**¿EXISTE ALGÚN COMPONENTE O COMPONENTES QUE CON UNA CANTIDAD CONCRETA DETERMINEN LA CALIDAD DE UN BUEN VINO?**

ANA CERRO GARCÍA Y ELENA SANTAMARÍA IZQUIERDO 3º GID MINERÍA DE DATOS

La calidad es aquella condición del producto ya realizado, la cual nos indica que tan bueno o malo puede ser[[1]](#endnote-1). Un vino de calidad es difícil de identificar, ya que no depende únicamente de elementos objetivos, también entra en juego una parte subjetiva, que es nuestro propio criterio. A pesar de esto, un buen vino debe de cumplir con las siguientes características; equilibrio (compone los 4 elementos esenciales: el tanino, la acidez, el alcohol y el azúcar), la profundidad de los elementos esenciales, longitud (hace referencia a que sabe a lo largo de todo el paladar), carácter (posee características propias y distintivas de su categoría), y complejidad (cantidad de matices que presenta)[[2]](#endnote-2). En este estudio nos centraremos en la parte objetiva, teniendo en cuenta que la calidad del vino es considerada buena si tiene una calificación de 7 o por encima, hemos estudiado los diferentes componentes del vino, analizándolos y combinándolos para intentar encontrar el componente o los componentes que determinan que la calidad de un vino tinto sea buena y en qué cantidad. Tras realizar el estudio, vimos que el pH, el azúcar residual y el dióxido de azufre libre no explican bien la calidad, pero no podemos eliminarlas, ya que aportan información valiosa si las con otras variables, y a su vez con la calidad. No hemos encontrado un componente que determine la calidad de un buen vino por sí solo, de hecho, probando el efecto de varios componentes con la calidad, y el mejor modelo que hemos encontrado ha sido el que relacionaba el alcohol y la densidad con la calidad, con una gran representatividad y una correlación del 50%. Obteniendo una precisión de 0,86 y una potencia de 0,89. La profundidad de estos elementos debería de ser de más de 12% de alcohol y una densidad mayor que 1 g /cm^3. La distribución de los errores del modelo global se asemeja a una normal, no es perfecta, pero es bastante buena en comparación con las demás realizadas

A pesar de que estos componentes nos indiquen cómo se puede mejorar en parte la calidad de un vino, el resultado que obtuvimos no fue satisfactorio, ya que no se puede asegurar que usando estos componentes con las cantidades adecuadas se encuentre un vino de calidad

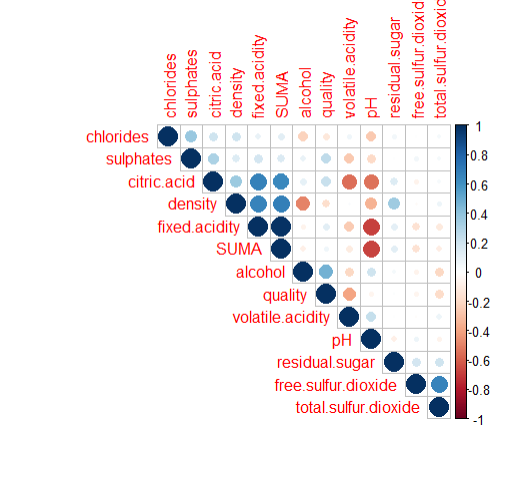
Como hemos dicho anteriormente, los vinos bien hechos (de calidad), son aquellos en los que se logra la armonía de los cuatro componentes y no destaca ninguno por encima de otro. A continuación, vamos a ir realizando diferentes pruebas con las variables de nuestra base de datos más representativas para la calidad, contrastando la información que hemos encontrado en internet con los resultados obtenidos de nuestro estudio e intentaremos responder a nuestra pregunta de la manera más concreta posible.

Antes de empezar explicando el trabajo, vamos a presentar las variables que hemos usado, con sus medidas[[3]](#endnote-3).

Medido en gramos/litro

1. Fixed acidity
2. Volatile acidity
3. Citric acid
4. Residual sugar
5. Chlorides
6. Free sulfur dioxide
7. Total sulfur dioxide
8. Density
9. Ph
10. Sulphates
11. Alcohol (% by volume) output variable (based on sensory data):
12. Quality (score between 0 and 10)

Empezamos el trabajo haciendo cosas simples, para, posteriormente, ir incorporando parte del material nuevo que hemos ido dando en clase.

Lo primero que hicimos fue mirar la variabilidad de nuestros datos haciendo un boxplot de la calidad con cada una de las variables.

Tras hacer esto vimos que los componentes más sensibles a la calidad fueron; acidez volátil, acidez cítrica, sulfatos y alcohol.

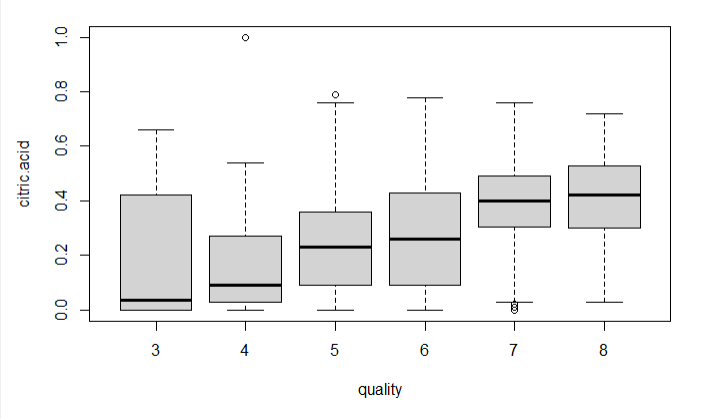
Otro gráfico que nos pareció muy útil hacerlo al inicio, fue el gráfico que mostraba las correlaciones entre todas las variables.

Ilustración 1: correlación de las variables de la base de datos

Gracias a esto, pudimos distinguir de forma muy clara y simple, qué variables podíamos usar posteriormente para realizar un análisis de componentes principales, que veremos más adelante.

***ESTUDIO DE ACIDEZ***

Los primeros componentes que decidimos analizar fueron la calidad con la acidez volátil y la acidez cítrica, elegimos estudiar estas dos últimas variables juntas ya que, según la información que encontramos en un informe acerca del vino tinto, están relacionadas entre sí y con la variable acidez fija, aunque esta última no presenta mucha variabilidad con la calidad, aun así, la usaremos más adelante.

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamenteEn nuestra base de datos podemos encontrar 3 variables para la acidez: acidez total, volátil y la cítrica.

Ilustración 2: boxplot calidad y ácido cítrico

Ilustración 3: boxplot calidad y ácido volátil

Resumidamente, lo que encontramos en el estudio sobre estas variables, es que miden la cantidad de ácido que hay en el vino. El pH lo que hace es medir lo fuerte o débil que es el ácido total, es con lo que se mide la acidez y existe una correlación inversa entre ellos. A mayor pH menor cantidad de ácido total[[4]](#endnote-4).

Teniendo en cuenta esto, nos vamos a centrar primero en el ácido cítrico y volátil, ya que son las variables que presentan mayor variabilidad respecto la calidad, más adelante usaremos el pH y la acidez fija para comprobar si lo que dice el estudio, se cumple para nuestra base de datos.

Primero, realizamos 3 modelos, estos sirven para ver la representatividad de nuestras variables estudiadas (ácido cítrico y ácido volátil) con la calidad, aunque, como hemos visto en clase, el p\_valor no es la mejor forma para medir la representatividad. Los modelos realizados son los siguientes:

1. MODELO GLOBAL (incluye los dos tipos de ácidos)
2. MODELO SIMPLE DE ÁCIDO CÍTRICO
3. MODELO SIMPLE DE ACIDEZ VOLÁTIL

El ácido cítrico no parece ser representativo en el modelo global, sin embargo, en los modelos simples ambos componentes son muy representativos. Por lo que puede que la acidez cítrica afecte a la calidad de manera indirecta

Tras realizar estos modelos, miramos la correlación. A pesar de que hay bastante variabilidad entre calidad y ácido cítrico y según el modelo parece ser una variable representativa para la calidad, tienen una correlación positiva y baja, de 0.23. En cambio, la correlación entre calidad y acidez volátil es negativa, pero algo más alta que la anterior, explicando el 40% de la calidad. Y el modelo que engloba las 3 variables tiene una correlación aún peor que las anteriores, de -0.15. Por lo tanto, fijándonos en esto, no son variables muy importantes a la hora de medir la calidad.

Como los gráficos nos ayudan a dar visibilidad a los cálculos que vamos haciendo, intentamos hacer uno para ver la correlación que habíamos calculado, pero, como era de esperar este gráfico no nos dice nada relevante, ya que, la calidad es una variable discreta (0 a 10). Ante esto, se nos ocurrió que si el ácido cambiaba bruscamente cuando la calidad del vino es buena podríamos graficarlo si hacíamos un gráfico logarítmico. Para esto, tuvimos que pasar la calidad a una variable binomial; el 0 contiene los números del 0 al 6 (mala/media calidad) y el 1 del 7 al 10 (buena calidad). Hicimos el gráfico logarítmico, pero no vimos ninguna diferencia en los ácidos cuando la calidad era buena, que cuando no lo era, por lo que no hay manera de hacer un único gráfico que contenga la calidad y estos ácidos con los métodos aprendidos.

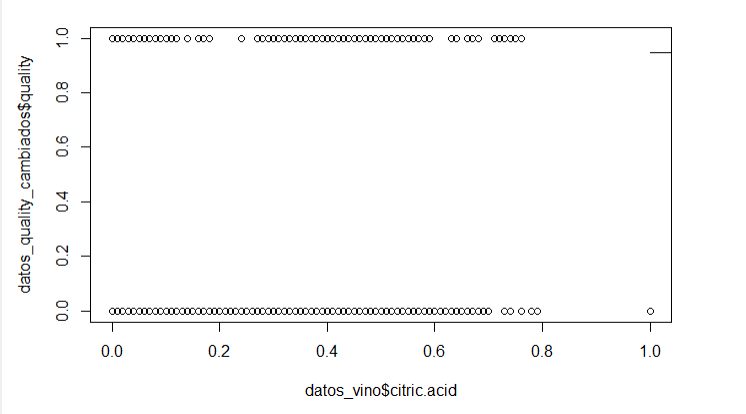


Ilustración 4: plot calidad y ácido cítrico

Sin embargo, si hay mucha correlación, (como podemos observar en el gráfico inicial que hemos mostrado las correlaciones de todas las variables), entre la acidez volátil y la cítrica, podemos aplicar la técnica de componentes principales, de esta manera si una variable explica en gran porcentaje a la otra, es que contiene información redundante y podríamos optar por omitir una de ellas de nuestro estudio.

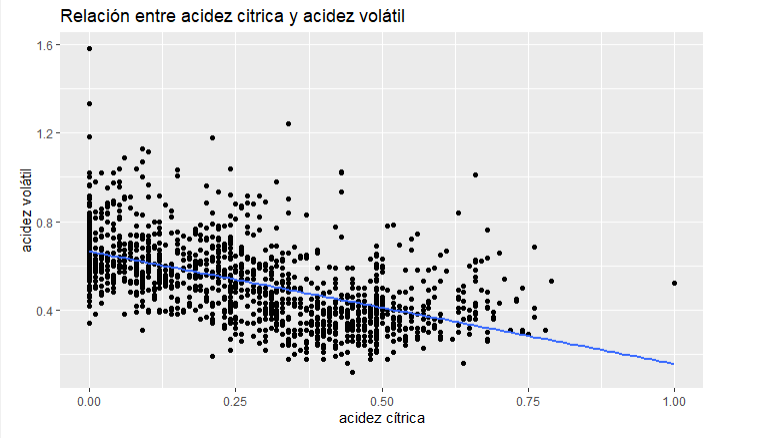


Ilustración 5: plot acidez volátil y acidez cítrica

La correlación es de 0.55, no es una correlación muy grande desde nuestro punto de vista, pero aun así hemos decidido aplicar este método para salir de dudas.

Creamos una matriz que contenía los datos de ambas variables y calculamos sus autovalores y autovectores, pero uno de los autovalores nos dio 0, no sabíamos que significaba esto por lo cual nos pusimos a investigar un poco sobre esto y encontramos que cuando un autovalor es igual a 0 significa que la matriz no tiene inversa, por lo cual no se puede continuar con este método.[[5]](#endnote-5)

A continuación, graficamos las distribuciones de los errores de los modelos, usando una parte de los datos como entrenamiento (80%) y el resto como para testear (20%). Al repetir esta secuencia muchas veces de manera aleatoria, nos saldrán las distribuciones de los errores de los modelos que estamos estudiando. Lo que buscamos con esto, es que, al hacer el gráfico, la mayor parte de los errores se encuentren cerca del 0, esto quiere decir que nuestro modelo se equivoca muy poco y, por lo tanto, es buen modelo para medir la calidad del vino.

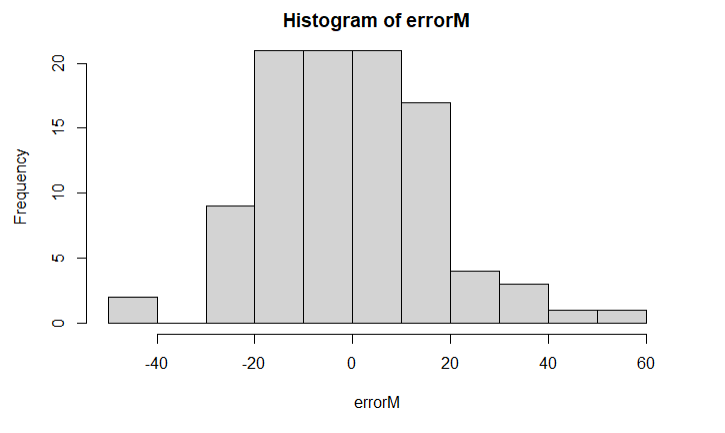
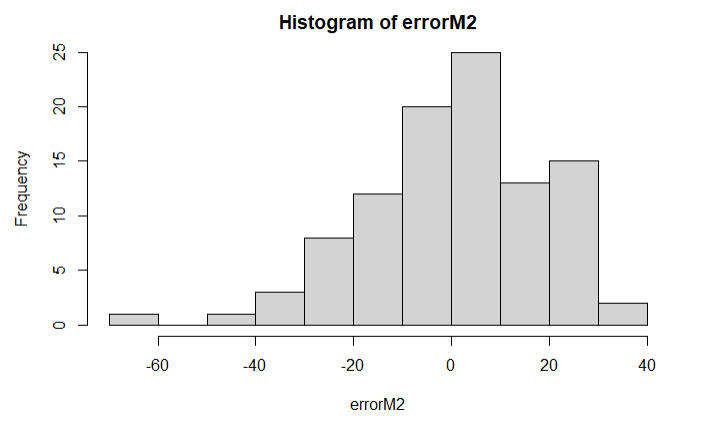


Ilustración 7: histograma del error del modelo ácido volátil

Ilustración 6: histograma del error del modelo global

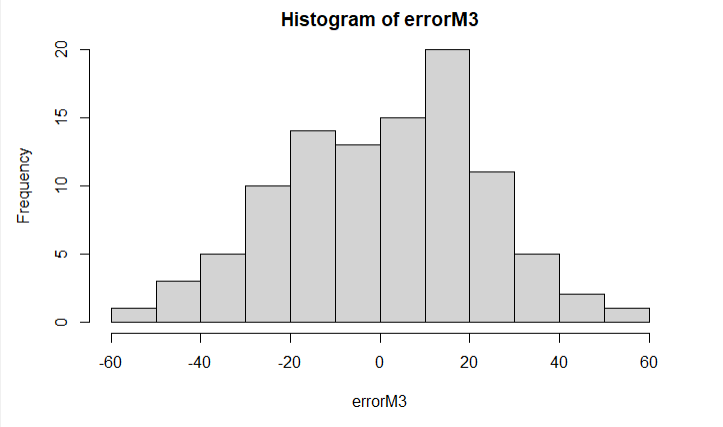


Ilustración 8: histograma del error del modelo ácido cítrico

Las 3 distribuciones de errores calculadas son bastante malas, si hay que elegir uno en función de la distribución de los errores, el mejor modelo es el que mide el efecto global

La matriz de confusión nos ayuda a medir cómo se comporta nuestro modelo cuando se lo aplicamos a data nueva, es decir nos dice de manera numérica cuánto se equivoca nuestro modelo. De la matriz de confusión nos interesa principalmente la potencia del contraste o sensibilidad, que es lo que queremos que sea más pequeño, ya que los falsos positivos son los peores errores que podemos cometer. Primero cambiamos la calidad a una variable cualitativa binaria (contiene 0 o 1), después hicimos una matriz de confusión por cada modelo que estábamos estudiando:

1. MODELO GLOBAL (incluye los dos tipos de ácidos)
2. MODELO SIMPLE DE ÁCIDO CÍTRICO
3. MODELO SIMPLE DE ACIDEZ VOLÁTIL

El modelo global presenta una potencia muy alta de 0.93 y una precisión de 0.74, lo cual indica que este modelo discrimina bien los falsos positivos. Además, se puede ver cómo, al cambiar la calidad por una variable binomial, el ácido cítrico es representativo en el modelo global, por lo que puede ser más interesante estudiar el modelo global cuando la variable calidad vale 0 o 1.

Por el contrario, los modelos simples tienen una potencia muy alta de 0.94, pero la precisión es peor que la del modelo global, de 0.6. Por lo tanto, los modelos simples son peores, ya que, a igual potencia, tienen peor precisión que el modelo global.

Lo siguiente que se nos ocurrió para mejorar nuestros modelos, fue probar a hacer un cluster para ver si una parte de los datos puede determinar la calidad del vino en mayor medida que el efecto global, o los efectos simples, en su totalidad

Una forma de hacer clusters que vimos, es el análisis cluster con método kmeans, pero esta forma de hacer clusters no nos interesa ya que no podemos saber las observaciones que conforman cada cluster, únicamente los centroides que nos interesa hacer y la dispersión de las observaciones de cada cluster. Pensamos en hacer dos clusters (calidad y acidez volátil, calidad y acidez cítrica) usando denogramas, de esta forma, se dividirán por grupos los datos que tuviesen relación. Así, podríamos estudiar cada grupo por separado, y ver si hay una correlación mayor con la calidad, pero al ser esta una variable discreta, no es útil este método.

Por último, una forma muy gráfica de ver las características de un vino de calidad es haciendo un árbol de clasificación con las variables que nos interesan. Podemos ver cómo para el 35% de los vinos, con una calidad de 6, tienen un ácido volátil menor que 0.59 pero mayor que 0.41, aunque también se encuentran el 28% de los vinos con una calidad de 6 cuando la acidez volátil es menor que 0.41. Por lo que los valores más aptos para que la calidad del vino sea mayor teniendo en cuenta el ácido cítrico y el volátil, es cuando el ácido volátil no supera el 0.59.

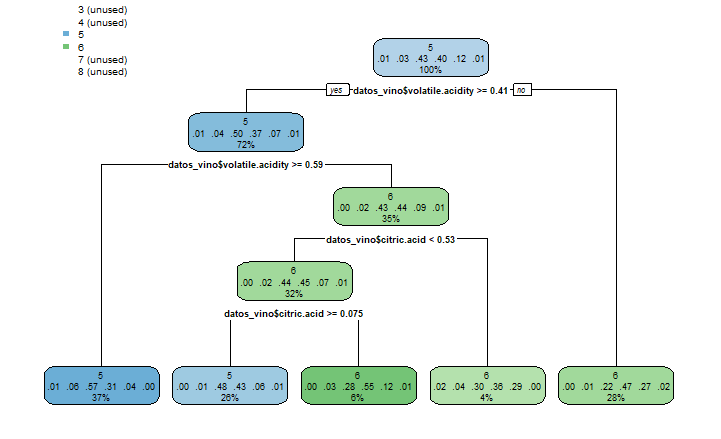


Ilustración 9: diagrama de árbol del ácido cítrico y acido volátil

**ESTUDIO DE ACIDEZ Y PH**

Por otro lado, vamos a estudiar el pH y todos los ácidos de nuestra base de datos, ya que, como hemos dicho anteriormente, deberían de tener relación entre sí y con la calidad, puede que un modelo global de todas estas variables sea más correcto y nos dé más información que los anteriores.

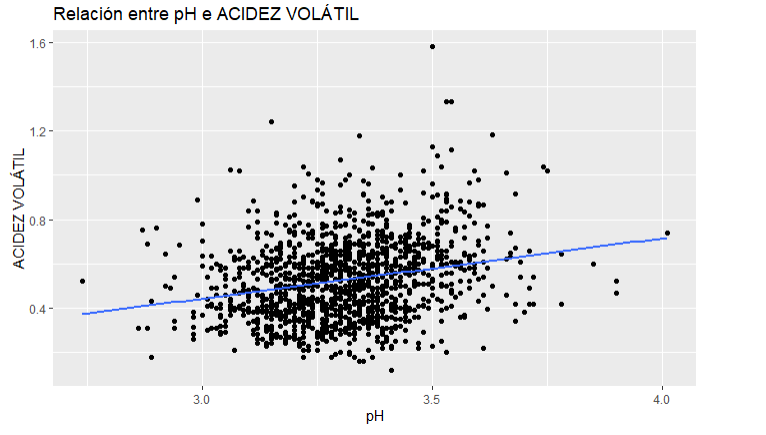
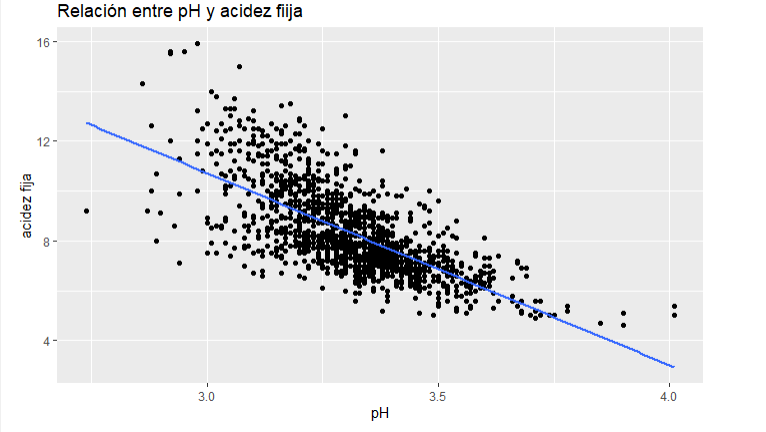
Lo primero que hicimos fue hacer gráficos y correlaciones.

1. MODELO SIMPLE ACIDEZ FIJA CON PH
2. MODELO SIMPLE ACIDEZ VOLÁTIL CON PH
3. MODELO SIMPLE ACIDEZ CÍTRICA CON PH

En los gráficos se puede ver cómo existe una correlación fuerte y negativa entre pH y acidez fija con un 0.68 y entre pH y acidez cítrica con un 0.54, pero, por el contrario, hay una correlación ligeramente positiva entre pH y acidez volátil con un 0.23, este fue el primer problema que nos encontramos, porque no le veíamos el sentido que la relación fuese positiva entre un ácido y el pH.

Para ver si esto tenía sentido hicimos una correlación entre la acidez total y acidez volátil y salió negativa, con una correlación entorno al 16%, y tras buscar información, vimos que este ácido se usa para corregir la acidez en mostos y vinos, esto significa que aumenta o disminuye en función de la cantidad de ácido

que contenga el vino, por eso daba una correlación negativa entre el ácido fijo y el volátil, puede ser que hubiese mucha cantidad de ácido en algunos vinos y esto hizo que el volátil disminuyese.

Para apoyar los cálculos, quisimos hacer los gráficos de estos modelos y nos salió lo siguiente:

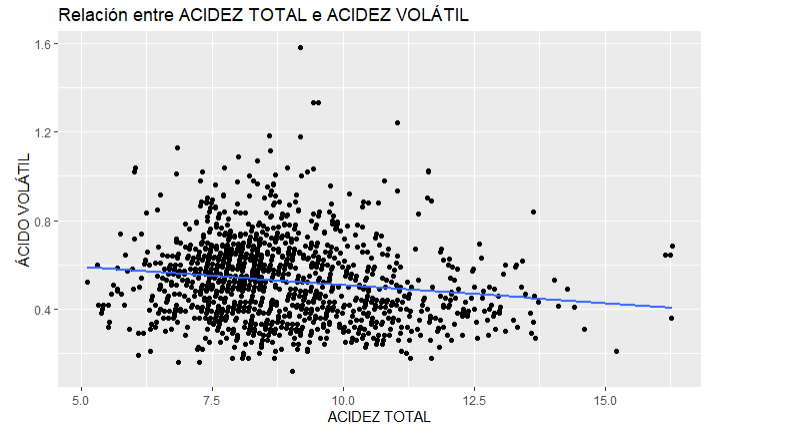
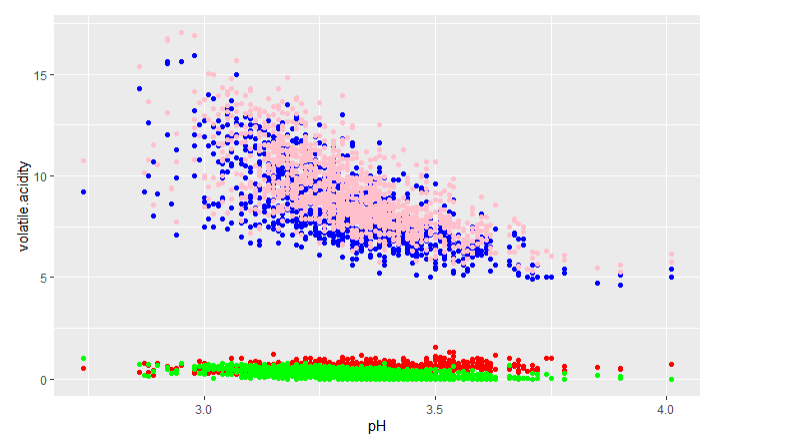


Ilustración 10: plot y recta de regresión entre PH y acidez volátil

Ilustración 11: plot y recta de regresión entre PH y acidez fija

Ilustración 12: plot y recta de regresión entre acidez total y acidez volátil

Aquí se puede ver la evolución de las 3 variables de la acidez, se ve como la acidez total disminuye cuando la volátil aumenta un poco.

Rojo: acidez volátil

Verde: Acidez cítrica

Rosa: Acidez total

Azul: Acidez fija

Ilustración 13: plot de las 4 variables de acidez

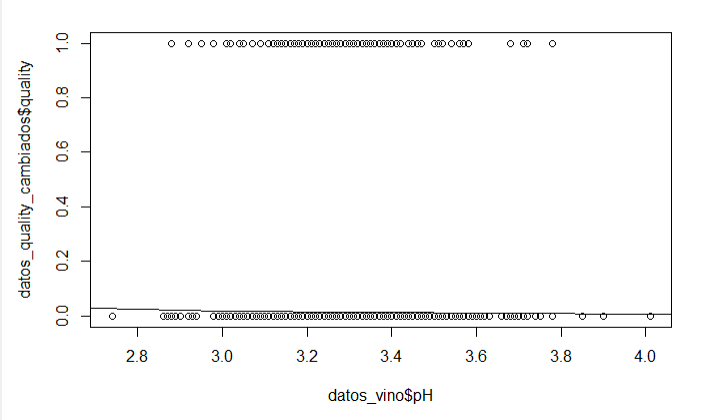
Hemos hecho 2 modelos múltiples para ver qué variables afectan de forma más directa o indirecta, medidas en función del p\_valor:

1. MODELO GLOBAL DE LA CALIDAD CON ACIDEZ Y PH
2. MODELO GLOBAL ENTRE DEL PH CON ACIDEZ

En el primer modelo, el ácido cítrico no es muy representativo, como se esperaba ya que este ácido es menos relevante que los demás tal y como se explica en el estudio, el ácido fijo es algo representativo y el pH es bastante representativo[[6]](#endnote-6).

Lo que nos llamó la atención al comparar los dos modelos fue que el ácido volátil es muy representativo para la calidad, pero no es representativa para el pH, lo que nos resultó raro ya que varía en función del pH para regular la cantidad de ácido de la muestra, y pensamos que debería de estar muy relacionado con este. Ante esto, pensamos que el ácido volátil afecta al pH de forma indirecta y por eso no se muestra su representatividad, además de que el p\_valor no es la mejor forma de ver lo relevantes que son las variables en un modelo

En un principio pensamos que tendría sentido que la acidez fuese una variable directa, mientras que pH fuese una indirecta, ya que, sin acidez no se puede medir el pH. Sin embargo, en nuestro modelo tiene más correlación la calidad con el pH que con la acidez, ya que el pH es cómo se mide la acidez, no es tan relevante para la calidad si hay mucha o poca acidez, sino lo fuerte que es en cada vino, ya que esto determina si es peligroso para la salud o no, lo cual afecta a la calidad. Además, tiene sentido que la representatividad del pH dependa de la del resto de ácidos, por lo que pensamos que su representatividad es una media de la representatividad de todos los ácidos del vino.



El gráfico entre la calidad y el pH no nos dice nada relevante, ya que la calidad es una variable cualitativa, intentamos hacer un gráfico logarítmico para ver si el pH cambiaba bruscamente cuando la calidad del vino es buena, pero no hubo ninguna variación.

Ilustración 14: plot de la recta logarítmica entre calidad y el PH

Un vino tiene que tener una acidez baja ya que, si no, pierde calidad y empieza a ser perjudicial para la salud. Para saber en qué niveles de pH se movían nuestras observaciones y en qué porcentaje, realizamos un árbol de clasificación entre pH y calidad, en este se puede ver cómo la mayor parte de los vinos tienen un pH entre el 3 y el 3.5 y ninguno sobrepasa el 4 ni es inferior al 2.8 que es el umbral máximo y mínimo en el que debería de estar el pH. Como regla general se intenta que los vinos no tengan un pH cercano o superior a 4 porque esta situación eleva el riesgo de oxidación y de aumentar el riesgo de problemas de salud. La mayor parte del porcentaje de los vinos con un 6 de calidad se encuentra cuando el pH es mayor que 3.3 y menor que 3.5 entorno al 44%, una cifra relativamente elevada, aunque no determinante.

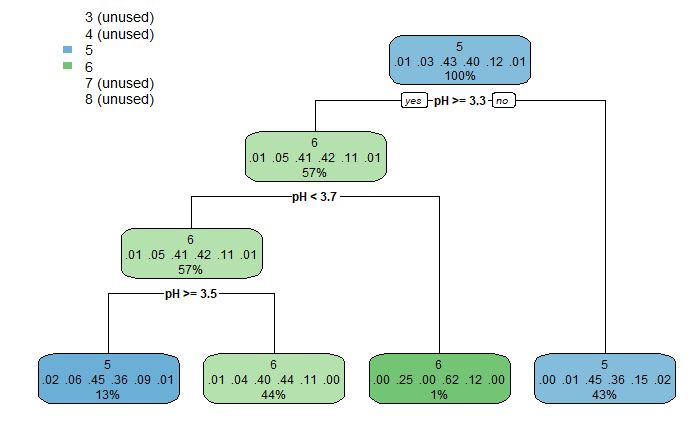


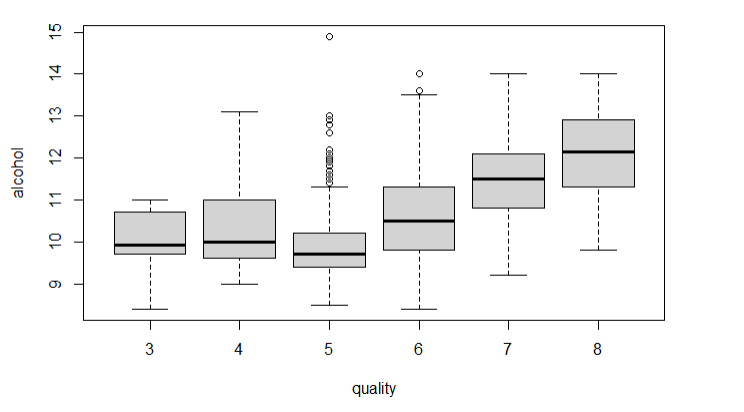
Ilustración 15: diagrama de árbol del PH

**ESTUDIO DEL ALCOHOL**

Otra de las variables que cambiaba bastante en función de la calidad del vino, es el alcohol.

Según el estudio que hemos realizado de la variable del alcohol, encontramos que es el elemento más abundante en el vino, que se produce por la transformación en los azúcares durante la fermentación[[7]](#endnote-7). En relación con la calidad hay que resaltar que, según las fuentes consultadas, es verdad que el alcohol asegura la buena conservación del caldo, defiende al líquido de ciertos microorganismos patógenos y resalta las virtudes de los otros componentes del vino, pero la calidad del vino no se ve condicionada por el grado alcohólico[[8]](#endnote-8). A continuación, veremos si esta teoría se cumple en nuestra base de datos, y en caso de ser así, si hay algún modelo que, usando el alcohol, pueda determinar la calidad.

La proporción de alcohol de una bebida, para un volumen dado de la misma, se denomina grado alcohólico. La graduación de los vinos varía entre un 7% y un 16% de alcohol por volumen, aunque la mayoría de los vinos embotellados oscilan entre 10 y 14 grados. Nuestra base de datos mide la variable alcohol en grados alcohólicos volumétricos (%).



Hemos observado el gráfico donde se muestran las correlaciones de todas las variables entre sí, y se puede observar como el alcohol tiene algo de relación con la densidad y con la calidad, por lo que realizaremos los siguientes modelos;

Ilustración 16: boxplot de la calidad y el alcohol

1. MODELO SIMPLE DE ALCOHOL CON CALIDAD
2. MODELO COMPUESTO DE ALCOHOL Y DESIDAD CON CALIDAD

Para ver la representatividad de nuestra variable de alcohol con la calidad, realizamos un modelo simple de regresión. Según este modelo podemos ver que el p-valor tiene un valor bajo por lo que la variable es bastante representativa y significativa para el modelo. Como ya sabemos no debemos fijarnos solo en el p-valor por lo que continuamos con nuestro estudio.

Tras realizar el modelo realizamos la correlación. La correlación entre la calidad y el alcohol es de 0,47, una relación positiva y relativamente alta en comparación con otras hechas anteriormente. Por lo tanto, ambas variables se correlacionan en sentido directo y el 47% de la variabilidad de la calidad se ve explicada por el alcohol. Teniendo en cuenta esto podríamos decir que los grados de alcohol podrían tener algo de influencia en la calidad.

Al hacer el modelo compuesto, ambas parecen representativas para la calidad, pero la correlación es negativa de -0,49. Pensamos que al añadir la densidad al modelo mejoraría la correlación con la calidad, pero apenas hay diferencia con el modelo simple, por lo que seguiremos estudiando ambos modelos, hasta ver cuál es mejor.

Para entender por qué en este caso la correlación salía negativa realizamos un modelo simple;

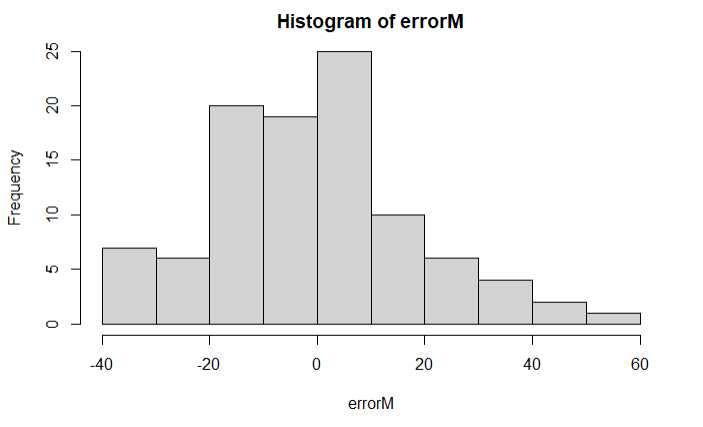
MODELO SIMPLE DE ALCOHOL Y DENSIDAD

Así vimos como el alcohol y la densidad tienen una relación negativa también de -0,49.

Para visualizar de mejor manera lo que estábamos calculando, intentamos realizar los gráficos entre estas dos variables para ver su correlación. Como ha pasado anteriormente, al hacer los gráficos, no se puede apreciar nada, ya que la calidad es una variable discreta del 0 al 10.

Una vez evaluado el impacto del grado de alcohol en la calidad, es hora de predecir las habilidades predictivas del modelo. Para ello emplearemos la validación cruzada en la que dividiremos nuestra muestra en una porción de entrenamiento y otra para testear el modelo. Para ello emplearemos la validación cruzada en la que dividiremos nuestra muestra en una porción de entrenamiento (80%) y otra para testear (20%) el modelo. De ese modo obtendremos la distribución de los errores del modelo.

Dibujamos la distribución de los errores del modelo simple, entre el alcohol y la calidad, se puede ver cómo no es un modelo muy bueno ya que acumula bastantes errores a un lado y en el 0. Mientras que, al hacer el gráfico del modelo compuesto, vemos que la distribución de los errores se asemeja más a una normal, acumulando la mayor parte de los errores cerca del 0. Usando este método nos interesa más quedarnos con el modelo compuesto.



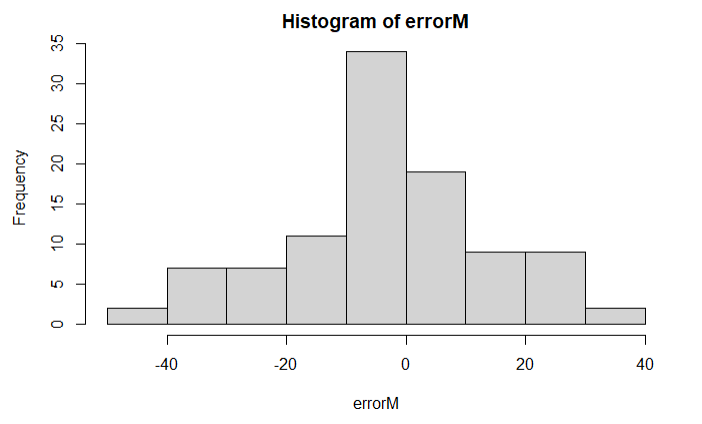


Ilustración 17: histograma del error del modelo alcohol

Ilustración 18: histograma del error del modelo alcohol y densidad

Para ver cuánto se equivoca nuestro modelo en número, emplearemos la matriz de confusión y nos permitirá trabajar con los distintos tipos de error del modelo. Como se ha mencionado con anterioridad nos fijamos en los falsos negativos, la sensibilidad ya que son los más importantes. En el caso del modelo del alcohol y la calidad hemos encontrado el valor 0.55 en el que la sensibilidad es de 86%, con los falsos negativos bastante bajos, aunque especificatividad es de un 50% (falsos positivos). El modelo global no mejora apenas, con una sensibilidad de 89% y especificatividad del 50%

Por último, realizaremos el árbol de clasificación para ver de otra forma más visual si un vino de calidad realmente depende del alcohol.

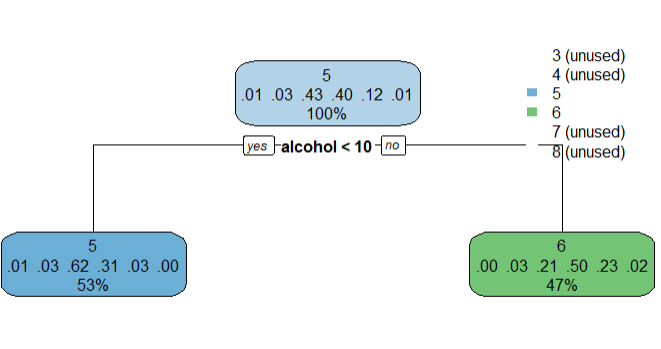
Para el primer árbol pudimos observar que cuando la calidad se encuentra en 6, los grados son mayores a 10 para el 47% de la muestra de los vinos, una cifra bastante elevada, fijándonos únicamente en el árbol de clasificación podríamos rechazar que los vinos que tienen un grado de alcohol menor que 10 tengan buena calidad.

Ilustración 19: diagrama de árbol del alcohol y densidad

Comparando el modelo simple de la calidad con el alcohol y el compuesto de la calidad con el alcohol y la densidad, únicamente se puede apreciar una diferencia notable en la distribución de los errores por lo que, elegiremos el modelo global como el mejor modelo entre los dos para medir la calidad, ya que este tenía mejor distribución, además, el árbol de clasificación nos muestra que los vinos de calidad 6 se encuentran cuando el alcohol supera los 10 grados.

Aun así, la correlación con la calidad es baja y los falsos positivos de la matriz de confusión son bastante altos, por lo que no podemos afirmar que usando únicamente el alcohol y la densidad podamos medir la calidad del vino.

**ESTUDIO DE DIÓXIDO DE AZUFRE**

Por último, hemos realizado el análisis del dióxido de azufre, que se añaden en el vino para evitar el crecimiento de microorganismos (bacterias y levaduras) indeseables, y en el acabado del vino aporta efectos beneficiosos sobre las características sensoriales, tanto aroma como color 9. Nos encontramos con dos variables para este componente en nuestra base de datos; dióxido de azufre total y dióxido de azufre libre. Ambas variables las estudiaremos juntas ya que el dióxido de azufre libre forma parte del total junto al dióxido de azufre combinado (variable que no aporta la base de datos).[[9]](#endnote-9)

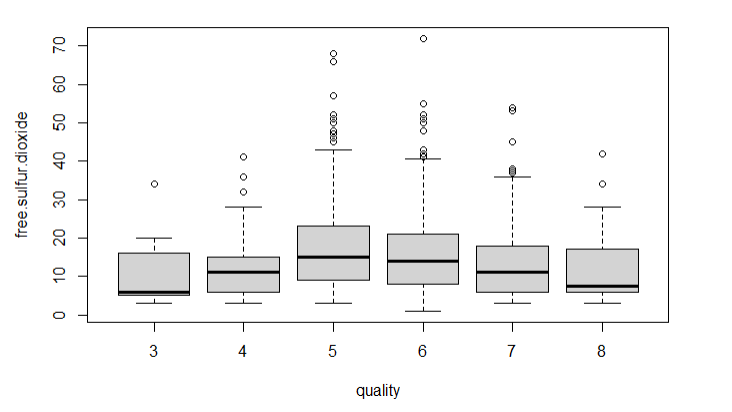
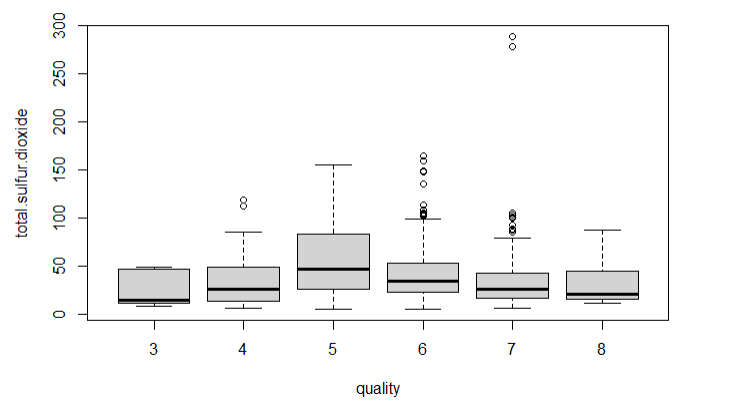
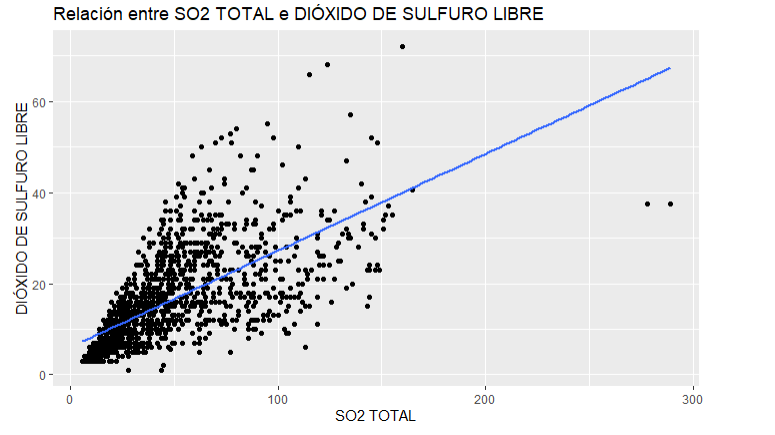


Ilustración 21: boxplot dióxido de azufre libre

Ilustración 20: boxplot dióxido de azufre total

Según nuestras investigaciones, el dióxido libre es el que realmente aporta propiedades antisépticas y antioxidantes al vino[[10]](#endnote-10). Por lo tanto, un buen vino sería aquel que con menos total consigue más libre. Ambos mantienen una relación positiva de correlación de 0,66 por lo que si aumenta uno aumenta el otro.



Para estudiar la representatividad de los dos tipos de dióxidos con la calidad realizaremos 3 modelos, dos individuales y uno global. Los modelos son los siguientes:

Ilustración 22: plot y recta de regresión entre dióxido de azufre total y dióxido de azufre libre

1. MODELO GLOBAL (incluye los dos tipos de dióxido de azufre)
2. MODELO DE DIÓXIDO DE AZUFRE TOTAL
3. MODELO DE DIÓXIDO DE AZUFRE LIBRE

Según el modelo individual de dióxido de azufre libre, este no es muy representativo. En cambio, en el modelo global si resulta representativo lo que quiere decir que puede que el dióxido de azufre libre afecte de manera indirecta a la calidad.

Tras analizar sus modelos, nos centraremos en sus correlaciones que muestran valores bastante bajos. En el caso de dióxido total, muestra un valor algo mayor pero negativo de -0,18. El dióxido de azufre libre tiene un valor bastante bajo y negativo lo cual tiene poco sentido ya que como hemos mencionado antes es el encargado de aportar las propiedades antioxidantes que aportan calidad al vino. Por lo tanto, si nos fijamos en esto no resultan variables importantes a la hora de medir la calidad del vino.

Como ya hemos mencionado, no podemos hacer gráficos con la calidad ya que no se puede sacar nada en claro.

En cambio, como ambas variables tienen bastante relación aplicaremos la técnica de los componentes principales. De ese modo veremos si deberíamos omitir el dióxido de azufre libre ya que está dentro del total y de ese modo quedarnos con la información más relevante. O al contrario quedarnos con el dióxido de azufre libre ya que es el que según los enólogos aporta la calidad al vino y eliminar el dióxido total.

Como la correlación de los sus autovectores es distinto de 0, pero muy cercano, Comparten información entre ellos, pero sus varianzas son muy diferentes. Por lo tanto, no podemos seguir investigando con el método de los componentes principales

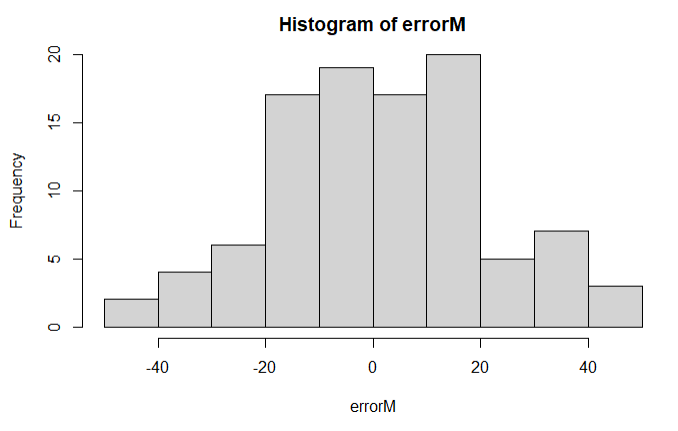
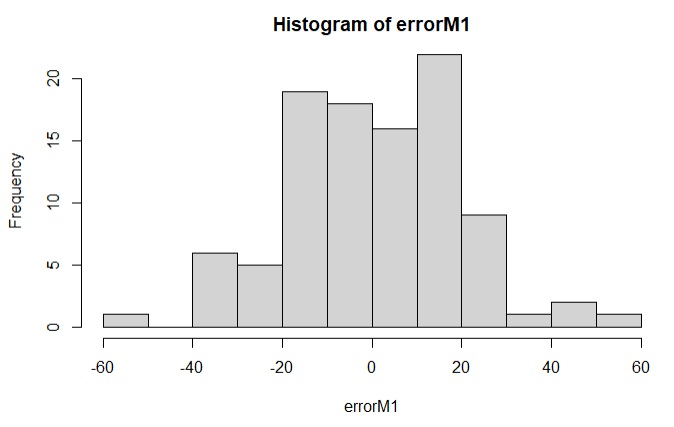
A continuación, analizaremos la distribución de los errores de los tres tipos de modelos planteados. Para ello emplearemos el histograma para graficar los resultados y poder sacar mejores conclusiones. Como ya sabemos consideraremos que la distribución de los errores será buena si se encuentran en torno a 0.

Ilustración 23: histograma del error del modelo dióxido total

Ilustración 24: histograma del erro del modelo dióxido libre

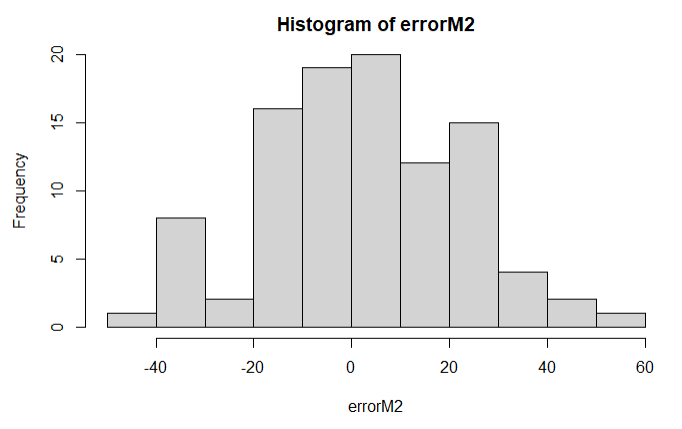


Ilustración 25: histograma del error del modelo global

En este caso, las tres distribuciones son bastante malas con valores alejados a 0 y con poca concentración de los valores entorno al 0. Si se tuviese que elegir un modelo será el global pero tampoco es bueno.

Ya que los errores en los modelos de estas variables son acentuados, emplearemos la matriz de confusión para poder diferenciar los distintos tipos de error. Nos centraremos en los falsos positivos como los peores errores que puede cometer el modelo y de ese modo veremos qué modelo tiene mejores niveles de sensibilidad y potencia de contraste. Para ello debemos realizar una matriz de confusión para cada modelo:

1. MODELO GLOBAL (incluye los dos tipos de dióxido de azufre)
2. MODELO DE DIOXIDO DE AZUFRE TOTAL
3. MODELO DE DIOXIDO DE AZUFRE LIBRE

Centrándonos en el modelo global, su predicción respecto a los falsos negativos es mala, por lo tanto, predice que tiene una calidad peor a la que realmente tiene. Su sensibilidad en cambio tiene un valor bastante bueno del 94%.

Por último, para un estudio más exhaustivo y grafico de las variables empleamos un gráfico de árbol en el que pudimos ver cómo el 15% de los valores más altos de dióxido de azufre total (mayor que 82) tienen una puntuación de calidad cercana a 5, un valor bajo. La puntuación más alta es de 6, que corresponde a los vinos con un dióxido de azufre total menor que 51, con un porcentaje del 66%. Sabiendo que el 85% de los vinos de calidad 6 tienen un dióxido de azufre total menor que 82, podemos intuir que los vinos de calidad 7 y 8 también tendrán un valor menor que 82

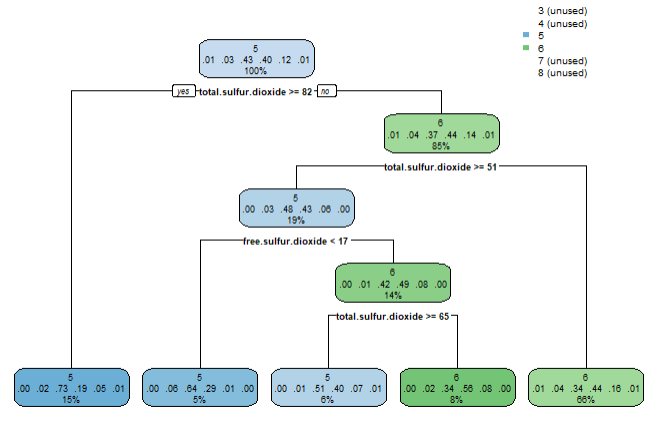
**Del estudio del dióxido de azufre podemos concluir que el modelo global es mejor que los modelos simples, aunque no gran diferencia entre ellos.

Ilustración 26: árbol de las variables de dióxido de azufre total y libre

**CONCLUSIÓN**

Como conclusión de este informe sobre la calidad del vino, podemos sacar en claro que, todas las variables que hemos relacionado con la calidad son relevantes, ya que cada una aporta información diferente. Además, después de haber realizado modelos simples y múltiples, estos últimos han demostrado ser mejores para medir la calidad del vino. Lo que nos lleva a pensar que hay variables que se relacionan de manera indirecta con la calidad, y que influyen en la variación de esta. Algo que nos ha llamado la atención, es que, a pesar de en los boxplot había variables que parecía que influían mucho en la calidad, y haciendo los modelos, aparecían como variables representativas para la calidad, la correlación con esta última era baja (por debajo del 50%), hemos pensado que esto puede ser porque al hacer la correlación solo se tiene en cuenta la relación entre 2 variables, por lo que no es muy útil mirar únicamente la correlación. Hacer las distribuciones de los errores no nos ha resultado de mucha ayuda ya que fijándonos en el gráfico ninguna parece ser buena para la calidad. Sin embargo, al mirarlo con números, haciendo la matriz de confusión, la precisión de cada modelo era relativamente buena, al igual que la potencia, pero todos fallaban en la especificatividad. En cuanto a los diagramas de árboles, hemos escogido los modelos múltiples para representarlos ya que contienen más información. Pero, ni en los múltiples ni en los simples se ven reflejados los vinos con una calidad mayor que 7 en las primeras ramificaciones, por lo que este es otro indicador de que no se puede determinar la calidad únicamente fijándonos en los componentes que tienen relación entre sí y con esta.

Teniendo en cuenta todo esto, no hemos encontrado un componente que determine la calidad de un buen vino por sí solo, de hecho, probando el efecto de varios componentes con la calidad, y el mejor modelo que hemos encontrado, ha sido el que relacionaba el alcohol y la densidad con la calidad, con una gran representatividad y una correlación del 50%. Obteniendo una precisión de 0,86 y una potencia de 0,89. La profundidad de estos elementos debería de ser de más de 12% de alcohol y una densidad mayor que 1 g /cm^3. La distribución de los errores del modelo global se asemeja a una normal, no es perfecta, pero es bastante buena en comparación con las demás realizadas. Esto nos ha sorprendido bastante ya que en las fuentes de información relacionadas con el alcohol se afirmaba que este no era determinante para la calidad del vino, sin embargo, en nuestra base de datos ha demostrado ser el componente que mejor determina la calidad, a pesar de que no sea el único elemento en el que tengamos que fijarnos para esto.

**REFERENCIAS**

Para este trabajo hemos usado gran parte de la información de un estudio sobre el vino tinto:

Estefanía Paz Salinas Flores, Dra. María Carolina Zúñiga López, Dr. Víctor Felipe Laurie Gleisner. “Efecto del ph sobre la capacidad antioxidante del vino modelo: interacción entre hierro y compuestos fenólicos”, Universidad de Chile Facultad de Ciencias Químicas y Farmacéuticas, Departamento de Química Inorgánica y Analítica, Laboratorio de Radicales Libres y Antioxidantes.  2017. [online]. Available: <http://repositorio.uchile.cl/bitstream/handle/2250/169810/Efecto-del-pH-sobre-la-capacidad-antioxidante-en-vino-modelo.pdf?sequence=1&isAllowed=y>

1. Adrián Yirda. “Definición de calidad”, ConceptoDefinición*.* 16 de noviembre de 2020. [online]. Available: <https://conceptodefinicion.de/calidad/> [↑](#endnote-ref-1)
2. “5 características de un buen vino”, *Bodegas Murillo Viteri,* [online]. Available: <https://www.bodegasmurilloviteri.com/aprende-de-vino/5-caracteristicas-de-un-buen-vino-tinto/> [↑](#endnote-ref-2)
3. Daria alekseeva. “Red and white wine quality”, Rpubs by RStudio, [online]. Available: https://rpubs.com/Daria/57835 [↑](#endnote-ref-3)
4. “Vino y PH”, *Aprender del vino*. 2020. [online]. Available: <https://www.aprenderdevino.es/ph-y-vino/> [↑](#endnote-ref-4)
5. Isabel Pustilnik y Federico Gómez, “Autovalores y autovectores: definiciones y propiedades”, [online]. Available:

   <https://aga.frba.utn.edu.ar/autovalores-autovectores-definiciones-propiedades/> [↑](#endnote-ref-5)
6. [↑](#endnote-ref-6)
7. “¿Qué supone el alcohol en el vino?”, blog Catavino. [online]. 05/12/2014. Available: <https://www.catadelvino.com/blog-cata-vino/que-supone-el-alcohol-en-el-vino>  [↑](#endnote-ref-7)
8. Mark O´Neill. “El alcohol en el vino”, blog de Mark O´Neill. [online]. Available: <https://markoneill.es/el-alcohol-en-el-vino> [↑](#endnote-ref-8)
9. “Qué son los sulfitos y porque hay sulfitos en el vino”, LocalBioMarket. 11/05/2020. [online].

   Available: <https://localbiomarket.es/que-son-los-sulfitos-y-porque-hay-sulfitos-en-el-vino/> [↑](#endnote-ref-9)
10. Dra. Montserrat Riu Aumatell. “Vino, seguridad alimentaria y salud del consumidor: el caso del azufre”, Departamento de Nutrición, Ciencias de la Alimentación y Gastronomía, Facultad de Farmacia y Ciencias de la Alimentación, Universidad de Barcelona. Instituto de Nutrición y Seguridad Alimentaria (INSA), Xarxa de Referència en Tecnologia dels Aliments (XaRTA).18/10/2018. [online]. Available: <https://www.interempresas.net/Vitivinicola/Articulos/227155-Vino-seguridad-alimentaria-y-salud-del-consumidor-el-caso-del-azufre.html> [↑](#endnote-ref-10)