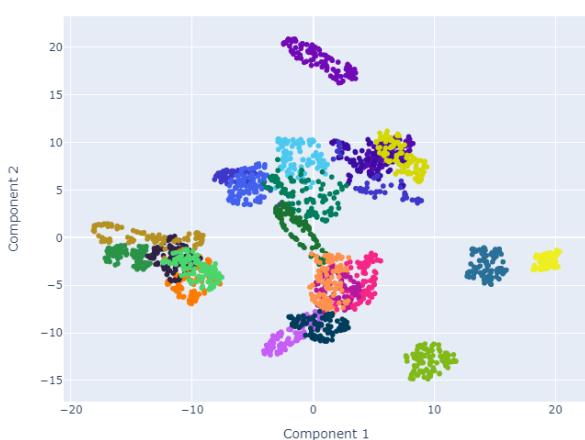
t-SNE (Iter: 1000, Comp: 3, Perp: 35)



t-SNE (Iter: 1000, Comp: 3, Perp: 40)

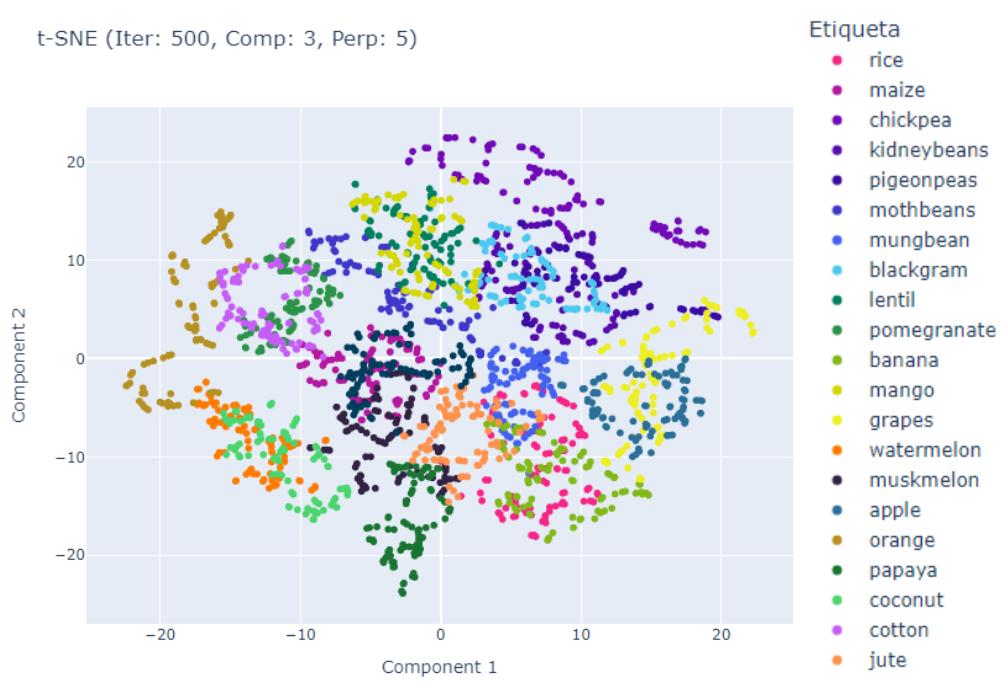


t-SNE (Iter: 1000, Comp: 3, Perp: 45)

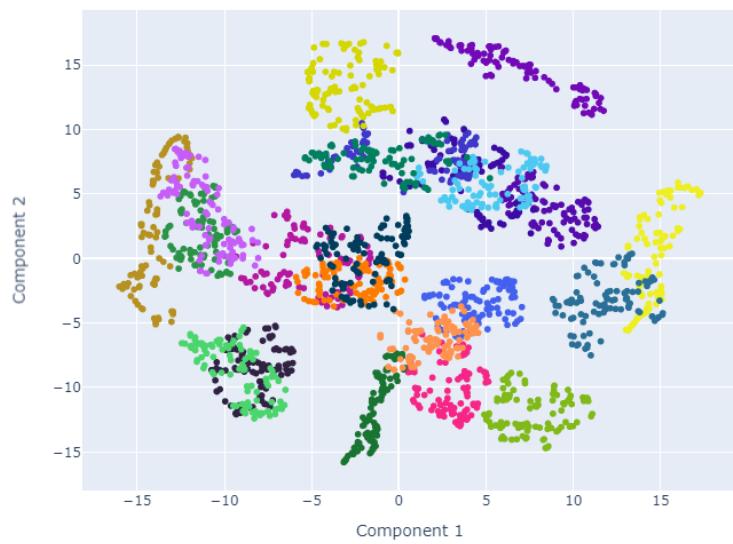


- 500 iteraciones, 3 componentes y distintas perplejidades:

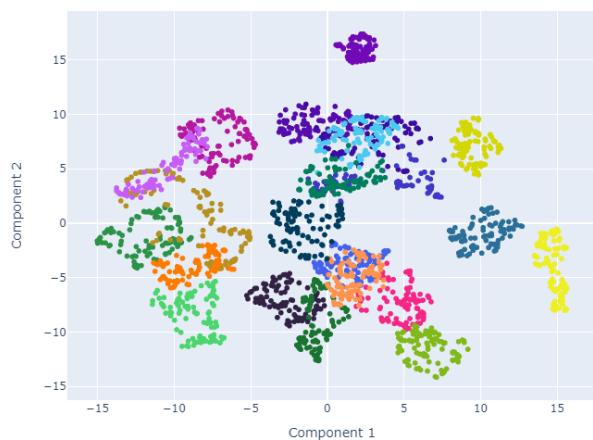
t-SNE (Iter: 500, Comp: 3, Perp: 5)



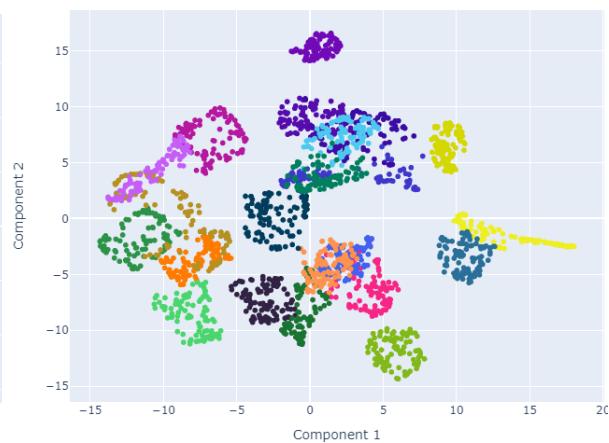
t-SNE (Iter: 500, Comp: 3, Perp: 15)



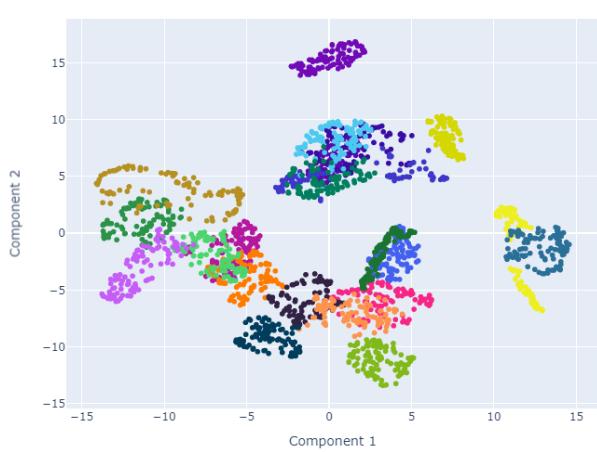
t-SNE (Iter: 500, Comp: 3, Perp: 25)



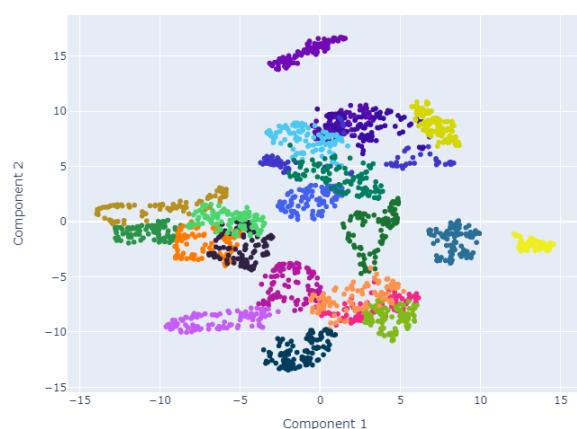
t-SNE (Iter: 500, Comp: 3, Perp: 30)



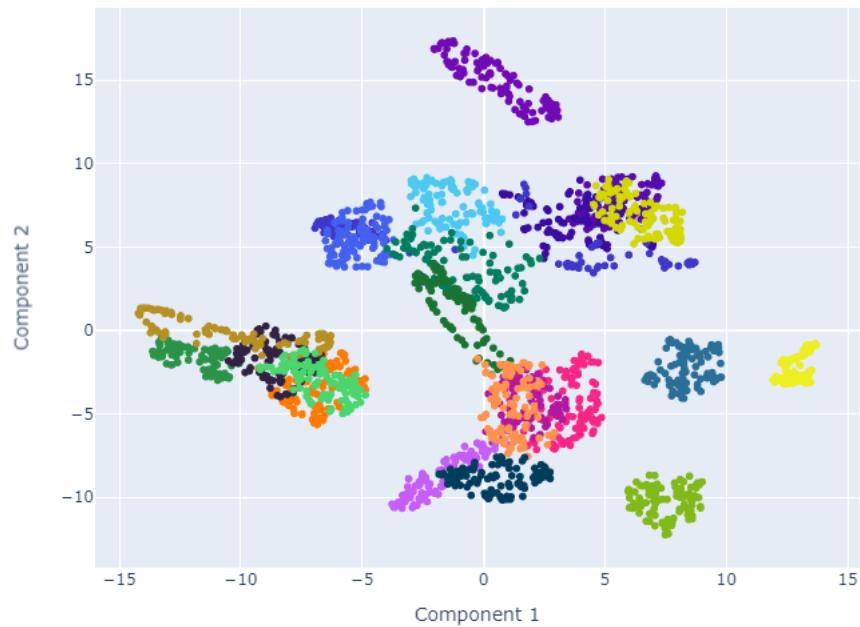
t-SNE (Iter: 500, Comp: 3, Perp: 35)



t-SNE (Iter: 500, Comp: 3, Perp: 40)

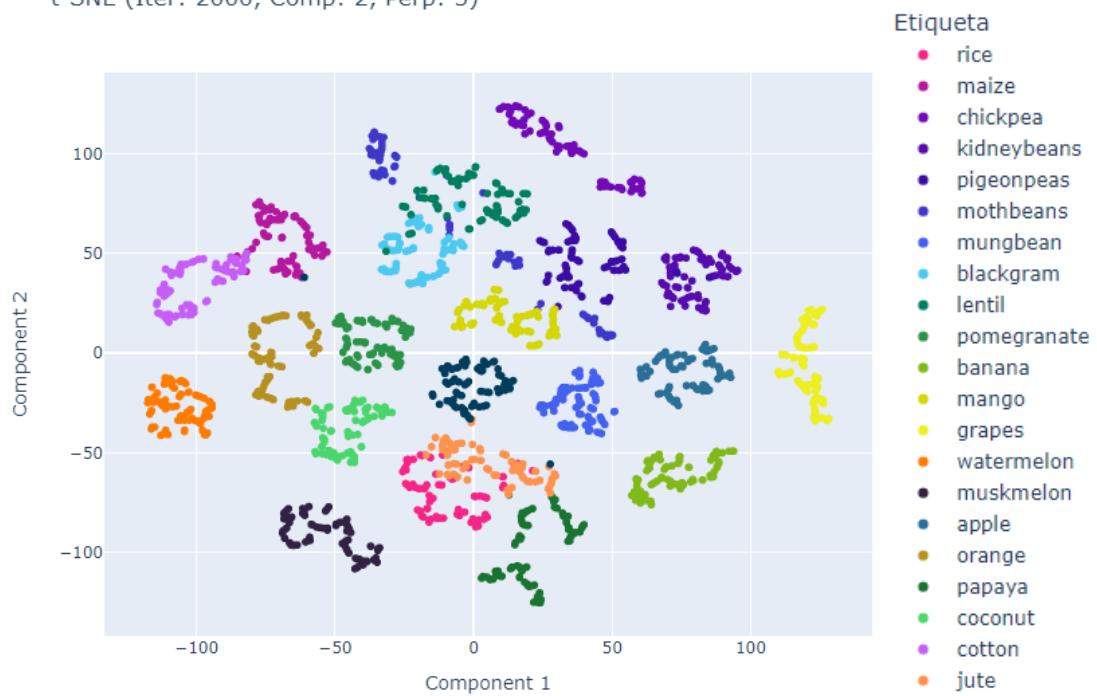


t-SNE (Iter: 500, Comp: 3, Perp: 45)

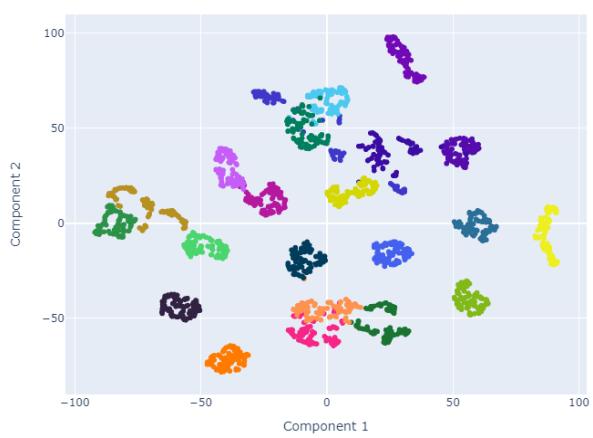


- 200 iteraciones, 2 componentes y distintas perplejidades:

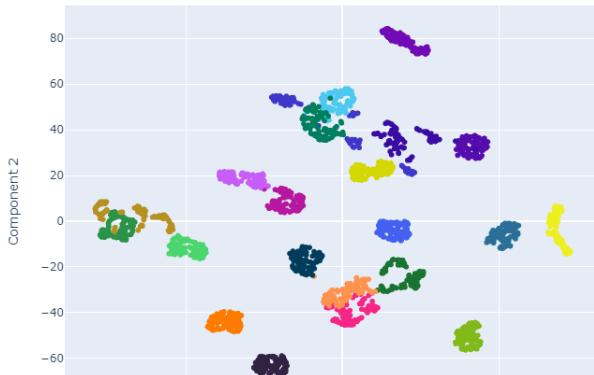
t-SNE (Iter: 2000, Comp: 2, Perp: 5)



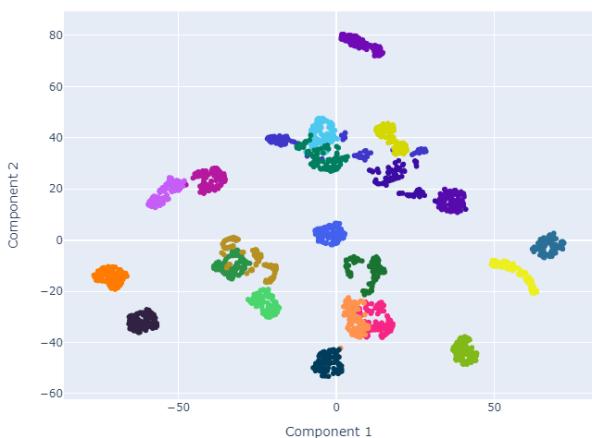
t-SNE (Iter: 2000, Comp: 2, Perp: 15)



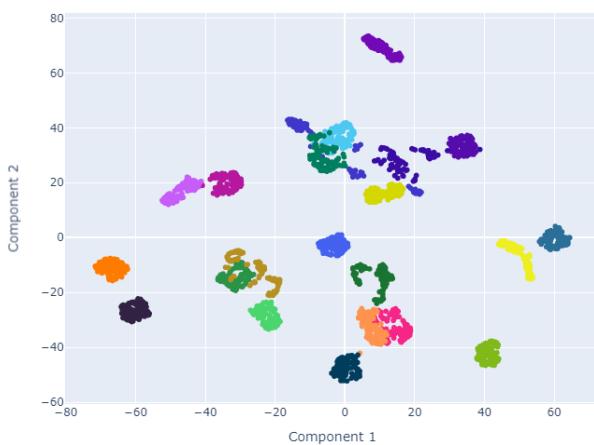
t-SNE (Iter: 2000, Comp: 2, Perp: 25)



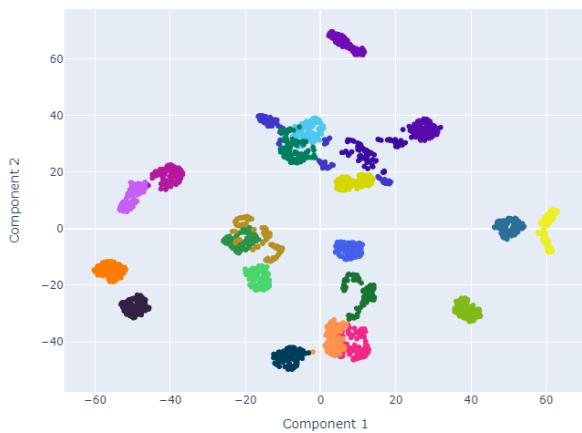
t-SNE (Iter: 2000, Comp: 2, Perp: 30)



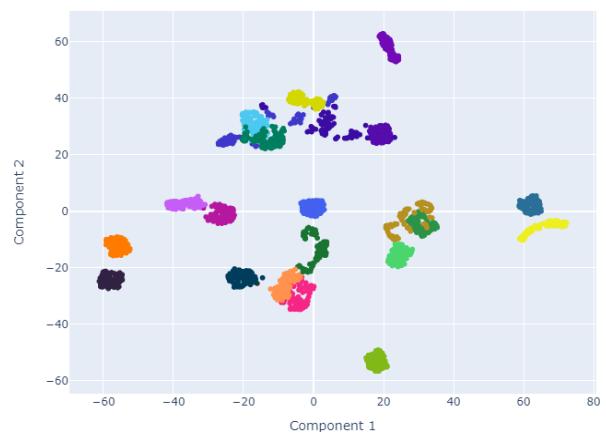
t-SNE (Iter: 2000, Comp: 2, Perp: 35)



t-SNE (Iter: 2000, Comp: 2, Perp: 40)

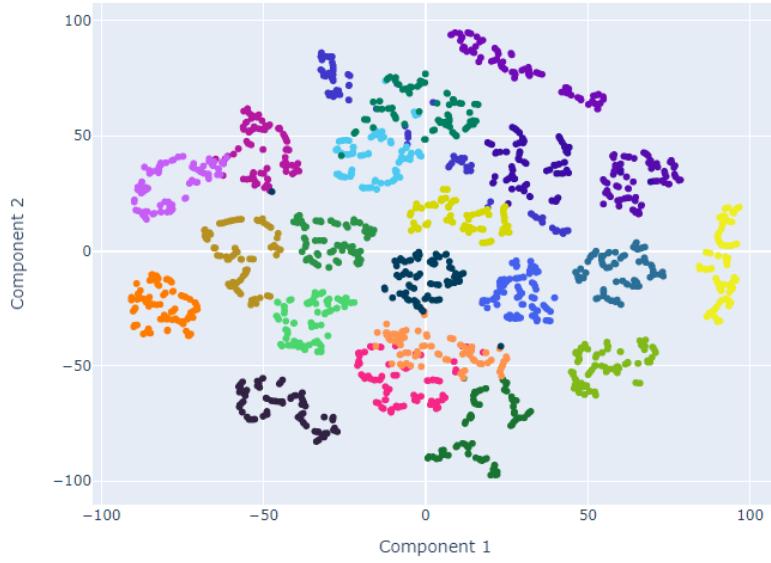


t-SNE (Iter: 2000, Comp: 2, Perp: 45)



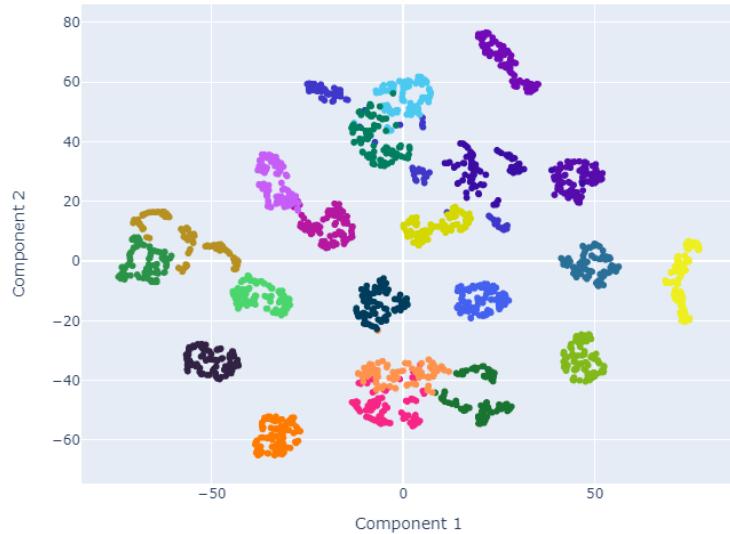
- 1000 iteraciones, 2 componentes y distintas perplejidades:

t-SNE (Iter: 1000, Comp: 2, Perp: 5)

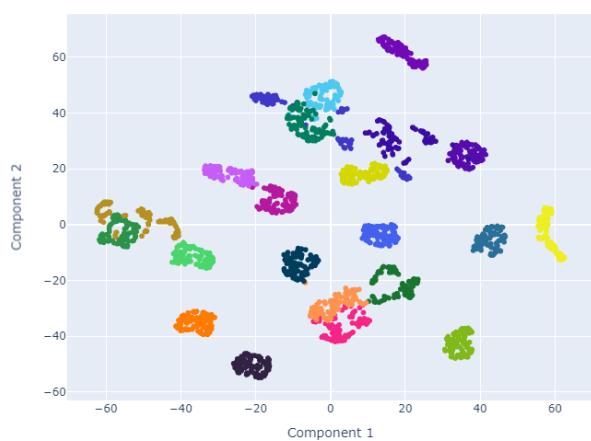
**Etiqueta**

rice
maize
chickpea
kidneybeans
pigeonpeas
mothbeans
mungbean
blackgram
lentil
pomegranate
banana
mango
grapes
watermelon
muskmelon
apple
orange
papaya
coconut
cotton
jute

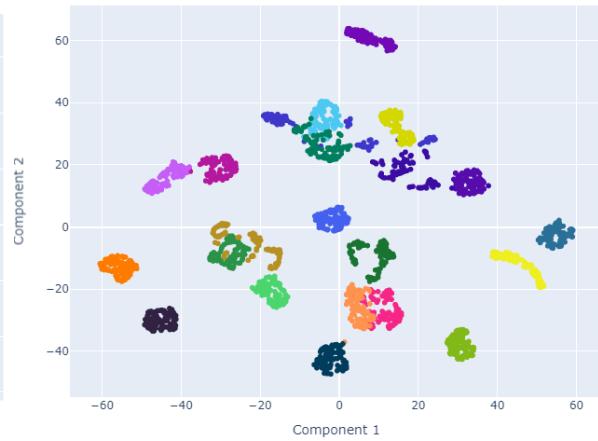
t-SNE (Iter: 1000, Comp: 2, Perp: 15)



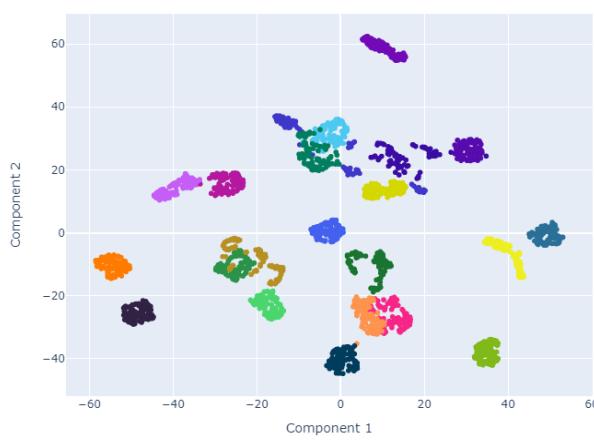
t-SNE (Iter: 1000, Comp: 2, Perp: 25)



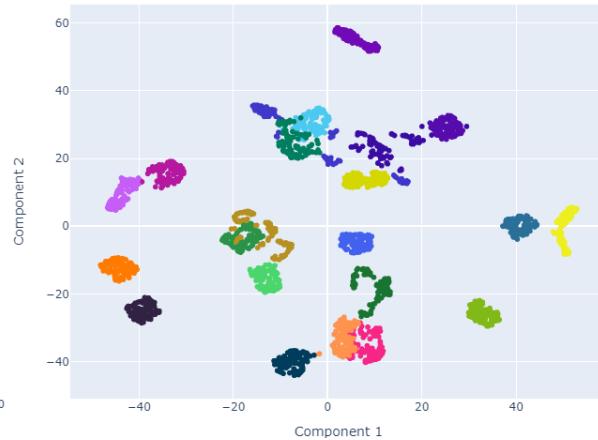
t-SNE (Iter: 1000, Comp: 2, Perp: 30)



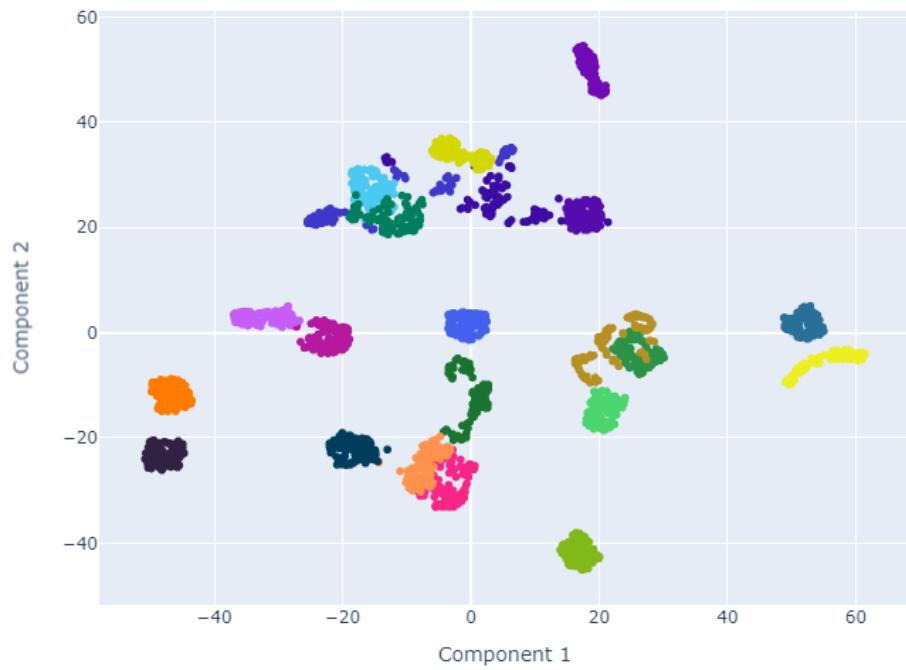
t-SNE (Iter: 1000, Comp: 2, Perp: 35)



t-SNE (Iter: 1000, Comp: 2, Perp: 40)

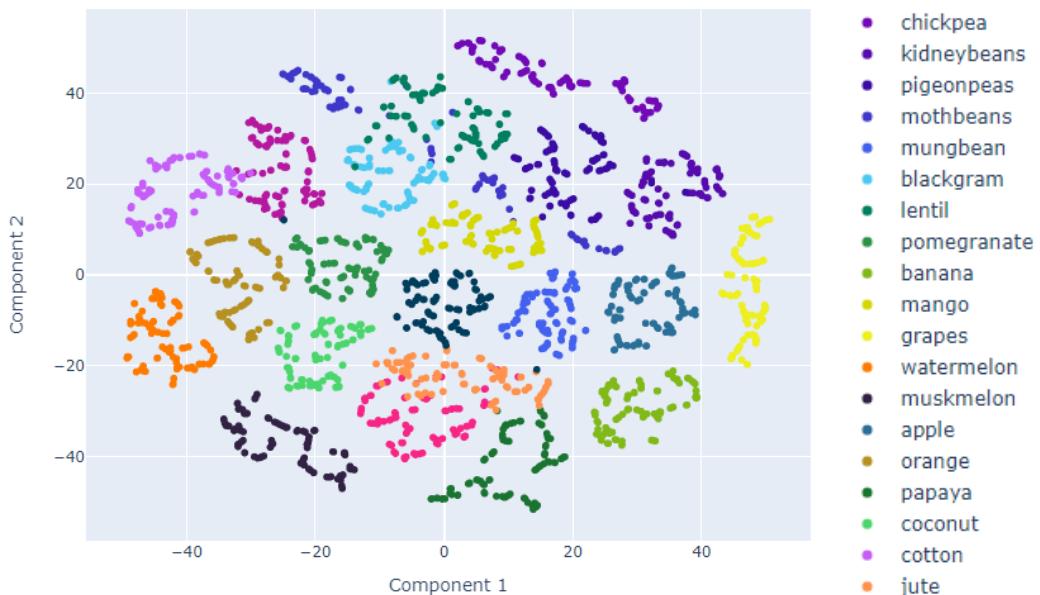


t-SNE (Iter: 1000, Comp: 2, Perp: 45)

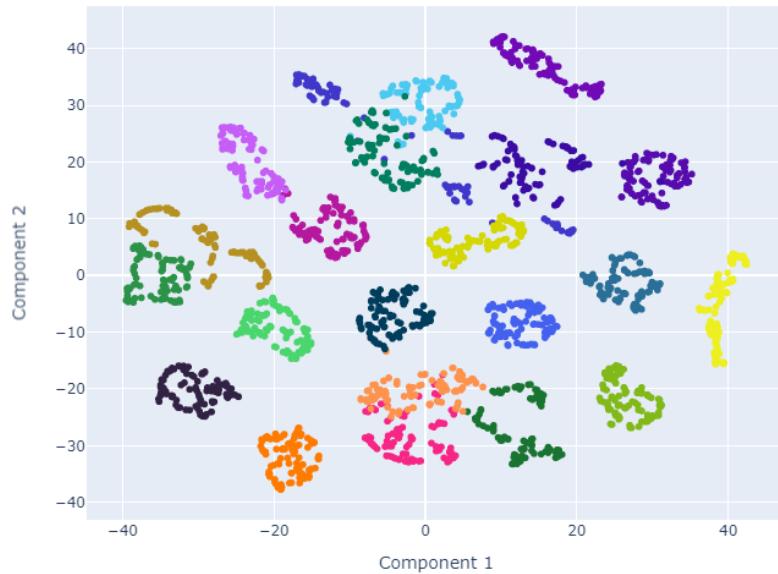


- 500 iteraciones, 2 componentes y distintas perplejidades:

t-SNE (Iter: 500, Comp: 2, Perp: 5)

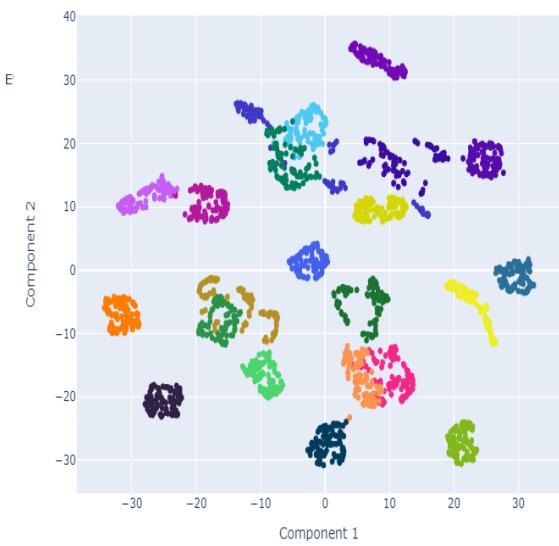
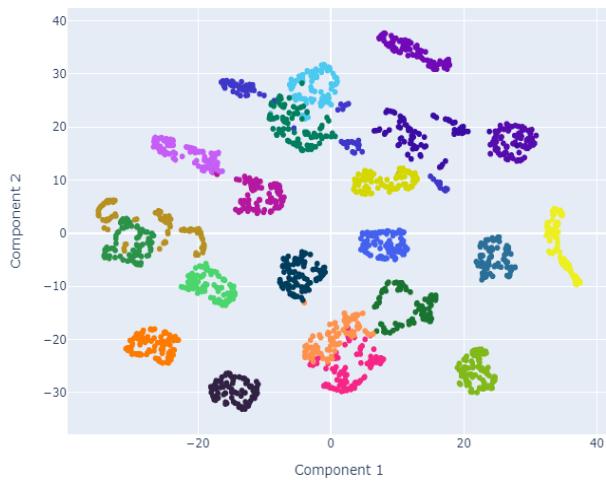


t-SNE (Iter: 500, Comp: 2, Perp: 15)

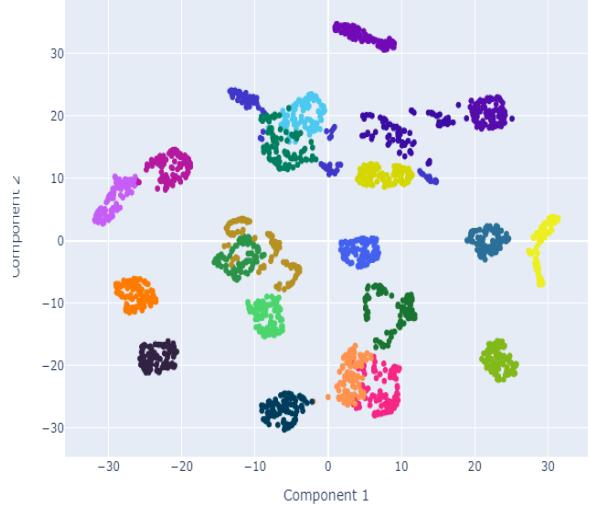


t-SNE (Iter: 500, Comp: 2, Perp: 35)

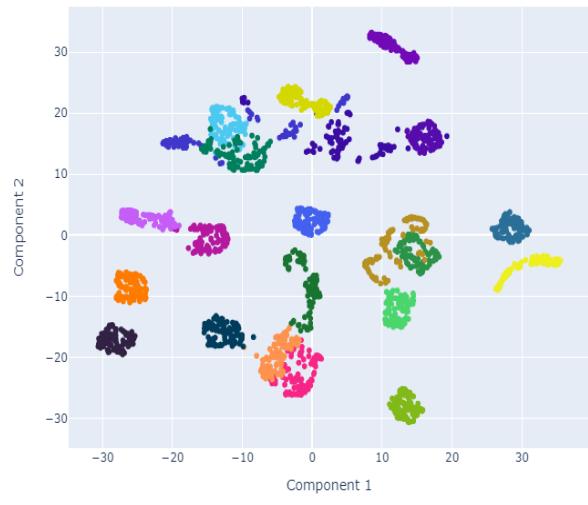
t-SNE (Iter: 500, Comp: 2, Perp: 25)



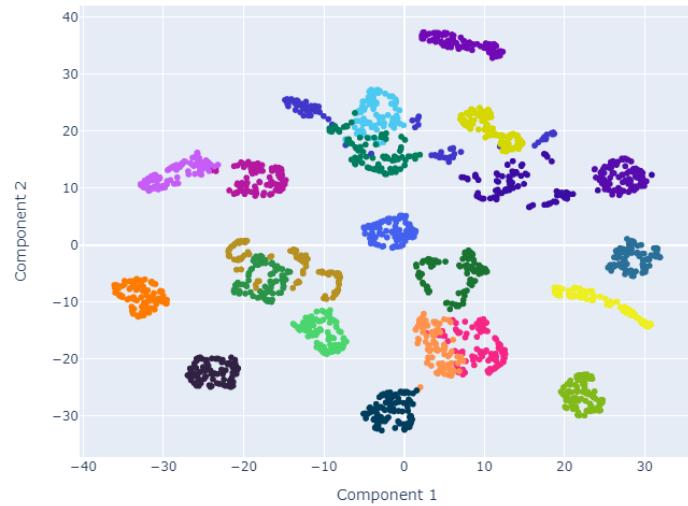
t-SNE (Iter: 500, Comp: 2, Perp: 40)



t-SNE (Iter: 500, Comp: 2, Perp: 45)



t-SNE (Iter: 500, Comp: 2, Perp: 30)



Al realizar los gráficos de los resultados al aplicar t-SNE se observa lo siguiente:

Observaciones para 2000 iteraciones, 3 componentes y distintas perplejidades:

- Perplejidad 5: los puntos se encuentran dispersos y no se identifica formación de grupos claros.
- Perplejidad 15: Comienzan a verse grupos de puntos más cercanos aunque dispersos dentro del grupo. Parecen dominar las variaciones locales.
- Perplejidad 25: Los grupos pequeños se compactan y los grandes se reacomodan, formando dos nuevos grupos grandes pero más claramente diferenciados. Estos grupos grandes siguen teniendo los datos dispersos dentro del grupo.
- Perplejidad 30: No hay cambios notables en la distribución de los puntos con respecto a la perplejidad 25.
- Perplejidad 35: Los grupos pequeños (que solo están formados por uno o 2 tipos de cultivos) se mantienen igual. Excepto por uno de ellos formado por dos cultivos. Donde los mismos se separan claramente pero sin encontrarse tan alejados. Los cultivos que formaban este grupo tenían características similares en sus variables. Los grupos grandes se reacomodan nuevamente y algunos cultivos se mueven de grupo. Ya no hay dos grupos grandes, sino 3, aunque dos de ellos antes formaban un mismo grupo y ahora se hallan separados pero muy cercanos.
- Perplejidad 40: La distribución se asemeja nuevamente a la de la perplejidad 30, aunque los cultivos del grupo pequeño siguen separados sin estar muy lejos entre ellos. Además, uno de los grupos pequeños formado por un solo cultivo que antes se hallaba separado de todos los demás puntos, ahora integra uno de los grupos grandes.
- Perplejidad 45: El grupo pequeño formado por un solo cultivo que se había integrado a un grupo grande ahora se distancia nuevamente. El grupo chico formado por dos cultivos con características similares continua separado y formando dos grupos. Los 2 grupos grandes se encuentran más compactados, más claramente definidos todos los grupos observados se encuentran alejados unos de otros. Hay un total de 6 grupos.

Observaciones para 1000 iteraciones, 3 componentes y distintas perplejidades:

Todas las perplejidades se comportan como su correspondiente para 2000 iteraciones y 3 componentes.

Observaciones para 500 iteraciones, 3 componentes y distintas perplejidades:

Todas las perplejidades se comportan en general como su correspondiente para 2000 iteraciones y 3 componentes. Pero los grupos formados nunca llegan a compactarse y los puntos dentro de los mismos se encuentran muy dispersos. En general, para las distintas perplejidades los grupos formados no solo tenían sus puntos muy dispersos, sino que la distancia entre dichos grupos era muy baja y, en ocasiones, a perplejidades más bajas costaba diferenciarlos.

Observaciones para 2000 iteraciones, 2 componentes y distintas perplejidades:

- Perplejidad 5: los puntos se encuentran dispersos y no se identifica formación de grupos claros.

- Perplejidad 15: Comienzan a verse grupos de puntos más cercanos y compactos aunque prácticamente hay un grupo por cada tipo de cultivo. Parecen dominar las variaciones locales.
- Perplejidad 25: Los grupos se compactan aún más y se identifican 13 grupos la mayoría de los cuales están formados por un único cultivo.
- Perplejidad 30: No hay cambios notables en la distribución de los puntos con respecto a la perplejidad 25.
- Perplejidad 35: Comienza a notarse mayor asociación entre grupos que antes se encontraban separados.
- Perplejidad 40: Sigue habiendo muchos grupos separados y formados por un único cultivo. Pero varios de ellos se acercan encontrándose a distancias pequeñas entre ellos, comenzando a formar grupos más grandes.
- Perplejidad 45: Los grupos más grandes que se comenzaban a ver anteriormente se compactan un poco más. Hay un total de 10 grupos, algunos de los cuales están formados por un único cultivo que se encuentran cercanos a otro también formado por un único cultivo.

Observaciones para 1000 iteraciones, 2 componentes y distintas perplejidades:
Todas las perplejidades se comportan como su correspondiente para 2000 iteraciones y 2 componentes.

Observaciones para 500 iteraciones, 2 componentes y distintas perplejidades:
Todas las perplejidades se comportan en general como su correspondiente para 1000 iteraciones y 2 componentes. Pero los grupos formados nunca llegan a compactarse y los puntos dentro de los mismos se encuentran dispersos. En general, para las distintas perplejidades los grupos formados no solo tenían sus puntos dispersos, pero la separación entre dichos grupos se encontraba mejor definida que cuando se usaron 500 iteraciones y 3 componentes.

Las observaciones anteriores sugieren:

- Perplejidad baja (5): Los puntos están dispersos y no se forman grupos claros. Esto sugiere que una baja perplejidad no es suficiente para capturar la estructura de los datos de manera significativa.
- Perplejidad moderada (15-30): Se empiezan a formar grupos de puntos más cercanos, aunque pueden estar algo dispersos dentro del grupo. Esto indica que una perplejidad moderada permite una mejor agrupación de los datos, pero puede haber aún cierta variabilidad local.
- Perplejidad alta (35-45): Se observa una mayor asociación entre grupos previamente separados. Los grupos tienden a compactarse y a estar más claramente definidos. Esto sugiere que una perplejidad alta ayuda a agrupar los datos de manera más coherente.
- Número de componentes (2 vs. 3): La reducción a dos componentes parece llevar a una mayor separación entre grupos. Sin embargo, también puede conducir a una mayor fragmentación de los grupos, especialmente cuando se utiliza una perplejidad alta.

- Número de iteraciones (500 vs. 1000 vs. 2000): Aumentar el número de iteraciones tiende a mejorar la formación de grupos y la coherencia de los mismos. Con 500 iteraciones, la separación entre grupos es menos definida.
- Para 3 componentes y perplejidades altas menos grupos: Esto indica que la representación tridimensional está capturando similitudes en mayor número de datos. Cada grupo representa una agrupación de cultivos que comparten similitudes en sus características.
- Para 2 componentes y perplejidades altas más grupos: Al tener más grupos, se indica una mayor fragmentación y separación de los datos.

Es posible que los datos tengan una estructura intrínseca que no se puede representar de manera efectiva en solo dos dimensiones. Esto puede llevar a una mayor fragmentación de los datos en múltiples grupos. Al reducir la dimensionalidad a solo dos componentes, es posible que se pierda información crucial sobre las relaciones entre los puntos.

Las observaciones anteriores sugieren:

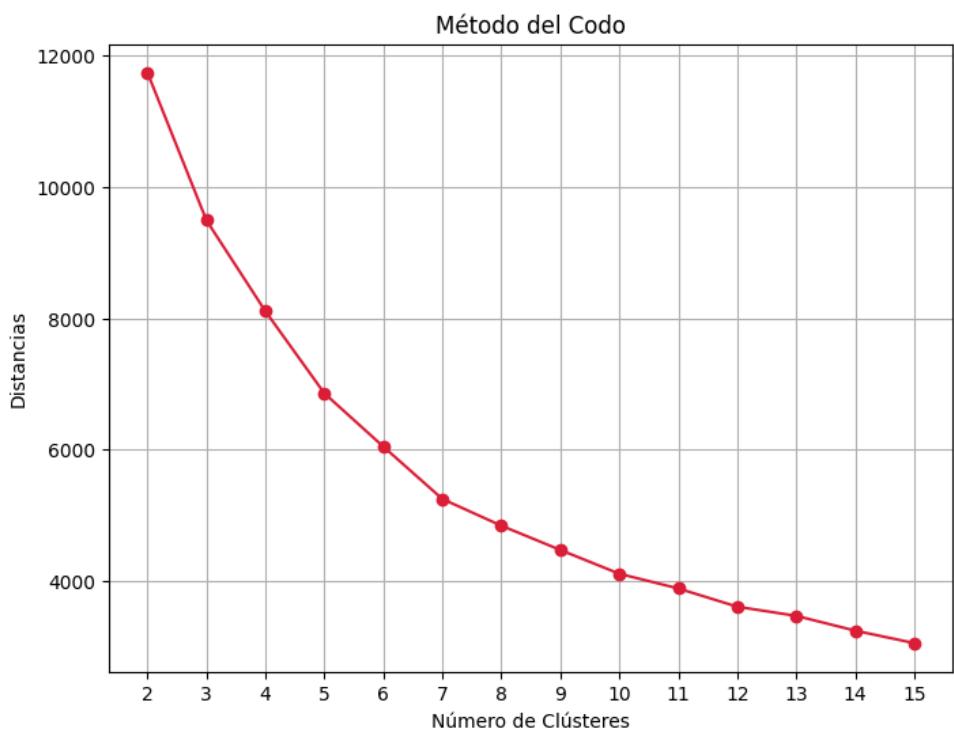
- Perplejidad baja (5): Los puntos están dispersos y no se forman grupos claros. Esto sugiere que una baja perplejidad no es suficiente para capturar la estructura de los datos de manera significativa.
- Perplejidad moderada (15-30): Se empiezan a formar grupos de puntos más cercanos, aunque pueden estar algo dispersos dentro del grupo. Esto indica que una perplejidad moderada permite una mejor agrupación de los datos, pero puede haber aún cierta variabilidad local.
- Perplejidad alta (35-45): Se observa una mayor asociación entre grupos previamente separados. Los grupos tienden a compactarse y a estar más claramente definidos. Esto sugiere que una perplejidad alta ayuda a agrupar los datos de manera más coherente.
- Número de componentes (2 vs. 3): La reducción a dos componentes parece llevar a una mayor separación entre grupos. Sin embargo, también puede conducir a una mayor fragmentación de los grupos, especialmente cuando se utiliza una perplejidad alta.
- Número de iteraciones (500 vs. 1000 vs. 2000): Aumentar el número de iteraciones tiende a mejorar la formación de grupos y la coherencia de los mismos. Con 500 iteraciones, la separación entre grupos es menos definida.
- Para 3 componentes y perplejidades altas menos grupos: Esto indica que la representación tridimensional está capturando similitudes en mayor número de datos. Cada grupo representa una agrupación de cultivos que comparten similitudes en sus características.
- Para 2 componentes y perplejidades altas más grupos: Al tener más grupos, se indica una mayor fragmentación y separación de los datos.

Es posible que los datos tengan una estructura intrínseca que no se puede representar de manera efectiva en solo dos dimensiones. Esto puede llevar a una mayor fragmentación de los datos en múltiples grupos. Al reducir la dimensionalidad a solo dos componentes, es posible que se pierda información crucial sobre las relaciones entre los puntos.

6. K-Means

Aplicar K-means y analizar los resultados obtenidos variando el número de clusters y obtener el número óptimo de clusters mediante GAP. Realizar un gráfico en 3D de utilizando tres atributos de los datos y donde los colores estén asociados a los clusters.

Se aplica la técnica del codo para elegir el número óptimo de clusters.

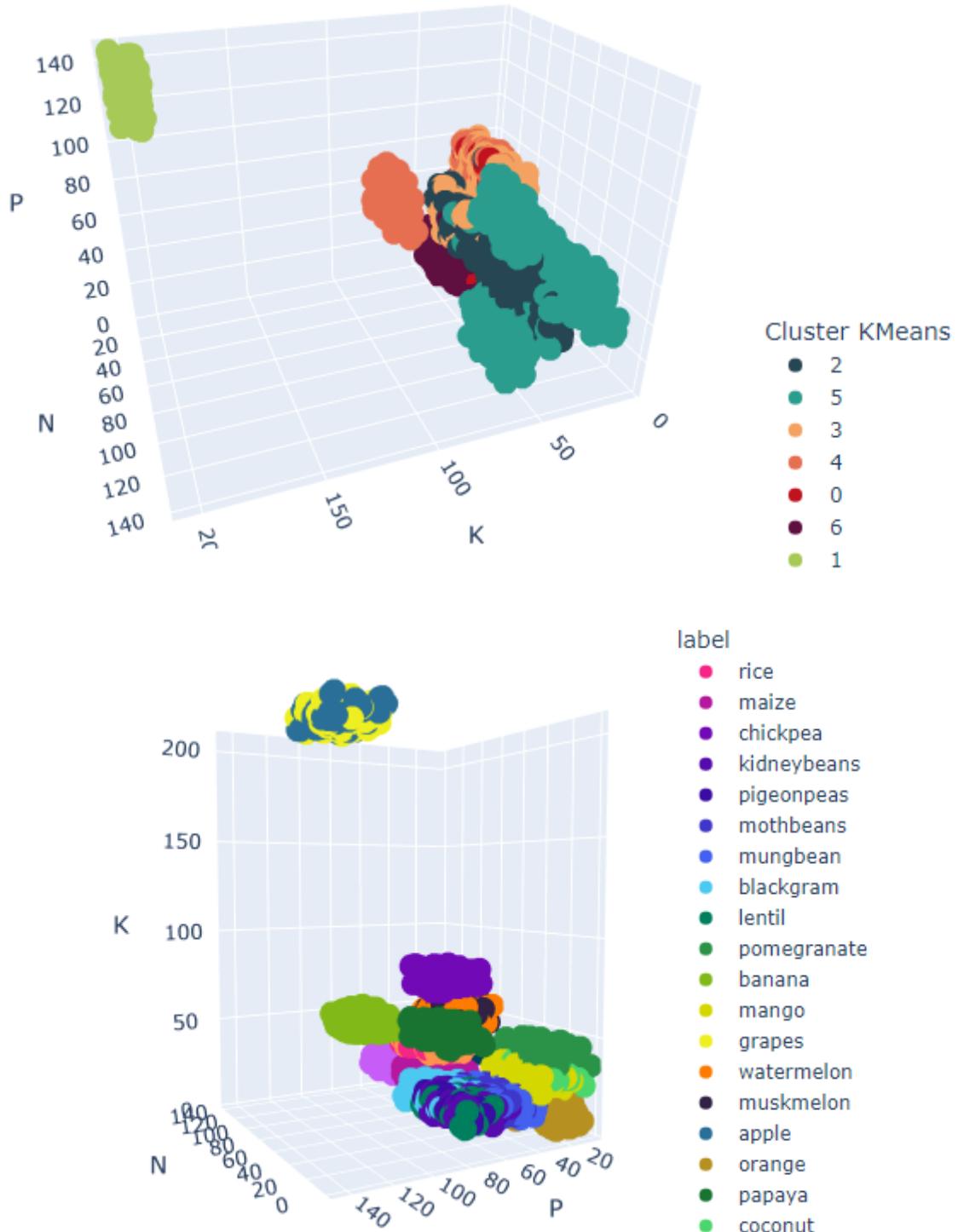


Se observa en el gráfico que a partir de 7 clusters, las distancias no disminuyen significativamente. Por lo cual, se elige dicho número para utilizar en el modelo KMeans.

Primeros 5 registros del nuevo dataframe:

	N	P	K	temperature	humidity	ph	rainfall	label	Cluster	KMeans
0	90	42	43	20.879744	82.002744	6.502985	202.935536	rice	2	
1	85	58	41	21.770462	80.319644	7.038096	226.655537	rice	2	
2	60	55	44	23.004459	82.320763	7.840207	263.964248	rice	2	
3	74	35	40	26.491096	80.158363	6.980401	242.864034	rice	2	
4	78	42	42	20.130175	81.604873	7.628473	262.717340	rice	2	

Se graficarán los clusters obtenidos y debajo otro gráfico 3D mostrando los cultivos correspondientes



Se obtendrá de manera más ordenada la lista de cultivos que pertenecen a cada cluster

- Cultivos en cluster 0 ['maize' 'banana' 'watermelon' 'muskmelon' 'papaya' 'cotton' 'coffee']

- Cultivos en cluster 1 ['maize' 'pigeonpeas' 'mothbeans' 'mungbean' 'blackgram' 'lentil' 'mango' 'orange' 'papaya']
- Cultivos en cluster 2 ['pigeonpeas' 'pomegranate' 'orange' 'papaya' 'coconut']
- Cultivos en cluster 3 ['maize' 'chickpea' 'kidneybeans' 'pigeonpeas' 'lentil']
- Cultivos en cluster 4 ['grapes' 'apple']
- Cultivos en cluster 5 ['rice' 'pigeonpeas' 'papaya' 'coconut' 'jute' 'coffee']
- Cultivos en cluster 6 ['pigeonpeas' 'mothbeans' 'lentil' 'mango']

Cantidades de observaciones por cluster:

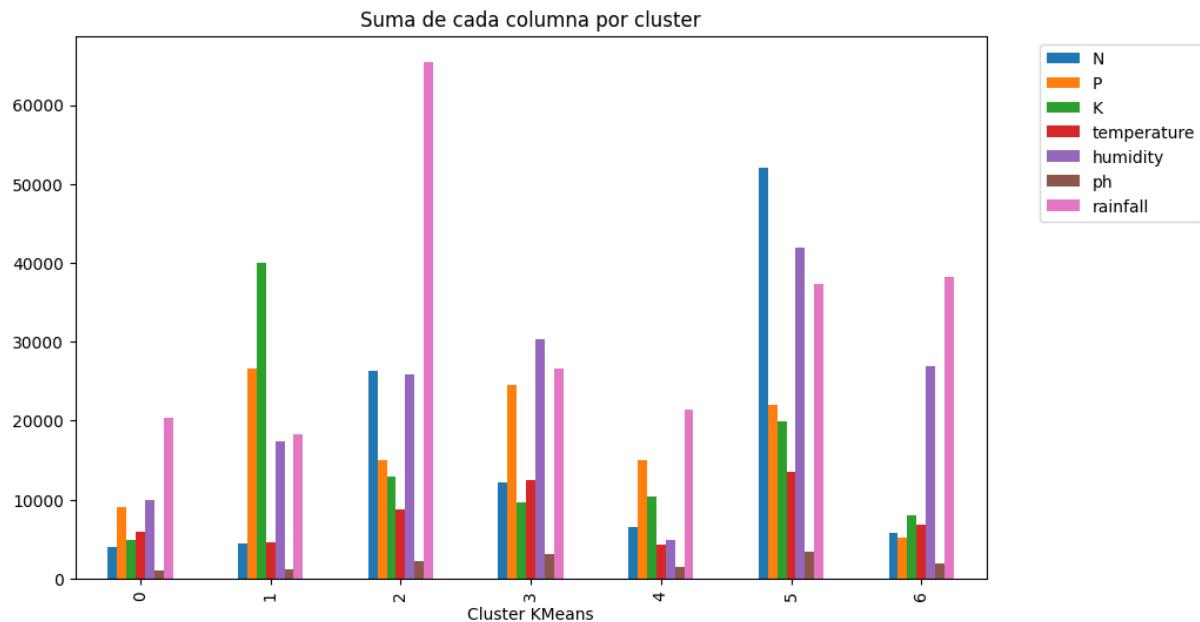
- 0 524
- 1 430
- 2 291
- 3 222
- 4 200
- 5 335
- 6 198

Clusters con mayor cantidad de observaciones (Cluster 0 y 1), sugieren que es un grupo relativamente grande y heterogéneo en comparación con los demás clusters. Puede representar una categoría de cultivos común o podría haber una mayor variabilidad en las características agronómicas de los cultivos en este grupo.

Clusters con menos observaciones sugieren que los cultivos en este grupo pueden tener características más específicas que los hacen menos comunes en comparación con otros grupos.

Se obtienen las medias de cada cluster por columna

	N	P	K	temperature	humidity	ph	rainfall
Cluster KMeans							
0	20.262626	45.545455	24.500000	29.824545	49.944578	5.415925	102.786330
1	21.990000	133.375000	200.000000	23.240259	87.104305	5.977800	91.133304
2	78.591045	44.602985	38.534328	25.987811	77.208122	6.650851	195.380596
3	28.153488	56.976744	22.567442	28.904633	70.685337	7.108328	61.908484
4	29.518018	67.617117	47.036036	19.527767	22.072149	6.502897	96.499741
5	99.204198	42.055344	38.057252	25.663235	80.026258	6.416158	71.166221
6	20.048110	17.920962	27.766323	23.659163	92.291528	6.441982	131.458755



- Columna N: cluster 0 supera ampliamente a los demás en valores de N.
- Columna K: El cluster 4 supera ampliamente a los demás en contenido de K. El resto, si bien presentan valores muy por debajo del cluster 4, tienen valores muy similares entre ellos. Excepto por el 2 y 6 que presentan las barras más bajas.
- Columna P: El cluster 4 supera a los otros en contenido de P y el cluster 2 presenta un valor muy bajo en relación a los demás.
- Columna rainfall: El cluster 5 supera ampliamente a los demás en rainfall. El resto tienen valores muy similares entre ellos.
- Columna ph: Todos los cluster presentan valores muy similares de ph.
- Columna temerature: Todos los cluster presentan valores muy similares de temperature.
- Columna humidity: La mayoría de los cluster muestran valores de humedad relativamente altos, siendo el 0 el que presenta la barra más alta y el 3 la más baja.

Las observaciones anteriores sugieren:

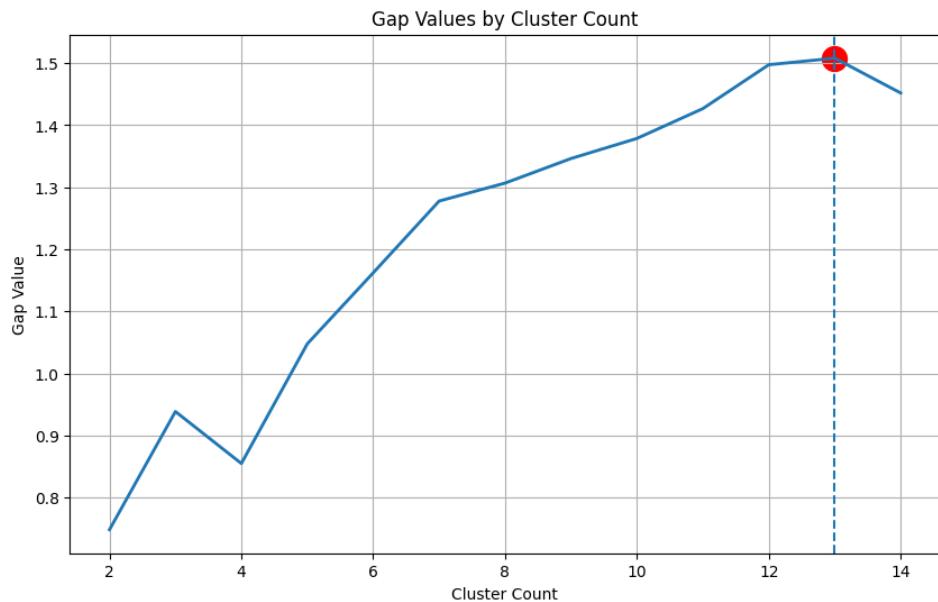
- En el dataset original, se notó que ciertos cultivos como ['pigeonpeas' 'pomegranate' 'orange' 'papaya' 'coconut'] compartían características similares, como el bajo contenido de K y P. Al aplicar k-means, estos cultivos fueron agrupados en el mismo cluster (cluster 2), lo que es coherente con las observaciones originales ya que estos cluster presentan bajo contenido de los mismos.
- En el dataset original, se observó que 'grapes' y 'apple' tenían características distintivas, superaban ampliamente a los demás cultivos en contenido de K y P. K-means también los separó en un clúster único (cluster 4), lo que respalda la observación.

- Se observó que algunos clusters tienen valores destacados en ciertas características agronómicas cantidad de lluvia (rainfall). El 5 es el que presenta la barra más alta y está compuesto por los cultivos ['rice' 'pigeonpeas' 'papaya' 'coconut' 'jute' 'coffee'] que eran los que se habían observado como aquellos que presentaban elevados valores de rainfall en el dataset originalmente.
- Se observó que el cluster 3 compuesto por los cultivos ['maize' 'chickpea' 'kidneybeans' 'pigeonpeas' 'lentil'] presenta la barra más baja para la característica humidity. Se había identificado en el dataset original a estos cultivos como aquellos que presentaban algunos de los valores más bajos de esa característica.
- Se notó que para la variable N, el cluster 0 compuesto por los cultivos ['maize' 'banana' 'watermelon' 'muskmelon' 'papaya' 'cotton' 'coffee'] es el que presenta la barra más alta para dicha característica. Los cultivos presentes en el cluster se habían identificado en el dataset original como aquellos con los mayores valores de N.

Se procede ahora a determinar el número óptimo de clusters por el método GAP.

Optimal clusters: 13		
	n_clusters	gap_value
0	2.0	0.748177
1	3.0	0.938509
2	4.0	0.854800
3	5.0	1.047672
4	6.0	1.162175
5	7.0	1.277849
6	8.0	1.306815
7	9.0	1.346396
8	10.0	1.378783
9	11.0	1.426880
10	12.0	1.497626
11	13.0	1.508163
12	14.0	1.452049

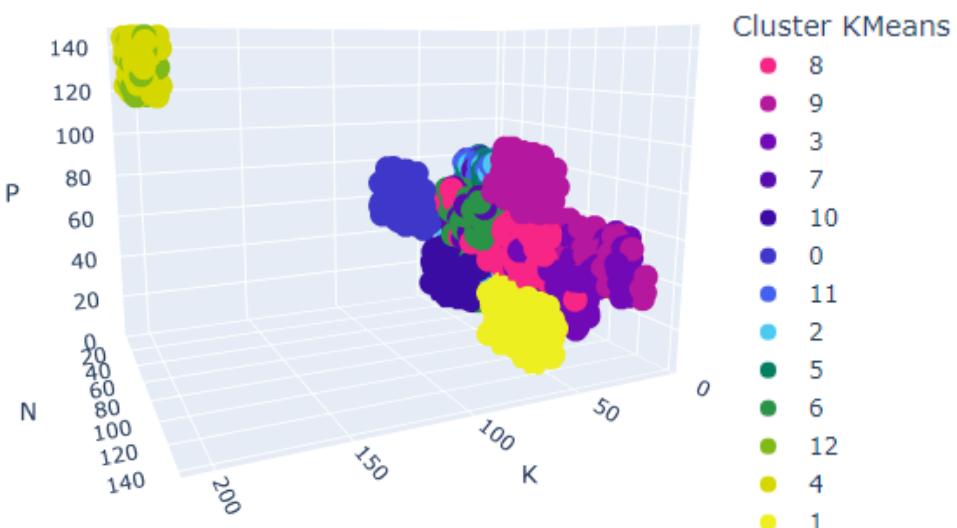
Se grafican los gap values con respecto al número de clusters:

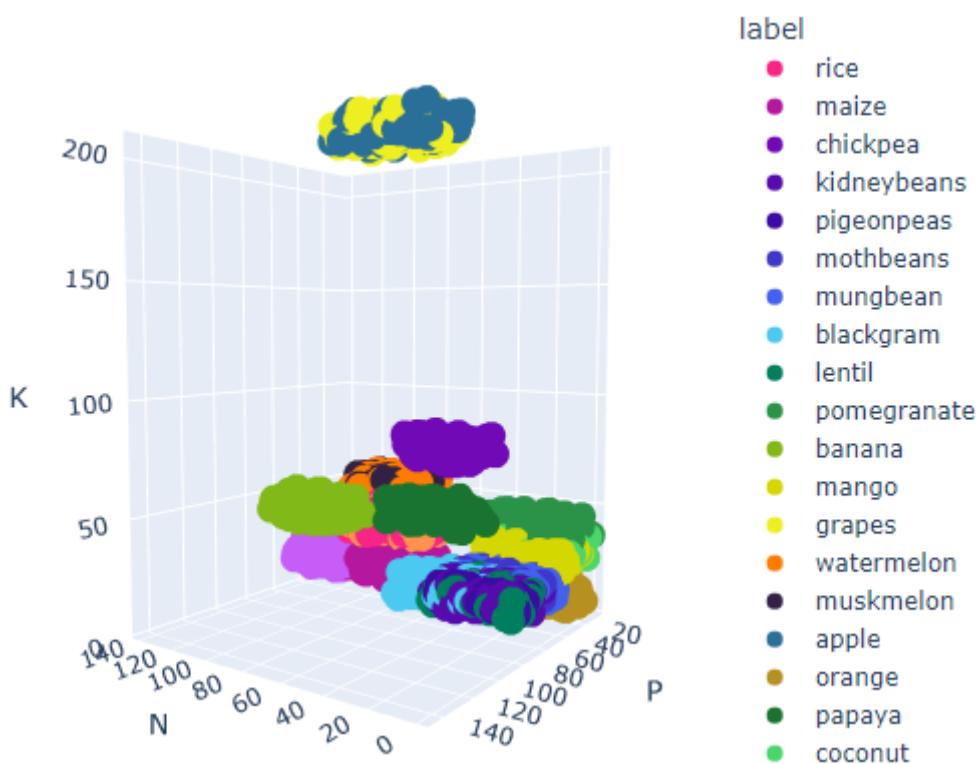


Se aplica KMeans para 13 clusters y se muestran los primeros 5 registros del nuevo dataframe:

	N	P	K	temperature	humidity	ph	rainfall	label	Cluster	KMeans
0	90	42	43	20.879744	82.002744	6.502985	202.935536	rice	11	11
1	85	58	41	21.770462	80.319644	7.038096	226.655537	rice	11	11
2	60	55	44	23.004459	82.320763	7.840207	263.964248	rice	11	11
3	74	35	40	26.491096	80.158363	6.980401	242.864034	rice	11	11
4	78	42	42	20.130175	81.604873	7.628473	262.717340	rice	11	11

Se graficarán los clusters obtenidos y debajo otro gráfico 3D mostrando los cultivos correspondientes.





Se obtendrá de manera más ordenada la lista de cultivos que pertenecen a cada cluster

- Cultivos en cluster 0 ['watermelon' 'muskmelon']
- Cultivos en cluster 1 ['pigeonpeas' 'mothbeans' 'mungbean' 'blackgram' 'lentil' 'mango' 'orange' 'papaya']
- Cultivos en cluster 2 ['rice' 'pigeonpeas' 'pomegranate' 'orange' 'papaya' 'coconut' 'jute']
- Cultivos en cluster 3 ['grapes' 'apple']
- Cultivos en cluster 4 ['chickpea' 'kidneybeans' 'pigeonpeas' 'lentil']
- Cultivos en cluster 5 ['maize' 'banana' 'papaya' 'cotton' 'coffee']
- Cultivos en cluster 6 ['pigeonpeas' 'pomegranate' 'orange']
- Cultivos en cluster 7 ['chickpea']
- Cultivos en cluster 8 ['pigeonpeas' 'mothbeans' 'blackgram' 'lentil' 'mango' 'orange']
- Cultivos en cluster 9 ['pigeonpeas' 'mothbeans' 'mango']
- Cultivos en cluster 10 ['orange' 'papaya']
- Cultivos en cluster 11 ['rice' 'pigeonpeas' 'papaya' 'jute' 'coffee']
- Cultivos en cluster 12 ['grapes']

Cantidad de observaciones por cluster:

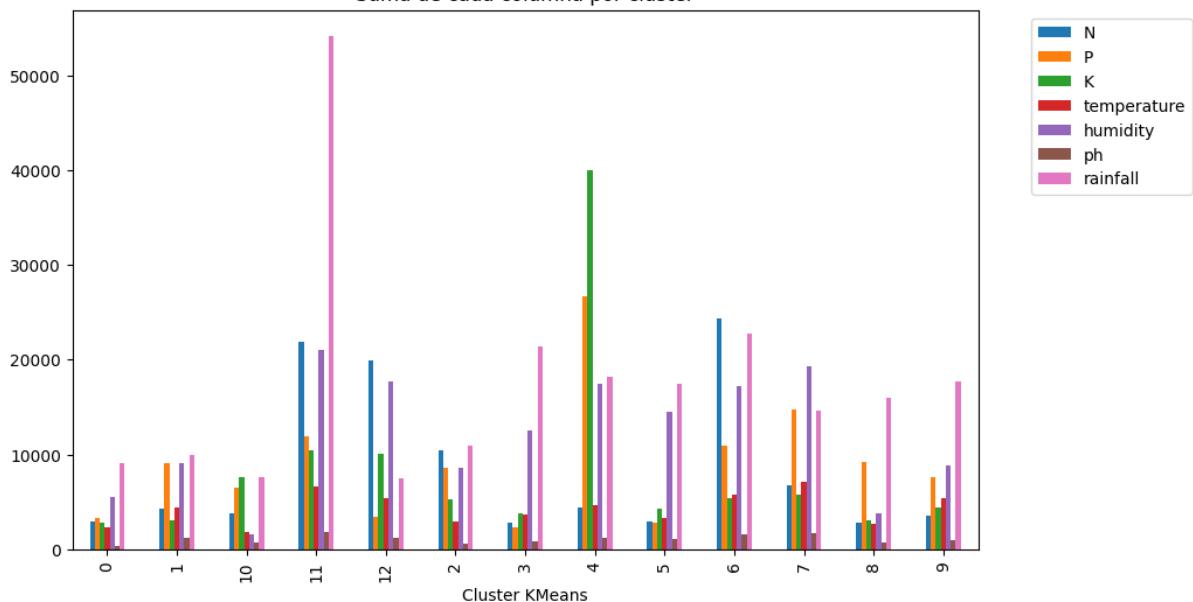
- 0 200
- 1 264
- 10 66
- 11 290
- 12 36
- 2 130

- 3 164
- 4 136
- 5 334
- 6 162
- 7 95
- 8 152
- 9 171

Se obtienen las medias de cada cluster por columna

	N	P	K	temperature	humidity	ph	rainfall
Cluster KMeans							
0	48.383333	55.300000	46.216667	38.168702	92.415234	6.815223	152.618459
1	28.333333	59.823529	19.732026	28.776288	59.833720	7.765935	65.326710
10	40.410526	67.989474	79.831579	18.882241	16.907768	7.403579	79.907281
11	81.036900	44.092251	38.424354	24.678294	77.553287	6.624173	199.791507
12	99.870000	17.360000	50.150000	27.127416	88.751589	6.427292	37.738086
2	98.924528	80.811321	49.603774	27.413722	80.696604	6.007302	103.752192
3	21.601504	17.285714	28.436090	27.759399	94.033224	6.087577	160.864079
4	21.990000	133.375000	200.000000	23.240259	87.104305	5.977800	91.133304
5	18.687500	17.968750	26.975000	20.392078	90.820656	6.742231	108.855418
6	98.473684	44.157895	21.951417	23.634185	69.740364	6.633165	92.361608
7	25.900000	57.003846	22.357692	27.635094	74.357047	6.705335	56.177346
8	20.291971	67.737226	22.291971	20.184459	27.667223	5.706106	117.014503
9	20.325843	42.831461	25.078652	30.668824	49.580945	5.435683	99.432831

Suma de cada columna por cluster



La aplicación de k-means con 13 clusters ha generado una división mucho más fina en comparación con el análisis anterior con 7 clusters. A continuación, se presentan algunas reflexiones sobre los resultados:

- Clusters altamente poblados:

Los clusters 1, 11 y 5 tienen una alta cantidad de observaciones (264, 290 y 334 respectivamente). Esto sugiere que estos grupos son relativamente grandes y podrían representar categorías comunes o grupos con una mayor variabilidad en las características de los cultivos.

- Clusters con pocas observaciones:

Los clusters 12 y 10 tienen muy pocas observaciones (36 y 66 respectivamente). Esto indica que los cultivos en estos grupos tienen características bastante distintivas y son menos comunes en comparación con otros grupos.

- Distribución Equitativa:

Los clusters 0, 2, 3, 4, 6, 8 y 9 tienen una distribución de observaciones relativamente equitativa (aproximadamente entre 130 y 200 observaciones cada uno). Esto puede indicar que los cultivos en estos grupos comparten características agronómicas similares, pero no son tan distintivos como los grupos con menos observaciones.

- Clusters Intermedios:

Los clusters 7 y 9 tienen una cantidad de observaciones intermedia (95 y 171 respectivamente). Esto sugiere que los cultivos en estos grupos tienen características que los diferencian de manera significativa de otros grupos, pero no son tan distintivos como los clusters con menos observaciones.

- Los clusters 3 y 12 contienen principalmente cultivos de frutas, como uvas y manzanas. Esto sugiere que estos cultivos comparten características agronómicas similares, como la necesidad de ciertas condiciones de temperatura, humedad y nutrientes. Ya se ha visto en el dataset original que poseen elevados valores de K y P.

- Los clusters 1, 4, 6, 8 y 9 incluyen principalmente legumbres como pigeonpeas, mothbeans, lentils y blackgram. Estos cultivos pueden requerir condiciones agronómicas específicas y, por lo tanto, se agrupan en función de sus necesidades compartidas.

- Los clusters 0, 2, 5, 7, 10 y 11 contienen una mezcla de cultivos que pueden tener necesidades agronómicas diversas. Por ejemplo, el cluster 5 incluye cultivos diversos, lo que indica que estos cultivos pueden tener una variabilidad en sus condiciones de crecimiento.

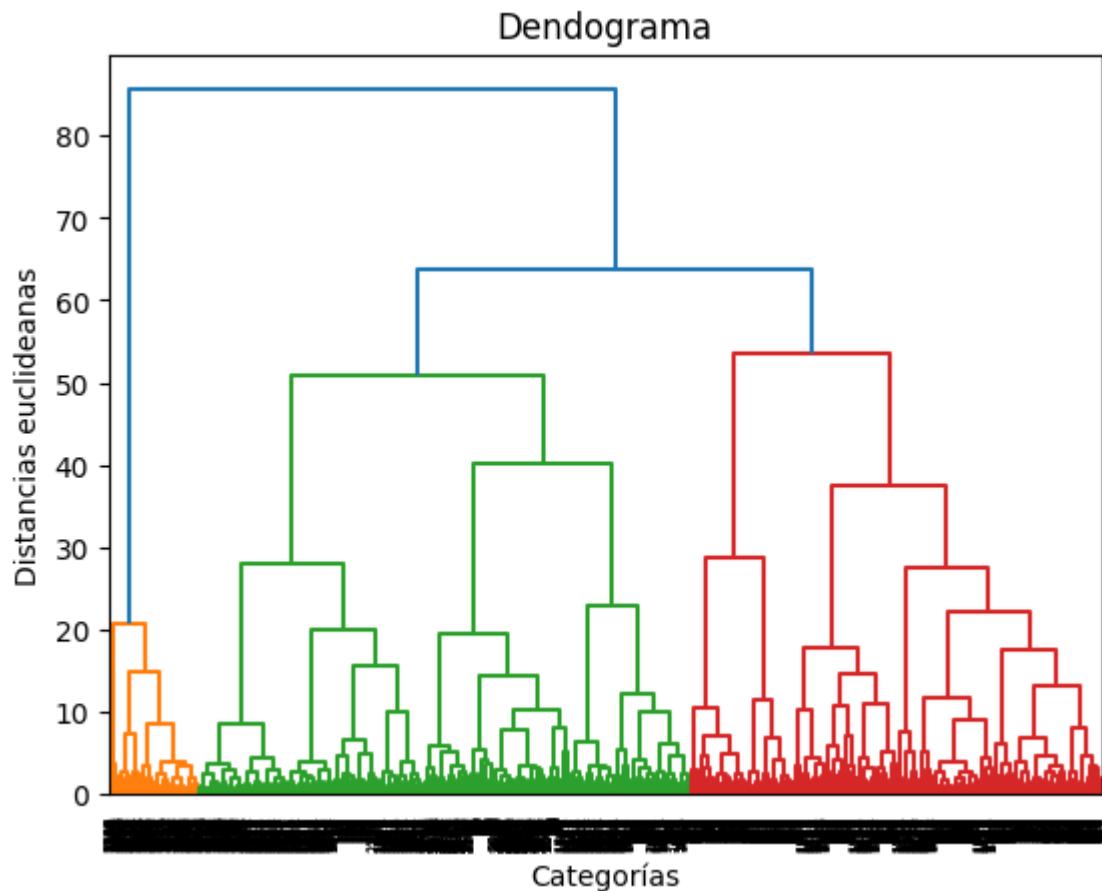
- Las medias de las variables agronómicas para cada cluster muestran diferencias significativas en las condiciones requeridas. Por ejemplo, el cluster 10 tiene altas medias de temperature, humidity y rainfall, lo que puede ser adecuado para cultivos específicos como oranges y papayas. Por otro lado, el cluster 8 tiene valores de pH más altos en promedio, lo que podría ser importante de controlar para el crecimiento de dichos cultivos, los cuales en el cluster 8 son mayormente legumbres.

En resumen, la elección de 13 clusters ha generado una mayor segmentación de los datos agronómicos. Esto puede ser útil si se requiere una mayor granularidad en la clasificación de los cultivos en función de sus características.

7. Clustering jerárquico

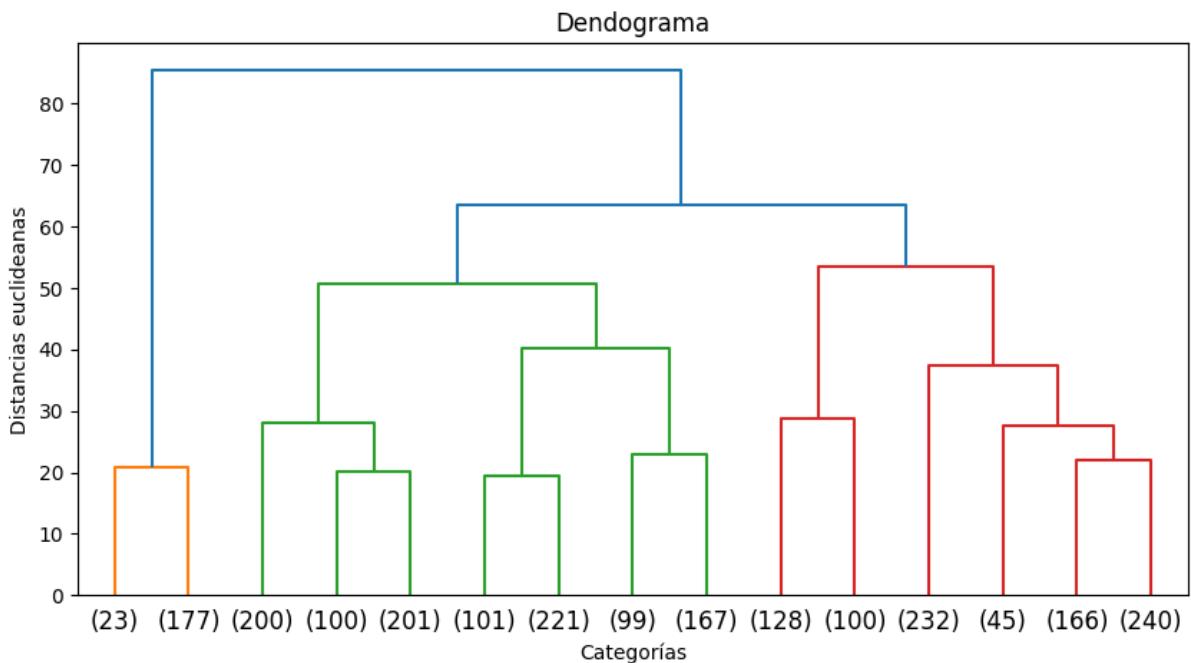
Aplicar clustering jerárquico y determinar cuál número sería el que mejor represente los datos. Utilizar el score de Silhouette y calcular el número óptimo de cluster por medio de GAP.

Se crea el dendrograma para encontrar el número óptimo de clusters



Como se dificulta visualizar el gráfico en su totalidad ya que sobre el final la cantidad de clusters es elevada se procede a recortarlo para una mejor visualización.

Para ello se observan las distancias registradas en el eje y y se establece un umbral de corte. Se observa que las distancias más largas y dónde se producen los mayores cambios en la altura (distancia) es alrededor del 15 por lo que se considera realizar el corte horizontalmente allí.



Se observa que a partir de 7 clusters las distancias no tienen demasiada variación por lo que se estima que dicho número de clústeres podría ser adecuado para el set de datos.

Se recuerda también que según la métrica GAP el número de clusters en los que deberían agruparse los datos es 13.

Se procederá entonces a modelar el clustering jerárquico con estas diferentes cantidades de clusters y se observarán los resultados.

Primeros 5 registros del dataframe luego de aplicar clustering jerárquico para 7 clusters:

	N	P	K	temperature	humidity	pH	rainfall	label	Cluster KMeans	Cluster_jerarquico
0	90	42	43	20.879744	82.002744	6.502985	202.935536	rice	11	5
1	85	58	41	21.770462	80.319644	7.038096	226.655537	rice	11	5
2	60	55	44	23.004459	82.320763	7.840207	263.964248	rice	11	5
3	74	35	40	26.491096	80.158363	6.980401	242.864034	rice	11	5
4	78	42	42	20.130175	81.604873	7.628473	262.717340	rice	11	5

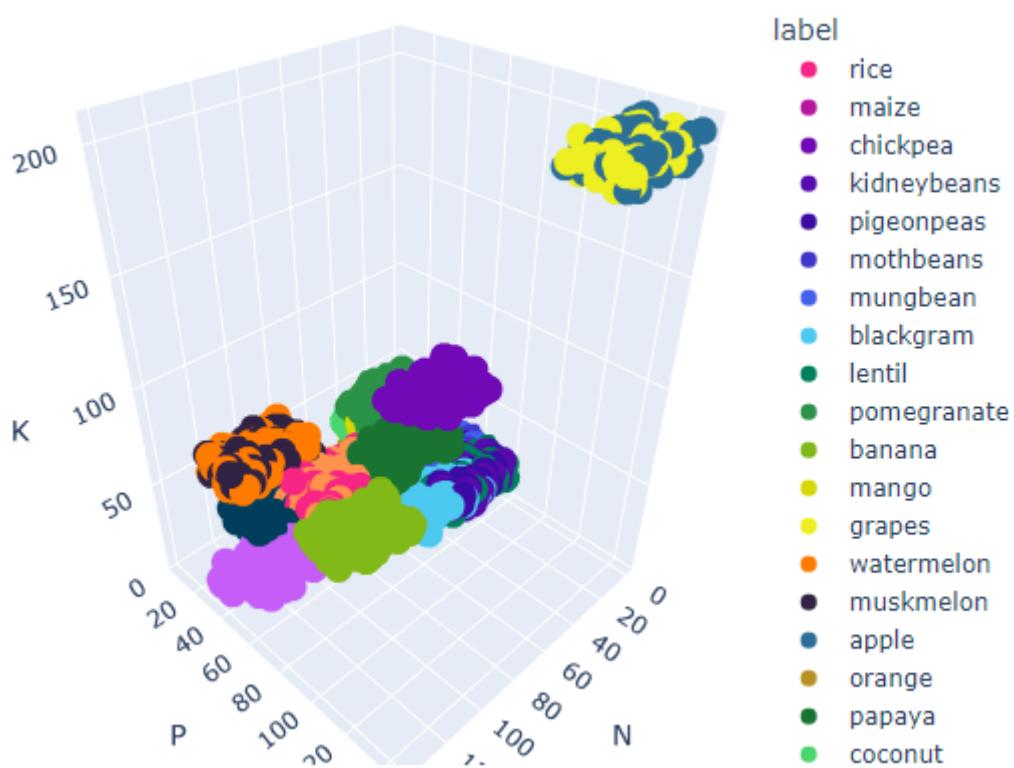
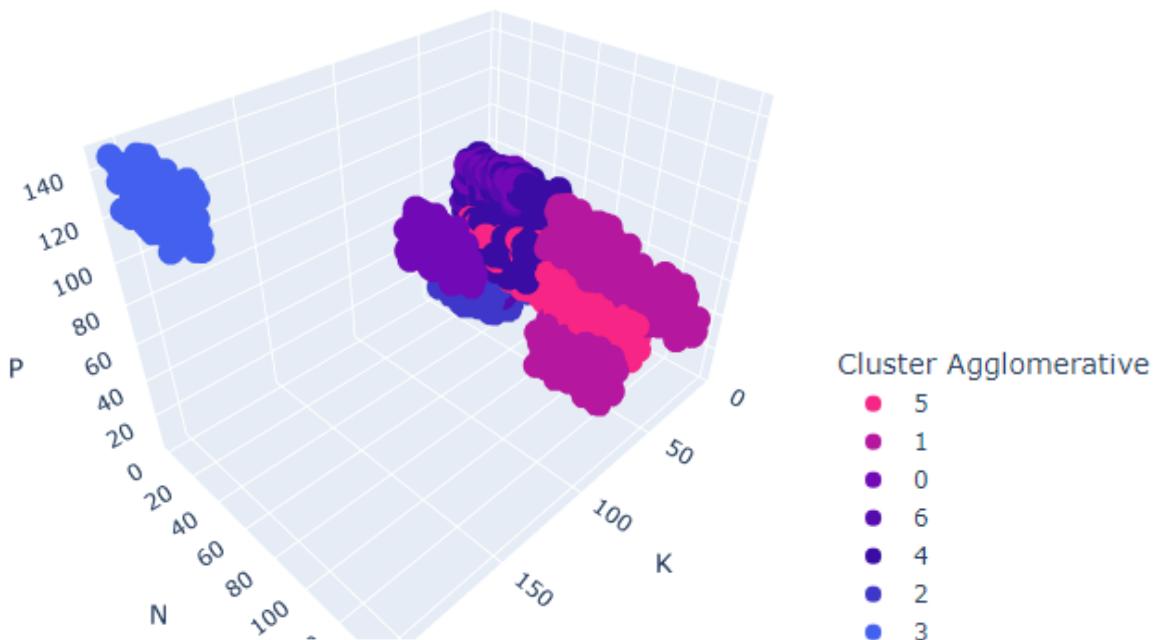
Silhouette:

Se utiliza la métrica de Silhouette para conocer la bondad de la técnica de agrupación:

- Silhouette score (n=6): 0.3169065328199988

Un Silhouette Score de 0.3169 es una indicación positiva de la calidad de los clusters obtenidos a partir del conjunto de datos con 7 clusters. Esta lejos de ser un 1, que sería lo ideal, pero aun así dicho valor sugiere que los cultivos están medianamente bien agrupados y definidos. Esto indica una estructura de clustering relativamente sólida en los datos.

Se graficarán los clusters obtenidos y debajo otro gráfico 3D mostrando los cultivos correspondientes.



Se obtendrá de manera más ordenada la lista de cultivos que pertenecen a cada cluster:

- Cultivos en cluster 0 ['chickpea' 'kidneybeans' 'pigeonpeas']
- Cultivos en cluster 1 ['maize' 'banana' 'watermelon' 'muskmelon' 'cotton' 'coffee']
- Cultivos en cluster 2 ['pomegranate' 'orange' 'coconut']
- Cultivos en cluster 3 ['grapes' 'apple']

- Cultivos en cluster 4 ['mothbeans' 'mungbean' 'blackgram' 'lentil' 'orange' 'papaya' 'coconut']
- Cultivos en cluster 5 ['rice' 'papaya' 'jute' 'coffee']
- Cultivos en cluster 6 ['pigeonpeas' 'mothbeans' 'lentil' 'mango']

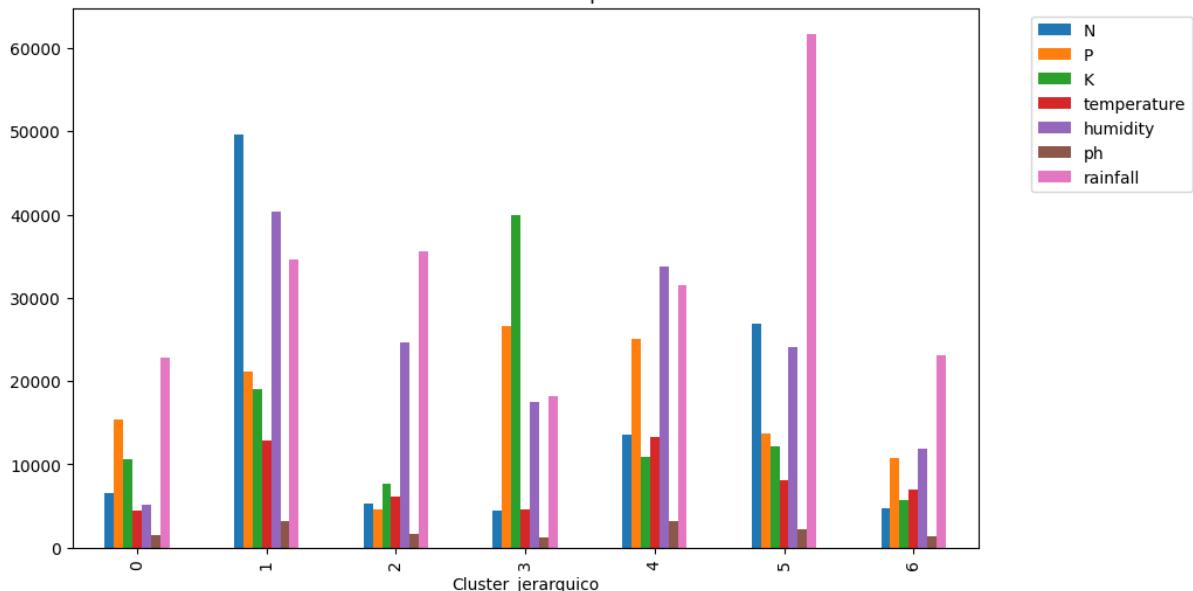
Cantidad de observaciones por clúster:

- 0 228
- 1 501
- 2 266
- 3 200
- 4 451
- 5 322
- 6 232

Se obtienen las medias de cada cluster por columna:

	N	P	K	temperature	humidity	ph	rainfall
cluster_jerarquico							
0	28.859649	67.688596	46.346491	19.774627	22.358286	6.365282	99.835046
1	99.101796	42.261477	37.916168	25.601225	80.513441	6.398558	69.163761
2	19.992481	17.454887	28.996241	23.091076	92.402176	6.420867	133.736210
3	21.990000	133.375000	200.000000	23.240259	87.104305	5.977800	91.133304
4	30.062084	55.534368	24.148559	29.365241	74.926192	7.085983	69.937252
5	83.664596	42.391304	37.599379	24.994155	74.975695	6.657743	191.647864
6	20.560345	46.456897	24.370690	29.908567	51.255451	5.744880	99.800867

Suma de cada columna por cluster



Se observa que el comportamiento y la distribución de los clusters y los cultivos es muy similar a la analizada y detallada para el modelo kmeans con 7 clusters.

Al basarse en las observaciones originales, parece que hay cierta correspondencia entre las características de los cultivos sobre los que superan o no la media en ciertos nutrientes y condiciones ambientales y la forma en que se agruparon en los clusters.

Por ejemplo, el Cluster 1 parece estar asociado con cultivos que requieren altos niveles de nutrientes (N, P, K) en el suelo, y el Cluster 3 agrupa cultivos que tienen altas necesidades de P y K, como apples y grapes.

Se señala también que los tamaños de los clusters varían, lo que indica que algunos grupos de cultivos son más homogéneos que otros.

En las columnas temperature y ph no se observa una coincidencia. Esto puede indicar que estos atributos no están teniendo un fuerte impacto en la agrupación.

Se realizará clustering jerárquico utilizando ahora 13 clusters, que es el número propuesto por el método GAP.

Primeros 5 registros del dataframe luego de aplicar clustering jerárquico para 13 clusters:

	N	P	K	temperature	humidity	ph	rainfall	label	Cluster	KMeans	Cluster_jerarquico
0	90	42	43	20.879744	82.002744	6.502985	202.935536	rice	11		2
1	85	58	41	21.770462	80.319644	7.038096	226.655537	rice	11		2
2	60	55	44	23.004459	82.320763	7.840207	263.964248	rice	11		2
3	74	35	40	26.491096	80.158363	6.980401	242.864034	rice	11		2
4	78	42	42	20.130175	81.604873	7.628473	262.717340	rice	11		2

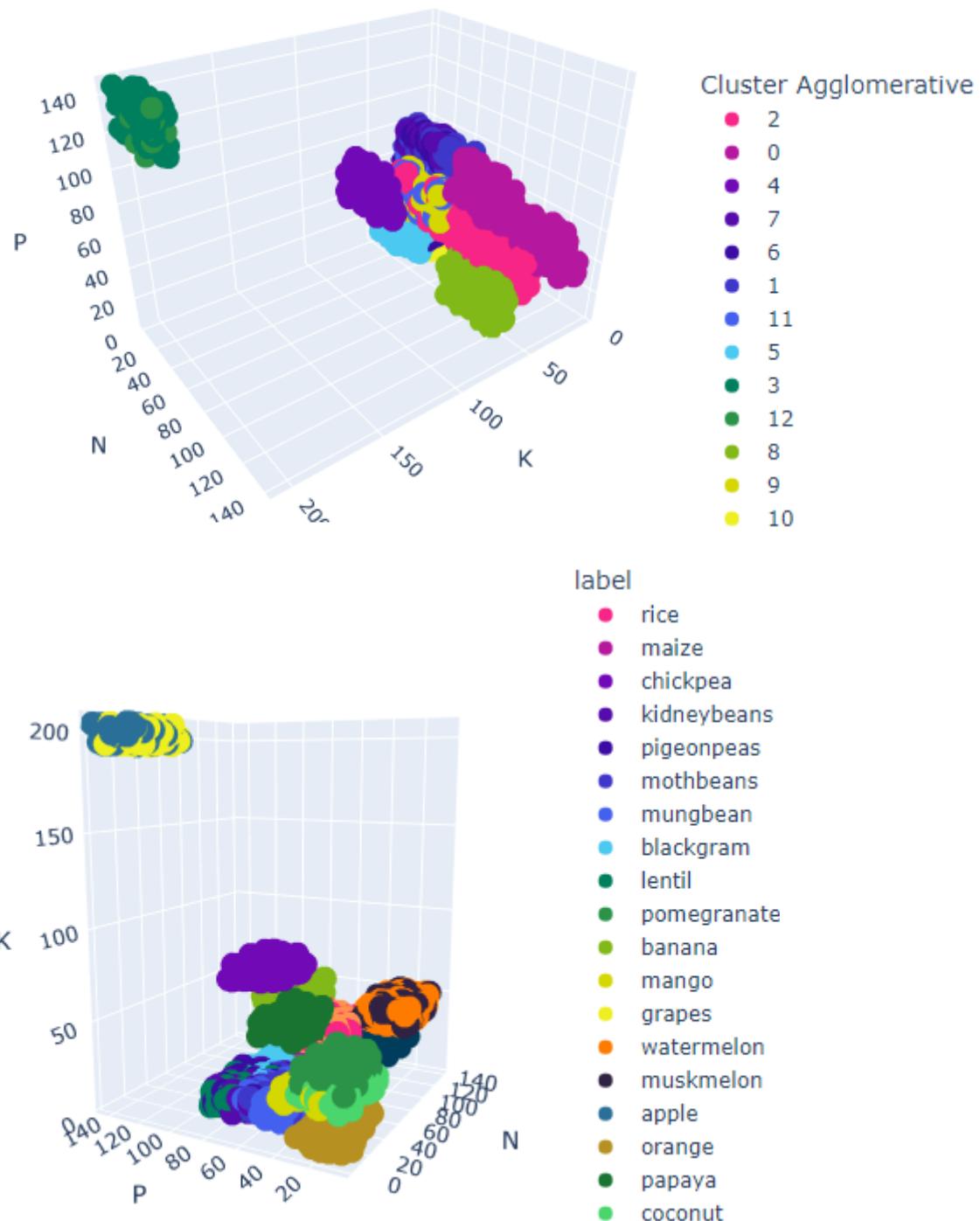
Silhouette:

Se utiliza la mérica de Silhouette para conocer la bondad de la técnica de agrupación:

- Silhouette score (n=13): 0.3197040718519583

El Silhouette score para 13 clusters es ligeramente mayor que para 7 clusters. Ha tenido una mejora mínima y la interpretación de dicho score es la misma que para 7 clusters.

Se graficarán los clusters obtenidos y debajo otro gráfico 3D mostrando los cultivos correspondientes:



Se obtendrá de manera más ordenada la lista de cultivos que pertenecen a cada cluster:

- Cultivos en cluster 0 ['maize' 'banana' 'cotton' 'coffee']
- Cultivos en cluster 1 ['mothbeans' 'blackgram' 'lentil']
- Cultivos en cluster 2 ['rice' 'papaya' 'jute' 'coffee']
- Cultivos en cluster 3 ['grapes' 'apple']
- Cultivos en cluster 4 ['chickpea']
- Cultivos en cluster 5 ['pomegranate' 'orange']
- Cultivos en cluster 6 ['pigeonpeas' 'mothbeans' 'lentil' 'mango']
- Cultivos en cluster 7 ['kidneybeans' 'pigeonpeas']
- Cultivos en cluster 8 ['watermelon' 'muskmelon']

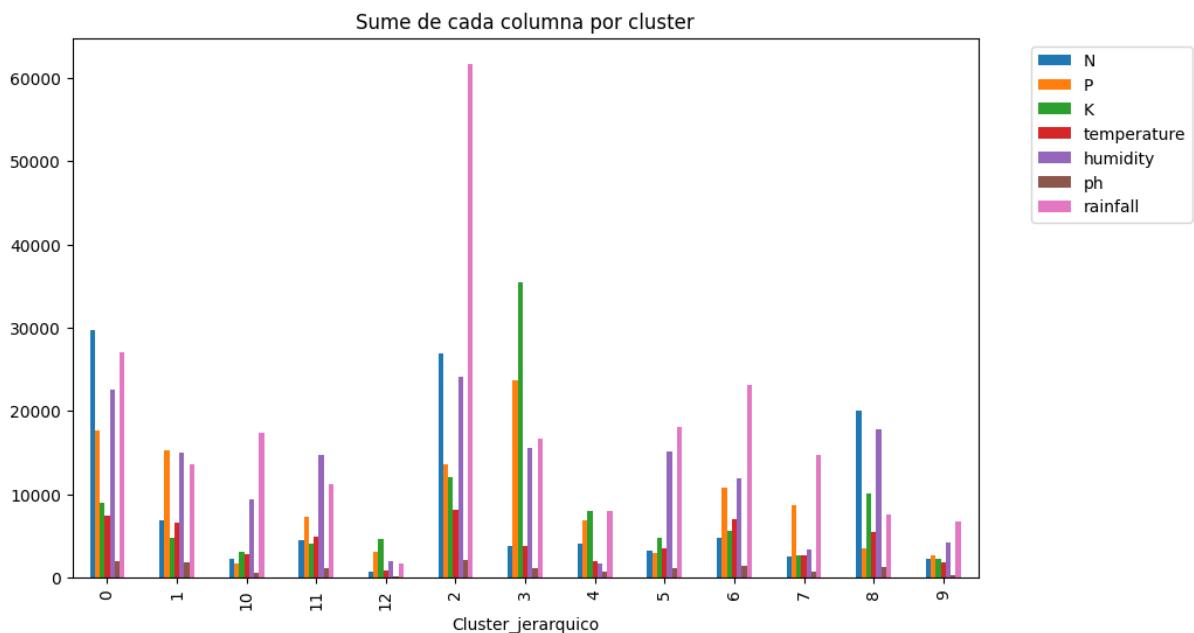
- Cultivos en cluster 9 ['papaya']
- Cultivos en cluster 10 ['coconut']
- Cultivos en cluster 11 ['mungbean' 'orange' 'papaya' 'coconut']
- Cultivos en cluster 12 ['grapes']

Cantidad de observaciones por clúster:

- 0 301
- 1 240
- 10 99
- 11 166
- 12 23
- 2 322
- 3 177
- 4 100
- 5 167
- 6 232
- 7 128
- 8 200
- 9 45

Se obtienen las medias de cada cluster por columna:

	N	P	K	temperature	humidity	ph	rainfall
Cluster_jerarquico							
0	98.591362	58.807309	29.787375	24.587145	75.039589	6.379466	90.044609
1	28.487500	63.320833	19.520833	27.559063	62.449205	7.374587	56.559170
10	22.040404	17.040404	30.606061	27.390929	94.873445	5.977656	176.126419
11	26.867470	43.451807	24.000000	29.202642	88.213750	6.760417	67.676882
12	27.130435	130.826087	199.782609	37.658380	81.992620	5.963237	70.509769
2	83.664596	42.391304	37.599379	24.994155	74.975695	6.657743	191.647864
3	21.322034	133.706215	200.028249	21.366718	87.768535	5.979692	93.813198
4	40.090000	67.790000	79.920000	18.872847	16.860439	7.336957	80.058977
5	18.778443	17.700599	28.041916	20.542062	90.937173	6.683608	108.606685
6	20.560345	46.456897	24.370690	29.908567	51.255451	5.744880	99.800867
7	20.085938	67.609375	20.117188	20.479143	26.653479	5.606160	115.285100
8	99.870000	17.360000	50.150000	27.127416	88.751589	6.427292	37.738086
9	50.244444	58.577778	49.377778	39.598005	92.453798	6.747736	149.625279



Nuevamente, los resultados que se observan son muy similares a los obtenidos al aplicar KMeans para 13 clusters.

Conclusiones:

Algunas observaciones y reflexiones basadas en los resultados:

PCA:

Se identificó que las componentes más influyentes son 4, lo cual indica que la reducción a 4 dimensiones mantiene una buena cantidad de información.

La distribución en 3 dimensiones es similar a la obtenida con las columnas K, N y P del conjunto de datos original. Esto sugiere que estas tres variables tienen un peso significativo en la estructura de los datos.

ISOMAP:

Se observó que a medida que aumenta el número de vecinos, los grupos forman asociaciones más en forma de nube y menos lineales. Esto puede indicar que la estructura intrínseca de los datos es más compleja de lo que puede ser capturado por un modelo lineal.

Se señaló también que hay dos cultivos que se encuentran sustancialmente alejados de los demás, lo que sugiere que estos cultivos pueden tener características únicas o inusuales.

t-SNE:

Se vio que los resultados varían significativamente con diferentes configuraciones de perplexidad e iteraciones. Esto es indicativo de la sensibilidad de t-SNE a estos parámetros y destaca la importancia de elegirlos cuidadosamente.

En general, t-SNE parece ser capaz de detectar estructuras complejas y agrupamientos no lineales en los datos.

Es útil considerar la aplicación de múltiples técnicas de reducción de dimensionalidad para obtener una comprensión más completa de la estructura subyacente de los datos.

Clustering:

Los agrupamientos realizados por k-means para 7 clusters parecen tener una correspondencia con las observaciones realizadas en el dataset original. Esto sugiere que el algoritmo ha identificado patrones subyacentes en los datos que reflejan similitudes en las características agronómicas de los cultivos.

La elección de 13 clusters ha generado una mayor segmentación de los datos agronómicos. Esto puede ser útil si se requiere una mayor granularidad en la clasificación de los cultivos en función de sus características. Si se pretende prestar especial atención a ciertos cultivos más que otros, por una cuestión de que son más difíciles de mantener o rentabilidad, usar 13 clúster puede devolver insights más detallados y precisos sobre las condiciones de cultivos más específicos.

Que ambos modelos (Agglomerative clustering y KMeans) produzcan clusters con tamaños y composición similares tanto para 7 como para 13 clusters, sugiere que la estructura de los datos tiene patrones de agrupación claros y consistentes que son detectados por ambos algoritmos.

Sin embargo, es importante tener en cuenta que esta es una interpretación inicial y que la validación con expertos en el campo agrícola sería crucial para confirmar la relevancia de los clusters.