Luceth Argote, Daniel Fonseca, Juan Sebastián Alvarado.
Ana María Beltrán Cortés.
Análisis Estadístico de Datos.
Universidad del Rosario
27 de noviembre de 2024

**Explorando el Mundo del Vino: *Técnicas de Clasificación y Clustering para su análisis*.**

## Justificación

Para este proyecto se optó por utilizar una base de datos sobre la calidad de vinos. Esta decisión se tomó por dos razones: primero, el interés que genera el tema, ya que permite observar cómo las propiedades físicas y químicas del vino ayudan a clasificarlo; segundo, este semestre uno de los integrantes del grupo se adentró en el mundo de los vinos, lo cual despertó interés por la base de datos, queriendo conocer como lo aprendido sobre vinos durante este semestre se puede comparar con los análisis aquí realizados.

## Origen de los datos.

Los dos datasets fueron resultado de la extracción de información de variantes de vino ’rojo’ y ’blanco’ del vino "*Vinho Verde*" portugués. Por motivos de privacidad se desconocen datos como la calidad de la uva, la marca, los precios de venta, etc.

## Método de recolección.

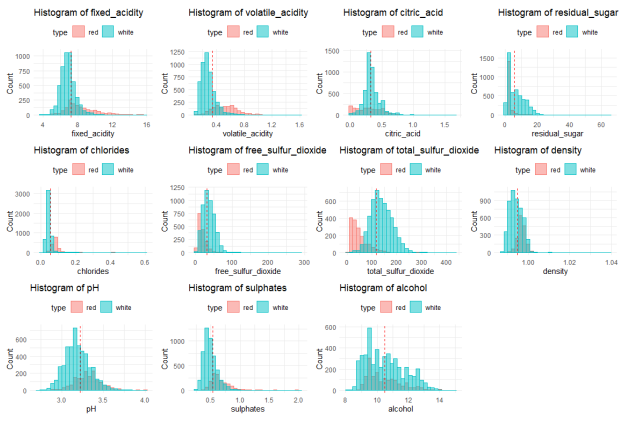
Para la recolección de los datos se hicieron pruebas de laboratorio, las cuales ayudaban a caracterizar a los vinos con cosas como su densidad, su porcentaje de alcohol, su pH, entre otras características. Estos datos, tal como lo dice en el artículo, pueden ser usados ya sea para una categorización o una regresión [1]. Para la calidad del vino, se usó la mediana de la calificación de al menos 3 expertos de vino [2].

## Reconocimiento de los datos.

En los dos datasets vamos a encontrar en conjunto 6497 registros, junto con 12 variables, 11 de las cuales consisten en pruebas físico-químicas realizadas a los vinos, y 1 que nos proporciona una calificación a la calidad del vino. El diccionario que contiene la información de las variables se encuentra anexo.

## Análisis descriptivo de las variables

Para el análisis descriptivo se crearon los histogramas de cada variable, junto con una tabla en donde se observan las medias de cada variable separadas por el tipo de vino.

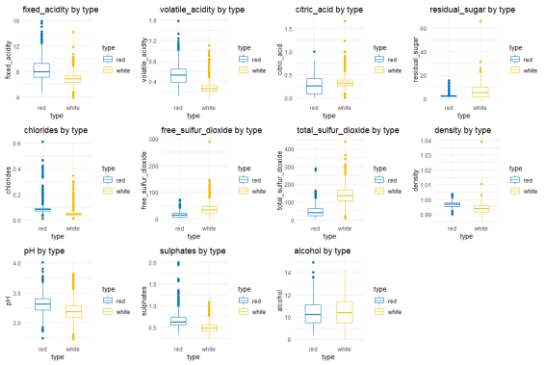


| | **red** | **White** |
|---|---|---|
| ***fixed acidity*** | 8.319637 | 6.854788 |
| ***volatile acidity*** | 0.5278205 | 0.2782411 |
| ***citric acid*** | 0.2709756 | 0.3341915 |
| ***residual sugar*** | 2.538806 | 6.391415 |
| ***chlorides*** | 0.08746654 | 0.04577236 |
| ***free sulfur dioxide*** | 15.87492 | 35.30808 |
| ***total sulfur dioxide*** | 46.46779 | 138.36066 |
| ***density*** | 0.9967467 | 0.9940274 |
| ***pH*** | 3.311113 | 3.188267 |
| ***sulphates*** | 0.6581488 | 0.4898469 |
| ***alcohol*** | 10.42298 | 10.51427 |

Para cada variable podemos evidenciar los siguientes comportamientos:

- Para *fixed_acidity* se evidencia una asimetría de cola derecha en ambos tipos de vinos; una media mayor en los vinos tintos. Los vinos tintos tienen una distribución más uniforme.
- Para la variable volatile_acidity, se observa un comportamiento similar a la variable anterior, con un ligero desplazamiento del agrupamiento de los vinos tintos hacia la derecha.
- En *citric_acidity* se puede evidenciar simetría en los vinos tintos mientras que en los vinos blancos se ve una ligera asimetría positiva. En cuanto a la media, no se puede evidenciar una diferencia clara.
- *Residual_sugar* posee asimetría positiva en ambos tipos de vinos, y no tenemos una diferencia significativa de medias entre ambos vinos.
- Para *chlorides* tenemos una asimetría fuerte en ambos tipos de vinos, y una media ligeramente mayor para los vinos tintos.
- Para las variables relacionadas con el dióxido de azufre, se evidencia una asimetría positiva en los vinos tintos y una media significativamente mayor en los vinos blancos.
- En *density* no se evidencia una asimetría en los datos y tampoco una diferencia entre la media de los tipos de vinos.
- El *pH* de los vinos parece ser en promedio el mismo y no parece haber una asimetría marcada.
- En *sulphate* podemos ver una media mayor para los vinos tintos y una asimetría positiva en ambos casos.
- En cuanto a la cantidad de alcohol en los vinos, en promedio, parece ser la misma. Ambos tipos de vinos tienen una asimetría positiva.

A continuación, se realizaron los boxplots de cada variable segmentados por cada tipo de vino.



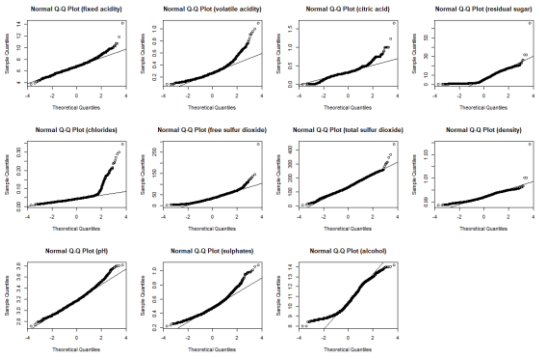

Con el uso de boxplots se puede evidenciar una cantidad considerable de outliers, lo cual a futuro puede ser problemático para los algoritmos que serán aplicados. Esto se puede deber a que los datos están muy agrupados respecto a la mediana (cajas pequeñas), pero la asimetría positiva es muy marcada. Respecto a las medias, se observa un comportamiento similar al ya descrito en los histogramas.

**Análisis de normalidad univariado**

Para las pruebas de normalidad univariadas, realizamos la prueba de Anderson-Darling, y obtenemos los Q-Q plots.

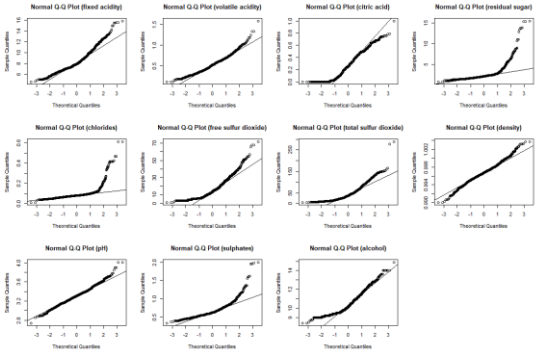
Estos son los resultados obtenidos para el vino blanco:

| | Test <S3: Asis> | Variable <S3: Asis> | Statistic <S3: Asis> | p value <S3: Asis> | Normality <S3: Asis> |
|---|---|---|---|---|---|
| 1 | Anderson-Darling | fixed acidity | 21.8347 | <0.001 | NO |
| 2 | Anderson-Darling | volatile acidity | 83.8841 | <0.001 | NO |
| 3 | Anderson-Darling | citric acid | 89.9292 | <0.001 | NO |
| 4 | Anderson-Darling | residual sugar | 160.9273 | <0.001 | NO |
| 5 | Anderson-Darling | chlorides | 406.7420 | <0.001 | NO |
| 6 | Anderson-Darling | free sulfur dioxide | 21.9327 | <0.001 | NO |
| 7 | Anderson-Darling | total sulfur dioxide | 11.9390 | <0.001 | NO |
| 8 | Anderson-Darling | density | 22.6354 | <0.001 | NO |
| 9 | Anderson-Darling | pH | 11.9717 | <0.001 | NO |
| 10 | Anderson-Darling | sulphates | 50.2427 | <0.001 | NO |




Y los siguientes son los resultados obtenidos para el vino tinto:

| | Test <S3: Asis> | Variable <S3: Asis> | Statistic <S3: Asis> | p value <S3: Asis> | Normality <S3: Asis> |
|---|---|---|---|---|---|
| 1 | Anderson-Darling | fixed acidity | 28.1430 | <0.001 | NO |
| 2 | Anderson-Darling | volatile acidity | 5.6831 | <0.001 | NO |
| 3 | Anderson-Darling | citric acid | 17.5421 | <0.001 | NO |
| 4 | Anderson-Darling | residual sugar | 188.0644 | <0.001 | NO |
| 5 | Anderson-Darling | chlorides | 210.4492 | <0.001 | NO |
| 6 | Anderson-Darling | free sulfur dioxide | 38.6099 | <0.001 | NO |
| 7 | Anderson-Darling | total sulfur dioxide | 52.4887 | <0.001 | NO |
| 8 | Anderson-Darling | density | 3.8676 | <0.001 | NO |
| 9 | Anderson-Darling | pH | 1.8641 | 1e-04 | NO |
| 10 | Anderson-Darling | sulphates | 46.9322 | <0.001 | NO |




Se usaron las pruebas de Anderson-Darling dado que tenemos más de 5000 datos, y la función que realiza el test de Shapiro en R no permite realizar pruebas sobre variables con una cantidad mayor a la mencionada. Los resultados de la prueba arrojaron que las variables tienen p-valores menores a 0.001, lo cual rechaza la normalidad de las variables individualmente. Adicionalmente, en la mayoría de variables se observa la influencia de los outliers en los gráficos

Q-Q, dado que las colas derechas se desvían notoriamente. Únicamente tenemos desviación en las colas de la izquierda en la variable alcohol.

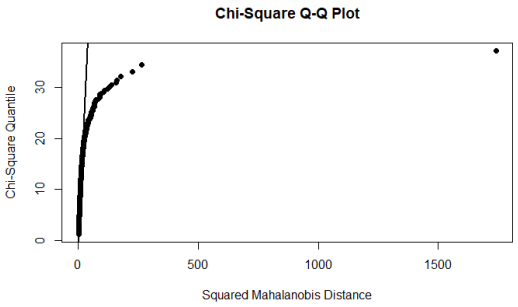
**Análisis de normalidad multivariado**

Se realizaron las pruebas de Mardia y HZ para probar la normalidad multivariada.

Se obtuvieron los siguientes resultados para el vino blanco:

| Test | Statistic | p value | Result |
|---|---|---|---|
| Mardia Skewness | 235569.169372725 | 0 | NO |
| Mardia Kurtosis | 1461.69831443977 | 0 | NO |
| MVN | NA | NA | NO |

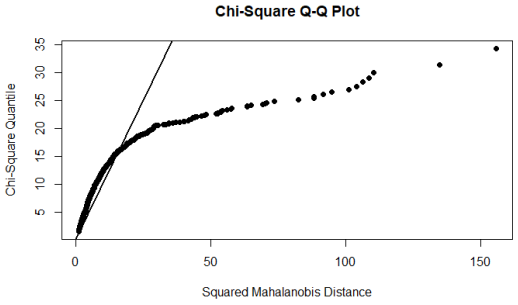
| Test | HZ | p value | MVN |
|---|---|---|---|
| Henze-Zirkler | 4.380657 | 0 | NO |




Y se obtuvieron los siguientes resultados para el vino tinto:

| Test | Statistic | p value | Result |
|---|---|---|---|
| Mardia Skewness | 23791.1436242531 | 0 | NO |
| Mardia Kurtosis | 161.631091429313 | 0 | NO |

| Test | HZ | p value | MVN |
|---|---|---|---|
| Henze-Zirkler | 4.575283 | 0 | NO |



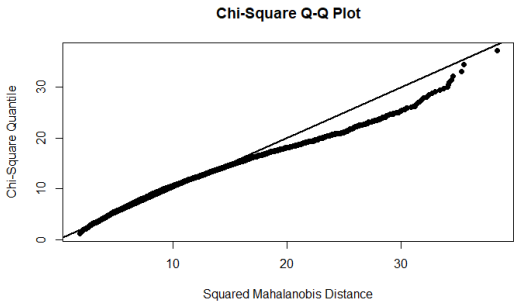

Al hacer pruebas de hipótesis sobre normalidad multivariada se obtuvieron valores de p-value iguales a 0, lo cual rechaza la normalidad multivariada tanto para el vino blanco como para el tinto, al revisar el gráfico Q-Q del vino blanco se observa que la cola derecha tiene una gran desviación, similarmente al QQ del vino tinto, lo que nos deja ver que la asimetría tiene una gran influencia sobre los datos.

Acto seguido, se realiza una transformación Box-Cox, que fue la que arrojó el mejor rendimiento a la hora de suavizar la asimetría. Después de esta transformación se le eliminaron los datos atípicos, para intentar obtener la menor asimetría posible.

Estos fueron los resultados para el vino blanco:

| Test | Statistic | p value | Result |
|---|---|---|---|
| Mardia Skewness | 6466.34591793927 | 0 | NO |
| Mardia Kurtosis | 20.5555313264803 | 0 | NO |

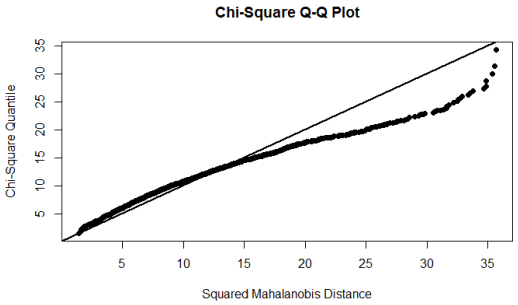
| Test | HZ | p value | MVN |
|---|---|---|---|
| Henze-Zirkler | 3.12227 | 0 | NO |




y los siguientes son los resultados para el vino tinto:

| Test | Statistic | p value | Result |
|---|---|---|---|
| Mardia Skewness | 3749.97820915015 | 0 | NO |
| Mardia Kurtosis | 21.0813865078339 | 0 | NO |

| Test | HZ | p value | MVN |
|---|---|---|---|
| Henze-Zirkler | 2.323498 | 0 | NO |




Aunque se puede ver una mejor significativa en el gráfico QQ, la asimetría derecha sigue siendo demasiado notoria, por lo que ambas pruebas siguen arrojando un p-valor igual a 0.

Dado que no fue posible obtener normalidad, se trabajará con el dataset original, con el fin de no perder la interpretabilidad de los datos.

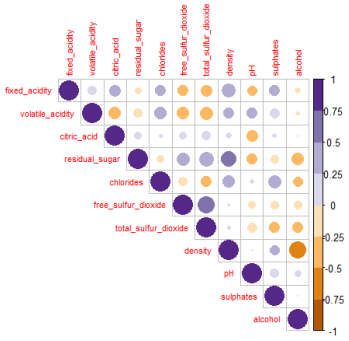
## Análisis de componentes principales (PCA)

Se realiza una prueba de Barlett para verificar si existe correlación de las variables y se revisa la matriz de correlación:

```
$chisq
[1] 35420.32

$p.value
[1] 0

$df
[1] 55
```

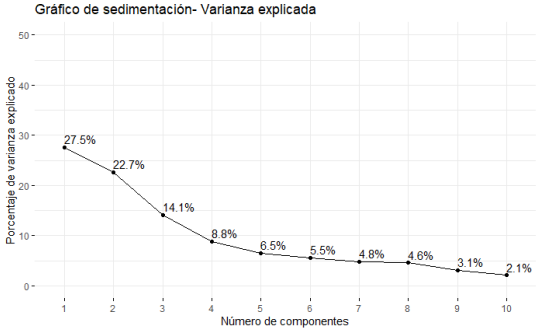


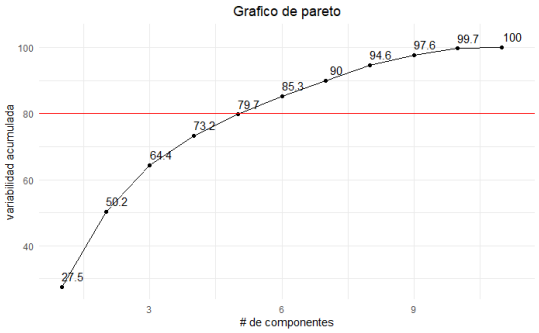

Como la prueba de Barlett da un p valor menor a 0.05, se rechaza la prueba, lo que significa que, si hay correlación significativa en las variables, cosa que se puede comprobar con la matriz de correlación, dado que se puede observar una cantidad modesta de correlaciones altas; se procede a verificar el KMO de los datos:

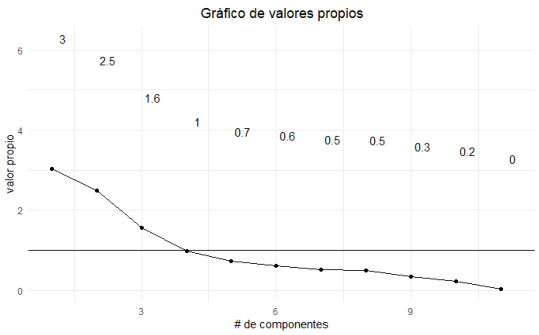
```
Kaiser-Meyer-Olkin factor adequacy
Call: KMO(r = wine_num)
Overall MSA =  0.42
MSA for each item =
       fixed_acidity     volatile_acidity          citric_acid       residual_sugar
                0.27                 0.67                 0.68                 0.29
           chlorides  free_sulfur_dioxide total_sulfur_dioxide              density
                0.70                 0.77                 0.76                 0.32
                  pH            sulphates              alcohol
                0.22                 0.51                 0.27
```

Aunque el KMO es alto en algunas variables y en otras es mediocre, se realizará el PCA, ya que los otros indicadores dejan ver que es posible reducir la dimensionalidad.

Al hacer graficas de sedimentación (*elbow*), de Pareto y valores propios se observa:

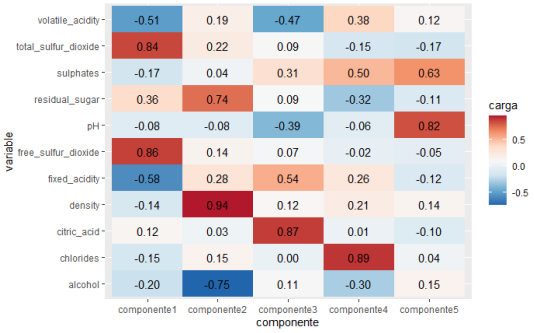





Se puede ver que en la gráfica de elbow el quiebre sucede en la 5 componente. En Pareto también se evidencia que en la 5 componente se alcanza casi el 80%. En la gráfica de valores propios el corte sucede en la 4 componente. Dados estos criterios, se eligió trabajar con 5 componentes. De igual forma se refuerza que el KMO no era tan significativo, ya que se resume más del 80% de la información con un número relativamente bajo de componentes.

Al hacer el análisis por componente y rotando la solución, usando la rotación *varimax*, se obtiene lo siguiente:




Junto con el valor de las comunalidades:

```
fixed_acidity    volatile_acidity         citric_acid      residual_sugar           chlorides
    0.7903418           0.6811080           0.7821311           0.7933024           0.8331696
free_sulfur_dioxide total_sulfur_dioxide             density                  pH           sulphates
    0.7634089           0.8169807           0.9775418           0.8429119           0.7730583
           alcohol
    0.7165441
```

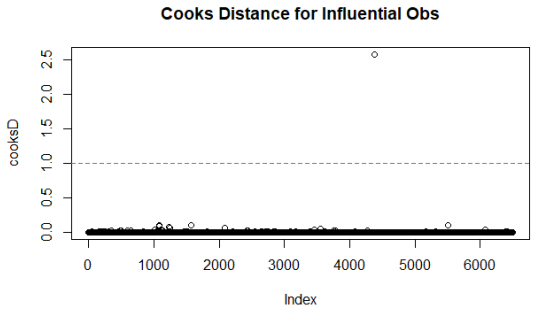
De aquí se puede ver que cada variable quedó bien representada en la solución factorial. Se evidencia que las variables que tuvieron mayor carga factorial sobre la primera componente son free_sulfur_dioxide y total_sulfur_dioxide lo que hace que la primera componente represente la oxidación del vino. Sobre la segunda componente fueron density, residual_sugar y alcohol, lo que representan la forma y el cuerpo del vino. Sobre la tercera fue citric_acid, lo que representa la acidez del vino. Sobre la cuarta fue chlorides, lo que representa la cantidad de sal en el vino. Y finalmente, sobre la quinta componente fueron el pH y sulphates, que representan la apariencia del vino.

**Regresión Logística.**

Al no tener supuesto de normalidad se implementó un modelo de Regresión Logística para hacer la clasificación de los vinos.

Previo a la creación final del modelo, se estandarizó el dataset y se convirtió la columna *type* a 0 y 1, 0 si el vino es blanco y 1 si es tinto. Se usó el dataset completo para entrenar el modelo y observar si se presentaban falencias en los supuestos, y se obtuvieron los siguientes resultados para las pruebas VIF y la distancia de Cook:

| | | |
|---|---|---|
| fixed_acidity | volatile_acidity | citric_acid |
| 3.745162 | 1.586839 | 1.625308 |
| residual_sugar | chlorides | free_sulfur_dioxide |
| 2.343361 | 1.447304 | 2.097711 |
| total_sulfur_dioxide | density | pH |
| 2.021329 | 10.240818 | 2.700131 |
| sulphates | alcohol | |
| 1.357355 | 5.089580 | |




Como se puede observar, la variable *density* posee un VIF mayor a 10, que según la regla [3], representa una colinealidad severa, por lo que se eliminará esa columna, junto con el dato que supera el umbral de 1 usado para la distancia de Cook, ya que puede afectar negativamente el modelo.

Posterior a esto se partió el dataset en *train* y *test* con la ley de Pareto, y se creó el modelo.

Los valores de los coeficientes fueron los siguientes:

| | | |
|---|---|---|
| (Intercept) | fixed_acidity | volatile_acidity |
| -4.1859100 | 1.8253259 | 1.8904148 |
| citric_acid | residual_sugar | chlorides |
| -0.1540146 | -0.5226947 | 1.2458997 |
| free_sulfur_dioxide | total_sulfur_dioxide | pH |
| 0.9946641 | -3.7783293 | 1.5113484 |
| sulphates | alcohol | |
| 1.3090202 | -0.4675617 | |

Se obtuvo que las variables *fixed_acidity*, *volatile_acidity*, *chlorides*, *free_sulfur_dioxide*, *pH y sulphates* al tener un valor positivo en sus coeficientes representan que a mayor valor de estas, se generan mayores probabilidades de que el vino sea vino tinto. Las variables *citric_acid*, *residual_sugar*, *total_sulfur_dioxide* y *alcohol* al tener valores negativos implican lo contrario, que a mayor valor de estas, mayores probabilidades hay de que el vino sea blanco.

Las probabilidades finales fueron las siguientes:

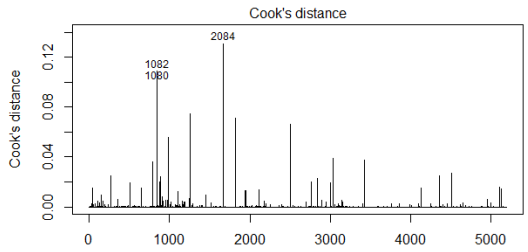
| | | |
|---|---|---|
| (Intercept) | fixed_acidity | volatile_acidity |
| 0.004940389 | 0.669489562 | 0.683729891 |
| citric_acid | residual_sugar | chlorides |
| 0.218664996 | 0.162173643 | 0.531570726 |
| free_sulfur_dioxide | total_sulfur_dioxide | pH |
| 0.468844299 | 0.007407863 | 0.596739969 |
| sulphates | alcohol | |
| 0.547251490 | 0.169804956 | |

Las variables que tienen una mayor probabilidad de ser vino tinto al aumentar 1 unidad (en sus respectivas unidades) de ellas son: *volatile_acidity* (68.37%), *fixed_acidity* (66.94%) y pH (59.67%), y las variables que tienen una menor probabilidad son:

*total_sulfur_dioxide* (0.74%), *residual_sugar* (16.21%) y *alcohol* (16.98%).

Se obtuvo un 1.231% de error en clasificación (*APER*) con el modelo implementado, lo que sugiere que es un buen modelo.

Posteriormente, se analizan el cumplimiento de supuestos:




```
       fixed_acidity     volatile_acidity          citric_acid
            1.770278             1.677962             1.736255
      residual_sugar            chlorides  free_sulfur_dioxide
            1.135566             1.437715             1.821000
total_sulfur_dioxide                   pH            sulphates
            2.491303             1.692084             1.289073
             alcohol
            1.486675
```

Podemos ver que no hay distancia de Cook tan desproporcionadas, y que los VIF's obtenidos son bastante bajos, por lo que se puede decir que el modelo cumple con los supuestos.

Adicionalmente, se usó la técnica de K-fold Cross Validation con *k=10* para comprobar la veracidad del APER obtenido, y se obtuvo el siguiente resultado:

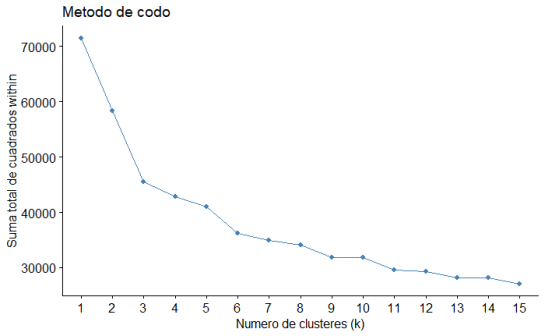
```
[1] "APER promedio: 0.0113"
```

Por tanto, verificamos que el modelo es bueno clasificando el tipo de vinos.

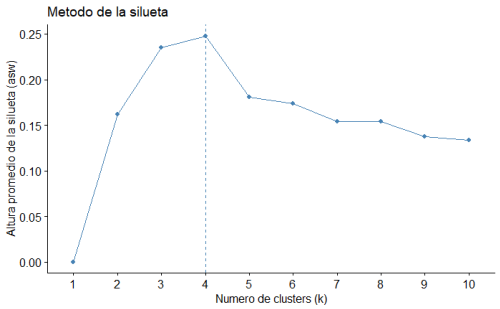
**Clustering.**

Debido a la gran cantidad de datos, no es recomendable ejecutar un algoritmo de clustering jerárquico, por lo cual, únicamente se hará el algoritmo de k-means.

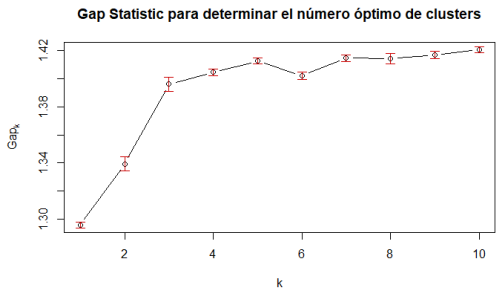
Primero analizaremos en número óptimo de clusters con distintos métodos.




El grafico de codo sugiere que el número óptimo de clusters se encuentra entre los 4 y los 6. Sin embargo, visualmente el análisis puede ser engañoso. Dado esto se decide a usar el método de la silueta para validar cual es la cantidad optima:



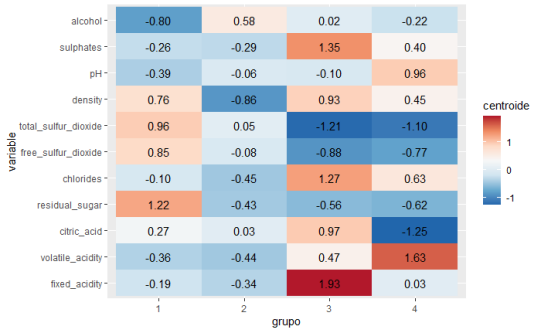

Esta medida indica que 4 clusters sería lo óptimo. Para confirmar, se ejecuta una tercera medida (GAP statistic), la cual también confirma que 4 clusters son lo óptimo.




Al realizarse el clustering, se analizan los tamaños de cada clúster. De esta información se puede ver una buena distribución en los clusters, un posible problema puede ser el cluster con solo 687 muestras, ya que, comparado con el número de observaciones, este cluster es pequeño.

```
[1] 1873 2930  687 1007
```

Se procede a hacer un análisis del factor latente de cada clúster, esto con el objetivo de identificar como se pueden agrupar diversos vinos.




Podemos identificar que en el primer cluster se encuentran los vinos que en promedio tienen menos cantidad de alcohol, dióxidos de sulfuro altos, azúcar residual y densidad alta, esto sugiere que se agrupan vinos más dulces. En el segundo cluster vemos vinos que en promedio poseen densidades más bajas y alcoholes poco más altos, lo cual se relacionaría con vinos ligeros, los cuales suelen tener bastantes características promedio. En el tercer cluster encontramos vinos con sulfatos, cloruros y densidades altas, dióxidos de sulfuro bajos y una acidez fija muy alta, este tipo de características se asocian con vinos blancos, más centrados en el sabor que en el alcohol presente. Por último, el cuarto clúster agrupa vinos con acidez volátil alta, acidez cítrica, dióxidos de sulfuro bajos, azúcar residual baja y pH alto, características de vinos más centrados en el aroma y en su estructura. Cabe recordar que al mencionar alto o bajo es respecto a la media general, dado que los datos están escalados.

Se usó la función *clusterboot* para verificar la calidad de los clusters, obteniendo los siguientes resultados:

```
[1] 0.5643055 0.7542696 0.8806795 0.6822231
```

Podemos ver que 3 los 4 clusters son buenos, mientras que solo uno de ellos es menor a 0.6, por lo cual es puede considerar como un cluster malo.

**Regresión Logística sobre los clusters como variable predictora.**

Previo a la creación del modelo, se añadió la columna cluster al dataset que usado previamente para la regresión logística, y se partió también usando la ley de Pareto.

Los valores de los coeficientes otorgados por el modelo fueron los siguientes:

| (Intercept) | fixed_acidity | volatile_acidity |
|---|---|---|
| -3.5254332 | 2.0285415 | 1.9120685 |
| citric_acid | residual_sugar | chlorides |
| -0.1725681 | -0.6139562 | 1.3033248 |
| free_sulfur_dioxide | total_sulfur_dioxide | pH |
| 1.1129206 | -3.8411428 | 1.5856950 |
| sulphates | alcohol | cluster |
| 1.2641931 | -0.2893095 | -0.2484405 |

Se obtuvo que las variables *fixed_acidity, volatile_acidity, chlorides, free_sulfur_dioxide, pH y sulphates* al tener un valor positivo en sus coeficientes representan que a mayor valor de estas, se generan mayores probabilidades de que el vino sea vino tinto. Las variables *citric_acid, residual_sugar, total_sulfur_dioxide* y *alcohol* al tener valores negativos implican lo contrario, que a mayor valor de estas, mayores probabilidades hay de que el vino sea blanco. Como el coeficiente de cluster es negativo, quiere decir que si cambia de cluster, la probabilidad de que sea vino tinto va a disminuir.

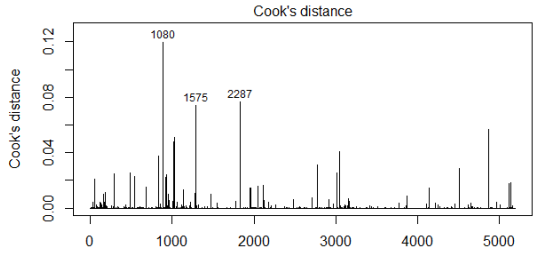
Las probabilidades finales fueron las siguientes:

| (Intercept) | fixed_acidity | volatile_acidity |
|---|---|---|
| 0.009519181 | 0.712814526 | 0.688393621 |
| citric_acid | residual_sugar | chlorides |
| 0.215511678 | 0.150152596 | 0.545839992 |
| free_sulfur_dioxide | total_sulfur_dioxide | pH |
| 0.498368022 | 0.006960001 | 0.614494961 |
| sulphates | alcohol | cluster |
| 0.536123099 | 0.196430027 | 0.202961114 |

Las variables que tienen una mayor probabilidad de ser vino tinto al aumentar 1 unidad (en sus respectivas unidades) de ellas son: *volatile_acidity* (68.83%), *fixed_acidity* (71.28%) y pH (61.44%), y las variables que tienen una menor probabilidad son: *total_sulfur_dioxide* (0.69%), *residual_sugar* (15.01%) y *alcohol* (19.64%). La variable cluster obtuvo un 20.29% de probabilidad de ser vino tinto si se cambia de un cluster a otro.

Se obtuvo un 1.308% de error en clasificación (*APER*) con el modelo implementado, lo que sugiere que es un buen modelo.

Posteriormente, se analizan el cumplimiento de supuestos:




| fixed_acidity | volatile_acidity | citric_acid |
|---|---|---|
| 1.973431 | 1.815913 | 1.830606 |
| residual_sugar | chlorides | free_sulfur_dioxide |
| 1.392447 | 1.703079 | 1.939732 |
| total_sulfur_dioxide | pH | sulphates |
| 2.615033 | 1.848729 | 1.360519 |
| alcohol | cluster | |
| 1.765747 | 1.904788 | |

Podemos ver que no hay distancia de Cook tan desproporcionadas, y que los VIF's obtenidos son bastante bajos, por lo que se puede decir que el modelo cumple con los supuestos.

Adicionalmente, se usó la técnica de K-fold Cross Validation con *k=10* para comprobar la veracidad del APER obtenido, y se obtuvo el siguiente resultado:

```
[1] "APER promedio: 0.0114"
```

Por tanto, verificamos que el modelo es bueno clasificando el tipo de vinos.

Respecto al clasificador anterior, en general se observa que no influye mucho añadir la columna de cluster, dado que el APER es muy similar y las probabilidades no varían significativamente.

**Consulta Bibliográfica**

El trabajo realizado en [2] se acerca al mismo problema abordado en el presente, que es clasificar el vino en blanco o tinto según sus características físico-químicas. En primera medida, usan dos métodos para partir el dataset, el primero es llamado *k-fold cross validation*, que consiste en partir el dataset en k bloques, con el fin de entrenar el modelo con k-1 bloques y testear con el bloque sobrante. Esto se repite k veces y el *accuracy* obtenido resulta del promedio de las 10 repeticiones de entrenamiento y testeo. El otro método usado es el *percentage Split*, que consiste en dividir la base de datos en dos bloques, cada uno con cierto porcentaje de la información. En el caso de este estudio, se usó *k=10* y la partición fue según la ley de Pareto (80% de train y 20% de test). Posterior a esto, usaron 3 modelos de *machine learning* (*KNN*, *Random Forests* y *Support Vector Machines*) para clasificar los vinos. El que tuvo mejor rendimiento fue el modelo *de Random Forests,* con un *accuracy* de 99.5229% en *cross validation* y 99.4611% en *percentage split.* También implementaron un clasificador para la calidad del vino, usando los 3 modelos y las particiones mencionadas previamente.

**Conclusión.**

El objetivo propuesto para este trabajo fue conocer cuales son las variables físico-químicas que nos ayudan a distinguir un vino tinto de uno blanco, y por medio del dataset usado con información de vinos de Portugal, obtuvimos resultados por con el modelo de regresión logística, que sugiere que las variables *fixed_acidity, volatile_acidity, chlorides, free_sulfur_dioxide, pH y sulphates* están relacionados con los vinos tintos, *mientras que* las variables *citric_acid, residual_sugar, total_sulfur_dioxide* están relacionados con vinos blancos. Además, se concluye que agrupar los vinos en clusters no marca una gran diferencia a la hora de clasificar los vinos. Ambos modelos tienen un gran rendimiento, con errores de clasificación menores al 1.5%. Respecto al trabajo consultado, los rendimientos de ambos modelos son muy similares, aunque el trabajo no concluye respecto a las variables que nos permiten distinguir los vinos en la clasificación.

**Referencias**


[1] P. Cortez, A. Cerdeira, F. Almeida, T. Matos y J. Reis, «Decision Support Systems,» 1 Noviembre 2009. [En línea]. Available: https://www.semanticscholar.org/paper/Modeling-wine-preferences-by-data-mining-from-Cortez-Cerdeira/bf15a0ccc14ac1deb5cea570c870389c16be019c

[2] Y. Er y A. Atasoy, «International Journal of Intelligent Systems and Applications in Engineering,» Diciembre 2016. [En línea]. Available: https://dergipark.org.tr/en/pub/ijisae/article/265954

[3] «Variance inflation factor,» Wikipedia, [En línea]. Available: https://en.wikipedia.org/wiki/Variance_inflation_factor#cite_note-5