Calidad del Vino Rojo según MCluster

Luis Orellana Altamirano, Fernando Cabrera Gajardo.

Universidad de Santiago de Chile, Chile
Avenida Ecuador #3659. Estación Central, Santiago de Chile., Chile
luis.orellana.a@usach.cl
fernando.cabrera.ga@usach.cl

Abstracto. El presente trabajo de investigación está centrado principalmente en la calidad del vino en su variedad tinto (rojo). La "calidad" cumple cualidades interesantes a mencionar en clusterizacion, las cuales son dadas según sus características fisicoquímicas como la impresión subjetiva que obtiene el catador al degustar el vino. Esto en efecto, tiene repercusiones directas entre estas dos medidas de calidad, las cuales en algunos casos se pueden ser contrastadas por grandes diferencias entre sí. Ante esto, se propone MCluster el cual permite realizar análisis sobre las variables correspondientes a cada vector, de esta manera contrastar la calidad del vino percibida (por catadores profesionales) y las medidas de calidad clásica propuestas de manera fisiológica.

Palabras Clave: Medidas de calidad, acidez, Ph, alcohol residual.

1 Introducción

De la misma manera que se disfruta una buena comida, el vino se ha convertido en un acompañante indiscutible para el paladar en numerosas situaciones rutinarias. Debido al gran aumento de las ventas en combinación al referente de esta área que representa Chile, se hace imperativo para tal industria el conocer cuáles son las relaciones entre el gusto percibido de manera subjetiva por los catadores, y medidas que se miden generalmente en laboratorios fisicoquímicos. Si vamos más a fondo en este tema, para el presente estudio se aborda la problemática utilizando un dataset de índole público, el cual es facilitado por el repositorio dispuesto por UCI [1]. Este dataset está compuesto por una muestra total de 1599 instancias, los cuales corresponden a impresiones subjetivas que obtuvieron catadores de distintos vinos de Portugal, pertenecientes a la variedad roja (vino tinto). Las variables que son presentadas están en el dataset se pueden ver en la tabla 1. Por otra parte, con la finalidad de conocer sobre los datos y sus distribuciones, como también el comportamiento y relación de las variables con sus respectivas etiquetas, se añade en el anexo un análisis de las estadísticas descriptivas e inferenciales. Además, el problema radica en determinar la relación que existe entre variables y sus respectivas etiquetas desde el punto de vista subjetivo presentado en el dataset y las principales variables medidas en laboratorios fisicoquímicos. Se ha abordado esta

temática por medio de MClust. Esta técnica es utilizada en el presente problema con la finalidad de obtener conocimiento a partir de un volumen de datos considerable como lo es este dataset.

Tabla 1. Estadísticas fisicoquímicas para vino tinto.

Atributos	Min	Max	Media
Fixed acidity (g(tartaric acid)/dm3)	4.6	15.9	8.3
Volatile acidity (g(acetic acid)/dm3)	0.1	1.6	0.5
Citric acid (g/dm3)	0.0	1.0	0.3
Residual sugar (g/dm3)	0.9	15.5	2.5
Chlorides (g(sodium chloride)/dm3)	0.01	0.61	0.08
Free sulfur dioxide (mg/dm3)	1	72	14
Total sulfur dioxide (mg/dm3)	6	289	46
Density (g/cm3)	0.990	1.004	0.996
pH	2.7	4.0	3.3
Sulphates (g(potassium sulphate)/dm3)	0.3	2.0	0.7
Alcohol (vol.%)	8.4	14.9	10.4

2 Método y Herramientas

2.1 Método de Minería de Datos

Con la finalidad de obtener conocimiento por medio de técnicas de minería de datos, se utilizará el método MClust proveniente de un paquete del software R. Este software es de código abierto, multiplataforma y de programación de alto nivel, el cual trata como caja negra operaciones vectoriales y matriciales. Posee un lenguaje estadístico llamado R. Específicamente, para el análisis de este paper, se utilizaron las librerías “plotly”, “modeest”, “ggplot2” y principalmente la librería “mclust” la cual contiene el método de clustering MCluster. Este método hace uso de distribución normal multivariada, la cual es expresada como:

$$f_X(x_1, \ldots, x_n) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^\top \Sigma^{-1}(\mathbf{x}-\mu)\right)$$

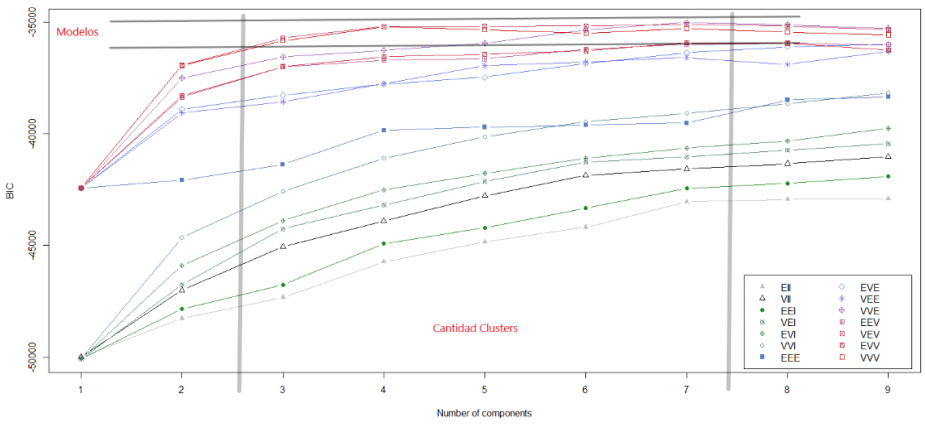
Donde “mu” es la media aritmética de la variable procesada y “sigma” es una matriz de varianza covarianza. Estos dos elementos proveen a la función gaussiana de tres parámetros importantes: alineación, forma y volumen. Estas tres variaciones son utilizadas por el método MClust de tal forma que la agrupación se efectúa por medio de la probabilidad de pertenencia que poseen los clusters al ser incluida las variables de cada

vector. Es por eso que los parámetros que recibe la función MClust son: modelo y cantidad de clusters. Por otra parte, la calidad de agrupación de los clusters se efectúa al analizar la medida BIC, la cual da una mayor ponderación a clusters que incluyen vectores con una mayor probabilidad de pertenencia y, por otro lado, disminuye su puntuación al procesar modelos complejos. Esta métrica también es contrastada con el error de agrupación de cada cluster. Por otra parte, se efectuó la normalización de los datos donde el rango de cada variable se modificó entre cero y uno.

2.2 Selección de Parámetros

Como primera instancia, el problema se abordó normalizando los datos sin emplear la etiqueta de cada vector. La selección de parámetros de MClust los cuales son el modelo y la cantidad de clusters, se realizó utilizando una técnica similar a la del Elbow empleada para K-Medias e ilustrada en la Imagen 2.

Imagen 2. Métrica de BIC para n clusters.



Debido a que las combinaciones de parámetros (problema combinatorio) pueden no ser eficientes en cuanto al tiempo que tardan en ser efectuadas, solo se seleccionan los mejores parámetros, los cuales para el caso de los datos normalizados son: (modelos) VVV – VEV – VVE. (Cantidad de clusters) 3 – 4 – 5 – 6 – 7. Con este procedimiento se obtienen los mejores BIC. Seguido de esto, se verifican los errores de cada uno de los clusters con mejores BIC, obteniendo para el mejor de ellos un error de 0.5866. Los parámetros para esta selección es VVV y 4 clusters.

Luego se efectuó el mismo procedimiento anterior, pero esta vez sin normalizar los datos. El tiempo de procesamiento aumento considerablemente en comparación al procedimiento anterior normalizando los datos. Con esto se obtuvo un error más bajo que lo anterior (0.5053), y los parámetros finales son:

- Volumen: Igual

- Forma: Variable
- Orientación: Variable
- Cantidad de clusters: 4

2.3 Variables Representativas

Recordando el problema inicial, el cual trata sobre determinar cuan relacionadas están las variables que conllevan el sabor y calidad del vino, con el etiquetado de las mismas, cada uno de los vectores de dataset corresponde a una métrica subjetiva propuesta por una serie de catadores de vinos. Esas métricas van desde el 0 al 10, donde 0 es un vino de muy mala calidad y 10 de muy buena calidad.

Por otra parte, es importante determinar las variables que representan la calidad del vino. Ante esto, se tienen dos aristas importantes. Una es la percepción del sabor, y las otras variables que son recurrentemente utilizadas por laboratorios fisicoquímicos. En este último, las variables que representan mayor sabor al vino según Doug Nierman (2004) son:

- Acidez Fija.
- Acidez Volátil.
- Ácido Cítrico.
- PH.
- Alcohol.

Como es posible observar en la matriz de correlación (anexo estadístico), algunas de las variables utilizadas de manera fisicoquímica para medir la calidad no tienen un grado de correlación positiva o negativa con la calidad del vino, pudiendo ser una de las razones causales del nivel de error en el cluster. El análisis sensorial es el menos comprendido de los sentidos humanos [3], y además no es bastamente conocido ni entendido aún [4]. Esto hace del dataset sin duda un problema complejo de abordar. Sin embargo, con la finalidad de analizar los clusters formados por MClust, se utilizarán las variables que tienen mayor correlación con sus etiquetas, y que coincidan con las variables planteadas por Gavin Sacks y David Jeffery (2016). Por lo tanto, las variables con las cuales se analizarán los clusters son:

- Ácido Cítrico.
- Ph.
- Alcohol.

2.4 Resultados Obtenidos (discusión)

Los resultados serán analizados por las variables más importantes que ayudan a definir el sabor y por lo tanto la calidad del vino.

- **Ácido cítrico:** Según los estudios realizados por Gavin Sacks and David Jeffery (2016), la acides de los vinos no debe superar los 1g / l en la cosecha, es decir, mientras más bajo es el nivel de acidez cítrica tiene el vino, de mejor calidad cera tanto para el nivel sensorial humano, como para métricas fisiológicas.

Imagen 3. Nivel de acidez según su calidad.

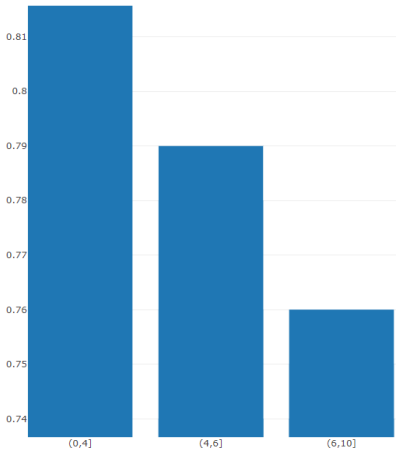


Tabla 2. Acidez de los Clusters.

Acidez/Clusters	1	2	3	4
0 – 0.25	429	37	93	38
0.25 – 0.5	249	38	286	42
0.5 – 0.75	23	23	136	2
0.75 – 1	0	2	3	1

Es posible observar en la imagen 3 que los datos del dataset corroboran las métricas fisiológicas. Por otra parte, en la tabla 2 muestra que el cluster número uno contiene la mayor cantidad de casos donde la acidez cítrica no supera los umbrales de calidad. Por otra parte, el cluster 3 posee los casos en donde se puede observar mayor acidez cítrica, no sobrepasando el umbral crítico.

- **Alcohol:** Según métricas fisicoquímicas, el alcohol en el vino no debe superar el 14%, ya que afecta al sabor de este, lo cual se ve reflejado en el dataset (imagen 4). Por otra parte, el sabor se ve afectado drásticamente si el porcentaje de alcohol es menor al 13,1%. Para este caso, el cluster 1 lleva un gran solapamiento de datos y contiene gran cantidad de muestras con bajo alcohol. Por otra parte, los casos de alcohol recomendado son observables en el cluster 3.

Imagen 4. Nivel de alcohol según su calidad.

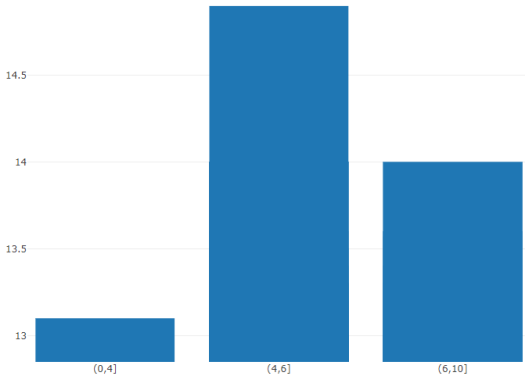


Tabla 3. Alcohol de los clusters.

Alcohol/Clusters	1	2	3	4
8 – 9	17	7	9	4
9 – 10	468	82	116	48
10 – 11	218	19	187	20
11 – 12	106	0	155	6
12 – 13	51	0	61	6
13 - 14	10	0	11	1

- **PH:** Según los estudios realizados por Gavin Sacks and David Jeffery (2016), el PH debe oscilar entre 3,1 y 3,6 lo cual se ve reflejado con valores cercanos a lo mencionado para los vinos de más alta calidad dentro del dataset. En este caso, el método MCluster tuvo resultados buenos, en los cuales se puede ver en la tabla 4 que el cluster y constituye los vinos de mejor calidad, junto con el cluster 3. Los vinos de peor calidad están repartidos entre el cluster 2 y 3.

Imagen 5. Nivel de pH según su calidad.

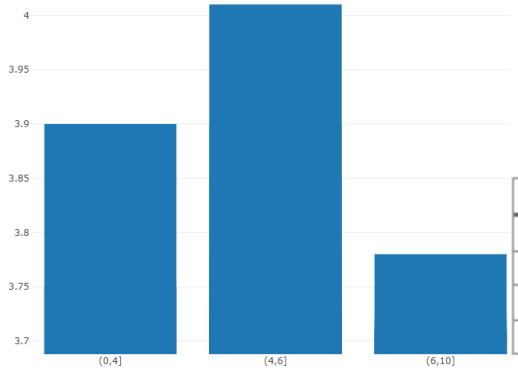


Tabla 4. Acidez de los Clusters.

pH/Clusters	1	2	3	4
2.5 – 3	4	13	16	2
3 – 3.5	708	92	515	81
3.5 – 4	152	3	9	2
4 – 4.5	2	0	0	0

3 Conclusiones

En los últimos años, la demanda de vino a crecido, y, en consecuencia, la calidad de este es crucial para suplir la demanda de este producto. Más aún, la minería de datos es un recurso valioso para obtener conocimiento al respecto. Es por ello que en esta etapa de análisis se ha abordado el método MClust con la finalidad de encontrar variables presentes en vinos las cuales son medidas por medio de indicadores fisicoquímicos, pero que, sin embargo, tales variables puedan ser indicadores de una buena calidad de vino al momento de degustarlos, es decir que afectan al sabor del producto. Ante este problema, se pudo observar que variables tales como PH, acidez cítrica y nivel de alcohol (de una manera más difusa) si son indicadores para determinar si un vino tiene un buen sabor y aroma. Por otra parte, el medidor empleado para determinar calidad en laboratorios fisicoquímicos, el cual es la acidez volátil no resulto ser una variable concluyente para este dataset y el método MCluster. Cabe destacar que la variable alcohol no resultó ser un elemento decisivo para estos datos y la técnica utilizada.

Las técnicas de minería de datos pueden ser de gran ayuda para la etapa de certificacion de vino, por lo que consideramos que analizar el presente dataset por medio de otras técnicas puede aclarar en gran medida la problemática propuesta y de esta manera, descrubrir patrones y relaciones entre variables y el sentido del gusto, el cual sin duda es un elemento subjetivo al momento de evaluar un producto como lo es el vino.

4 Referencias

[1] A. Asuncion, D. Newman, UCI Machine Learning Repository, University of California, Irvine, 2007 http://www.ics.uci.edu/~mlearn/MLRepository.html.

[2] Gavin Sacks and David Jeffery, Understanding Wine Chemistry, University of California, Irvine, 2016 *http://waterhouse.ucdavis.edu/*.

[3] Cortez,P.,Teixeira,J.,Cerdeira,A.,Almeida,F.,Matos,T.,Reis,J., *Using Data Mining for Wine Quality Assessment*, Springer-Verlag Berlin Heidelberg (2009) 66-79.

[4] D. Smith, R. Margolskee, *Making sense of taste*, Scientific American, Special issue 16 (3) (2006) 84–92.

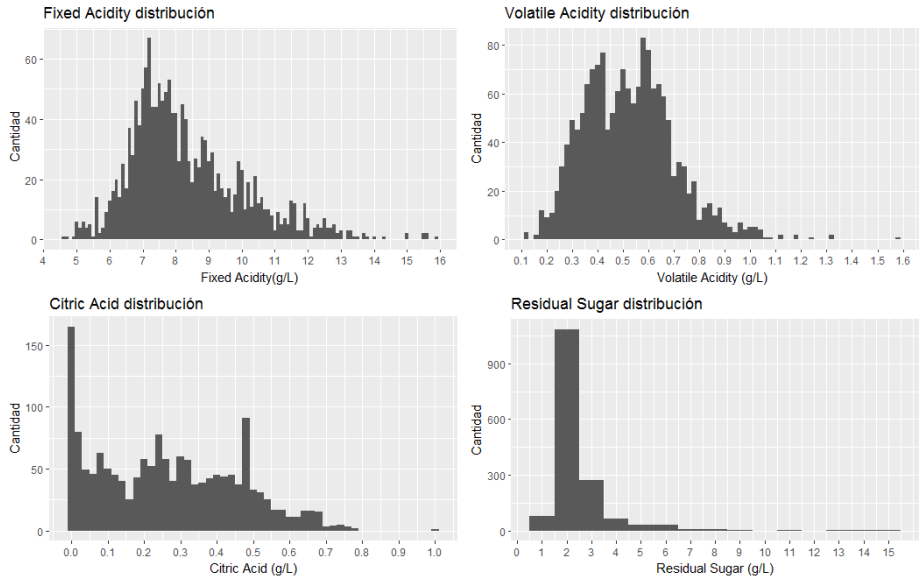
5 Anexo

5.1 Estadística Descriptiva e Inferencial:

5.1.1 Análisis Univariable:

La primera acción que realizamos sobre del dataset de “Wine Quality” correspondió a un mirada de cada característica en particular, con el principal objetivo de poder observar la forma en que se encontraban distribuidos los datos. Como ejemplo, podemos apreciar en la imagen 5 el gráfico de los primeros cuatro atributos del dataset.

Imagen 6. Distribución de cuatro primeros atributos “Wine Quality”.



Las observaciones respecto a la distribución de los datos de todos los atributos son presentadas a continuación:

Tabla 5. Distribución de atributos del dataset “Wine Quality”.

Atributo	Observación
Fixed acidity	Se distribuye como la distribución normal (presencia de outliers)
Volatile acidity	Presenta un comportamiento como una distribución normal multimodal (presencia de outliers).

Citric acid	Se puede apreciar que los datos no se distribuyen como la normal.
Residual sugar	Presenta sesgo a la derecha, llegando a alcanzar valores como 15.5 g/L que son muy superiores al tercer cuartil(2.6 g/L).
Chlorides	Presenta un rango muy pequeño, los datos se distribuyen como la normal aunque se presencian algunos outliers.
Free SO2	Los datos no se distribuyen como la normal, se presencia un sesgo hacia la derecha, siendo el valor máximo (72 mg/L) muy superior al tercer cuartil(21 mg/L).
Total SO2	Los datos no se distribuyen como la normal, se presencia un sesgo hacia la derecha.
Density	Los datos se distribuyen de forma normal dentro de un rango muy pequeño.
Sulphates	Los datos se comportan como la distribución normal (presencia de outliers).
PH	Los datos se distribuyen como la distribución normal.
Alcohol	Se puede apreciar que la distribución está sesgada a la derecha.

Resultados:

El ácido cítrico presenta un comportamiento de los datos distinto a la distribución normal.

Algunas de las variables tenían un ligero sesgo a la derecha (azúcar residual, sulfatos) pero en general el resto de los datos bastante normales. El dióxido de azufre libre, el dióxido de azufre total y el alcohol presentan un sesgo más pronunciado a la derecha.

5.2.2 Estadística Bivariable:

En este apartado tiene como finalidad analizar los resultados obtenidos a partir de la relación entre dos variables obtenidas a partir de la matriz de correlación. Para el caso de los datos que se comportan como la distribución normal usamos el coeficiente de correlación de Pearson y para el caso de datos que no presenten un comportamiento similar a la distribución normal, usamos el coeficiente de correlación de Spearman.

Resultados de matriz de correlación:

Considerando el análisis univariable que nos mostró la distribución de los datos, podemos apreciar que dos variables que se comportan como la distribución normal y además presentan una correlación lineal positiva son:

Fixed Acidity/ Density: 0.67

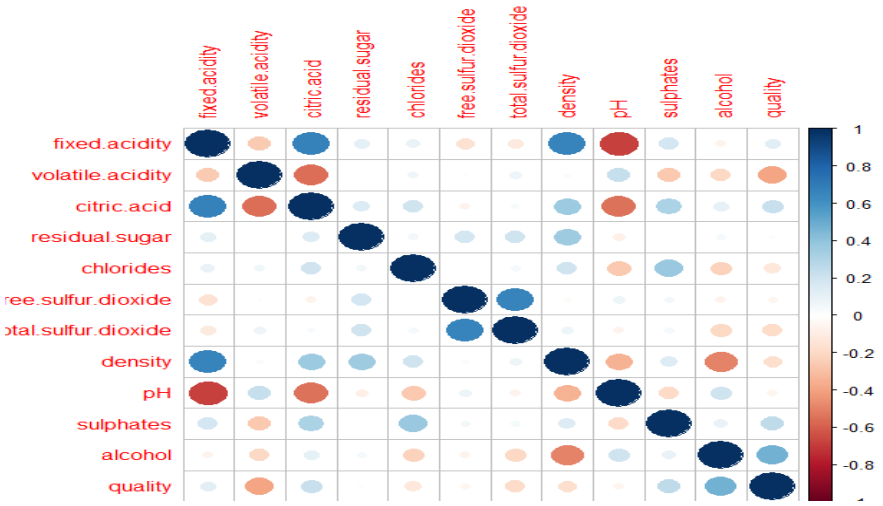
Si está información la llevamos al problema que estamos tratando, significaría que al añadir ácido tartárico provocaría un aumento en la densidad del vino.

Para el caso de correlación lineal negativa tenemos:

Fixed Acidity/PH: -0.68

Lo que nos indica este resultado es que al añadir ácido tartárico disminuye el PH, provocando un aumento en el grado de acidez del vino, lo que llevado a la información que tenemos del problema, es verdadero.

Imagen 7. Matriz de correlación de los atributos.



Los resultados obtenidos a partir de la matriz de correlación de Spearman son:

Free SO2 / Total SO2: 0.79 (correlación positiva que indica que si aumenta la cantidad de S02 libre también existirá un aumento del total de SO2).

Citric Acid/PH: -0,55(Nos indica este resultado es que al añadir ácido cítrico disminuye el PH, provocando un aumento en el grado de acidez del vino).

La variable que presenta mayor correlación con la calidad del vino es la acidez, aunque es baja (0.48).