UNIVERSIDAD DE SANTIAGO DE CHILE

FACULTAD DE INGENIERÍA

Departamento de Informática

Análisis de datos

Laboratorio 2: Agrupamiento de k-medias

Richard Torti

Profesor: Max Chacón

Ayudante: Ignacio Ibañez

Santiago – Chile

2018

TABLA DE CONTENIDO

ÍNDICE DE TABLAS

ÍNDICE DE ILUSTRACIONES

CAPÍTULO 1. INTRODUCCIÓN

Una de las bebidas más conocidas mundialmente es el vino, pero, ¿Cuáles son los procesos involucrados para hacer este producto?, ¿Será importante la procedencia de los vinos segun el territorio para evaluar si uno es mejor que otro?. Estas respuestas son muy complejas, ya que dependen del gusto de cada persona para evaluar si le gusta o no, pero lo que sí se puede analizar es los componentes que traen los vinos. Durante este laboratorio, se presenta la base de datos y se aplican procesos para limpiar los datos, luego procesarlos con el algoritmo de las k medias. Para poder aplicar este último, es necesario estudiar los tipos de distancia que existen para realizar una buena deducción y conclusión de datos al final del proyecto. Una vez realizado los procesos de clustering, se hace un análisis para ver la procedencia de los vinos y sus compuestos, en caso de que la base de datos tenga alguna relación entre éstos.

CAPÍTULO 2. MARCO TEÓRICO

En este informe se presentan términos que no son tan conocidos, por lo que en esta sección se da una explicación de lo que son los temas a abordar en este laboratorio.

2.1 CLUSTERING

Wikipedia (2003) Un algoritmo de agrupamiento (en inglés, clustering) es un procedimiento de agrupación de una serie de vectores de acuerdo con un criterio. Esos criterios son por lo general distancia o similitud. La cercanía se define en términos de una determinada función de distancia, como la euclídea, aunque existen otras más robustas o que permiten extenderla a variables discretas. La medida más utilizada para medir la similitud entre los casos es la matriz de correlación entre los nxn casos. Sin embargo, también existen muchos algoritmos que se basan en la maximización de una propiedad estadística llamada verosimilitud.

Generalmente, los vectores de un mismo grupo (o clusters) comparten propiedades comunes. El conocimiento de los grupos puede permitir una descripción sintética de un conjunto de datos multidimensional complejo. De ahí su uso en minería de datos. Esta descripción sintética se consigue sustituyendo la descripción de todos los elementos de un grupo por la de un representante característico del mismo.

En algunos contextos, como el de la minería de datos, se lo considera una técnica de aprendizaje no supervisado puesto que busca encontrar relaciones entre variables descriptivas pero no la que guardan con respecto a una variable objetivo.

2.1.1 Aplicaciones

Las aplicaciones que tiene el clustering son variadas, pero las más comunes son:

- **Biología:** para clasificar plantas y animales.

- **Medicina:** para identificar enfermedades
- **Marketing:** para identificar personas con hábitos de compras similares.
- **Biometría:** para identificación del locutor o de caras

2.1.2 Algoritmos

Existen dos grandes técnicas para el agrupamiento de casos:

- **Agrupamiento jerárquico:** que puede ser aglomerativo o divisivo.
- **Agrupamiento no jerárquico:** en los que el número de grupos se determina de antemano y las observaciones se van asignando a los grupos en función de su cercanía. Existen los métodos de k-mean y k-medoid.

El paquete clúster de R-lenguaje implementa una serie de algoritmos de particionamiento como agnes, mona y diana, jerárquicos, y pam, clara y fanny, de particionamiento.

2.2 ALGORITMO K-MEANS

Wikipedia (2005) K-means es un método de agrupamiento, que tiene como objetivo la partición de un conjunto de n observaciones en k grupos en el que cada observación pertenece al grupo cuyo valor medio es más cercano. Es un método utilizado en minería de datos.

La agrupación del conjunto de datos puede ilustrarse en una partición del espacio de datos en celdas de Voronoi.

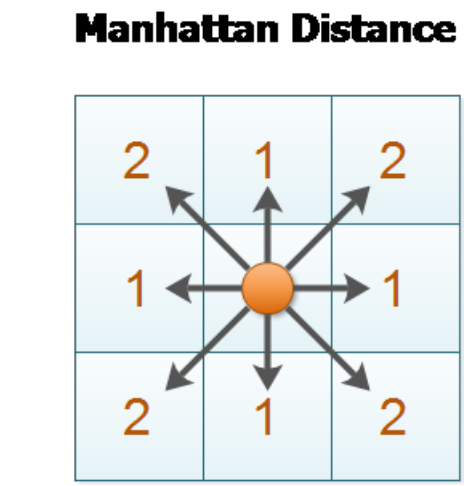
El problema es computacionalmente difícil (NP-hard). Sin embargo, hay eficientes heurísticas que se emplean comúnmente y convergen rápidamente a un óptimo local. Estos suelen ser similares a los algoritmos expectation-maximization de mezclas de distribuciones gausianas por medio de un enfoque de refinamiento iterativo empleado por ambos algoritmos. Además, los dos algoritmos usan los centros que los grupos utilizan para modelar los datos, sin

embargo k-means tiende a encontrar grupos de extensión espacial comparable, mientras que el mecanismo expectation-maximization permite que los grupos tengan formas diferentes.

2.3 DISTANCIAS UTILIZADAS

Dado que el conjunto de datos tiene una gran variedad de dimensiones, se escoge la distancia de Manhattan al tener mejores resultados en las pruebas experimentales Aggarwal (2001).

La distancia de Manhattan es una forma de geometría en la que la métrica usual de la geometría euclidiana es reemplazada por una nueva métrica en la que la distancia entre dos puntos es la suma de las diferencias (absolutas) de sus coordenadas. En la figura 2.1 se puede ver de forma numérica cual es la distancia de los vecinos. Note que para ir diagonal es necesario sumar los valores absolutos de las distancias en el eje x y en el eje y.



$$|x_1 - x_2| + |y_1 - y_2|$$

Figura 2.1: Distancia de manhattan para los vecinos del cuadro central

CAPÍTULO 3. PREPROCESAMIENTO

3.1 ELIMINACIÓN DE DATOS

Para poder hacer estimaciones de mejor calidad, es necesario eliminar datos que son .ºutliers". Wikipedia (2018) Un outlier es un dato que dista de los demás datos, normalmente debido a la variabilidad de la medida o un error experimental y puede causar serios problemas en análisis estadístico. Debido a que no existe una forma absoluta para detectar outliers, existen varios métodos. Uno de los más utilizados es el "Tukey's fences"(Cercas de Tukey). Este método puede detectar a los outliers tomando en cuenta la distancia entre el primer y tercer quintil. Tukey dice que los datos normalmente pertenecen a un intervalo y lo que esta fuera de éste, puede ser un potencial outlier. El intervalo propuesto es el siguiente:

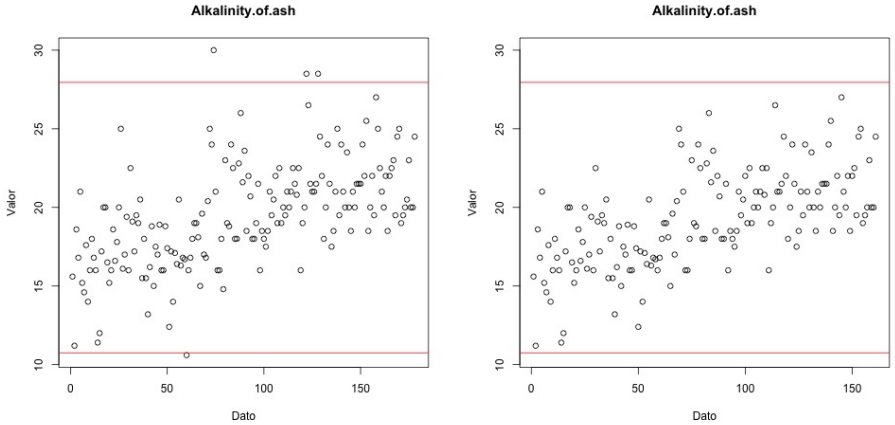
$$[Q_1 - k(Q_3 - Q_1), Q_3 + k(Q_3 - Q_1)]$$

Donde k representa un factor multiplicativo y se utilizan dos valores y cambia su concepto.

- $k = 1{,}5$**:** Indica un outlier.
- $k = 3$**:** Indica que el dato esta exageradamente alejado.

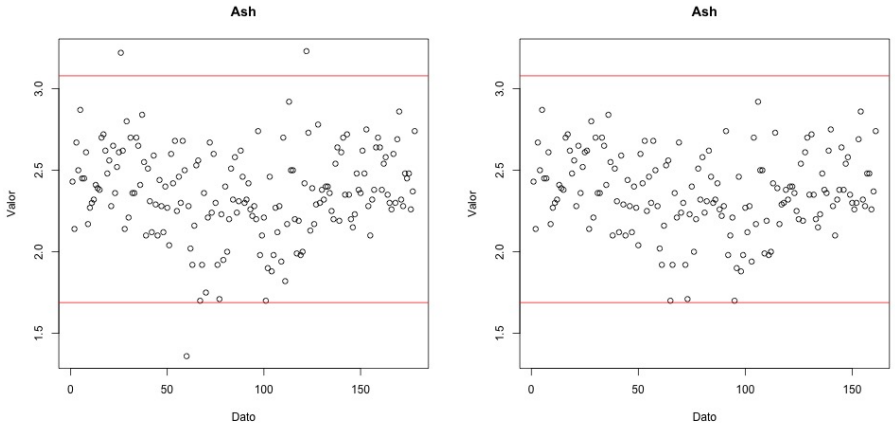
Por lo tanto, a través del lenguaje de programación R, se grafican los datos separados por cada variable y se muestran en la figura 3.1 3.2. Éstas muestran también las líneas divisorias para considerar un outlier (a) y se análogamente los gráficos después de haber eliminado esos outliers (b).

Antes de haber eliminado los outliers, la cantidad de datos muestreados era de 178 instancias, sin embargo, durante el proceso de limpiado esto se redujo pasando a ser 161. Esto da un resultado total de 17 instancias eliminadas durante el proceso.



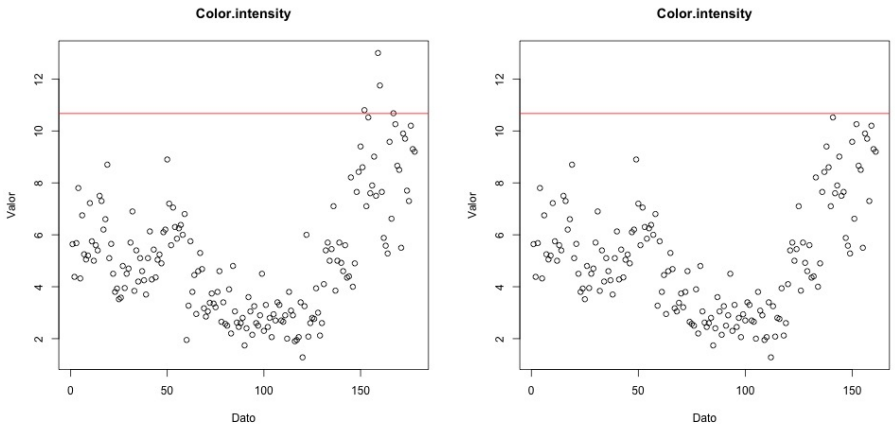
(a) Original

(b) Outliers eliminados



(c) Original

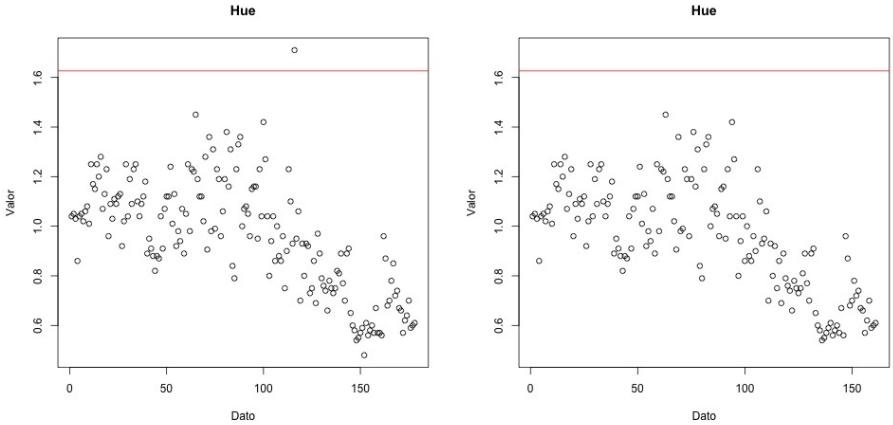
(d) Outliers eliminados



(e) Original

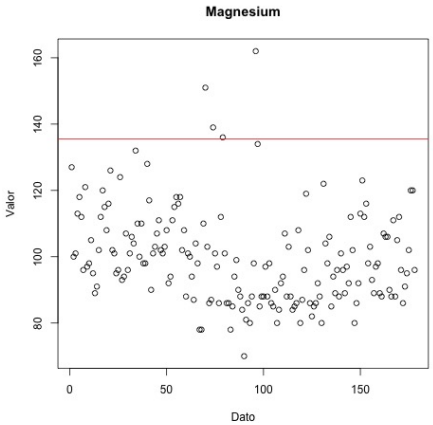
(f) Outliers eliminados

Figura 3.1: Eliminación de outliers (parte 1)

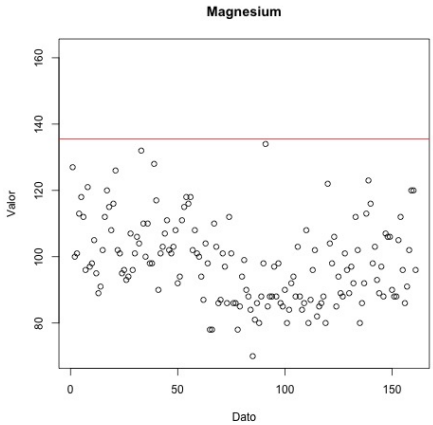


(a) Original

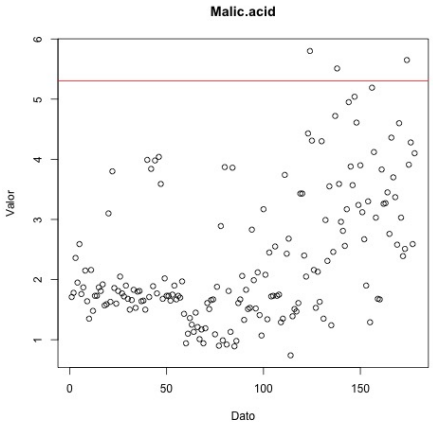
(b) Outliers eliminados



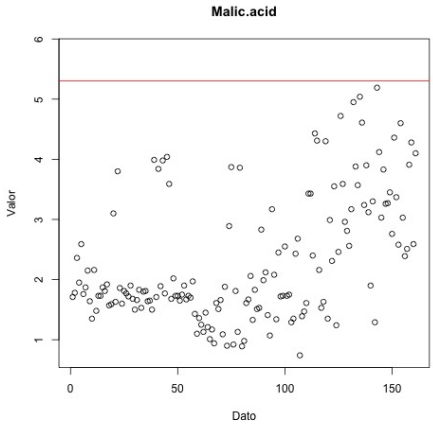
(c) Original



(d) Outliers eliminados



(e) Original



(f) Outliers eliminados

Figura 3.2: Eliminación de outliers (parte 2)

CAPÍTULO 4. OBTENCIÓN DEL CLUSTER

4.1 NÚMERO DE GRUPOS

Para poder realizar el agrupamiento, es necesario haber definido la cantidad de grupos a formar, ya que el algoritmo de las k-medias requiere el parámetro k ingresado. Una forma de obtener el mejor valor de k, es probar iterativamente cada una de las posibilidades (fuerza bruta), sin embargo, el resultado que entrega debe ser analizado por un experto en el tema que pueda asegurar a vista cual es el mejor k. Este problema es una falencia que tiene la búsqueda del k, pero una solución es utilizar el método de las siluetas, en el que entrega un coeficiente que puede ser usado para medir la calidad del k, donde entre más alto es mejor esa cantidad k de grupos.

La figura 4.1 muestra los resultados que da la variación del k, haciéndolo variar desde $k = 2$ hasta $k = 160$, el cual es la variación máxima permitida.

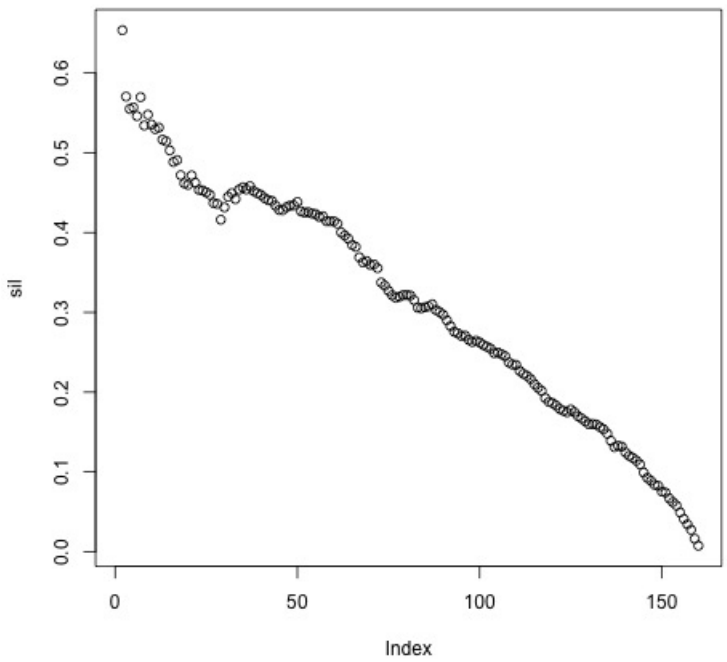


Figura 4.1: Resultados de la variación del k

El mayor valor es de 0.929167 con $k = 2$, por lo que se considera que ese es el mejor k para el posterior análisis.

4.2 CLUSTERING

Ahora que ya se tiene el valor del k, se procede a ejecutar el algoritmo de las k-medias con $k = 2$. Este algoritmo se ejecuta llamando a la función $dviz_cluster()$ y se obtiene la figura 4.2.

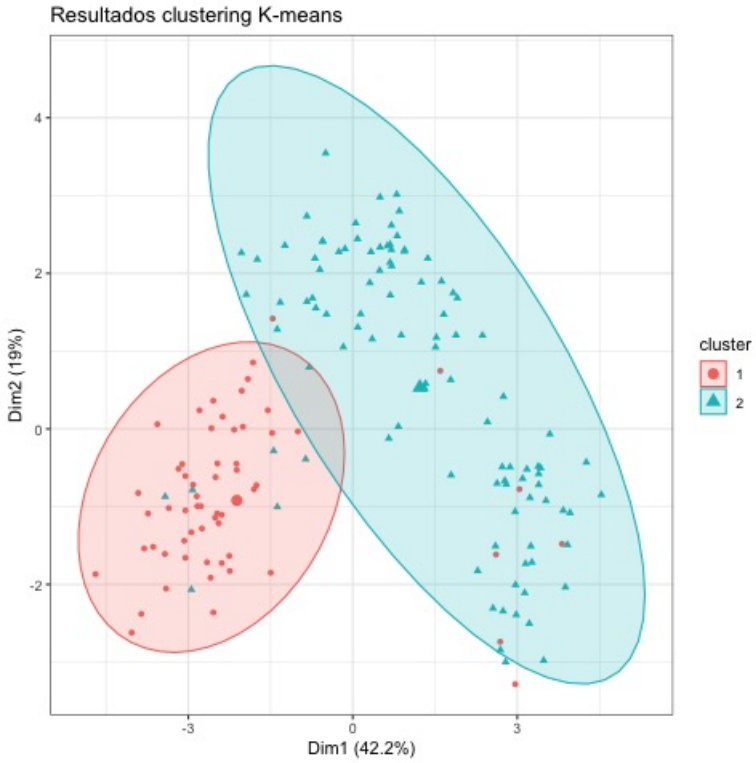


Figura 4.2: Clustering con k=2

CAPÍTULO 5. ANÁLISIS DE LOS RESULTADOS

Una vez obtenido los clusters, es necesario verlos de otra perspectiva. Para esto, se realiza en primer lugar el summary para ambos clusters. En la tabla 5.1 y 5.2 se muestran estos valores.

Tabla 5.1: Tabla summary para el cluster 1.

Variable	Min	1Qu	Mediana	Promedio	3Qu	Max
Class.identifier	1.000	1.000	1.000	1.203	1.000	3.000
Alcohol	11.96	13.29	13.73	13.65	14.08	14.83
Malic.acid	1.090	1.655	1.770	2.016	1.935	4.280
Ash	2.040	2.270	2.410	2.422	2.605	2.840
Alkalinity.of.ash	11.20	16.00	17.10	17.38	19.05	27.00
Magnesium	89.0	98.0	103.0	105.3	112.0	132.0
Total.phenols	1.100	2.490	2.800	2.719	3.000	3.880
Flavanoids	0.550	2.570	2.920	2.754	3.235	3.930
Nonflavanoid.phenols	0.1300	0.2600	0.2900	0.3002	0.3250	0.6300
Proanthocyanidins	1.140	1.460	1.760	1.825	2.035	2.960
Color.intensity	3.050	4.650	5.600	5.729	6.675	10.200
Hue	0.590	0.930	1.050	1.029	1.125	1.280
OD280.OD315	1.560	2.800	3.000	2.993	3.365	4.000
Proline	830.0	987.5	1095.0	1126.7	1280.0	1680.0

Tabla 5.2: Tabla summary para el cluster 2.

Variable	Min	1Qu	Mediana	Promedio	3Qu	Max
Class.identifier	1.000	2.000	2.000	2.304	3.000	3.000
Alcohol	11.41	12.18	12.62	12.69	13.21	14.22
Malic.acid	0.740	1.480	2.350	2.475	3.353	5.190
Ash	1.700	2.203	2.320	2.330	2.495	2.920
Alkalinity.of.ash	13.2	18.5	20.0	20.3	22.0	26.5
Magnesium	70.00	86.00	91.50	94.64	101.75	134.00
Total.phenols	0.980	1.620	1.990	2.042	2.420	3.520
Flavanoids	0.340	0.830	1.595	1.633	2.240	3.750
Nonflavanoid.phenols	0.1400	0.2900	0.4000	0.3917	0.4800	0.6600
Proanthocyanidins	0.410	1.022	1.390	1.388	1.655	2.910
Color.intensity	1.280	2.800	3.920	4.601	5.638	10.5
Hue	0.5400	0.7400	0.9050	0.9186	1.0775	1.4500
OD280.OD315	1.270	1.750	2.450	2.418	3.072	3.710
Proline	278.0	441.0	550.0	546.5	657.5	795.0

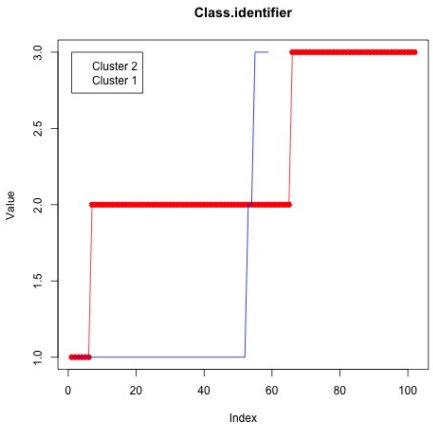
Al entrar en detalle entre las diferencias de los atributos, hay algunas que son de alguna forma distantes, por lo que se entra en detalle:

- **Class identifier:** El promedio entre ambas variables es algo bastante alejado. Mientras que

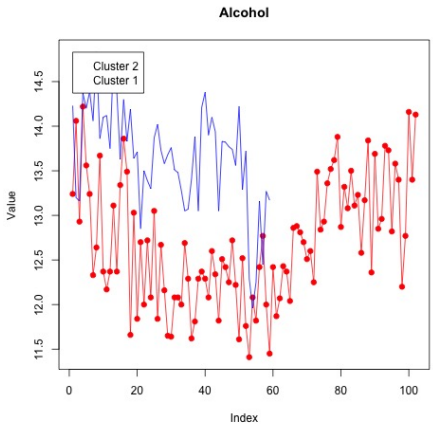
en el primer cluster es de 1.2, en el segundo es de 2.3; esto da una señal de que el cluster uno contiene en su mayoría vinos de la clase uno, mientras que el cluster dos tiene una gran parte de dos y tres. Esto se analiza más adelante con gráficos para ver los datos de mejor manera.

- **Alcohol:** A simple vista, no muestra una gran diferencia más allá de que el cluster 1 tiene una leve proporción mayor de alcohol, pero nada desproporcional.
- **Alkanility of ash:** El cluster 1 muestra que tiene una alcanilidad de cenizas bastante menor en comparación con el cluster 2.
- **Magnesium:** Se puede notar una pequeña variación de que el cluster 1 tiene mayores niveles de magnesio.
- **Total phenols:** Cluster 1 sigue mostrando mayores niveles de este componente.
- **Flavanoids:** Cluster 1 muetsra nuevamente que tiene mayores niveles de Flavanoids.
- **Proanthocyanidins:** Cluster 1 muestra mayor cantidad de este componente.
- **Color intensity:** En este dato, por los valores entregados por el summary, se aprecia que existe una fuerte distinción en la intencidad de color segun el cluster.
- **OD280 OD315:** Este elemento también esta presente en mayor cantidad en el cluster 1.
- **Proline:** Esta diferencia podría considerarse abismal, siendo el cluster 1 con una proporción mucho mayor al cluster 2.

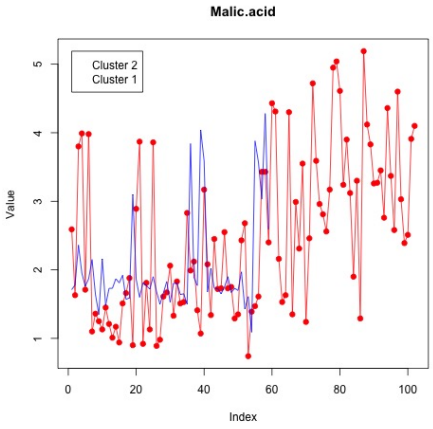
Dado estos datos, pareciera ser que los vinos del cluster 1 tienen más componentes analizados que el cluster 2, pero ¿Habría alguna forma de relacionarlos?. Quizás en el identificador se encuentre la respuesta... Es importante también otra forma de ver los datos, ya que solamente con un dato numérico no se puede realizar buenas conclusiones, por lo que en las figuras 5.1, 5.2 y 5.3 se muestran los datos evidenciando cada variable para ver si concuerda con las conclusiones hechas con el summary. Note que el cluster 1 aparece en azul, mientras que el cluster 2 en rojo.



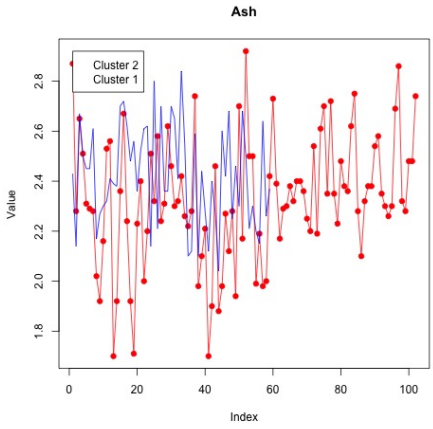
(a) Class identifier



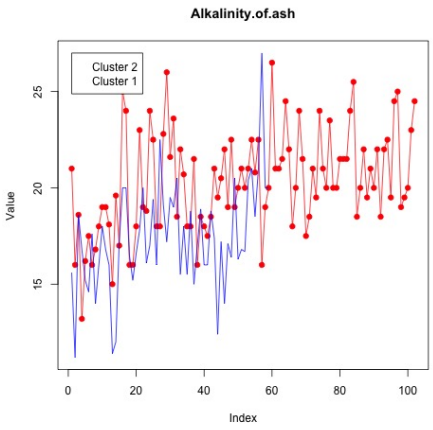
(b) Alcohol



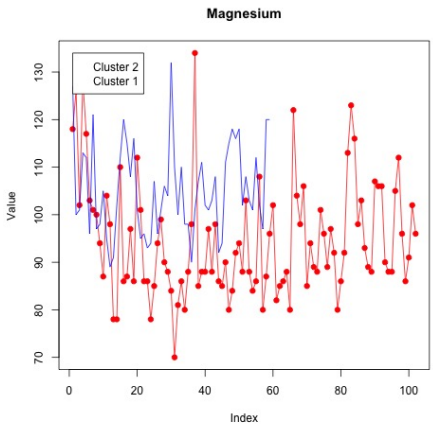
(c) Malic acid



(d) Ash

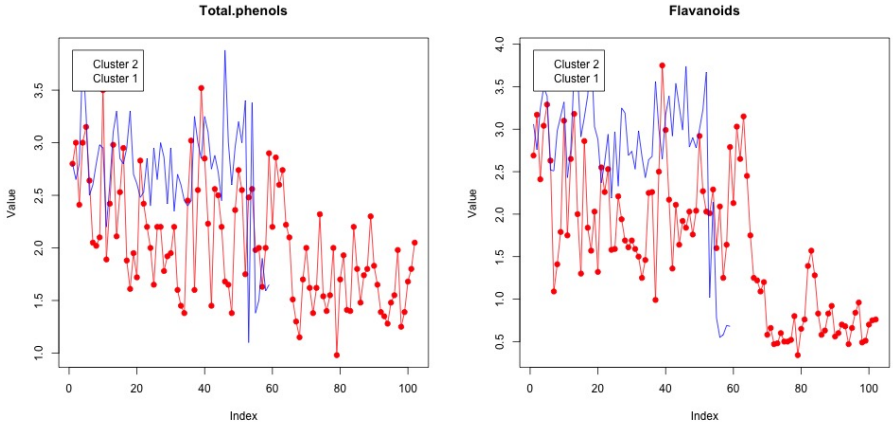


(e) Alkanility of ash



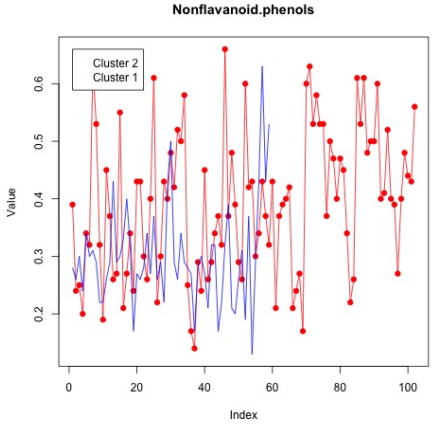
(f) Magnesium

Figura 5.1: Grafico de los cluster para cada variable (parte 1)

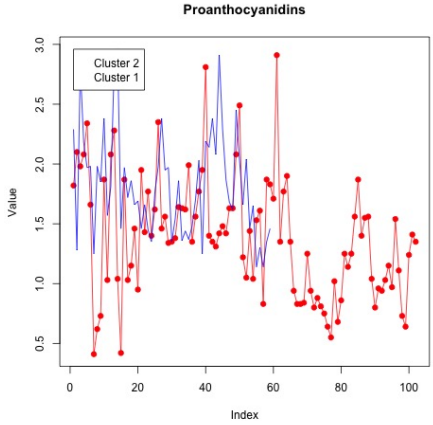


(a) Total phenols

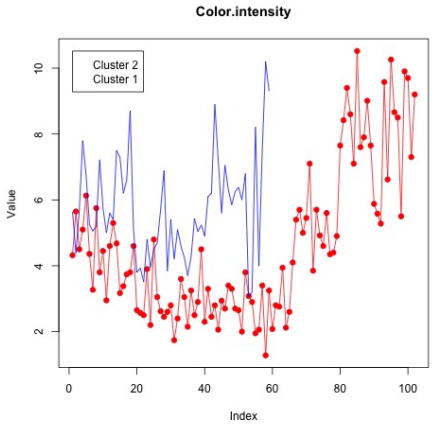
(b) Flavanoids



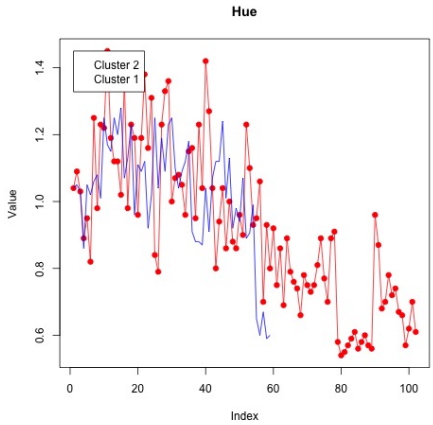
(c) Non flavanoids phenols



(d) Proanthocyanidins

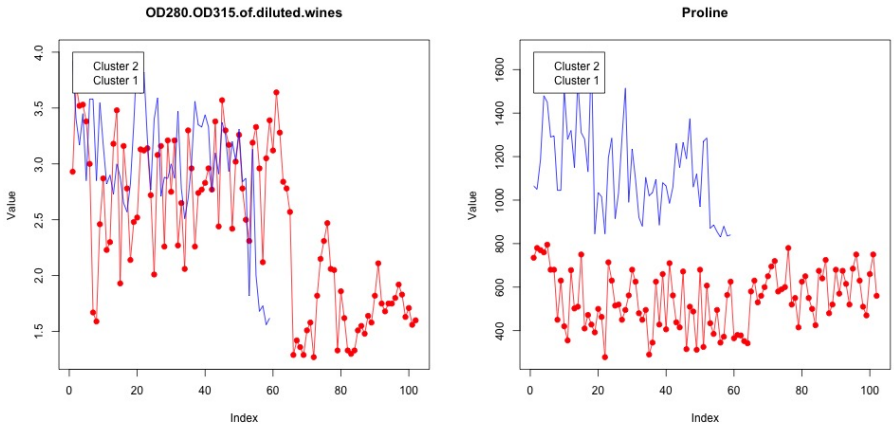


(e) Color intensity



(f) Hue

Figura 5.2: Grafico de los cluster para cada variable (parte 2)

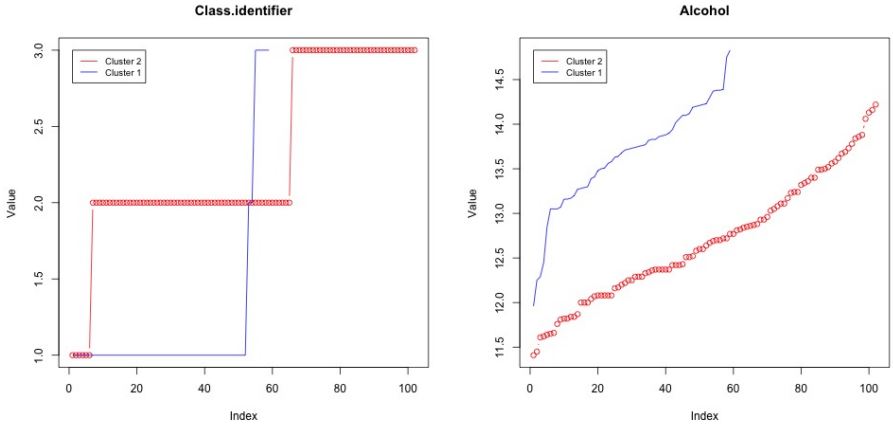


(a) OD280 OD315 of diluted wines (b) Proline

Figura 5.3: Grafico de los cluster para cada variable (parte 3)

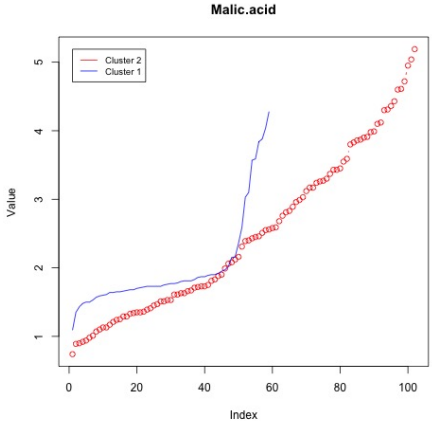
Puede ser que haya poca visualización en algunos gráficos que bajan y suben muy rapido solamente por orden, por lo que se muestra a continuación en las figuras 5.4, 5.5 y 5.6 los mismos datos pero antes ordenados según su valor.

Con esta mejor visualización, se puede ver que efectivamente el cluster 1 agrupa mayormente los vinos con un identificador 1, mientras que el cluster 2 agrupa a los identificados con 2 y 3. La variación en Malic acid en el cluster 1 es bastante significativa respecto al cluster 2. También hay algo interesante que muestra el gráfico y es la variable Proline, la cual en el cluster 1 hay una fuerte tendencia a ser mayor, al nivel de que el menor del cluster 1 es similar al mayor del cluster 2.

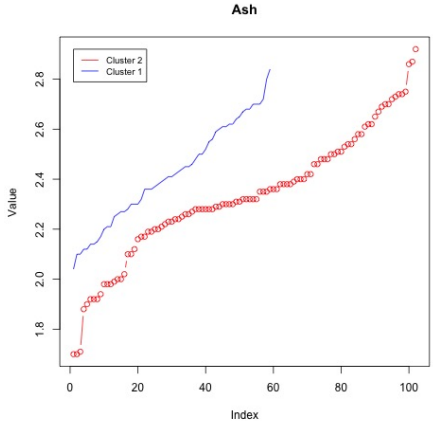


(a) Class identifier

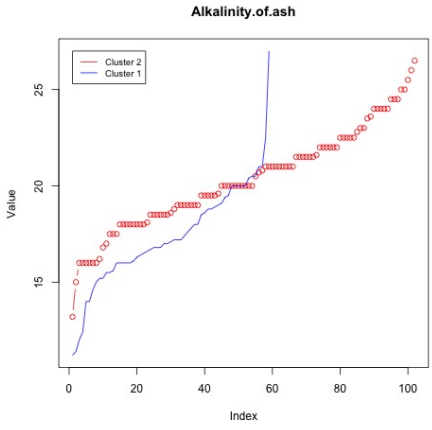
(b) Alcohol



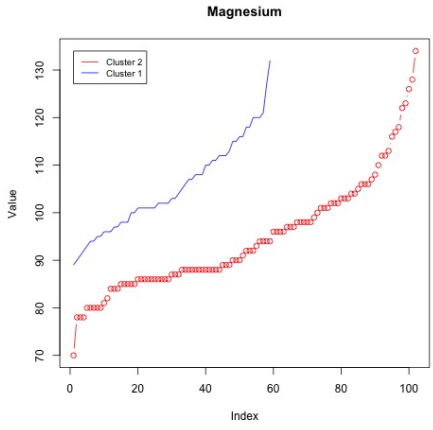
(c) Malic acid



(d) Ash

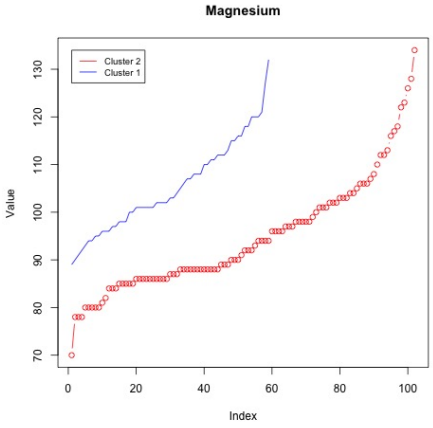


(e) Alkanility of ash

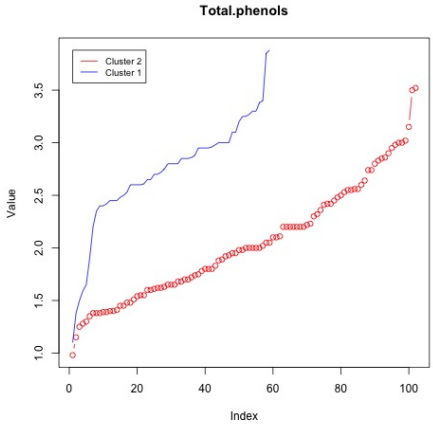


(f) Magnesium

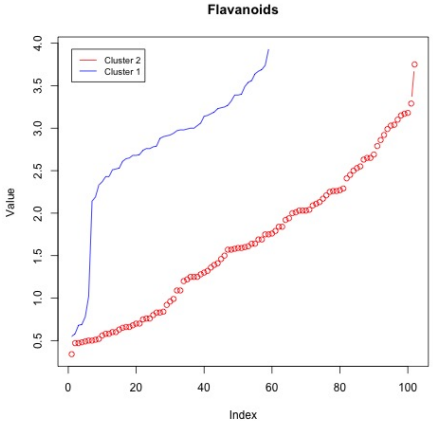
Figura 5.4: Grafico de los cluster ordenados para cada variable (parte 1)



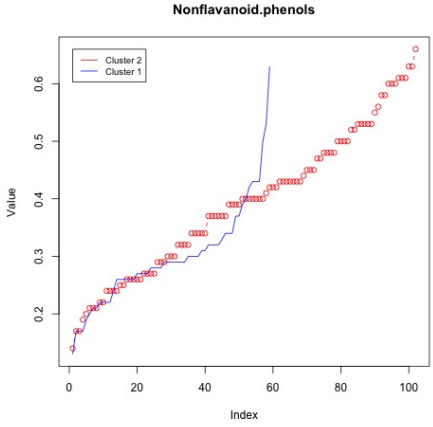
(a) Class identifier



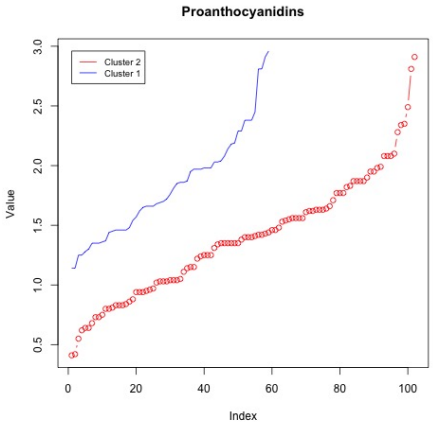
(b) Alcohol



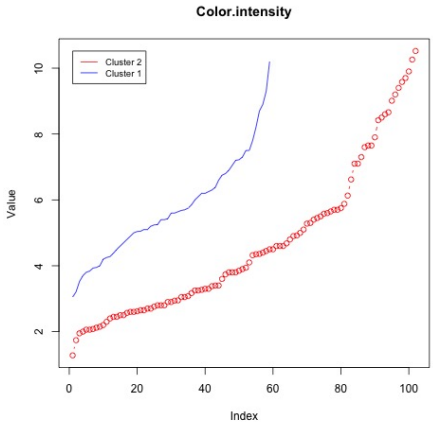
(c) Malic acid



(d) Ash

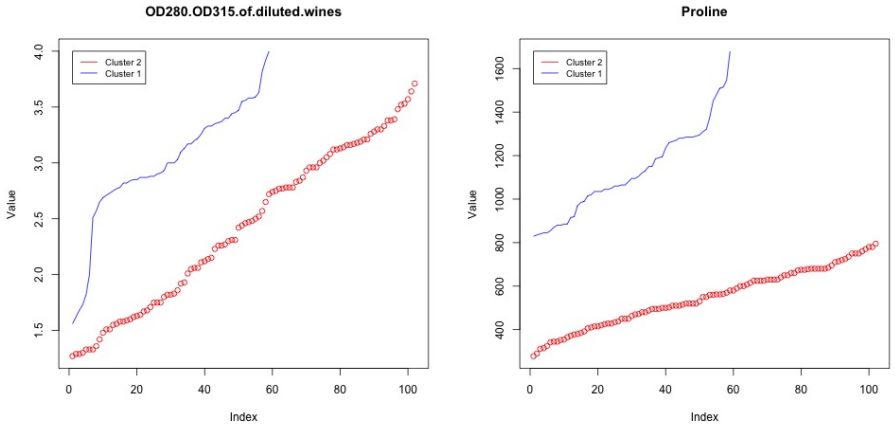


(e) Alkanility of ash



(f) Magnesium

Figura 5.5: Grafico de los cluster ordenados para cada variable (parte 2)



(a) OD280 OD315 of diluted wines

(b) Proline

Figura 5.6: Grafico de los cluster ordenados para cada variable (parte 3)

CAPÍTULO 6. CONCLUSIONES

En esta experiencia fue posible aprender de la aplicación del agrupamiento de datos mediante el uso del algoritmo k-medias, sin embargo, a pesar de la correcta implementación del algoritmo, el análisis de datos ha sido difícil debido a la complejidad para leer y comprender los datos presentados ante un agrupamiento en el plano bidimensional. Entre las principales dificultades, se tiene el poco conocimiento y especialización del tema, ya que al no tener bastos conocimientos de vinos y sus procesos de fabricación, se hace complejo el manejo de datos, especialmente para establecer límites y decidir que es un dato normal y que es un dato malo (o outlier). Esto es un proceso meticuloso que depende exclusivamente de la especialización de quién hace el estudio. Aún así, hay algoritmos conocidos para la limpieza pero no siempre son totalmente seguro, por lo que para descartar un dato es necesario ser experto en el tema. Para el caso del pre-procesamiento de datos, se puede decir que su principal dificultad también se encuentra en la especialización del tema a estudiar, esto porque es necesario tener amplios conocimientos para poder decidir de manera correcta y eficiente, cuáles son los datos que realmente aportan información relevante, cuáles no condicionan el agrupa- miento, entre otros. Por otro lado, es importante mencionar que este paso sigue siendo importante para realizar un buen análisis de datos, esto ya que se requiere de un estudio sólido que utilice información de calidad para no obtener conclusiones erróneas, ocasionadas por un mal empleo con los datos. De acuerdo a los resultados que fueron obtenidos y posteriormente analizados, en primera instancia el estudio realizado trajo como resultados que con los datos, el algoritmo realizó el agrupamiento de acuerdo a la clase en que estos vinos pertenecen, siendo los gráficos los que ayudaron a realizar esta tarea. Un problema que ocurrió durante el desarrollo del laboratorio, es que se usó una función que estaba siendo mal utilizada, ya que entregaba valores extremadamente raros. Este problema fue principalmente por la escasa documentación oficial del lenguaje R, por lo que se llegaron a funciones buscando en la web, sin saber bien su funcionamiento interno. Otros problemas fueron la decisión al tomar las distintas distancias que pueden ser utilizadas, y optando finalmente después de leer bastante literatura, la que se utilizó, que se asemeja de mejor forma al laboratorio realizado. Como comentario, no solo existe el metodo de la silueta,

si no que hay otros que también podría haber ayudado en el desarrollo del laboratorio pero se habría extendido demasiado, por lo que se optó por dejarlo como un futuro análisis. Por último, haciendo una breve comparación entre la experiencia anterior y la actual, se puede deducir que el actual estudio entrega mayor información, ya que a diferencia de la experiencia pasada, ahora fue posible agrupar las observaciones en base a las similitudes entre las medidas de cada variable. Esto puede ser de gran ayuda para estudios que necesitan diferenciar las características que se encuentran dentro de una muestra, donde las observaciones de éstas pueden ser agrupadas y posteriormente clasificadas.

BIBLIOGRAFÍA

Aggarwal, C. C. (2001). On the surprising behavior of distance metrics in high dimensional space. Recuperado desde https://bib.dbvis.de/uploadedFiles/155.pdf".

Wikipedia (2003). Algoritmo de agrupamiento. Recuperado desde https://es.wikipedia.org/wiki/Algoritmo_de_agrupamiento".

Wikipedia (2005). K-means. Recuperado desde https://es.wikipedia.org/wiki/K-means".

Wikipedia (2018). Outliers. Recuperado desde https://en.wikipedia.org/wiki/Outlier".