UNIVERSIDAD DE SANTIAGO DE CHILE
FACULTAD DE INGENIERÍA
DEPARTAMENTO DE INGENIERÍA INFORMÁTICA

Laboratorio 2 - Clustering

Integrantes:	Gustavo Aguilar Morita
	Sandra Hernández Yamunaqué
Curso:	Inteligencia Computacional
Programa:	Magíster en Ingeniería Informática
Profesor Cátedra:	Max Chacón Pacheco
Profesor Laboratorio:	Héctor Rojas Pescio

25 de Mayo de 2021

Tabla de contenidos

1. Introducción

La siguiente entrega corresponde a la segunda del curso de Análisis de Datos, la cual consiste en extraer conocimiento del problema escogido, es decir, de la base de datos "hepatitis"por medio del uso del algoritmo de clustering K-means. Cabe destacar que este procedimiento se logra gracias al desarrollo del laboratorio anterior, en el cual se estudiaron los atributos de la base de datos de manera detallada, permitiéndonos conocer el comportamiento de los datos así como el tipo de datos presente en la base de datos y de esta manera trabajar los datos de manera adecuada. Luego los grupos o clusters que son obtenidos por el algoritmo de clustering son analizados para encontrar las características más relevantes y obtener así conocimiento de las mismas así como de los grupos formados. Para el desarrollo de la experiencia se usaron diferentes técnicas de estadística descriptiva e inferencial mediante el uso del lenguaje de programación R.

1.1. Objetivos

- 1. Extraer conocimiento del dataset hepatitis, mediante el uso del lenguaje de programación R, utilizando el algoritmo de clustering K-means y realizar un análisis respectivo.
- 2. Comparar los resultados con lo expuesto en la literatura encontrada y ver si se sustenta el conocimiento obtenido.
- 3. Analizar los grupos e identificar aquellas características más relevantes y cuales clasifican mejor a una clase con respecto a otra. Además de inferir conocimiento respecto a los resultados entregados del aplicar los procedimientos anteriormente mencionados.

2. Marco Teórico

2.1. Clustering

Un algoritmo de agrupamiento o clustering, es un proceso usado muy recurrentemente en la mineria de datos o en Machine learning, el cual dado un conjunto de datos y un criterio, se realizan agrupaciones de datos de los cuales cada uno comparte caracterısticas o propiedades similares con el fin de obtener similitudes (entre elementos del mismo grupo) o diferencias (entre elementos de distintos grupos). Dentro de los algoritmos que existen para realizar esta técnica están el algoritmo de Mean-shift (cambio de medias) o el algoritmo de K-means (1) .

2.2. Distancias

Los métodos y algoritmos de Clustering requieren encontrar patrones de similitud para identificar los grupos, la más conocida es la medida de las distancias. Dentro de las mas importantes está la distancia de Gower, la cual tiene como caracterıstica generar una medida de distancia con datos de distintos tipos como por ejemplo en donde se tienen valores continuos y binarios. Su fórmula es la siguiente:

$$s_{ij} = \frac{\sum_{h=1}^{p_1} \left(1 - \left|x_{ih} - x_{jh}\right| / G_h\right) / + a + \alpha}{p_1 + (p_2 - d) + p_3} \tag{1}$$

- p_1, p_2, p_3 corresponden al número de variables cuantitativas continuas, variables binarias y cualitativas (que no sean binarias) respectivamente.
- a y d corresponden al número de coincidencias de las variables binarias $(1, 1)$ en el caso de a y $(0, 0)$ en el caso de d.
- α es el número de coincidencia en las variables cualitativas.
- G_h corresponde al recorrido o rango de h-ésima variable.

2.3. K-means

Es uno de los algoritmos de agrupamiento más conocido y más usado, su objetivo es particionar los datos en un conjunto de k grupos generados en base a la distancia de

los datos con respecto a un centroide. Este algoritmo es iterativo y cuenta con 5 pasos de ejecución que se describen a continuación:

- 1. Formar k agrupaciones con los datos de forma aleatoria.
- 2. Obtener el centroide de cada grupo mediante el cálculo de la media.
- 3. Calcular la distancia de cada elemento con respecto a todos los centroides y asignarlo al grupo cuyo centroide sea el más cercano.
- 4. Recalcular los centroides de cada grupo.
- 5.Repetir los pasos 3 y 4 hasta que los centroides dejen de variar y los datos dejen de cambiar de grupo.

Una vez realizado este procedimiento se obtiene un conjunto de k grupos de datos los cuales poseen una relación entre sı y de la cual se puede obtener información.

3. Desarrollo

3.1. Preprocesados relevantes

3.1.1. Imputación y conversión

La base de datos cuenta con diversas proporciones de datos no encontrado (2), por lo que se aplica Multivariate Imputation by Chained Equations (MICE) para imputar datos perdidos.
Para ello se consideran 3 grupos acorde a la naturaleza de las variables y sus datos perdidos:

1. Grupo 1: variables que no tienen datos vacios y que fueron descartados de MICE, correspondientes a: *class*, *age*, *sex*, *antivirals*, *histology*.

2. Grupo 2: aquellos atributos categóricos (factor) que tienen dos niveles o tipos de respuesta y que se le aplicó una regresión logística.

3. Grupo 3: aquellos atributos categóricos (factor) que tienen 3 o más níveles o tipos de respuesta, y que se les aplicó regresión logística politómica.

3.1.2. Outliers

Según la naturaleza de los datos, se analizaron atributos que no sean dicotómicos, de las cuales se detectó de acuerdo al test de Rosner (3) un 4,5 % de los datos que presentan una anomalía, de las cuales se desglosan en la tabla (1). Por otra parte, es importante destacar que la *bilirubin* para valores entre 4 y 5, se detecta algunos outliers, sin embargo, estos registros presentaban su clase como paciente fallecido, por lo que, debido a la poca cantidad de este tipo de casos, se deja como parte del dataset.

Variable	**N Outliers**	**Rango**
bilirubin	3	$[5 - 8{,}00]$
sgot	3	$[420{,}00 - 648{,}00]$
albumin	1	$[6, 40]$

Cuadro 1: Variables numéricas con outliers, que fueron eliminadas del dataset.

3.1.3. Normalización de los datos

Para este laboratorio, se considero la función (2) ya que presentó mejores resultados expuestos en los capítulos posteriores.

$$s = normalize(datos, method = "standardize") \tag{2}$$

Donde todos los atributos quedan entre 0 y 1, donde 1 se corresponde a su máximo valor para el caso continuo/discreto y para aquellos dicotómicos, se expresa 0 para la negación o muerte y el 1 como la afirmación o vida, acorde a las definiciones en (2).

3.2. Criterio de proximidad

Una parte importante del proceso de agrupamiento, es seleccionar la distancia con la que se trabajarán los algoritmos, de tal forma que se pueda ajustar de la mejor forma posible, acorde a la naturaleza de los datos. En este laboratorio, se probaron diversas estrategias para poder obtener los mejores resultados posibles, esto es, obtener la mejor calidad de los agrupamientos.

De la sección anterior, vemos que al normalizar los datos, cuyos valores pertenecientes a los reales, contenidos entre 0 y 1, de acuerdo a la literatura, nos sugería utilizar la distancia euclideana o bien la de manhattan, también se anallizó la de mahalanobis, sin embargo no se obtuvieron resultados relevantes. En consecuencia, al observar los resultados para cada uno de estos criterios, en el presente experimento se utiliza la distancia euclideana.

3.3. Obtención del clúster

3.3.1. Cantidad de clusters

Para encontrar un k apropiado para aplicar kmeans, se utilizaron diversos métodos y estrategias. Para efectos de este laboratorio, es importante tener en cuenta que el objetivo intrínseco es poder lograr la mejor clusterización posible, dado ello, las figuras que se presentarán a continuación, son las mejores representaciones obtenidas en todas las iteraciones realizadas, donde se consideraron los siguientes aspectos:

1. Tipo de normalización.
2. Eliminación de outliers y el tipo de técnica o test.
3. Distancias
4. Método de obtención del K.
5. Representaciones de los clusters y los coeficientes de silueta.

De acuerdo a la figura (1) se puede notar, que es sencillo concluir algún tipo de k apropiado para aplicar kmeans, por lo que se procede a analizar otros métodos, que nos permitan otorgar mayor información. Aún así, podemos notar que hay ciertos puntos de ïnflexión"que nos permiten dar ciertas pistas, como por ejemplo para k=4, k=6 y k=8.

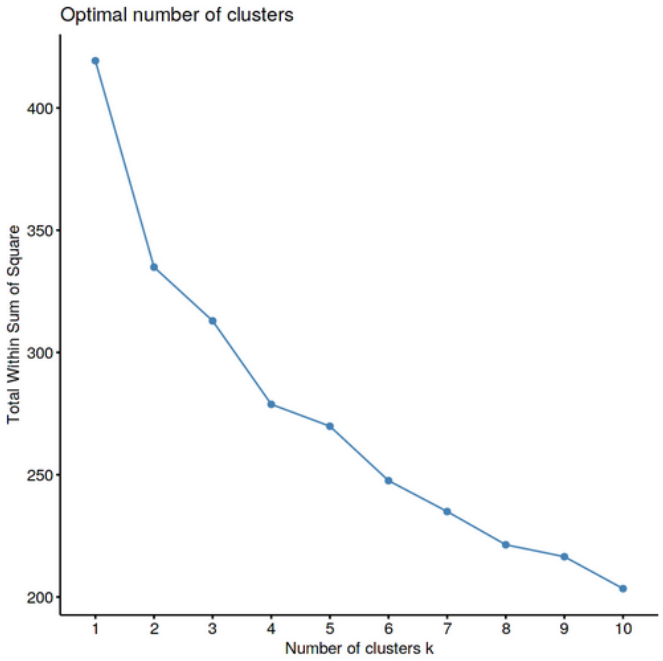


Figura 1: Método del codo con puntos de inflexión relevantes en k=4,k=6 y k=8.

En la figura (2) vemos que el k sugerido es 2, lo cual en primera instancia es bastante interesante dado que es el mismo número de clases de la base de datos (pacientes vivos/fallecidos). Sin embargo se descarta este k, dado que los centros de los clusters arrojan valores que no logra identificar la clase apropiadamente.

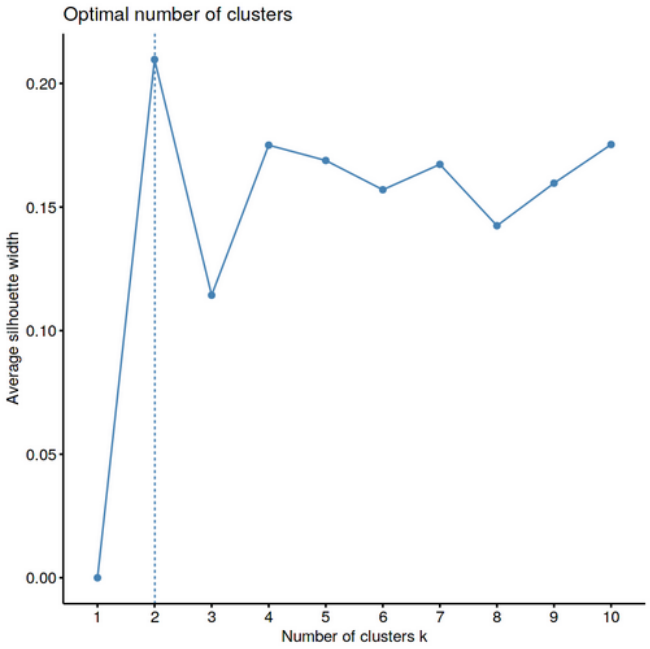


Figura 2: Método de silueta, cuyo sugerencia es k=2.

Con el método de la brecha estadística, se obtiene un k=5 que posteriormente será descartada, aunque haya obtenido una mejor media de anchuras de siluetas.

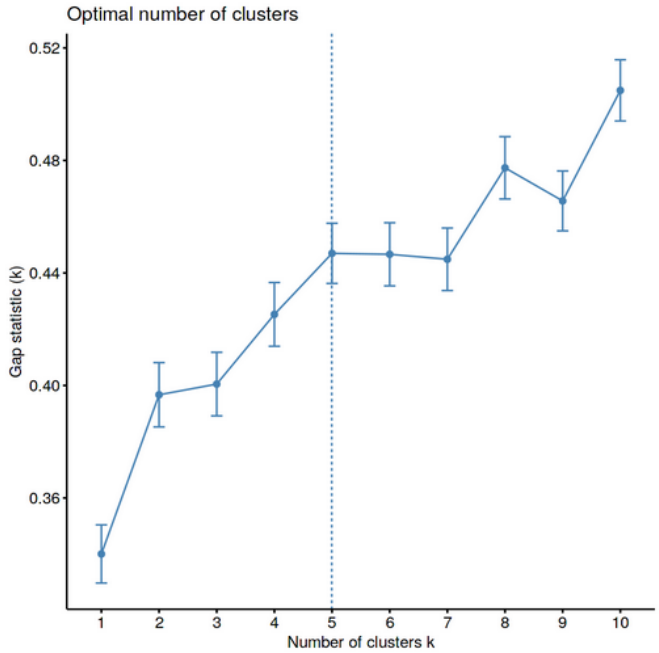
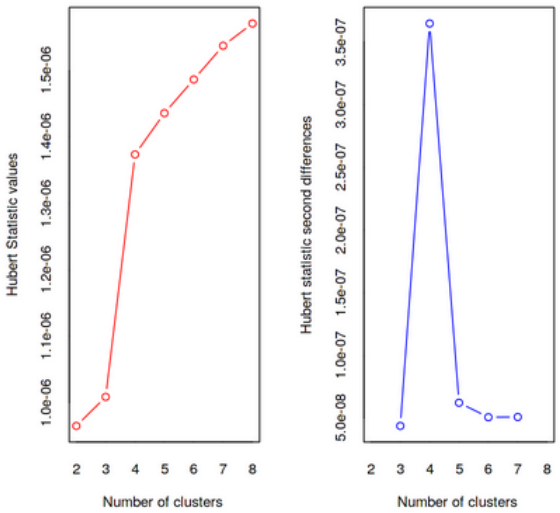


Figura 3: Método de la brecha estadística, sugiere un k=5.

Finalmente de acuerdo a la figura (4), nos sugiere un k=4 para aplicar el análisis de agrupamiento.




```
*** : The D index is a graphical method of determining the number of clusters.
                In the plot of D index, we seek a significant knee (the significant peak in Dindex
                second differences plot) that corresponds to a significant increase of the value of
                the measure.

*******************************************************************
* Among all indices:
* 3 proposed 2 as the best number of clusters
* 4 proposed 3 as the best number of clusters
* 16 proposed 4 as the best number of clusters
* 1 proposed 6 as the best number of clusters

                    ***** Conclusion *****

* According to the majority rule, the best number of clusters is  4
```

Figura 4: Método gráfico de Hubert, donde se sugiere un k=4.

4. Análisis de los resultados

4.1. Análisis de todos los datos

Como primera instancia se utiliza el método de clustering o agrupamiento K-means a los datos previamentes trabajados, es decir, sin outliers y normalizados. La cantidad de cluster a utilizar corresponde a la entregada por la función NbClust de R (Sección 3.3.1), la cual nos sugiere que el número de clusters apropiado para este conjunto de datos corresponde a cuatro clusters. Analizando la figura 5 la cual es la representación gráfica del resultado de realizar el algoritmo de k-means al conjunto de datos, se puede observar que a simple vista que, los clusters 2 y 3 se superponen casi en su totalidad así como a sus clusters vecinos. Lo anterior no sucede con los clusters 1 y 4 los cuales están totalmente separados entre si, luego se puede decir que aunque el número sugerido como la cantidad de clusters óptimo, este no lo es, pues no se lograr clusterizar de manera que todos los clusters queden separados entre si, el cual es el objetivo del método.

Figura 5: Gráfico resultante con k=4.

4.1.1. Evaluación de calidad

Aunque ya se pudo evidenciar de manera gráfica que la cantidad de cluster no es óptima, se busca validar esta conclusión realizando una evaluación de calidad para el clustering de los datos con un K=4. Es por lo anterior que se utilizó el método Silhouette o Silueta para validar la coherencia de los grupos formados, el resultado se aplicar este método nos entrega un gráfico donde se puede observar que tan similar son los datos con respecto a su propio cluster en comparación a los otros, así como entregar el valor de la silueta para cada uno de los clusters y el promedio entre ellos. Cabe señlar que los valores varían entre -1 y 1 siendo 1 una agrupación de excelente calidad y entre 0 y -1 lo contrario. A continuación en la figura 6 se pueden observar los resultado de aplicar el método Silhouette a nuestro cluster con un K=4.

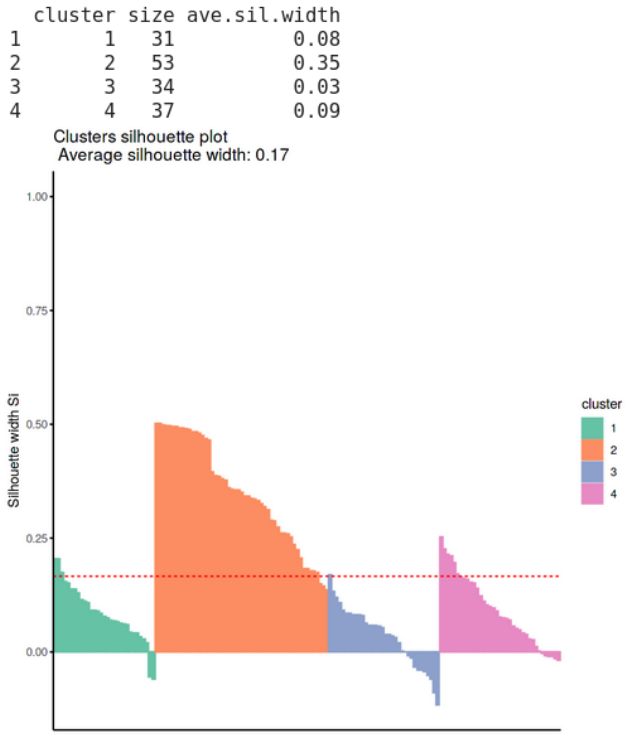


Figura 6: Coeficiente de silueta: Calidad del agrupamiento con K=4.

Analizando la figura 6 se puede observar que todos los clusters alcanza o superan el valor promedio de silueta lo cual se considera como un requisito mínimo de calidad pero luego vemos que tres de los cuatro clusters presentan datos mal ubicados o mal agrupados, lo

cual se representa como las barras bajo el eje horizontal. Luego los valores correspondiente de silueta para cada cluster se observan la última columna de la tabla superior del gráfico, en el cual se ven valores entre 0.03 y 0.35. Aunque los valores sea superiores a 0 esta muy alejados valor ideal (1), el conclusión la clusterización con un K=4 no resulta ser un cluster de poca calidad como se observó en la representación gráfica del mismo. Por otra parte de acuerdo a lo evidenciado con el método de la brecha estadística (k=5), se realiza una inspección de la calidad, la cual se obtiene una media de anchos de silueta de un total de 0.18, sin embargo, se descarta dado que el perfil es inferior a la media (ver anexo).

Dado lo anterior, se agrega a este análisis otras variantes respecto al agrupamiento, donde el método PAM por ejemplo, no se evidencia una mejoría en la calidad de la clusterización, arrojando un coeficiente de silueta promediado de 0.13 para k=9. Sin embargo, en el siguiente apartado, podemos ver que al aplicar PCA antes de emplear kmeans con el dataset y todas sus dimensiones, se logra obtener una leve mejoría en el agrupamiento.

4.2. Análisis KMeans con PCA

De acuerdo al estudio realizado en (4) se logra identificar que algunos atributos pudiesen ser descartados, según el análisis de componentes principales, por lo que a continuación se realizará un contraste de los resultados obtenidos aplicando kmeans con reducción de la dimensionalidad.

Las variables que se descartaron para este análisis fueron *sex*, *steroid*, *anorexia*, *spleen − palpable*, *ascites* y *varices*.

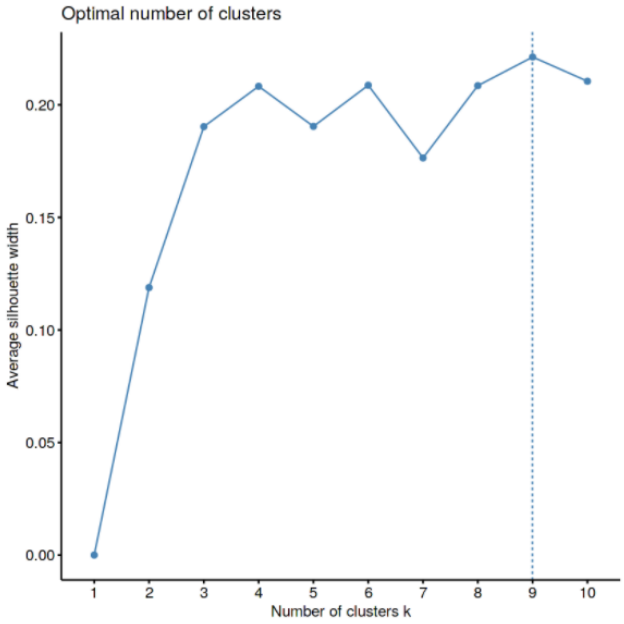


Figura 7: Obtención del k óptimo (k=9) para kmeans con PCA, con método Silhouette.

De acuerdo a la figura (7) vemos que el k apropiado para aplicar kmeans es de 9, y cabe mencionar que los otros métodos fueron descartados, dado que se obtenía un k=2, que es igual al número de opciones del atributo class.

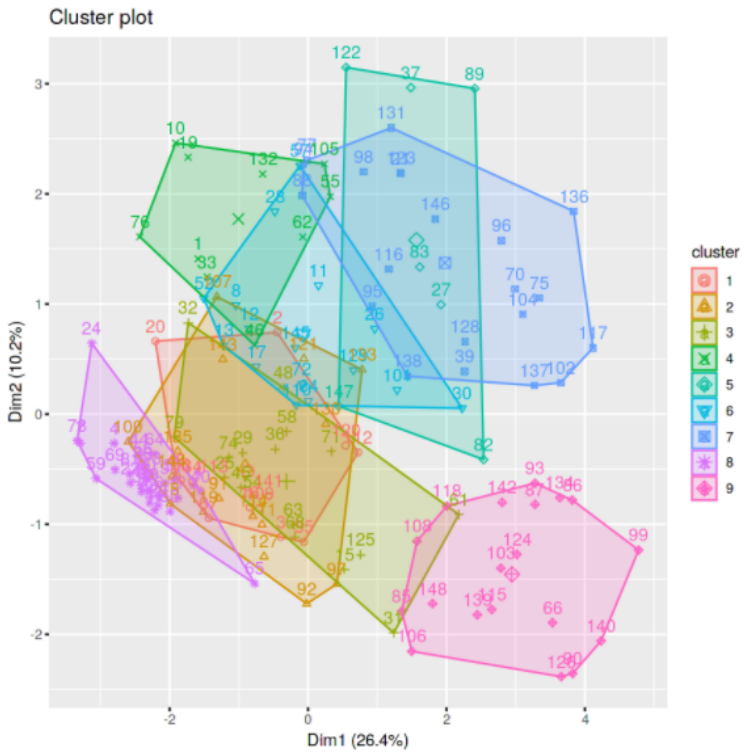


Figura 8: Representación 2d de los 9 clusters aplicando PCA y kmeans.

Se puede notar que existe cierto solapamiento entre los clusters acorde a la figura (8), por lo que las primeras conjeturas apuntan a que no se obtendrá una buena calidad en el agrupamiento. Para argumentar de mejor manera lo anterior, veamos los resultados obtenidos de acuerdo a la figura (9). Se puede notar, que los clusters 4 y 5 están bajo la media, y además se presentan clusters con anchos cercano a 0 (0.04 y 0.08), por lo tanto no se logra obtener una calidad que pueda satisfacer el análisis de kmeans. Si bien el PCA, mejora en la calidad del agrupamiento respecto a lo presentado en las secciones anteriores, podemos notar que no es sustantivo.

	cluster	size	ave.sil.width
1	1	14	0.13
2	2	16	0.21
3	3	17	0.28
4	4	10	0.04
5	5	7	0.08
6	6	16	0.09
7	7	21	0.11
8	8	28	0.50
9	9	19	0.24

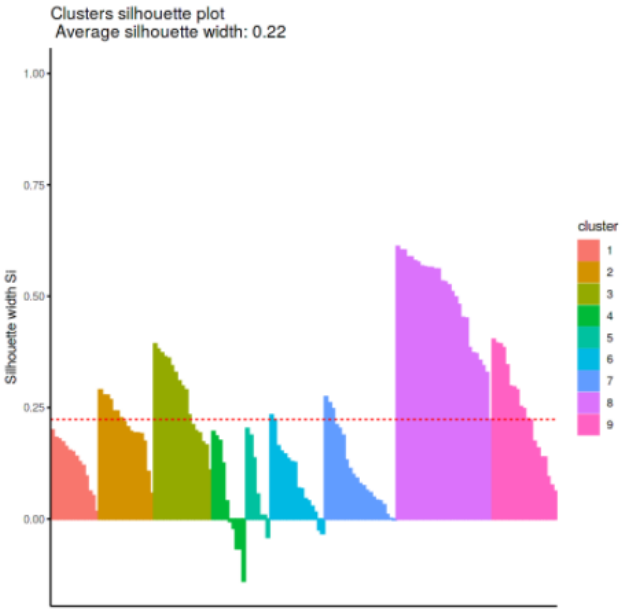


Figura 9: Coeficiente de silueta. Calidad del agrupamiento con K=9, para kmeans con PCA.

Ahora bien, de acuerdo a la figura (10) podemos observar algunos detalles interesantes que aportan a la investigación. Para el caso de la class=0, correspondiente a los pacientes que fallecieron, podemos notar que la *bilirubin* es significativamente más alta que en el resto de los clusters (alrededor del 50 %), *albumin* y *protime*, presentan niveles bajo respecto a los otros clusters y finalmente la *histology*, se acentúa la presencia de la condición en pacientes fallecidos.

```
pca.cluster.9k$centers[,c("class", "bilirubin", "albumin", "protime", "histology")]
```

A matrix: 9 × 5 of type dbl

	class	bilirubin	albumin	protime	histology
1	0.92857143	0.1964371	0.6435412	0.5916728	0.35714286
2	0.87500000	0.1960530	0.5848602	0.6474557	1.00000000
3	0.94117647	0.2180378	0.5810926	0.6010520	0.05882353
4	1.00000000	0.1684813	0.5673466	0.5932625	0.20000000
5	1.00000000	0.2625044	0.4598214	0.4458228	0.42857143
6	0.93750000	0.2350404	0.5585938	0.6009279	0.18750000
7	0.76190476	0.3462854	0.5090757	0.4962645	0.80952381
8	1.00000000	0.1726154	0.6626365	0.7085838	0.00000000
9	0.05263158	0.4300473	0.2970027	0.2935407	0.94736842

Figura 10: Centros de los clusters, filtrados por atributos relevantes para pacientes fallecidos.

4.3. Contraste con otras investigaciones

Antes de contrastar los resultados obtenidos del clusterizar con un K=4 los datos sin aplicar PCA, es necesario interpretar lo clusters formados y para ello se observaron los centros de cada uno de los clusters e identificaron aquellos parámetros que los representan. Se descartaron los siguientes parámetros **antivirales**, **hígado grande**, **bazo palpable** y **varices** dados que estos se encuentran presentes en todos los cluster formados por lo que no son representativos de su cluster.

- Cluster 1: Paciente vivo que presenta durezas en el hígado, , arañas vasculares, líquido seroso en la zona abdominal.
- Cluster 2: Paciente vivo que presenta esteroides, fatiga, malestar, anorexia, durezas en el hígado, arañas vasculares, líquido seroso en la zona abdominal y tendencia de la sangre a coagularse.
- Cluster 3: Paciente vivo que presenta malestar, anorexia, durezas en el hígado, arañas vasculares, líquido seroso en la zona abdominal y tendencia de la sangre a coagularse.
- Cluster 4: Paciente fallecido que presenta un estudio histológico.

	class	age	sex	steroid	antivirals	fatigue	malaise	anorexia	liver_big	liver_firm
1	0.9354839	0.4675148	0.22580645	0.4516129	0.7419355	0.0000000	0.0000000	0.4193548	0.8064516	0.58064516
2	0.9622642	0.4429976	0.07547170	0.7169811	0.8679245	0.7735849	1.0000000	1.0000000	0.9811321	1.00000000
3	0.9411765	0.4917150	0.11764706	0.4411765	0.7647059	0.4117647	1.0000000	0.9705882	0.6176471	0.05882353
4	0.2972973	0.5397792	0.02702703	0.3243243	0.9729730	0.0000000	0.1891892	0.6216216	0.7567568	0.45945946

Figura 11: Cuadro detalle de los centros resultantes del clustering con K=4.

spleen_palpable	spiders	ascites	varices	bilirubin	alk_phosphate	sgot	albumin	protime	histology
0.8387097	0.6774194	0.9677419	0.9032258	0.1426825	0.4065682	0.15396357	0.4273584	0.4541147	0.1612903
0.9622642	0.9622642	0.9811321	0.9811321	0.1100806	0.2996843	0.07375349	0.4903366	0.5523657	0.3018868
0.7352941	0.6176471	1.0000000	0.9411765	0.1247320	0.4229842	0.11809488	0.3947285	0.5146199	0.4411765
0.5945946	0.2162162	0.4594595	0.6216216	0.2923903	0.4875191	0.14274624	0.2631139	0.2954825	0.9189189

Figura 12: Continuación de la figura 11.

Se puede decir muy poco respecto al clustering realizado y de las características de los mismos pues si nos enfocamos en el objetivo del estudio el cual pretende determinar o encontrar aquellas características que presentan aquellos pacientes con hepatitis y con riesgos de morir, solo el parámetro *histology* sería de importancia según los resultados del cluster 4 el cual agrupa pacientes fallecidos. Ahora bien si nos enfocamos en los parámetros restantes entregados por el PCA, *bilirubin*, *histology*, *albumin* y *protime*, se tiene que los pacientes fallecidos (Cluster 4) presentan niveles superiores de bilirrubina con respecto a los pacientes que sobrevivieron y bajos niveles de *albumin* y *protime* con lo cual se puede decir que aunque el cluster no sea de la mejor calidad los parámetros considerados relevantes según el PCA de igual manera representa una diferencia con respecto a los cluster de los pacientes que sobrevivieron.

De acuerdo al estudio recogido en (5), se evidencia que los pacientes que alto grado de bilirrubina o bien tasas de cambios espontáneas de la misma, podrían significar en un daño severo en el hígado debido a la hepatitis, lo cual de alguna manera se relaciona de manera tímida en nuestros análisis de agrupamiento. Por otra parte, de acuerdo a las figuras (10) y (12), notamos que el *protime* para pacientes fallecidos es bajo respecto a los pacientes vivos, lo que se contradice con la investigación encontrada en (5) y (6), lo que se podría relacionar con la calidad de los agrupamientos obtenidos.

Finalmente el siguiente estudio (7) hace mención de como bajos niveles de *albumin* evidencian problemas hepáticos graves y que este es una variables que va disminuyendo a medida que el daño hepático es más severo, el mismo estudio se refiere también al estudio histológico el cual es representado en el dataset como *histology* y como este es usado para conocer el estado del paciente y confirmar el diagnóstico de hepatitis crónica. Luego podemos ver que la importancia de el parámetro *histology* se corresponde con lo entregado por nuestro cluster y que el paciente se haya realizado o no un estudio histológico es poco concluyente para determinar si un paciente sobrevivirá o no.

5. Conclusión

De acuerdo al análisis realizado, se observa que las variables *bilirubin*, *protime* y *albumin* son condiciones que presentan una cierta relevancia al momento de predecir si un paciente tiene probabilidades de fallecer o vivir, dada sus condiciones bioquímicas analizadas. Kmeans, nos indica que el dataset, se puede agrupar de mejor forma para un k=4 para un análisis sin reducción de la dimensionalidad y para un k=9 con PCA aplicado previamente a la clusterización. En ambos casos se obtuvieron los clusters con centros diferenciados en la *bilirubin* y bajos índices para *protime* y *albumin*, respecto a pacientes fallecidos y vivos. Sin embargo de acuerdo a la literatura, *protime* es una condición relevante que contradice nuestro agrupamiento.

Por otra parte, se obtiene que la histología es un variable fuertemente presente en pacientes fallecidos en ambos análisis propuestos.

La calidad del agrupamiento para k=4, se obtuvo una media de anchos de siluetas de 0.17, por lo que se sugiere buscar otras técnicas y estrategias que permitan encontrar un mejor agrupamiento de los datos, como por ejemplo el haber aplicado PCA, mejora a 0.22 en el presente laboratorio. Sin embargo, el significado de los clusters, en algunos puntos se contradice con las investigaciones presentadas en el documento, por lo que no se pueden obtener conjeturas acertadas, respecto a si un paciente que cumpla ciertas condiciones dado los clusters generados, pueda implicar en una alta probabilidad de fallecer o no.

Otros aspectos relevantes que mejoraron la calidad del agrupamiento, fue considerar diversas formas de preprocesar los datos, considerando el tipo de normalización, eliminación de outliers, método de la obtención del k apropiado y las distancias.

Finalmente se propone para este laboratorio, encontrar otras formas de agrupar los datos o bien utilizar otras técnicas de aprendizaje no supervisado o supervisado, con el fin de predecir de mejor forma la problemática.

Bibliografía

[1] J. A. Hartigan and M. A. Wong, "Algorithm as 136: A k-means clustering algorithm. applied statistics, 28s," 1979.

[2] D. Dua and C. Graff, *UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science*. 2019.

[3] A. Soetewey, *Outliers detection in R [https://statsandr.com/blog/outliers-detection-in-r/]*. 2020.

[4] G. Aguilar and S. Hernández, *Laboratorio 1, Inteligencia Computacional [https://github.com/naotoam/magister/blob/master/Inteligencia%20Computacional/Informe%201%20Laboratorio.pdf]*. 2021.

[5] C. Y. K. S. K. D. e. a. Lee M, Kim W, *Spontaneous Evolution in Bilirubin Levels Predicts LiverRelated Mortality in Patients with Alcoholic Hepatitis*. 2014.

[6] N. S. E.-F. Myron J. Tong and M. I. Grew, *Clinical Manifestations of Hepatitis A: Recent Experience in a Community Teaching Hospital*. 1995.

[7] S. Alegría, "Hepatitis crónica," *Revista chilena de pediatría*, vol. 73, no. 2, pp. 176–180, 2002.