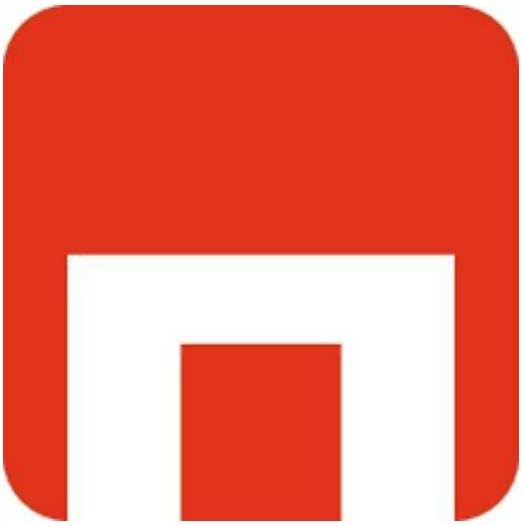
Memoria EDA

27 de julio de 2021



1. Introducción

Para el proyecto he utilizado un dataset con 572 muestras de aceite de oliva obtenido de la web de la comisión europea de consumidores con 9 columnas, 8 con la composición sobre 10.000 ('palmitic','palmitoleic','stearic','linoleic','linolenic','arachidic','eicosenoic') y una mas con el ID de cada muestra.

Se trata de un ejercicio que se realiza de manera real en el cual la empresa certifica que está vendiendo varios tipos de aceite de oliva como por ejemplo, aceite de oliva virgen extra, aceite de oliva virgen, aceite de oliva o aceite de orujo de oliva por ejemplo. Siendo la diferencia de precios notable, la comisión pide un análisis de composición de muestras aleatorias de aceite y comprueba si realmente la composición varía entre grupos y podemos validar los productos de la empresa. Recibiendo en un primer momento el dataset:

ID	palmitic	palmitoleic	stearic	oleic	linoleic	linolenic	arachidic	eicosenoic
1	1075	75	226	7823	672	36	60	29
2	1088	73	224	7709	781	31	61	29
3	911	54	246	8113	549	31	63	29
4	966	57	240	7952	619	50	78	35
5	1051	67	259	7771	672	50	80	46

2. Limpieza del Dataset

Observando la tabla de arriba observamos que tenemos variables demasiado diferentes en términos de unidades, esto podría ser un problema si trabajamos en términos de variabilidad de la muestra. Vamos a hacer una descripción de los datos actuales:

	palmitic	palmitoleic	stearic	oleic	linoleic	linolenic	arachidic	eicosenoic
count	572.000000	572.000000	572.000000	572.000000	572.000000	572.000000	572.000000	572.000000
mean	1231.741259	126.094406	228.865385	7311.748252	980.527972	31.888112	58.097902	16.281469
std	168.592264	52.494365	36.744935	405.810222	242.799221	12.968697	22.030250	14.083295
min	610.000000	15.000000	152.000000	6300.000000	448.000000	0.000000	0.000000	1.000000
25%	1095.000000	87.750000	205.000000	7000.000000	770.750000	26.000000	50.000000	2.000000
50%	1201.000000	110.000000	223.000000	7302.500000	1030.000000	33.000000	61.000000	17.000000
75%	1360.000000	169.250000	249.000000	7680.000000	1180.750000	40.250000	70.000000	28.000000
max	1753.000000	280.000000	375.000000	8410.000000	1470.000000	74.000000	105.000000	58.000000

Lo que primero llama la atención de los datos es que no vamos a poder comparar bien unos datos con otros ya que el promedio es demasiado distinto entre unas variables y otras. El ácido Oleico acapara gran parte de la composición (tiene una media de 7312) y concentra gran parte de la variabilidad de la muestra (desviación típica de 405), comparar esta variable con otras como el ácido arachidic que toma como valor máximo 105 y tiene una variabilidad muy pequeña en comparación (desviación típica de 22).Además no recibimos datos negativos y la suma de las composiciones suele estar cerca de los 10.000 con lo cual el fichero está en principio limpio, solo tendremos que tratarlo.

Vamos a proceder a tratar el fichero, para ello, le aplicamos una función estrictamente creciente que ‘’compacte” los datos y así reduzca es variación tan desmesurada en algunas variables. Elijo la función logaritmo del fichero sumándole 1 (evitamos los puntos singulares).

$$x \to Log(x+1)$$

Tenemos que sumar una unidad debido a que el cero logaritmo es 1, y log(0) es -∞. Ahora nuestra data tendrá una forma mas razonable y podemos empezar a realizar un estudio más descriptivo.

ID	palmitic	palmitoleic	stearic	oleic	linoleic	linolenic	arachidic	eicosenoic
1	6.981006	4.330733	5.424950	8.964951	6.511745	3.610918	4.110874	3.401197
2	6.993015	4.304065	5.416100	8.950273	6.661855	3.465736	4.127134	3.401197
3	6.815640	4.007333	5.509388	9.001346	6.309918	3.465736	4.158883	3.401197
4	6.874198	4.060443	5.484797	8.981304	6.429719	3.931826	4.369448	3.583519
5	6.958448	4.219508	5.560682	8.958283	6.511745	3.931826	4.394449	3.850148

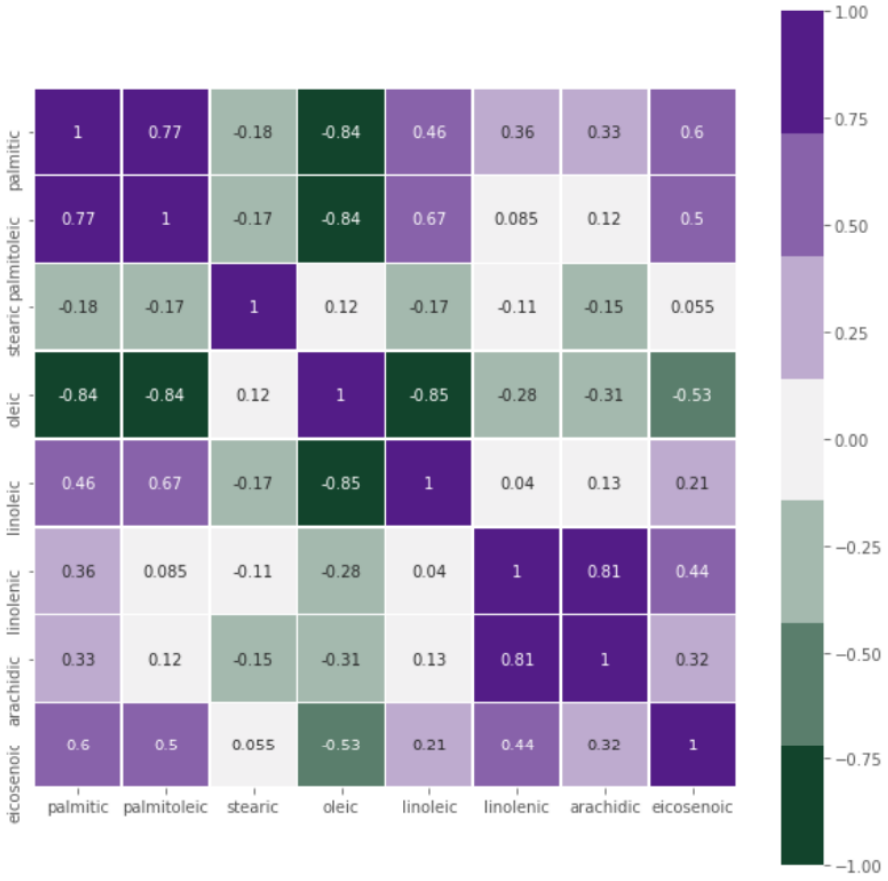
Donde la tabla descriptiva nos queda:

	palmitic	palmitoleic	stearic	oleic	linoleic	linolenic	arachidic	eicosenoic
count	572.000000	572.000000	572.000000	572.000000	572.000000	572.000000	572.000000	572.000000
mean	7.107708	4.752890	5.425494	8.895837	6.855507	3.301672	3.870479	2.321076
std	0.136553	0.445578	0.153238	0.055521	0.266213	0.859640	0.946396	1.150338
min	6.415097	2.772589	5.030438	8.748464	6.107023	0.000000	0.000000	0.693147
25%	6.999422	4.485811	5.327876	8.853808	6.648661	3.295837	3.931826	1.098612
50%	7.091742	4.709530	5.411646	8.896109	6.938284	3.526361	4.127134	2.890372
75%	7.215975	5.137265	5.521461	8.946505	7.074751	3.719596	4.262680	3.367296
max	7.469654	5.638355	5.929589	9.037296	7.293698	4.317488	4.663439	4.077537

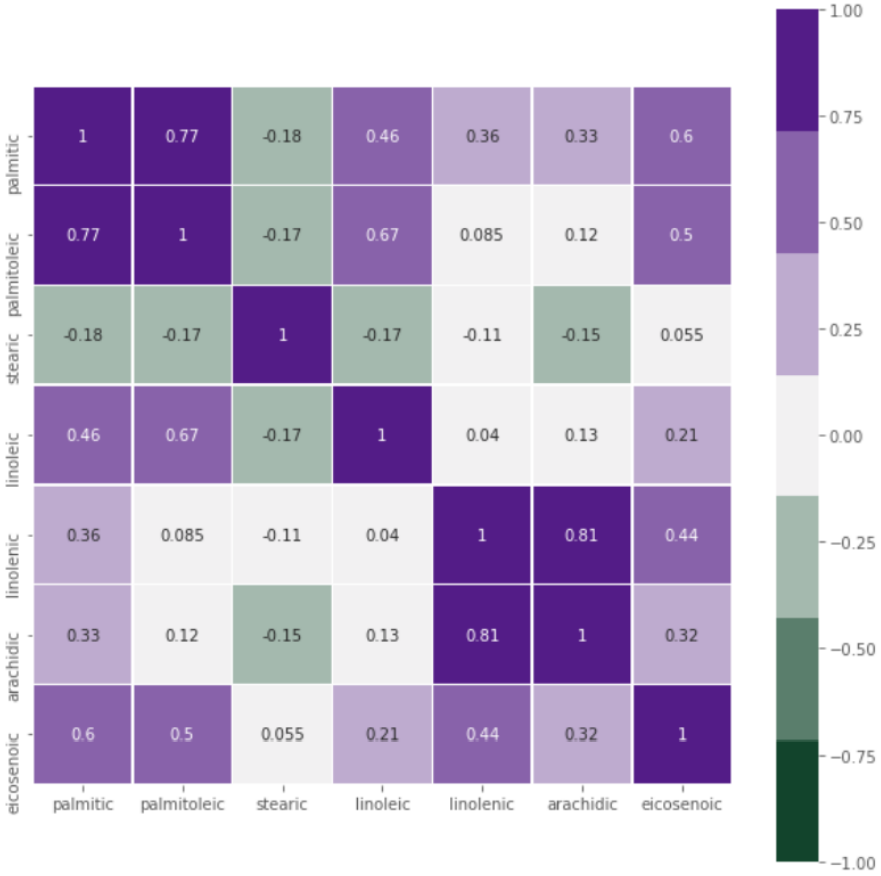
El cambio en el fichero es mas que notable, la variable que presumiblemente nos podía dar mas problemas, el ácido oleico, pasa de ser la variable que acumula mayor variabilidad de la muestra a ser la que menos. Además, al ''compactar" la muestra variables como arachidic en valores absolutos y relativos no se aleja tanto de los datos del ácido oleico, el mayor valor de arachidic pasa de ser 63 veces menor que el menor dato de oleico a ser la mitad.

3. Análisis dimensional

Comenzaremos con el analisis más tipico a la par que útil observando si existen correlación lineal entre variables con los coeficientes de Pearson, de manera gráfica:

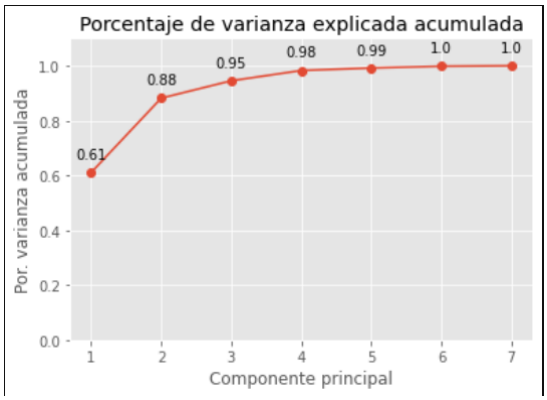


Observamos que existen variables con un alto nivel de correlación, sin embargo queremos ser cautos y solo eliminaríamos de forma tajante si el coeficiente está por encima de 0.9. Sin embargo como la variable 'Oleic' está relacionada fuertemente con 3 variables a la vez (-0.84,-0.84,-0.85) procederemos a eliminarla del dataset y repetiremos el coeficiente de Pearson:

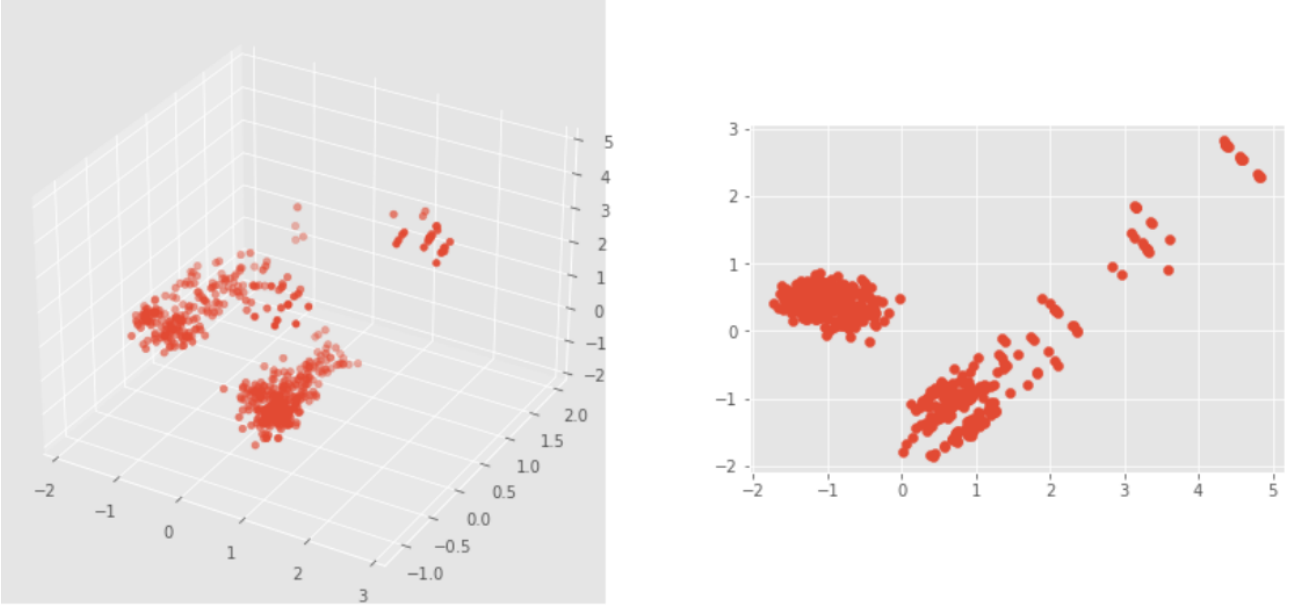


Vemos como la eliminación de la variable ha estabilizado linealmente el DataSet, siendo ahora la única correlación fuerte la de 'arachidic' con 'linoleic', pero como ya dijimos antes, ni es por encima del 0.9 en valor absoluto, ni sabríamos cual de las dos nos beneficia eliminar, con lo cual el análisis por Pearson finaliza aquí.

Gracias a que hemos normalizado la data con una transformación podemos realizar un analisis de la variacion con componentes principales para no solo acabar de pulir nuestro dataset quedándonos con las variables mas importantes, si no que podríamos incluso visualizar un criterio en el mejor de los casos. Una vez realizado el criterio con las 7 componentes principales de nuestro espacio en los datos, pasamos a escoger con cuantas nos vamos a quedar, para ello vemos la cantidad de variabilidad explicada:

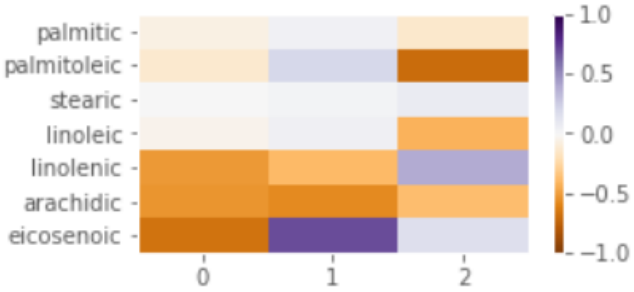


Con 2 componentes principales explicamos el 88 % de la muestra con lo cual trabajaremos de manera 2-dimensional, por mera curiosidad graficaremos nuestros datos también en la tercera componente. Vamos a proyectar nuestros puntos para poder observar como se comportan:



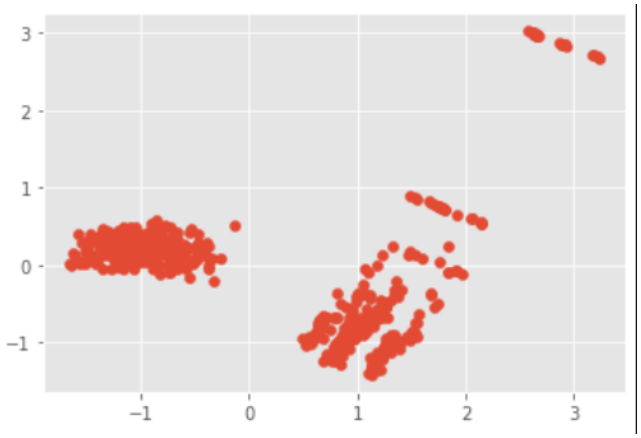
Desde luego, la conclusión empieza a ser favorable a la empresa de aceite, como mínimo tiene sentido hablar de 2 grupos muy claramente diferenciados de aceite, aunque quizás podamos extenderlo incluso a 3 o 4 grupos. En un futuro trabajaremos con 2 dimensiones y con la hipotesis de si tiene sentido 3 o 4 grupos de aceite de oliva.

Antes de ponernos a clasificar, acabaremos este analisis dimensional sobre la data original viendo que variables tienen impprtancia sobre estas 3 componentes principales y eliminando aquellas que no aporten nada a la variabilidad de la muestra, de manera gráfica:



Observamos como la variable 'stearic' apenas tiene ninguna importancia en las 3 componentes principales por lo tanto la eliminamos de facto de nuestro analisis, y en las dos primeras las variables importantes que forman gran parte de la variabilidad son 'linolenic', 'arachidic' y 'eicosenoic'. Quizás en una futura clasificación copen toda la importancia.

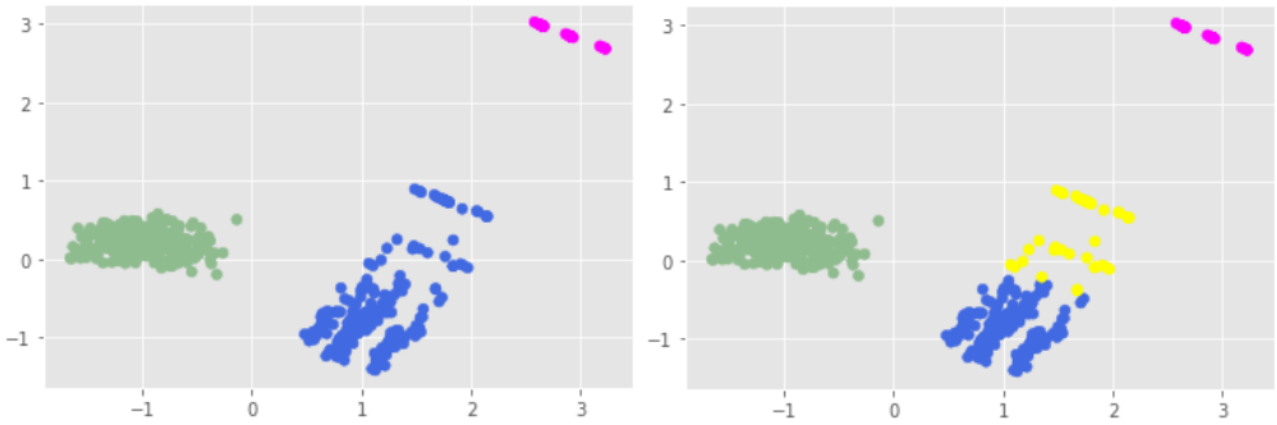
Con nuestra nueva data sin la variable 'stearic' volvemos a hacer componentes principales saliéndonos igual que antes pero ahora ya solo trabajaremos con la data limpia:



Observar que ahora que hemos quitado una variable residual, se diferencian incluso mejor nuestros tipos de aceites.

4. Clasificación

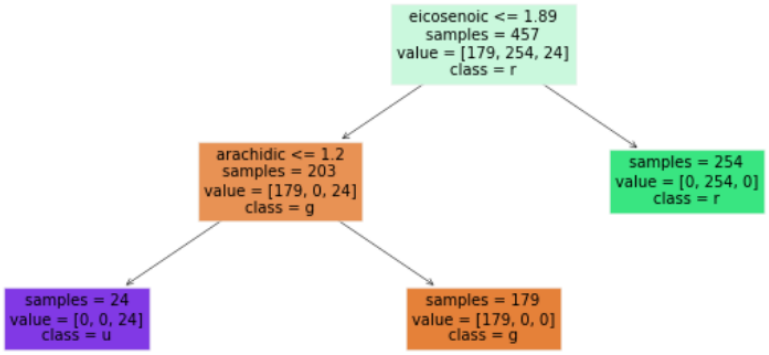
Tras la gráfica anteriormente ploteada donde proyectamos en dos dimensiones nuestros datos, nos surgió la duda de si podríamos hacer 3 o 4 grupos y sería correcto en el análisis. Para ello vamos a realizar una agrupación en 3 o 4 grupos con un algoritmo k-means basado en la distancia euclidea puesto que nuestros datos están normalizados:



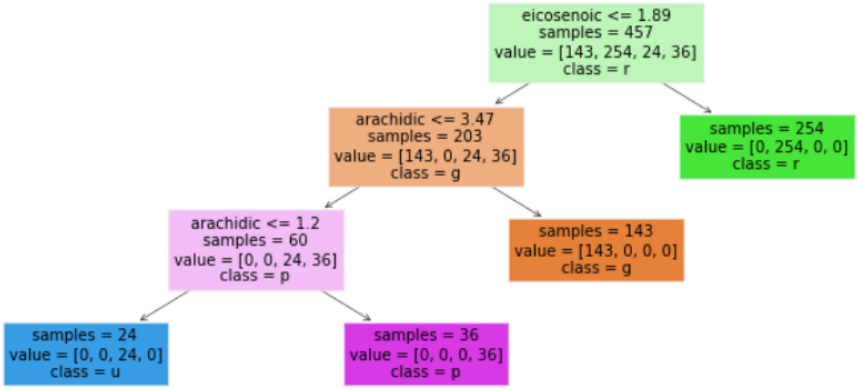
El método lo que hace es clasificar en n-grupos preseleccionados dependiendo de la distancia euclidea entre puntos.

Lo que tenemos que hacer es testear si tienen sentido estas divisiones, para ello guardaremos estas clasificaciones y sobre la data limpia (No sobre las proyecciones) si un árbol de decisión por ejemplo puede clasificar de manera correcta los distintos tipos de muestras. Para ello dividiremos la data limpia en muestra de entrenamiento y muestra test y realizaremos arboles de regresión en estos dos grupos midiendo su capacidad de acierto:

```
Profundidad del árbol: 2
Número de nodos terminales: 3
```




```
Profundidad del árbol: 3
Número de nodos terminales: 4
```



Para el primer grupo, llegamos a un acierto del 100 % con un árbol de profundidad 2, mientras que para el segundo grupo necesitaremos un árbol con una profundidad un grado mayor pero el acierto sigue siendo del 100 %. Por lo tanto la conclusión es clara, tiene sentido separar tanto en 3 como en 4 grupos distintos de aceite y ahora nuestro interés pasa por caracterizar cada uno de esos grupos. Para ello decidiremos trabajar con uno u otro modelo en función de la visualización de sus tablas.

5. Caracterización de las clases de aceite

Ahora que tenemos nuestra data definitiva y con los tipos de aceite que hemos definido, vamos a caracterizar de manera gráfica cada uno de ellos, primero visualicemos nuestros datos con los grupos, sin las variables residuales y sin la transformación:

ID	palmitic	palmitoleic	linoleic	linolenic	arachidic	eicosenoic	grupos1	grupos2
1	1075	75	672	36	60	29	1	1
2	1088	73	781	31	61	29	1	1
3	911	54	549	31	63	29	1	1
4	966	57	619	50	78	35	1	1
5	1051	67	672	50	80	46	1	1
...	...	...	...	...	...	...	...	...
568	1280	110	790	10	10	2	0	3
569	1060	100	810	10	10	3	0	3
570	1010	90	970	0	0	2	2	2
571	990	120	870	10	10	2	0	3
572	960	80	740	10	20	2	0	3

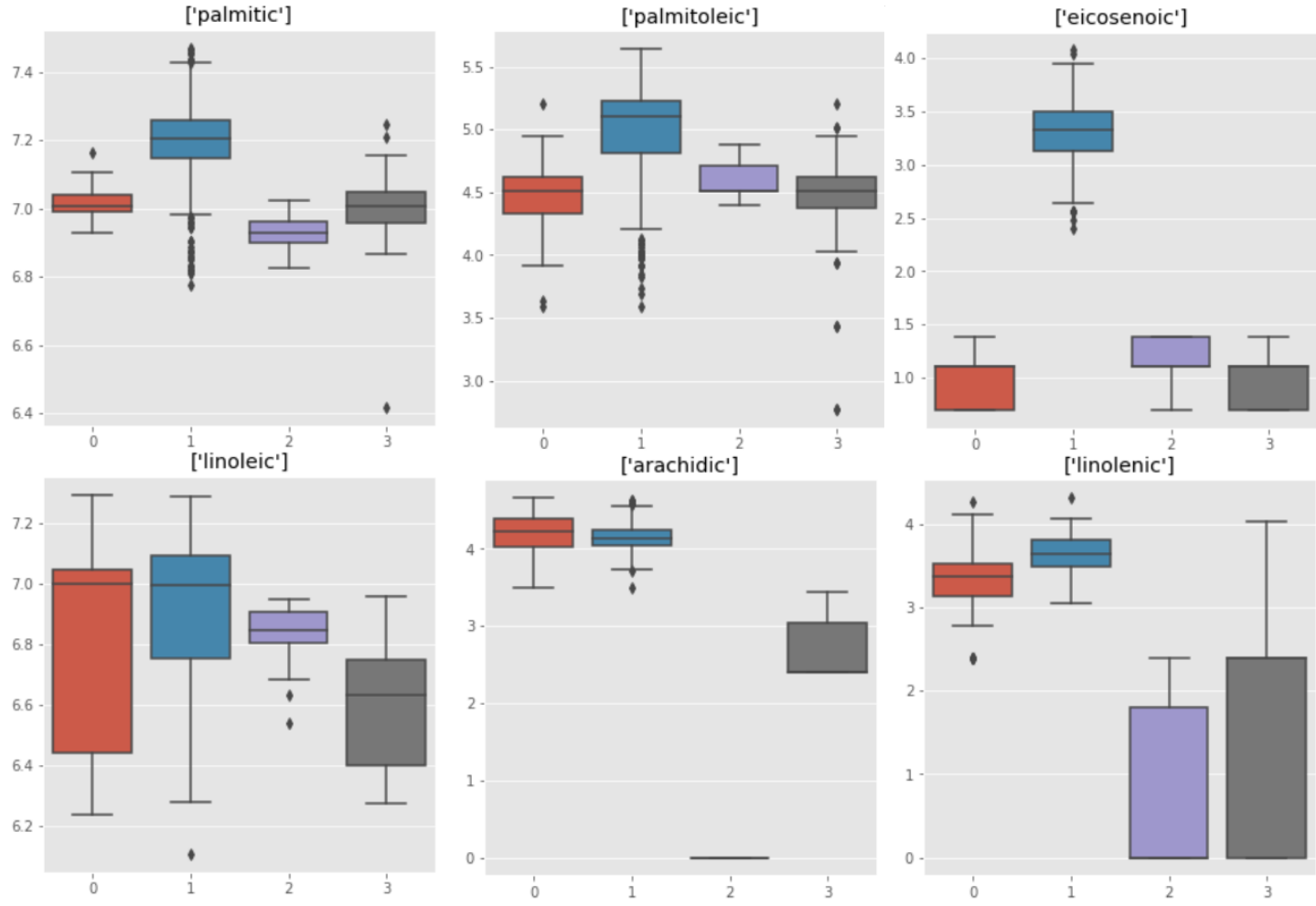
Donde podemos agruparlos y hacer una descripción en media para comparar ambos criterios:

grupos1	palmitic	palmitoleic	linoleic	linolenic	arachidic	eicosenoic
0	1110.614350	87.645740	909.829596	26.345291	57.600897	1.928251
1	1332.287926	154.801858	1033.498452	38.065015	63.117647	27.321981
2	1021.538462	99.230769	928.846154	2.692308	0.000000	2.230769

grupos2	palmitic	palmitoleic	linoleic	linolenic	arachidic	eicosenoic
0	1112.703297	87.016484	945.615385	29.054945	66.923077	1.923077
1	1332.287926	154.801858	1033.498452	38.065015	63.117647	27.321981
2	1021.538462	99.230769	928.846154	2.692308	0.000000	2.230769
3	1101.341463	90.439024	750.975610	14.317073	16.219512	1.951220

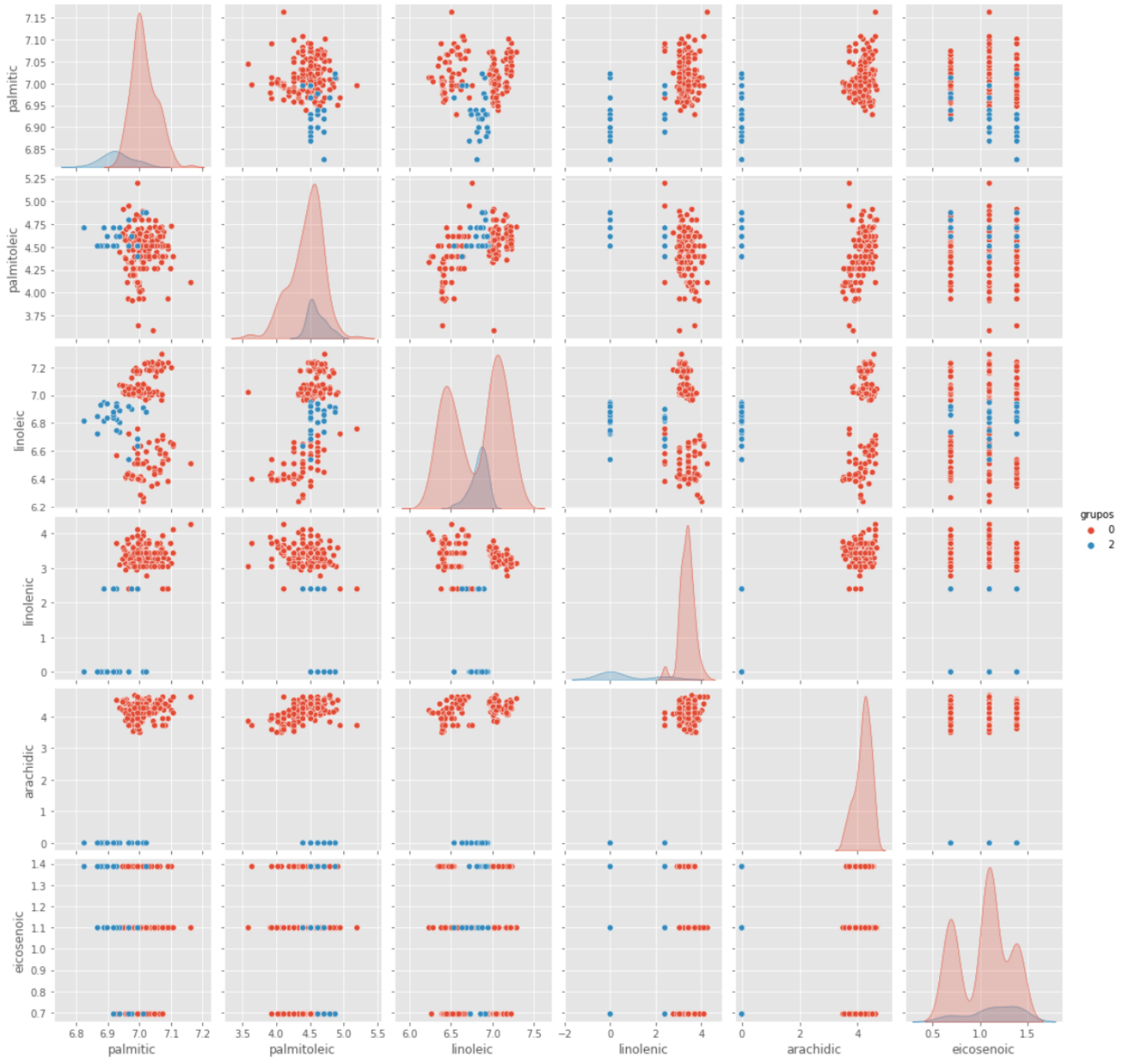
Como podemos observar, el nuevo grupo lo hemos despegado del grupo 0 en el primer modelo, y aunque en 3 de las 6 composiciones se mantiene muy parejo, en las varaibles 'linoleic' y 'arachidic' su composición varía notablemente, siendo este ultimo módelo el que optaremos por describir y trabajar con el.

Primero observaremos como se comportan sus variables:



De aquí ya podemos extraer muchas conclusiones, el grupo 1 tiene la mayor tasa de ácido palmitico sin embargo tiene algunos outliers, lo que si está claro es que es el que mayor ácido eicosenoico tiene y por tanto será su distinción mas clara del resto de grupos. Pese a que antes habíamos afirmado que el nuevo grupo se diferenciaba del grupo cero por su composición de ácido linoleico, en esta gráfica no queda muy claro y deberíamos como poco realizar más pruebas, sin embargo confirmamos que la distribución del ácido arachidico si que es muy diferente siendo este considerablemente menor de igual manera que el linoleico. La diferencia entre el grupo 2 y el resto es que no contiene nada de ácido arachidico.

Realizaremos a continuación una gráfica que nos deje mas claro si realmente el grupo 0 y el 3 tienen también una diferencia significativa en el ácido linolieico:



Observamos como en la variable ácido arachidico no existe discusión, la diferencia es significativa y vemos como sin embargo en Linoleico es más sútil, en el grupo2 se toman valores centrales de este ácido mientras que en el grupo 0, se toman valores extremos.

6. Conclusión

Tiene sentido adoptar 4 clases:

Una primera clase cuya carácterística principal es un alto contenido en ácido linoleico y araquídico, pero con niveles normales de ácido palmítico, este será el segundo aceite con menor calidad puesto que el araquídico es tóxico a niveles altos y a mas caliadad menos de este ácido.

Una segunda clase cuya carácterística principal es un alto contenido en ácido eicosenoico en comparación al resto. Es un ácido que se suele encontrar mucho en cacahuetes no sabemos por que tiene mucha mas concentración que el resto debería ser investigado, además tiene mucho ácido araquídico como el anterior e e incluso es el que más tiene en ácido palmítico teniendo una composición parecida al aceite de palma, es el de peor calidad.

Estos dos tienen un alto contenido en acido linolénico y araquídico que reduce considerablemente su calidad.
Una tercera clase cuya principal característica es un nulo contenido en ácido araquídico. Es un agente tóxico cuanto menos tenga mayor es la calidad del aceite.

Una cuarta clase con carácterísticas parecidas a la primera pero con menos concentración en ácido linolenico y arachídico, con lo que nos da entender que es un tipo de aceite ligeramente superior en calidad al primer aceite puesto que el ácido linoleico es mas común en el aceite de girasol que en el de oliva.