Centro de Investigación en Matemáticas (CIMAT). Unidad Monterrey

Tarea 1. Italian Olive Oils

Marcelo Alberto Sanchez Zaragoza

8 de febrero de 2021

1. Introducción

El siguiente trabajo tiene como objetivo encontrar ciertas caracteristicas que nos puedan ayudar a identificar las regiones y también las áreas dentro de cada región de una muestra de aceites de oliva proviniente de Italia. Dichas muestras cuentan con sus respectivos porcentaje de ácidos grasos. Se desea hacer distinción entre las siguientes regiones: Norte, Sur e Islas de Sardina. Las áreas son las siguientes: Liguria, Umbria, Sardinia, Apulia, Calabria y Sicily. Para poder encontrar estas caracteristicas se va a realizar una descripción de los datos, despues se realizara una exploración de los datos y finalmente una conclusión.
Para realizar la tarea antes mencionada nos apoyamos del lenguaje de programación Python y GGobi, al final se anexan las evidencias.

2. Descripción de los Datos

La base de datos proporcionada cuenta con 10 variables y tiene un total de 572 registros. Cada registro cuenta con la información necesaria para saber si pertenece a determinada región y área, entre las regiones que se muestran estan las siguientes: Norte, Sur e Islas de Sardina, donde cada una de ella cuenta con sus respectivas áreas. Cada registro nos proporciona el porcentaje de cada ácido graso(caracteristicas) presente en ese tipo de aceite de oliva.

Región Norte

- Liguria
- Umbria

Región Sur

- Apulia
- Calabria
- Sicily

Islas de Sardina

- Sardina

Entre las caracteristicas que se presentan en las distintas muestras de aceite son las siguientes: palmitic, palmitoleic, stearic, oleic, linoleic, linolenic, arachidic, eicosenoic, Area y Region.

3. Exploración de los Datos

Se realizo una primera consulta con los datos para encontrar la cantidad de registros en cada región, dichas cantidades vienen en el cuadro 3.1:

Región	Número de registros
Norte	151
Sur	323
Islas de Sardina	98

Cuadro 3.1: Registros de la base de datos por Región.

Se observa que cada una de las regiones presento una número casi o superior a las 100 unidades, despues se encontro que cada región presento los siguientes cantidades en sus respectivas áreas:

Región Norte

- Liguria - 100
- Umbria - 51

Región Sur

- Apulia - 231
- Calabria - 56
- Sicily - 36

Islas de Sardina

- Sardina - 98

En el cuadro 3.2 se muestra como se proporciono la información y su estado antes de empezar a trabajar con ella . En el cuadro 3.3 se muestra cual es el valor de la media de cada uno de los ácidos grasos en las distintas regiones y en el cuadro 3.4 muestra el valor de la media pero por área.

	Unnamed: 0	Region	Area	palmitic	palmitoleic	stearic	oleic	linoleic	linolenic	arachidic	eicosenoic
0	1	1	North-Apulia	1075	75	226	7823	672	36	60	29
1	2	1	North-Apulia	1088	73	224	7709	781	31	61	29
2	3	1	North-Apulia	911	54	246	8113	549	31	63	29
3	4	1	North-Apulia	966	57	240	7952	619	50	78	35
4	5	1	North-Apulia	1051	67	259	7771	672	50	80	46
...	...	...	...	...	...	...	...	...	...	...	...
567	568	3	West-Liguria	1280	110	290	7490	790	10	10	2
568	569	3	West-Liguria	1060	100	270	7740	810	10	10	3
569	570	3	West-Liguria	1010	90	210	7720	970	0	0	2
570	571	3	West-Liguria	990	120	250	7750	870	10	10	2
571	572	3	West-Liguria	960	80	240	7950	740	10	20	2

Cuadro 3.2: Registros de la base de datos.

En la figura 3.1 se muestra el gráfico de dispersión ocupando todas las caracteristicas pero tomando en cuenta de que región provienen los registros, en la figura se notan 3 distintos colores que se les fue asignado a cada región. Para la región del Norte(North) se le fue asignado el color morado y los registros de esta región son rombos, la región de Islas de Sardina(Island) viene de color guinda y sus registros vienen representados por cuadrados y finalmente la región del Sur(South) tiene el color amarillo y sus registros son circulos.

Region1	Region	palmitic	palmitoleic	stearic	oleic	linoleic	linolenic	arachidic	eicosenoic
Islad	2.0	1111.346939	96.744898	226.183673	7268.020408	1196.530612	27.091837	73.173469	1.938776
North	3.0	1094.801325	83.735099	230.801325	7793.052980	727.033113	21.788079	37.576159	1.973510
South	1.0	1332.287926	154.801858	228.773994	7100.009288	1033.498452	38.065015	63.117647	27.321981

Cuadro 3.3: Medias por cada región.

Area	Region	palmitic	palmitoleic	stearic	oleic	linoleic	linolenic	arachidic	eicosenoic
Calabria	1.0	1302.232143	121.357143	262.517857	7307.178571	819.000000	45.571429	63.625000	28.321429
Coast-Sardinia	2.0	1138.181818	101.060606	243.818182	7085.787879	1337.272727	23.757576	71.969697	1.878788
East-Liguria	3.0	1145.400000	84.200000	241.400000	7746.000000	689.400000	26.400000	63.600000	1.880000
Inland-Sardinia	2.0	1097.723077	94.553846	217.230769	7360.538462	1125.076923	28.784615	73.784615	1.969231
North-Apulia	1.0	1027.000000	61.600000	234.840000	7820.400000	705.840000	42.560000	71.960000	34.560000
Sicily	1.0	1228.361111	104.916667	273.888889	7357.833333	834.722222	42.472222	75.555556	38.444444
South-Apulia	1.0	1395.669903	183.922330	210.980583	6911.208738	1166.310680	34.708738	59.733010	24.228155
Umbria	3.0	1086.372549	59.882353	194.333333	7955.705882	597.098039	34.117647	42.431373	1.980392
West-Liguria	3.0	1052.800000	107.600000	257.400000	7674.200000	897.200000	4.600000	6.600000	2.060000

Cuadro 3.4: Medias por cada área.

Al observar el gráfico de dispersión en la figura 3.1 si se presta atención detenidamente podemos ver que en algunos cuadros una región se separa mucho de las otras dos y en otros donde dos regiones se separan y por detras de ellas queda la tercer región. La caracteristica que nos ayuda por lo menos visualmente a notar cierta distinción entre la región del Sur con las demas es la eicosenoic.
En cada una de las tablas donde esta presente la caracteristica eicosenoic, la región del Sur se encuentra en una pequeña nube de puntos alejada de las demás regiones. La región del Norte e Islas de Sardina se ven más sobre una recta, que en algunas gráficas esta de forma vertical y en otras horizontal, además cuesta trabajo poder hacer cierta distinción entre ellas. Cabe recalcar que para este caso estamos tomando en cuenta las ocho ca-

racteriticas que nos proporciona al principio pero podemos encontrar cierta combinación entre las ocho o menos caracteristicas, tal que encontremos una mejor vista de las 3 regiones.

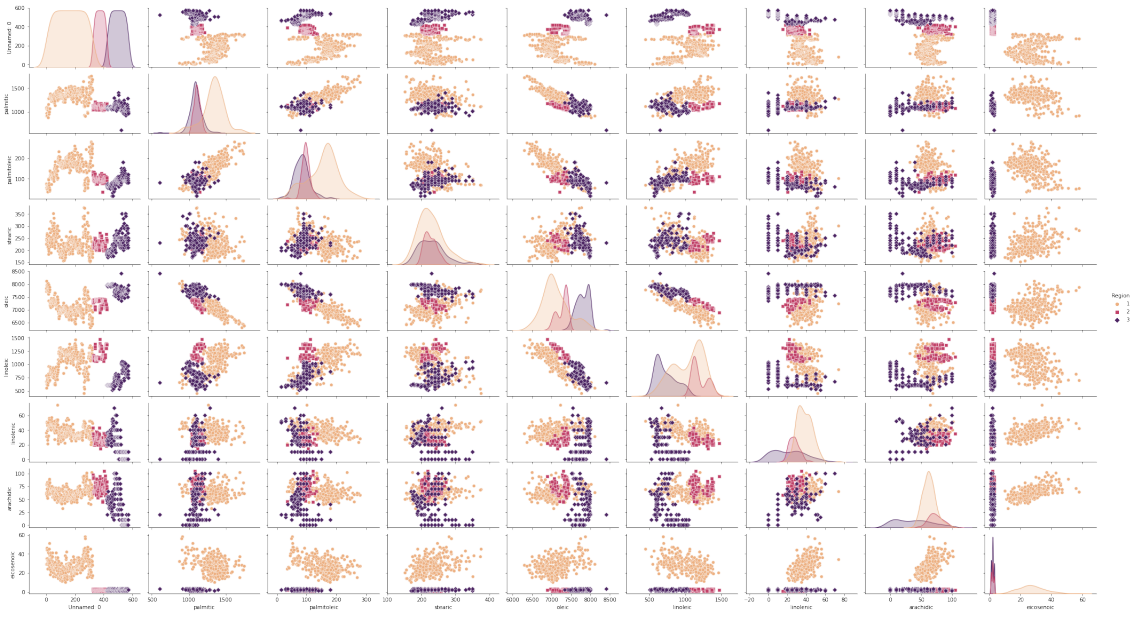

Figura 3.1: Gráfico de dispersión de las caracteristicas.

En la figura 3.3 podemos notar que al hacer uso de alguna combinación de las caracteristicas encontramos una mejor vista de las 3 regiones, en ella podemos distinguir mejor cada una de ellas, incluso poder decir que se agrupan en un determinado lugar los resultados. En la figura 3.3 la región del Norte lleva por color el amarillo, la región del Sur tiene el color verde y las Islas de Sardina color azul.
En apoyo a lo antes mencionado en la figura 3.2 se muestra una tabla con los coeficientes de correlación entre cada una de las caracteristicas.

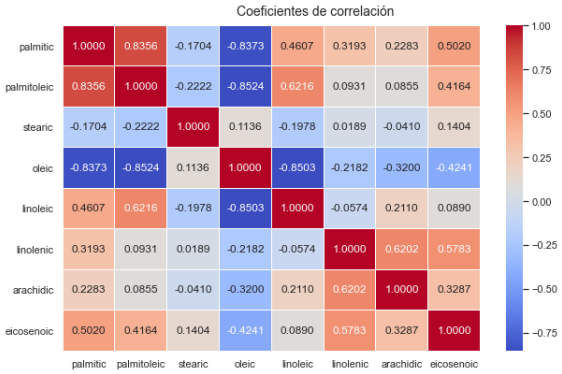

Figura 3.2: Coeficientes de correlación de las caracteristicas.

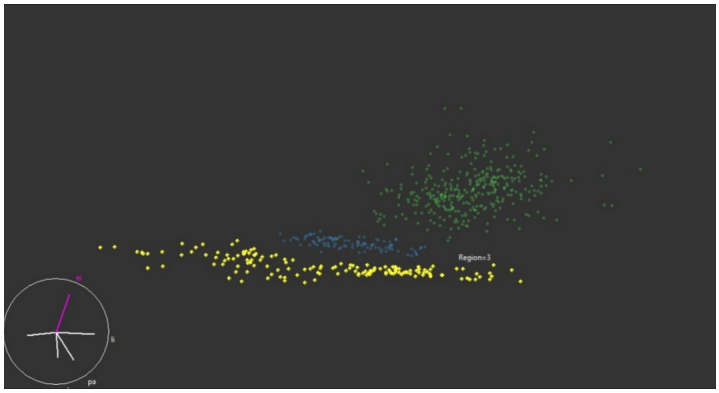

Figura 3.3: Gráfico de dispersión tomando ciertas caracteristicas.

También se agregan las gráficas para cada caracteristica por separado en la figura 3.4.

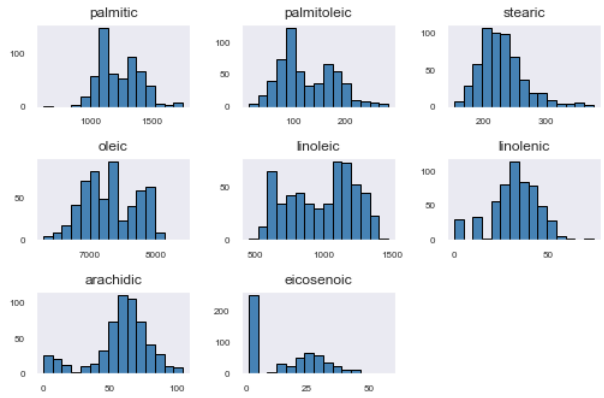

Figura 3.4: Visualización de los datos en una dimensión.

Las caracteristicas que ocupamos para hacer la combinación lineal fueron las siguientes: palmitic, stearic, oleic, linolenic y eicosenoic. Los respectivos valores de la matriz que nos ayudo a realizar dicha combinación estan en el cuadro 3.5.
La figura 3.5 nos ayuda a notar la diferencia que hay entre las regiones del Norte e Islas de Sardina, como se menciono antes solo se omitió la región del Sur para poder apreciar estas pequeñas diferencias entre estas dos regiones. La región del Norte lleva por color el naranja y la región de Islas de Sardina el azul.

Caracteristica	X	Y
Palmitic	0.330	-0.517
Palmitoleic	0.0	0.0
Stearic	-0.545	-0.061
Oleic	0.036	-0.471
Linoleic	0.0	0.0
Linolenic	0.725	-0.039
Arachidic	0.0	0.0
Eicosenoic	0.257	0.711

Cuadro 3.5: Valores de la matriz.

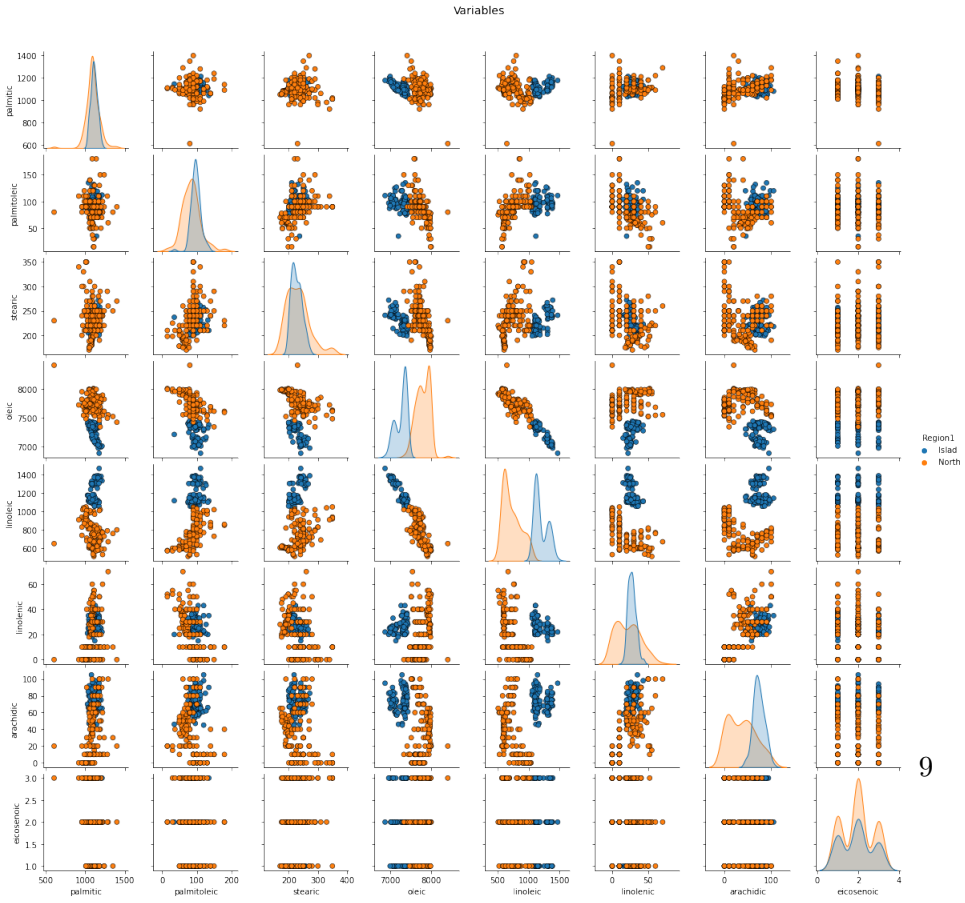

Figura 3.5: Gráfico de dispersión de la región Norte con Islas de Sardina (scatterplots).

En la figura 3.6 ahora intentamos encontrar las áreas que hay dentro de la región del Norte. El área de Liguria tiene el color naranja y el área de Umbria azul. En algunos cuadros podemos ver que si hay cierta distancia entre ambos grupos de puntos.

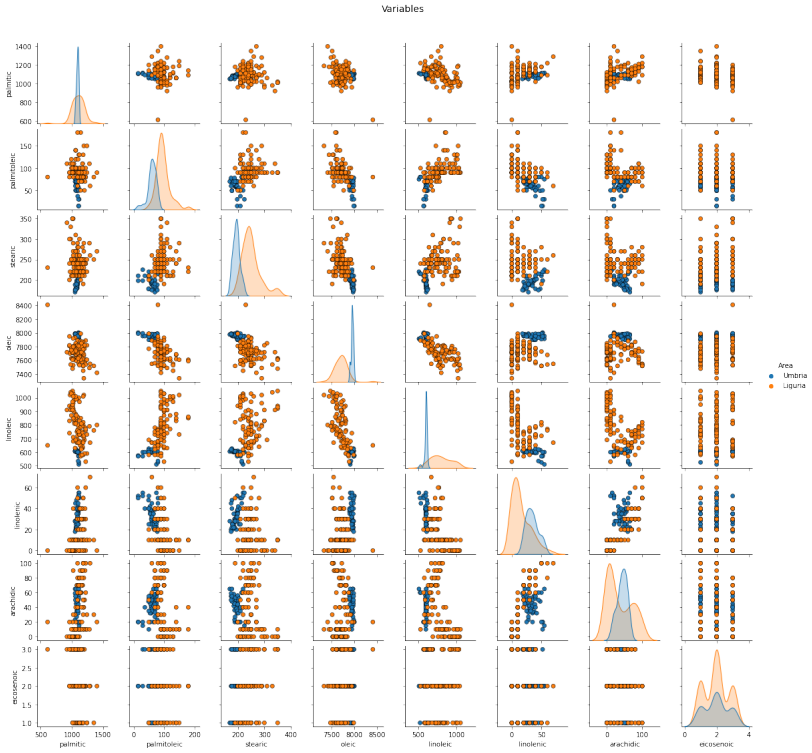

Figura 3.6: Gráfico de dispersión de la región Norte (scatterplots).

Volvemos a tomar una combinación de las caracteristicas, de los registros que sean de la región del Norte y encontramos una que nos ayudo a distinguir entre ambas áreas que tiene esta región, la figura 3.7 nos muestra los dos grupos donde el área de Umbria tiene el color azul y el área de

Liguria tiene el color amarillo.

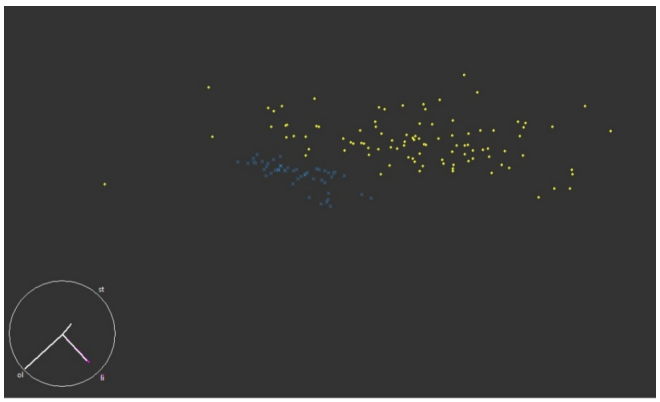

Figura 3.7: Gráfico de dispersión de la región Norte II.

Las caracteristicas que ocupamos para hacer la combinación lineal fueron las siguientes: stearic, oleic, linoleic y linolenic. Los respectivos valores de la matriz que nos ayudo a realizar dicha combinación lineal estan en el cuadro 3.6.

Caracteristica	X	Y
Palmitic	0.0	0.0
Palmitoleic	0.0	0.0
Stearic	0.174	0.211
Oleic	-0.705	-0.655
Linoleic	0.505	-0.530
Linolenic	0.466	-0.495
Arachidic	0.0	0.0
Eicosenoic	0.0	0.0

Cuadro 3.6: Valores de la matriz.

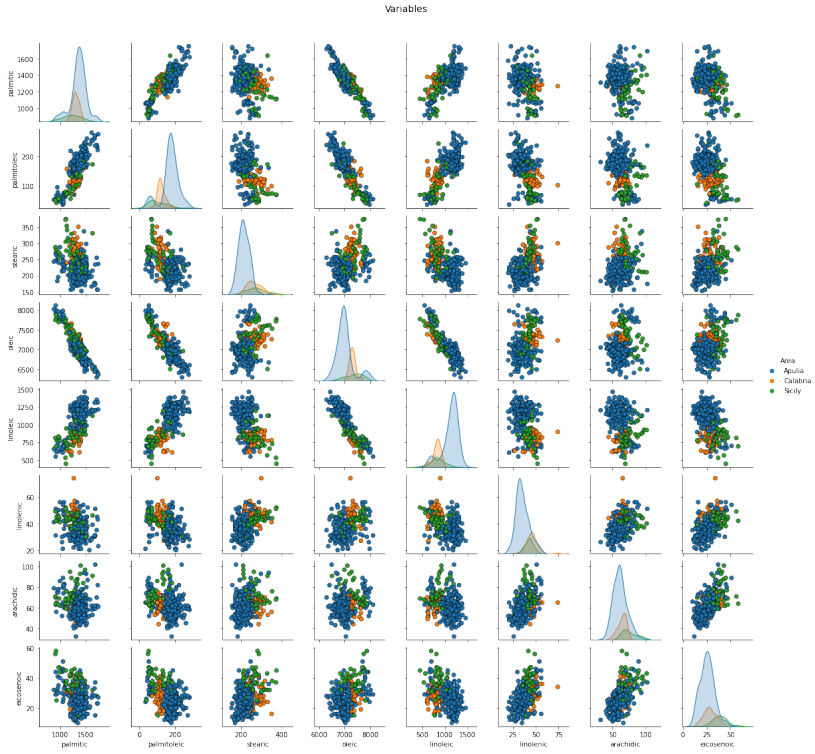

Figura 3.8: Gráfico de dispersión de la región del Sur (scatterplots). 12

Para la región del Sur tenemos la figura 3.8 que nos muestra gráficas de dispersión, donde el área de Apulia tiene el color azul, el área de Calabria tiene el color naranja y finalmente Scily tiene el color verde. No es muy fácil de hacer distinción entre estas áreas ya que es evidente que se traslapan bastante en varias gráficas.
Al intentar realizar el gráfico de dispersión y encontrar los valores de la matriz, con la región del sur, no encontramos valores que nos ayuden a realizar una distinción entre las áreas. Por lo que no se agrego una figura y la matriz.

4. Conclusión

Al terminar el análisis que al principio se planteo podemos llegar a la conclusión que dada una correcta combinación de las caracteristicas se puede encontrar una gráfica que nos ayuden a observar cierta distinción entre las regiones y en el caso de solo trabajar con una región sucede los mismo, en las figuras mostradas arriba nos dan un ejemplo de como encontrar esta combinación. Se puede agregar que con las caracteristicas brindadas podemos hacer distinción entre las 3 regiones que nos presentan y que para algunas regiones podemos hacer los mismo con sus respectivas áreas. En el último caso no pudimos llegar a una gráfica que nos ayudara visualizar mejor las áreas de la región del Sur pero no ocupamos un método en especifico solo se buscaron las combinaciones al azar por lo que no se descarta que se pudiera mejorar las gráficas para este caso.

Bibliografía Consultada

- T. Hastie, R. Tibshirani, J. Friedman. The elements of statistical lear- ning. Data mining, inference and prediction. 2nd. edition. Springer, 2009.