



Trabajo Práctico Especial

Análisis de la calidad del vino por medio de técnicas de ciencia de datos

Alumnos:

Ortiz, Matías

Ramirez, Gonzalo Ezequiel

Leon, Nicolas

Profesores a cargo:

Prof. Mg. Cepeda, Rosana

Prof. Dra. Godoy, Daniela

Prof. Dra. Gonzales Císaro, Sandra

JTP. Dr. Orlando, Juan Ignacio

Ay. Dip. Prof. Perez Fernández, Débora

Índice

Introducción.....	3
Hipótesis.....	3
Metodología y Resultados.....	4
Recursos.....	4
Análisis univariado.....	5
Planteo de Hipótesis.....	6
Mientras más ácido volátil peor es la calidad de un vino.....	6
Los vinos de buena calidad tienen menos cloruros.....	7
Los vinos de buena calidad tienen más alcohol.....	9
Los vinos de buena calidad tienen menos densidad.....	10
Mientras más alto es el ácido cítrico, mejor es la calidad del vino.....	11
Mientras más alta es la cantidad de sulfitos libres, mejor es la calidad de un vino.....	13
Los vinos de uva Garnacha, tienen mejor calidad que los vinos de uva Riesling.....	15
No hay una relación entre el tipo de uva y la calidad.....	16
La calidad de un vino está determinada, en general, por el alcohol, los cloruros, densidad y ácido volátil.....	17
Existe una relación lineal entre la densidad, el grado de alcohol y los azúcares residuales en los vinos Riesling.....	19
Existe una relación lineal entre la acidez fija, el pH, la acidez cítrica y densidad en los vinos Garnacha.....	20
Existen determinados atributos que determinan la calidad de los vinos Riesling.....	21
Conclusiones.....	23
Referencias.....	24

Introducción

La calidad del vino se suele definir mediante ciertos criterios obtenidos a través de la cata del vino. Los catadores califican el vino según sus sentidos del gusto y olfato, siendo este un medio subjetivo para valorar un vino, ya que podemos desconfiar de los sentidos del catador, como también puede verse sesgada su opinión. Por eso mismo, en este informe, se tiene como objetivo analizar la calidad del vino en relación a sus propiedades químicas, variables obtenidas con mediciones directas de sus valores.

En esta investigación, se realizó un análisis exploratorio a un conjunto de datos sobre la calidad de los vinos (de tipo Riesling y Garnacha) producidos por la bodega “El Refugio”, junto a diversos métodos de la Ciencia de Datos, para extraer información y patrones de estos, y así lograr resolver las hipótesis planteadas.

Hipótesis

Se desean comprobar las siguientes hipótesis:

1. *Mientras más ácido volátil peor es la calidad de un vino.*
2. *Los vinos de buena calidad tienen menos cloruros.*
3. *Los vinos de buena calidad tienen más alcohol.*
4. *Los vinos de buena calidad tienen menos densidad.*
5. *Mientras más alto es el ácido cítrico, mejor es la calidad del vino.*
6. *Mientras más alta es la cantidad de sulfitos libres, mejor es la calidad de un vino.*
7. *Los vinos de uva Garnacha, tienen mejor calidad que los vinos de uva Riesling.*
8. *No hay una relación entre el tipo de uva y la calidad.*
9. *La calidad de un vino está determinada, en general, por el alcohol, los cloruros, densidad y ácido volátil.*
10. *Existe una relación lineal entre la densidad, el grado de alcohol y los azúcares residuales en los vinos Riesling.*
11. *Existe una relación lineal entre la acidez fija, el pH, la acidez cítrica y densidad en los vinos Garnacha.*
12. *Existen determinados atributos que determinan la calidad de los vinos Riesling.*

Metodología y Resultados

Recursos


Para realizar la investigación, se posee un archivo en formato *Word* que describe el conjunto de datos y sus variables, junto al correspondiente dataset, brindados por la cátedra. Se ha utilizado una notebook de *Jupyter* para realizar el código, explicando cada paso que se siguió para completar la investigación, junto a varias librerías de *Python* para manipular los datos, realizar gráficos, hacer tests, entre otros.

Contamos con 3232 muestras de vino obtenidas mediante pruebas físicoquímicas en la bodega “El Refugio”, que tienen las siguientes variables:

- **Quality** (Calidad): Puntuación del vino, con un rango del 1 al 10.
- **Type** (Tipo): Tipo de uva con el que se elaboró el vino.
- **Citric acid** (Ácido cítrico): Es uno de los ácidos que componen la *acidez fija* y está presente naturalmente en la uva. Puede ser agregado durante el proceso de vinificación. Se utiliza para la prevención de la oxidación y modificar el sabor. En exceso provoca un sabor amargo. En el dataset está medida en gramos por litro. En el dataset está medido en gramos por litro.
- **Residual sugar** (Azúcar residual): Es el azúcar que no se convierte en alcohol durante la fermentación y queda en el producto final. El vino principalmente obtiene su dulzor del azúcar de la uva (glucosa y fructosa). Según su cantidad el vino puede ser seco (0 - 12g/l), semiseco (12 - 18g/l), semidulce (18 - 45g/l) y dulce (> 45g/l). En el dataset está medido en gramos por litro.
- **Chlorides** (Cloruros): Concentración de cloruros (sales) en el vino. El cloruro puede tener un impacto en las características organolépticas del vino. En concentraciones bajas, puede contribuir al sabor salado y mejorar la percepción de otros sabores, como el dulzor. En el dataset está medido en gramos por litro.
- **Density** (Densidad): Es la relación entre la masa y el volumen del vino, lo que puede dar una idea de la concentración de azúcares, alcohol y otros compuestos disueltos en el líquido. A medida que el vino fermenta, las levaduras consumen los azúcares y los convierten en alcohol y dióxido de carbono, lo que hace que la densidad disminuya a lo largo del proceso de fermentación. En el dataset está medido en gramos por litro.
- **pH**: medida de la acidez o alcalinidad del vino. Se puede considerar que, a mayor pH, menor acidez, y viceversa.
- **Sulphates** (Sulfatos): Concentración de sales de sulfato en el vino. Son generados por la oxidación del dióxido de azufre. Ayudan a mantener la frescura y el sabor del vino, y prolongan su conservación. En el dataset está medido en gramos por litro.
- **Alcohol**: Contenido alcohólico del vino. La graduación alcohólica puede influir en el aroma, apariencia y sabor. En el dataset está medido en porcentaje de volumen.
- **Free sulfur dioxide** (Dióxido de azufre libre): El SO₂ libre es un compuesto formado por oxígeno y azufre que protege el vino de la oxidación y del crecimiento microbiano, por lo que es el principal responsable de la acción antioxidante y conservante de los sulfitos. La mayor parte del dióxido de azufre reacciona con algunas de las sustancias presentes, la parte que no se combina es libre. En el dataset está medido en miligramos por litro.
- **Total sulfur dioxide** (Dióxido de azufre total): Es la suma del *dióxido de azufre libre* y *combinado*. El segundo se refiere a cuando el sulfuroso se une a otros compuestos como el azúcar, siendo una forma que acarrea la pérdida casi total de los efectos buscados con el SO₂.

En el dataset está medido en miligramos por litro.

- ***Volatile acidity*** (Acidez volátil): La acidez volátil se refiere al ácido acético del vino. Se utiliza para determinar el estado del vino. Se dice que está picado (no recomendable para consumo) a partir de un 1 g/l. En el dataset está medido en gramos por litro.
- ***Fixed acidity*** (Acidez fija): La acidez total de un vino finalizado se obtiene de la suma del valor de la acidez fija más el valor de la acidez volátil. La acidez fija se puede componer del ácido tartárico, málico, cítrico, y succínico. Afecta al color, sabor, aroma y a la conservación del vino. En el dataset está medido en gramos por litro.

Para la investigación, y el planteo de ideas en general del trabajo, estuvimos utilizando un tablero en Miro  TPE FCD, donde realizamos comentarios, planteos de hipótesis, escrituras y análisis de cada una de las variables (buscando información en diversas páginas y papers).

Análisis univariado

A modo de resumen, y para no tener qué poner un análisis extenso de cada variable, usamos este espacio para hablar a modo general de todas las distribuciones del dataset de vinos. Si se quiere ahondar en el tema, en el tablero proporcionado de *Miro* se encuentra una explicación dedicada a cada *feature*.

En general, la mayoría de las distribuciones estaban sesgadas a derecha, y presentaban outliers. El principal problema de estos, fue que en todos los casos tomaban un rango muy grande de valores, lo que dificulta el análisis. Muchos de estos outliers eran valores posibles dentro del mundo real y no fueron modificados o eliminados. Otros, en cambio, como en el caso de *alcohol* y *density*, se debían a errores de carga.

Había otras que, además, parecían tener una bimodalidad como *citric acid* o *total sulfur dioxide*, y otras que sobresalían por sus características como es *density* por su rango tan ínfimo, *chlorides* por su Kurtosis de 41,1 o *residual sugar* por su coeficiente de variación casi del 100%.

Respecto a la proporción de vinos, había casi la misma cantidad de muestras de uno que de otro. Los vinos de media calidad eran los más abundantes, y de los de buena calidad, había más del tipo Riesling. Más adelante, se encontrará información sobre qué se considera bueno, medio o malo.

Planteo de Hipótesis

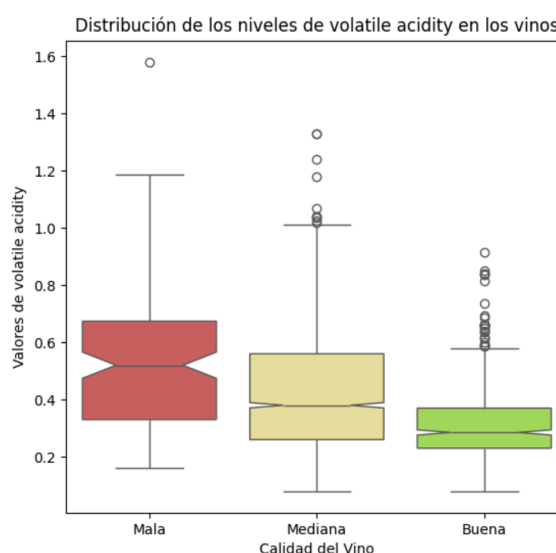
Mientras más ácido volátil peor es la calidad de un vino

Esta hipótesis fue extraída de la información del dominio en un principio, debido a que según información que extrajimos desde diversas páginas, el ácido volátil es una variable la cual afecta a la calidad del vino en cierto punto. Los valores normales de esta variable rondan entre los 0.3-0.6 g/l. Cuando un vino está picado (debido a una alteración o a un envejecimiento excesivo en los barriles) presenta una acidez volátil por encima de 1 g/l, lo que genera unos aromas similares al vinagre y al barniz. Siempre y cuando los valores de acidez volátil no sobrepasen los 0.6 g/l el sabor del vino no se verá demasiado afectado. Según diversas fuentes, la acidez volátil más baja se asocia a vinos de mejor calidad.

Durante el análisis bivariado, generamos matrices de Spearman sobre los datos de Riesling y Garnacha (no separar por tipo de uva podría afectar al resultado de las correlaciones o se podrían perder) para conocer qué variables podrían llegar a estar correlacionadas con la calidad de los vinos. En ellas, pudimos observar que existen diferentes variables que presentan una asociación moderada con la calidad, lo que es de esperar porque sería raro que esta dependa mayormente de unas pocas variables o de ninguna. En este caso, la acidez volátil tenía una correlación de -0.23 con la calidad en vinos de uva Riesling y de -0.38 en vinos de uva Garnacha.

Al ver el resultado, nos decantamos por un boxplot para conocer los efectos y comportamiento de la acidez, y antes de graficar, realizamos un agrupamiento de los vinos según su calidad, es decir, generamos tres grupos (mala calidad: 0-4, mediana calidad: 5-6, buena calidad: 7-10) para poder estudiar cómo se distribuían los valores en estos grupos, y pudimos observar que los vinos con buena calidad tendían a tener niveles más bajos de acidez volátil y que, por contraparte, aquellos con mayor nivel de la variable tenían peor calidad.

Para comprobar nuestra hipótesis, necesitábamos testear normalidad y homocedasticidad de los datos, para saber si podíamos realizar un *test t*. Realizamos el test de Shapiro-Wilk el cual nos dio que los tres grupos no eran normales (mala: p-valor = 0.000, mediana: p-valor = 0.000, buena: p-valor = 0.000), y el test de Levene, el cual no resultó efectivo para ningún par de grupos (mala/mediana: p-valor = 0.000, mala/buena: p-valor = 0.000, mediana/buena: p-valor = 0.000). Por la negación de los supuestos, aplicamos Kruskal-Wallis para comprobar la hipótesis, que resultó en que hay diferencias significativas entre las medias de los grupos de mala y mediana calidad (p-valor = 0.000), mediana y buena calidad (p-valor = 0.000) y mala y buena calidad (p-valor = 0.000). Estos resultados implican que, como vimos en el gráfico, los vinos de buena calidad tienen significativamente menores niveles de ácido volátil en promedio que los vinos de mediana calidad, y estos últimos, que los de mala calidad. Sin embargo, al analizar un comportamiento poblacional no estamos teniendo en cuenta si es que este se cumple para ambas uvas. Además, otro problema que surgió es que al separar a los vinos en Riesling y Garnacha, se reduce bastante la cantidad de muestras de los grupos de buena y mala calidad. Por estas razones, decidimos comprobar si es que los vinos de buena/mediana (de 6 a 10) calidad tienen una menor media de niveles de ácido volátil que los de mala/mediana (de 1 a 5). En



otras palabras, separamos a los vinos Riesling y Garnacha en dos grupos, lo que nos da la ventaja de trabajar con muchas muestras.

Realizando el estudio por cada tipo de uva, utilizando los mismos procedimientos antes explicados, Shapiro-Wilk dio que las distribuciones no eran normales para ningún grupo de ambas uvas (Riesling: mediana/baja - p -valor = 0.000, mediana/alta - p -valor = 0.000. Garnacha: mediana/baja - p -valor = 0.000, mediana/alta - p -valor = 0.000), pero el test de Levene arrojó que las distribuciones para ambos vinos eran homocedásticas (Riesling: p -valor = 0.056, Garnacha: p -valor = 0.064), por lo cual procedimos a realizar Mann-Whitney en ambas, resultando en que se rechazan las hipótesis nulas (Riesling: p -valor = 0.000, Garnacha: p -valor = 0.000) y dando por entendido que existen diferencias significativas entre los grupos de mediana/baja y mediana/alta para ambos tipos de uva, lo cual indica que nuestra hipótesis planteada estaba en lo cierto, que los vinos de buena calidad tienen significativamente menos cantidad de ácido volátil, sin importar el tipo de uva, y por tanto podemos decir que un vino se ve afectado por la cantidad de acidez volátil que tenga.

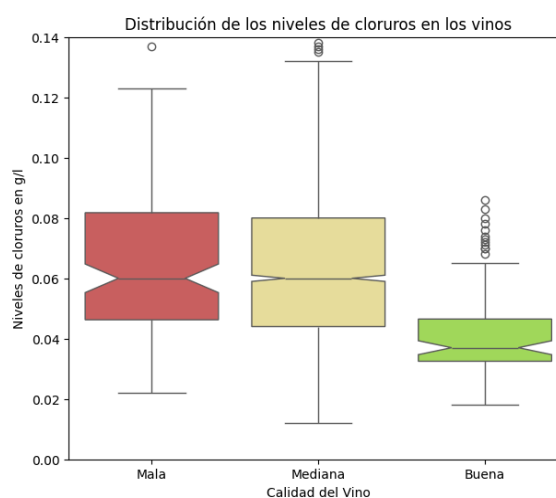
Los vinos de buena calidad tienen menos cloruros

Esta hipótesis fue extraída del análisis previamente explicado, donde el valor de la correlación en la matriz de Spearman con la variable calidad también fue moderado. En este caso, los cloruros tenían un nivel de correlación de -0.29 en los vinos de uva Riesling y -0.19 en los vinos de uva Garnacha.

Para comprobar nuestra hipótesis, necesitábamos testear normalidad y homocedasticidad de los datos, para saber si podíamos realizar un *test t*. Realizamos el mismo planteo que en la hipótesis anterior, por tanto iniciamos buscando normalidad en los datos utilizando el test de Shapiro-Wilk, el cual dio que no existía normalidad en los datos (mala: p-valor = 0.000, mediana: p-valor = 0.000, buena: p-valor = 0.000). Luego aplicamos el test de Levene buscando homocedasticidad, la cual existía para uno de los tres test realizados, debido a que el grupo mala/mediana calidad se aceptó la hipótesis (p-valor = 0.277), pero mediana/buena calidad (p-valor = 0.000) y mala/buena calidad (p-valor = 0.000) no. Por tanto, como existe homocedasticidad en el par mala/mediana calidad, procedimos a realizar un test de Mann-Whitney y dos test de Kruskal-Wallis para los otros grupos, dando como resultado que el grupo de mala/mediana calidad (p-valor = 0.875) no presentaba diferencias significativas como mediana/buena (p-valor = 0.000) y mala/buena calidad (p-valor = 0.000) que sí.

Con estos resultados, es factible decir que existen diferencias significativas entre el grupo de buena con respecto a mediana y baja calidad, aunque estos datos se explican en la población. Para comprobar que realmente se cumpliera la hipótesis para ambos tipos de uvas (lo lógico sería decir que sí debido a que la correlación de Spearman dió moderada para ambos, lo cual indicaría que si se cumple para la población se debería cumplir para los grupos que la representan), decidimos realizar el análisis por separado, agrupando la calidad en dos partes: mediana/baja y mediana/alta.

Realizando el estudio por cada tipo de uva, utilizando los mismos procedimientos antes explicados, Shapiro-Wilk dio que las distribuciones no eran normales (Riesling: Mediana/Baja - p-valor = 0.000, Mediana/Alta - p-valor = 0.000. Garnacha: Mediana/Baja - p-valor = 0.000, Mediana/Alta - p-valor = 0.000), pero el test de Levene arrojó que las distribuciones para ambos vinos eran homocedásticas (Riesling: p-valor = 0.056, Garnacha: p-valor = 0.064), por lo cual procedimos a realizar Mann-Whitney a ambas, arrojando en ambos que se rechazan las hipótesis nulas (Riesling: p-valor = 0.000, Garnacha: p-valor = 0.000) dando por entendido que existen diferencias significativas entre los grupos de mediana/baja y mediana/alta para ambos tipos de uva, lo cual indica que nuestra hipótesis planteada estaba en lo cierto. Esto indica que, los vinos de buena calidad tienen significativamente menos cloruros, y por tanto podemos decir que un vino se ve afectado por la cantidad de cloruros que tenga.

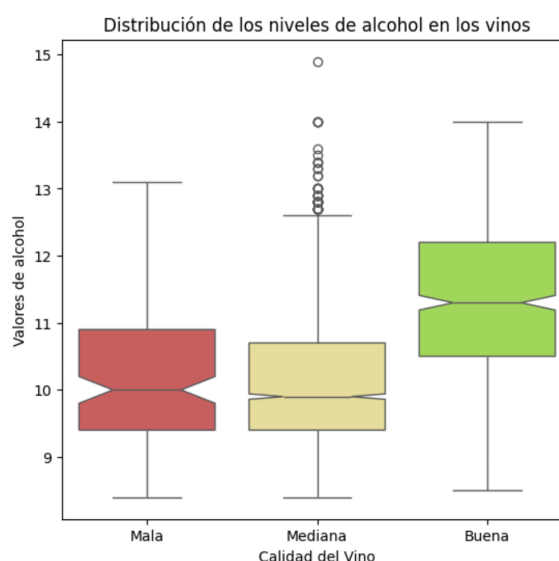


Los vinos de buena calidad tienen más alcohol

Esta hipótesis fue extraída del análisis previamente explicado, donde el valor de la correlación en la matriz de Spearman con la variable calidad también fue moderado. En este caso, el alcohol tenía un nivel de correlación de 0.42 en los vinos de uva Riesling y 0.48 en los vinos de uva Garnacha.

Para comprobar nuestra hipótesis, necesitábamos testear normalidad y homocedasticidad de los datos, para saber si podíamos realizar un *test t*. Realizamos el mismo planteo que en la hipótesis anterior, por tanto iniciamos buscando normalidad en los datos utilizando el test de Shapiro-Wilk, el cual dio que no existía normalidad en los grupos (mala: p-valor = 0.001, mediana: p-valor = 0.000, buena: p-valor = 0.000). Luego aplicamos el test de Levene buscando homocedasticidad, la cual existía para uno de los tres test realizados, debido a que el grupo mala/mediana calidad se aceptó la hipótesis (p-valor = 0.826), pero mediana/buena calidad (p-valor = 0.000) y mala/buena calidad (p-valor = 0.008) no. Por tanto, como existe homocedasticidad en el par mala/mediana calidad, procedimos a realizar un test de Mann-Whitney y dos test de Kruskal-Wallis para los otros grupos, dando como resultado que el grupo mala/mediana calidad (p-valor = 0.993) no representaba una diferencia significativa como mediana/buena (p-valor = 0.000) y mala/buena calidad (p-valor = 0.000). Esto indica que, es factible decir que existen diferencias significativas entre el grupo de buena con respecto a mediana y baja calidad, aunque estos datos se explican en la población. Para comprobar que realmente se cumpliera la hipótesis para ambos tipos de uvas (lo lógico sería decir que sí debido a que la correlación de Spearman dió moderada para ambos, lo cual indicaría que si se cumple para la población se debería cumplir para los grupos que la representan), decidimos realizar el análisis por separado, agrupando la calidad en dos partes: mediana/baja y mediana/alta, como en la hipótesis anterior.

Realizando el estudio por cada tipo de uva, utilizando los mismos procedimientos antes explicados, Shapiro-Wilk dio que las distribuciones no eran normales (Riesling: mediana/baja - p-valor = 0.000, mediana/alta - p-valor = 0.000. Garnacha: mediana/baja - p-valor = 0.000, mediana/alta - p-valor = 0.000), pero el test de Levene arrojó que las distribuciones para ambos vinos no eran homocedásticas (Riesling: p-valor = 0.000, Garnacha: p-valor = 0.000), por lo cual procedimos a realizar Kruskal-Wallis a ambas, arrojando en ambos que se rechazan las hipótesis nulas (Riesling: p-valor = 0.000, Garnacha: p-valor = 0.000) dando por entendido que existen diferencias significativas entre los grupos de mediana/baja y mediana/alta para ambos tipos de uva. Esto indica que nuestra hipótesis planteada estaba en lo cierto, que los vinos de buena calidad tienen significativamente más porcentaje de alcohol, y por tanto podemos decir que un vino se ve afectado por la cantidad del mismo que tenga.

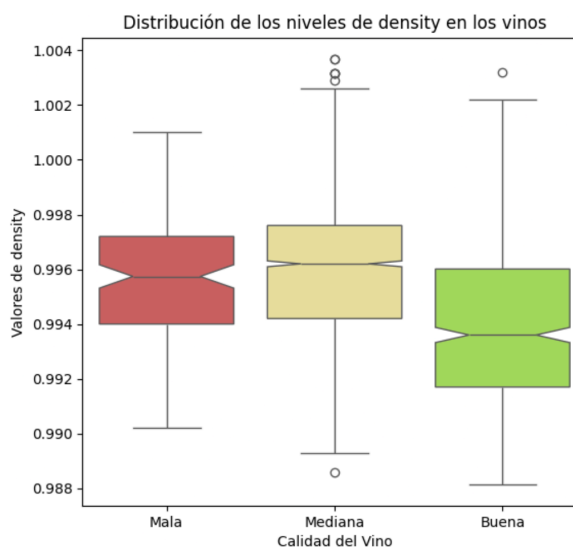


Los vinos de buena calidad tienen menos densidad

Esta hipótesis fue extraída del análisis previamente explicado, donde el valor de la correlación en la matriz de Spearman con la variable calidad también fue moderado. En este caso, el alcohol tenía un nivel de correlación de -0.32 en los vinos de uva Riesling y -0.18 en los vinos de uva Garnacha.

Para comprobar nuestra hipótesis, necesitábamos testear normalidad y homocedasticidad de los datos, para saber si podíamos realizar un *test t*. Realizamos el mismo planteo que en la hipótesis anterior, por tanto iniciamos buscando normalidad en los datos utilizando el test de Shapiro-Wilk, el cual dio que existía normalidad para una de las tres distribuciones (mala: p-valor = 0.562, mediana: p-valor = 0.000, buena: p-valor = 0.000). Luego aplicamos el test de Levene buscando homocedasticidad, la cual existía para uno de los tres test realizados, debido a que el grupo mala/mediana calidad se aceptó la hipótesis (p-valor = 0.507), pero mediana/buena calidad (p-valor = 0.000) y mala/buena calidad (p-valor = 0.000) no. Casi podemos realizar un *test t*, pero como la normalidad solo se probaba para un conjunto debimos, por tanto decantarnos a realizar otro test, y como existía homocedasticidad en el par mala/mediana calidad, procedimos a realizar un test de Mann-Whitney y dos test de Kruskal-Wallis, dando como resultado que el grupo de mala/mediana calidad (p-valor = 0.156) no tenía una diferencia significativa a comparación de mediana/buena (p-valor = 0.000) y mala/buena (p-valor = 0.000) que si. Por tanto, es factible decir que existen diferencias significativas entre el grupo de buena con respecto a mediana y baja calidad, aunque estos datos se explican en la población. Para comprobar que realmente se cumpliera la hipótesis para ambos tipos de uvas (lo lógico sería decir que si debido a que la correlación de Spearman dió moderada para ambos, lo cual indicaría que si se cumple para la población se debería cumplir para los grupos que la representan), decidimos realizar el análisis por separado, agrupando la calidad en dos partes: Mediana/Baja y Mediana/Alta, como en la hipótesis anterior.

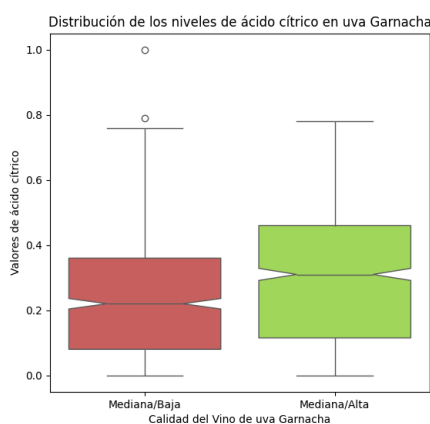
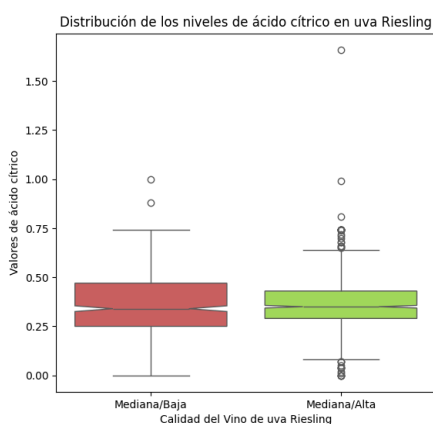
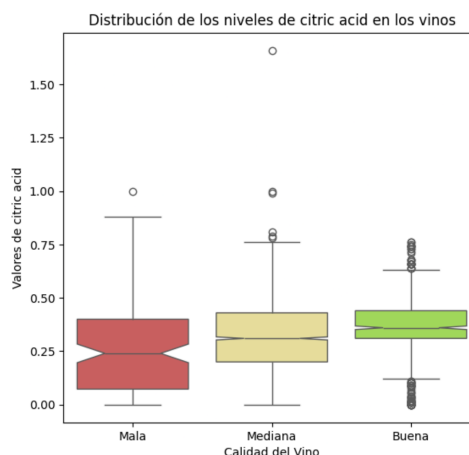
Realizando el estudio por cada tipo de uva, utilizando los mismos procedimientos antes explicados, Shapiro-Wilk dio que las distribuciones no eran normales (Riesling: mediana/baja - p-valor = 0.000, mediana/alta - p-valor = 0.000. Garnacha: mediana/baja - p-valor = 0.000, mediana/alta - p-valor = 0.005), y el test de Levene arrojó que las distribución de los vinos Riesling eran homocedástica (p-valor = 0.292), en cambio en vinos Garnacha no (p-valor = 0.000), por lo cual procedimos a realizar Mann-Whitney para la uva Riesling y Kruskal-Wallis para la uva Garnacha, arrojando en ambos que se rechazan las hipótesis nulas (Riesling: p-valor = 0.000, Garnacha: p-valor = 0.000) dando por entendido que existen diferencias significativas entre los grupos de mediana/baja y mediana/alta para ambos tipos de uva. Esto indica que nuestra hipótesis planteada estaba en lo cierto, que los vinos de buena calidad tienen significativamente menos porcentaje de densidad y por tanto podemos decir que un vino se ve afectado por la densidad que tenga.



Mientras más alto es el ácido cítrico, mejor es la calidad del vino

Esta hipótesis fue extraída del análisis, pero no de la matriz de correlación. En este caso, vimos las distribuciones de los boxplots, los cuales parecían brindar una diferencia significativa en los valores de acidez cítrica en los distintos grupos de calidad. En cambio, los valores de la correlación en la matriz de Spearman con la variable calidad resultaron moderados para la uva Garnacha (0.21), pero no para la uva Riesling (0.02), aunque observando las gráficas nos planteamos realizar el análisis igualmente.

Para comprobar nuestra hipótesis, necesitábamos testear normalidad y homocedasticidad de los datos, para saber si podíamos realizar un *test t*. Realizamos el mismo planteo que en la hipótesis anterior, por tanto iniciamos buscando normalidad en los datos utilizando el test de Shapiro-Wilk, el cual dio que no existía normalidad para ninguna de las distribuciones (mala: p-valor = 0.000, mediana: p-valor = 0.000, buena: p-valor = 0.000). Luego aplicamos el test de Levene buscando homocedasticidad, la cual no existía para ninguna distribución (mala/mediana calidad - p-valor = 0.008, mediana/buena calidad - p-valor = 0.000, mala/buena calidad - p-valor = 0.000). Como no teníamos homocedasticidad ni normalidad, planteamos realizar tres test de Kruskal-Wallis, dando como resultado que los tres grupos rechazaban las hipótesis (mala/mediana calidad - p-valor = 0.001, mediana/buena calidad - p-valor = 0.000, mala/buena calidad - p-valor = 0.000), por tanto es factible



decir que existen diferencias significativas entre los tres grupos de calidad, aunque estos datos se explican en la población. Para comprobar que realmente se cumpliera la hipótesis para ambos tipos de uvas (analizando gráficas de las distribuciones según

cada tipo de uva, se observaba que esto no se cumplía para las dos), decidimos realizar el análisis por separado, agrupando la calidad en dos partes: mediana/baja y mediana/alta, como en la hipótesis anterior.

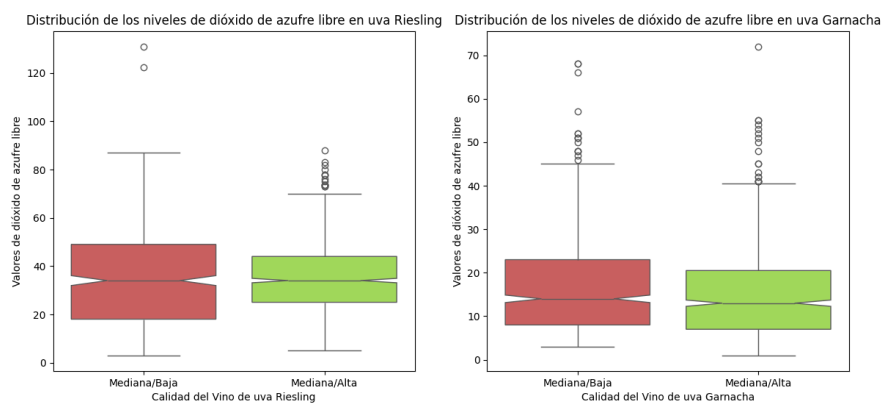
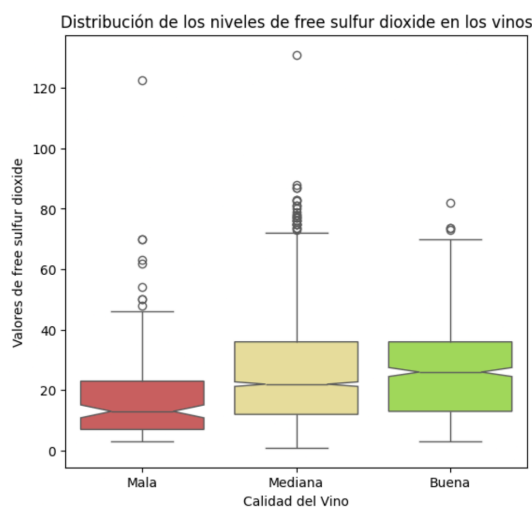
Realizando el estudio por cada tipo de uva, utilizando los mismos procedimientos antes explicados, Shapiro-Wilk dio que las distribuciones no eran normales (Riesling: mediana/baja - p-valor = 0.000, mediana/alta - p-valor = 0.000. Garnacha: mediana/baja - p-valor = 0.000, mediana/alta - p-valor = 0.000), y el test de Levene arrojó que las distribuciones para ambos vinos no eran homocedásticas (Riesling: p-valor = 0.000, Garnacha: p-valor = 0.000), por lo cual procedimos a realizar Kruskal-Wallis para ambas, arrojando que se rechazaba la hipótesis nula para la uva Garnacha (p-valor = 0.000) pero no para la uva Riesling (p-valor = 0.133) lo cual implica que la visualización era cierta, y por tanto los vinos Riesling no se ven afectados por los niveles de acidez cítrica. Lo que se concluye es que nuestra hipótesis planteada es falsa, dando por entendido de que aunque en uno de

los dos vinos puede verse influido por la variable, como no es así para ambos, no es posible decir que los vinos se ven afectados por los niveles de ácido cítrico.

Mientras más alta es la cantidad de sulfitos libres, mejor es la calidad de un vino

Esta hipótesis fue extraída del análisis, pero no de la matriz de correlación. Utilizando información del dominio vimos que al aumentar la cantidad de sulfitos libres mejora la acción conservante del vino, evitando el crecimiento microbiano, lo que está relacionado con acidez volátil. El aumento de los sulfitos totales no acompañado del aumento de los sulfitos libres es una mala señal, ya que se incrementan los niveles de dióxido de azufre combinado, lo que anula los efectos del otro. En este caso, vimos las distribuciones de los boxplots, los cuales parecían brindar una diferencia significativa en los valores de sulfitos libres en los distintos grupos de calidad. En cambio, los valores de la correlación en la matriz de Spearman con la variable calidad resultaron bajos tanto para la uva Riesling (0.03) como para la uva Garnacha (-0.05).

Para comprobar nuestra hipótesis, necesitábamos testear normalidad y homocedasticidad de los datos, para saber si podíamos realizar un *test t*. Realizamos el mismo planteo que en la hipótesis anterior, por tanto iniciamos buscando normalidad en los datos utilizando el test de Shapiro-Wilk, el cual dio que no existía normalidad para ninguna de las distribuciones (mala: p-valor = 0.000, mediana: p-valor = 0.000, buena: p-valor = 0.000). Luego aplicamos el test de Levene buscando homocedasticidad, la cual existía para el grupo de mediana/buena calidad (p-valor = 0.112) pero para mala/mediana calidad (p-valor = 0.010) y Mala/Buena calidad (p-valor = 0.035) no. Realizamos un test de Mann-Whitney y dos de Kruskal-Wallis, dando como resultado que mala/mediana (p-valor =

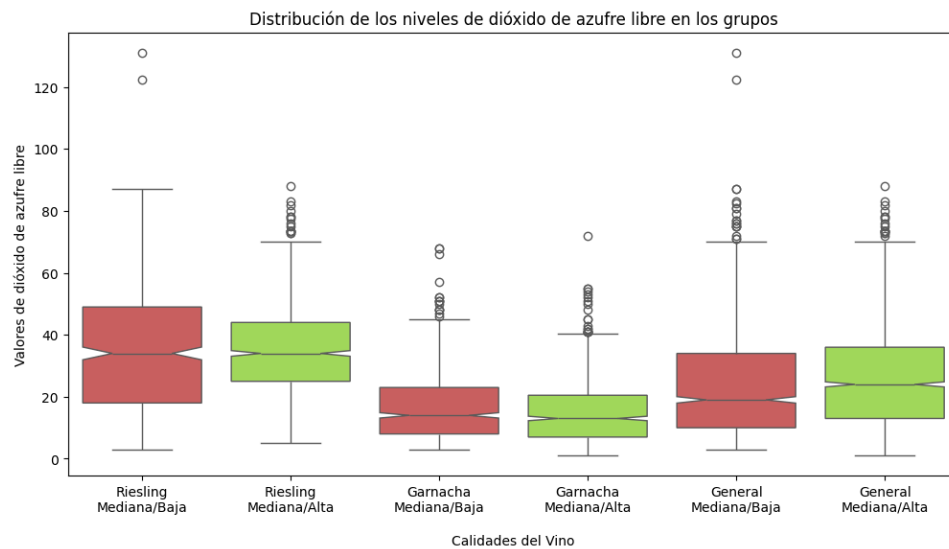


0.000), mediana/buena calidad (p-valor = 0.038) y mala/buena calidad (p-valor = 0.000) rechazarán las hipótesis nulas, por tanto es factible decir que existen diferencias significativas entre los tres grupos de calidad, aunque estos datos se explican en la población. Para

comprobar que realmente se cumpliera la hipótesis para ambos tipos de uvas (analizando gráficas de las distribuciones según cada tipo de uva, se observaba que esto no se cumplía para las dos), decidimos realizar el análisis por separado, agrupando la calidad en dos partes: mediana/baja y mediana/alta, como en la hipótesis anterior.

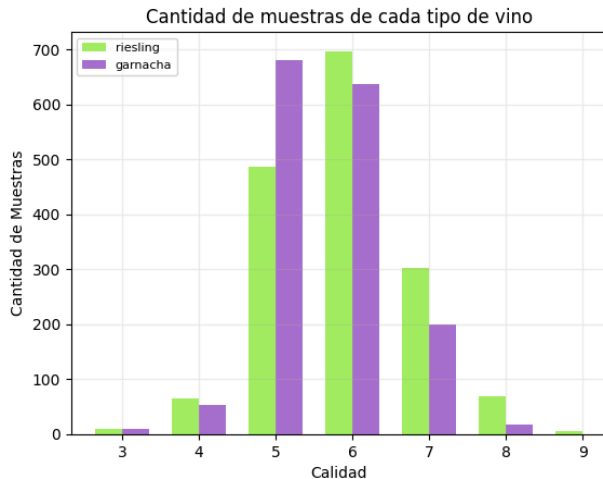
Realizando el estudio por cada tipo de uva, utilizando los mismos procedimientos antes explicados, Shapiro-Wilk dio que las distribuciones no eran normales (Riesling: mediana/baja - p-valor = 0.000, mediana/alta - p-valor = 0.000. Garnacha: mediana/baja - p-valor = 0.000, mediana/alta - p-valor = 0.000), y el test de Levene arrojó que las distribuciones para ambos vinos no eran homocedásticas (Riesling: p-valor = 0.000, Garnacha: p-valor = 0.040), por lo cual procedimos a realizar Kruskal-Wallis para ambas, arrojando que se rechazaba la hipótesis nula para la uva Garnacha

(p-valor = 0.033) pero no para la uva Riesling (p-valor = 0.154), lo cual implica que la visualización era cierta, y por tanto los vinos Riesling no se ven afectados por los niveles de sulfitos libres. Lo que se concluye es que nuestra hipótesis planteada es falsa, dando por entendido de que aunque en uno de los dos vinos puede verse influido por la variable, como no es así para ambos, no es posible decir que los vinos se ven afectados por los niveles de sulfitos libres. Aunque antes de cerrar con la hipótesis, cabe aclarar que nos surgió un debate con esta variable, porque cuando agrupamos los dos tipos de uva, gráficamente se puede observar que la tendencia de la población es distinta a la de los grupos que la conforman. Esto lo dejamos a libre interpretación del lector a modo de debate. De nuestra parte, investigamos al respecto y concluimos que puede haber una relación con la “Paradoja de Simpson”, que a modo general describe a las tendencias poblacionales que luego desaparecen al separar al grupo total de individuos o muestras en clases. Igualmente, dejamos abierta cualquier posibilidad al respecto.



Los vinos de uva Garnacha, tienen mejor calidad que los vinos de uva Riesling

Esta hipótesis resulta bastante intuitiva a la hora de analizar un conjunto de datos de vinos porque plantea si es que hay un vino de un tipo de uva que tiene significativamente mejor media que otro. En un inicio, comparamos las calidades de cada vino según la uva, y parecía haber una tendencia en los vinos Riesling a obtener mejores resultados en promedio y de hecho calculamos las medias de ambos grupos, siendo la de Riesling de 5.88 y la otra de 5.64, lo que nos resultó bastante parecido.



Posteriormente, cuestionamos si realmente tenía sentido formular esta hipótesis, considerando que el tipo de uva podría ser independiente de la calidad. Dicho de otra forma, ¿qué utilidad tendría saber que el promedio de calidad es mayor para una uva en comparación con otra si el tipo de uva no influye en la calidad final? Si se demostrara que el tipo de uva afecta la calidad, podríamos priorizar la producción de la uva que ofrezca mejores resultados en promedio. Por este motivo, antes de realizar cualquier prueba estadística, necesitábamos formular la hipótesis sobre si existe o no una relación entre

el tipo de uva y la calidad. Esta segunda, es la misma que planteamos abajo.

Una vez que determinamos que el tipo de uva sí afectaba a la calificación resultante del vino, es que investigamos al respecto de qué podíamos hacer para probar la significancia de una media respecto de otra. Claramente, no tenía sentido aplicar una prueba de hipótesis que tenga como supuestos la normalidad, homocedasticidad o que las variables sean continuas porque nosotros teníamos una variable cualitativa ordinal. Encontramos de diversos lugares que, la prueba de Kruskal-Wallis obtenía buenos resultados cuando se trataba de variables de este tipo. Por lo que, así fue cómo planteamos el test que resultó en el rechazo de la hipótesis nula ($p\text{-valor} = 0.000$). Este resultado se interpretaría como que la media calculada de Riesling es significativamente mayor a la de Garnacha (contra intuitivamente), por lo que, en general, los vinos Riesling obtienen mejores resultados.

No hay una relación entre el tipo de uva y la calidad

Esta hipótesis fue extraída del análisis de los datos. En un principio, cuando recién mirábamos las distribuciones de los datos, comparábamos cuál era la proporción de vinos de Riesling y Garnacha según la calidad, nos surgió la duda de si era posible que la uva influyera en la calidad realmente. Nuestra idea principal es que no debería influir porque la calidad se evaluaba según los atributos químicos de las uvas, pero no teniendo en cuenta la uva misma.

Dado que no podemos corroborar alguna asociación como si podemos hacerlo con las correlaciones de Pearson por dar un ejemplo, es que buscamos otros métodos. Si no existiera una relación entre ambas variables, entonces la probabilidad de encontrar un vino de una determinada calidad y uva, dependería únicamente de la probabilidad de encontrar un vino de tal calidad multiplicada por la probabilidad de encontrar un vino de tal uva.

Para comprobar la hipótesis, realizamos una prueba de Chi cuadrado, cuyo resultado fue el rechazo de la misma ($p\text{-valor} = 0.000$). Esto significa que hay diferencias significativas entre la frecuencia esperada de cada par de tipo de uva y calidad (la probabilidad esperada multiplicada por la cantidad de muestras) y la realmente encontrada. También, recuperamos los residuos del test, y comprobamos que para las calidades 7, 8 y 9, los vinos Riesling fueron más frecuentes (residuos iguales a 3.76 para las primeras dos, e igual a 1.57 para la tercera), mientras que los Garnacha lo fueron para la calidad 5 (residuo igual a 4.31). La interpretación del resultado es que, contrario a lo que pensábamos, el tipo de uva aumenta la probabilidad de obtener más ciertos puntajes que otros de forma significativa. En este caso, los Riesling están asociados a mejores calificaciones.

La calidad de un vino está determinada, en general, por el alcohol, los cloruros, densidad y ácido volátil.

Habiendo planteado lo anterior, sabemos que en la población general de vinos hay tendencias que nos ayudan a explicar en parte que la calidad puede estar determinada por ciertos factores como lo son alcohol, el ácido volátil, la densidad o los cloruros, pero para tener vinos de calidad... ¿únicamente miramos estas variables generales o realmente debemos tener en cuenta cuáles son los atributos químicos que afectan la calidad de forma específica en cada tipo de uva?

Sabemos, por ejemplo, cuando estudiamos el comportamiento del ácido cítrico y de los sulfitos libres que, los grupos de vinos Garnacha de media/buena calidad tienen más concentraciones ácido cítrico que los de media/baja, lo que no ocurre con los Riesling. También probamos que, el tipo de uva afecta al puntaje obtenido, por lo que, suena lógico pensar que esto se debe a las características químicas que distinguen a Riesling de Garnacha. Sin embargo, esto no es suficiente como para hacer alguna afirmación porque podrían haber relaciones entre las variables que no permitan, por ejemplo, que haya niveles de cloruros, ácidos volátiles y ácidos cítricos bajos al mismo tiempo (en el caso de Garnacha). Por este motivo, es que optamos por un análisis multivariado donde todas las variables entren en juego y con el cual determinemos si el alcohol, densidad, cloruros y ácidos volátiles marcan o no un camino común donde muestras de ambas uvas converjan en vinos buenos.

Lo primero que hicimos fue visualizar los datos con PCA sobre todos los vinos, para ver si hay un patrón de calidad del cual nos podríamos basar para afirmar que todo vino que siga la dirección de ese patrón resulte bueno. Sin embargo, si bien PCA fue útil para ver que Garnacha y Riesling se encuentran separados en el espacio (por sus propiedades que los distinguen), no contaba con una varianza explicada confiable (menor a 0.6 con 3 componentes principales, probablemente por la poca linealidad entre las variables). Como alternativa, probamos con t-SNE, con el cual pudimos ver ciertos patrones de calidad no muy evidentes y notar nuevamente la distancia espacial entre los tipos de uva, y con UMAP que no dió resultados muy distintos con la diferencia de que parecían verse grupos dentro de Riesling y Garnacha. Lo más importante fue ver que parecía que los vinos Garnacha de buena calidad se concentraban en puntos distintos a los de Riesling apoyando el rechazo de la hipótesis.

Posteriormente, probamos visualizar las distribuciones de ambas uvas por separado con los mismos métodos. Se hicieron más evidentes los patrones de calidad y la división en UMAP. PCA volvió a fallar en términos de confiabilidad.

En este punto, ya teníamos algo pero no tan claro, por lo que, aplicamos algoritmos de Clustering para ver cómo se podrían llegar agrupar los datos y si había una relación entre las calidades y las características químicas resultantes de cada cluster. Para ello, aplicamos K Means y clustering jerárquico aglomerativo y los comparamos con el índice de Davies-Boulding y el coeficiente de Silueta quedándonos mejor el primero. Para determinar los clústers en K Means utilizamos el Elbow Plot, mientras que en el clustering jerárquico nos quedamos con los parámetros que nos conseguían los resultados más óptimos.

Por último y teniendo ya los clusters, vimos que cada uno de estos coincidía (espacialmente) con los grupos vistos en la reducción de la dimensionalidad. Analizando las calidades promedio de cada uno, también notamos que el clúster de mejor media era mayormente Riesling y que encerraba, parcialmente, al patrón de calidad antes mencionado. Por otro lado, el segundo clúster con mejor media encerraba parcialmente al patrón calidad de los vinos Garnacha y estaba compuesto en gran medida por vinos de este tipo de uva. Luego, procedimos a graficar los boxplots de cada una de las variables agrupando por clúster y es de esta forma que notamos que el clúster con más media era el que más se acomodaba a los niveles de alcohol, densidad, cloruros y ácidos volátiles vistos durante las pruebas de hipótesis anteriores. Sin embargo, su calificación media, no estaba dentro de lo que

considerábamos el grupo de los vinos de buena calidad, así que, tenía que ser porque había otras variables en juego que también determinaban la calidad de los Riesling en cierto punto.

Cerrando todas las ideas e intentando explicar por qué el cluster con segunda mejor media tenía una calificación promedio similar al primero siendo que este no cumplía muy bien los parámetros enunciados del alcohol, densidad, cloruros y ácidos volátiles, vimos, mirando los boxplots (los de la sección de “visualización de grupos”) que sus propiedades químicas tenían bastante similitud con la de los vinos Garnacha de buena calidad (como en el caso de ácido cítrico). Reforzando, de nuevo, la idea de que no podemos determinar una combinación química que sirva de criterio general de calidad para todos los vinos, sino que, pareciera haber más de una combinación válida para lograr la calidad. Además, graficamos con t-SNE para ver dónde estaban los vinos buenos y notamos no sólo que parecía haber una coincidencia con los dos clústers de mejor media, sino que, había algunos vinos Garnacha que tendían al patrón de calidad de los Riesling y que en general, la calidad de estos últimos parecía dispersarse más, de nuevo, apoyando el rechazo de nuestra hipótesis original.

Existe una relación lineal entre la densidad, el grado de alcohol y los azúcares residuales en los vinos Riesling

Esta hipótesis fue extraída en el dominio, debido a que durante la elaboración del vino, la levadura consume el azúcar de la uva y produce etanol, lo que se traduce a que se genera más porcentaje de alcohol. A su vez, los vinos que tienen menos graduación de alcohol, presentan una densidad más baja. En cambio, aquellos que tienen un nivel de azúcar más alto, presentan también una mayor densidad. Por lo tanto, el grado de alcohol es un factor del vino que 'tira' de la densidad para abajo y el azúcar residual es otro elemento que 'tira' de él para arriba.

Si tenemos en cuenta que los vinos de más calidad tienen más graduación alcohólica y menos densidad, como aclara el dominio, entonces, si encontramos alguna relación lineal entre el alcohol y la densidad, podríamos tener información importante sobre cómo mejorar la calidad de un vino.

Sabemos que los vinos de buena calidad tienen menos densidad y más alcohol sin importar el tipo de uva. Por tanto realizamos el análisis sobre la uva Riesling y viendo que la correlación entre el alcohol y la densidad era alto (0.76), aplicamos una regresión lineal, teniendo como variable objetivo densidad y predictora alcohol. A pesar de tener una buena R^2 (0.584), los residuos no cumplían los supuestos de normalidad (Omnibus: p-valor = 0.000, Jarque-Bera: p-valor = 0.000), implicando que la regresión planteada no era útil para describir la densidad.

Como la correlación con la azúcar residual y la densidad también era importante (0.8), realizamos también la regresión lineal, teniendo como variable predictora la azúcar y objetivo a densidad. El test tampoco dió un buen resultado a pesar de tener una buena R^2 (0.746), debido a que los supuestos de normalidad de los residuos no se cumplían (Omnibus: p-valor = 0.000, Jarque-Bera: p-valor = 0.000).

Por último, probamos realizar la regresión lineal con ambas variables, intentando así ver si ambas podían predecir mejor que por separado. Este test tampoco dió buenos resultados, a pesar de tener una muy buena R^2 (0.908) que explica casi toda la variable, pero el modelo no es útil debido a que los residuos no cumplían los supuestos de normalidad (Omnibus: p-valor = 0.000, Jarque-Bera: p-valor = 0.000).

Luego de realizar las regresiones lineales, podemos decir que la hipótesis que planteamos no tiene validez, debido a que no se puede predecir la variable densidad con ninguna de los dos factores con los cuales estaba más correlacionado. Aun así, quisimos probar si podíamos predecir esta variable utilizando casi todos los atributos que tiene el vino. Por lo tanto, realizamos una última regresión lineal, con densidad como nuestra variable dependiente nuevamente, y utilizando todas las variables que podía tener una muestra, exceptuando la calidad, y el tipo de vino. Este análisis nos brindó el mismo resultado que los anteriores, siendo que teníamos una muy buena R^2 (0.968), el modelo sigue sin ser útil porque los residuos no cumplían los supuestos de normalidad (Omnibus: p-valor = 0.000, Jarque-Bera: p-valor = 0.000).

Existe una relación lineal entre la acidez fija, el pH, la acidez cítrica y densidad en los vinos Garnacha

Esta hipótesis fue extraída de todo lo que vimos durante el análisis multivariado, el análisis de correlaciones y los test de hipótesis, donde sabemos que hay relación entre los vinos de buena calidad y las cantidades de ácido cítrico en Garnacha. También vimos que la acidez fija está correlacionada con el pH de forma inversa (-0.71) y con la acidez cítrica (0.66) y la densidad (0.62) proporcionalmente, y además, cuando realizamos la visualización de grupos, estos gráficos arrojaban que cuando una crecía (en términos de niveles) las otras parecían bajar o subir respectivamente. En el clustering observamos lo mismo. Aunque, como en la hipótesis anterior lo explicamos, la densidad también depende de los valores de alcohol y azúcar residual, y puede ser por esta razón que la densidad se mantenga baja en los vinos de calidad a pesar de los niveles de acidez fija y ácido cítrico.

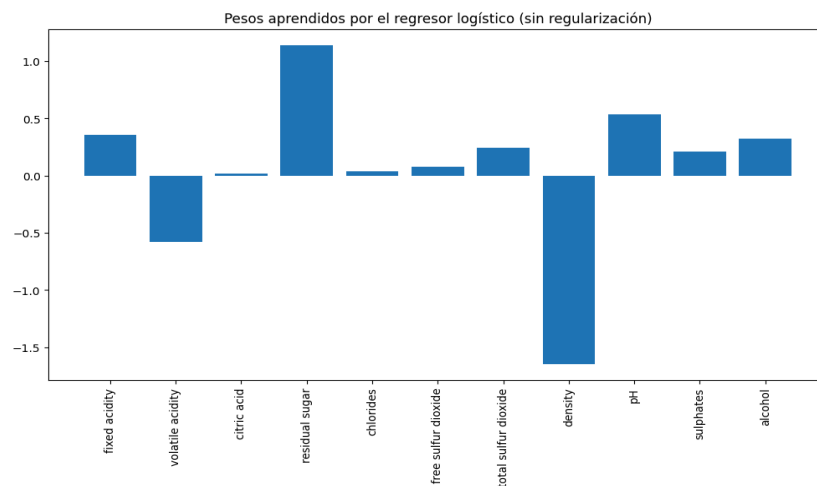
Saber si existe algún tipo de proporcionalidad lineal puede llegar a ser muy útil para los fabricantes de vinos porque podría ser más fácil predecir la calidad resultante y controlar los niveles de cada compuesto. Por esto, iniciamos probando realizar una regresión lineal de variables únicas como predictoras y la acidez fija como objetivo. Los tres test realizados, los cuales las variables eran ácido cítrico con $R^2 = 0.451$, pH con $R^2 = 0.466$ y densidad con $R^2 = 0.466$ dieron negativos (Omnibus: p-valor = 0.000, Jarque-Bera: p-valor = 0.000), lo cual indica que la acidez fija no se puede predecir únicamente por una sola variable.

Luego realizamos una regresión lineal, utilizando como objetivo la acidez fija y como predictoras ácido cítrico, pH y densidad. Este test tampoco dio positivo, aunque la variable era bastante explicada por las otras tres ($R^2 = 0.466$), los residuos volvían a incumplir los supuestos de normalidad (Omnibus: p-valor = 0.000, Jarque-Bera: p-valor = 0.000). Por lo tanto, decidimos, como la hipótesis pasada, realizar una regresión con todas las variables que pueden entrar en juego para predecir la acidez fija, y aún así no se explicaba en su totalidad ($R^2 = 0.871$), sino en gran parte, y tampoco se cumplían los supuestos de normalidad en los residuos (Omnibus: p-valor = 0.000, Jarque-Bera: p-valor = 0.000). Por tanto, la hipótesis que planteamos es falsa, debido a que no es posible predecir la acidez fija.

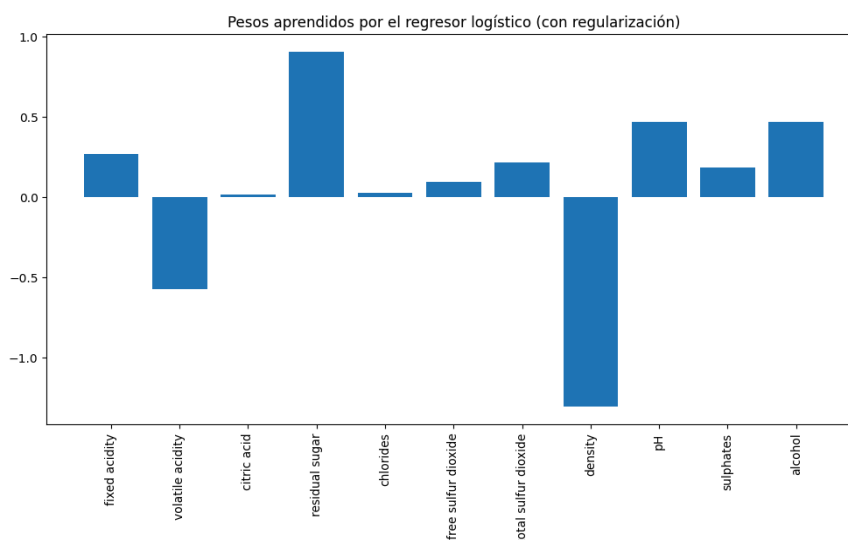
Existen determinados atributos que determinan la calidad de los vinos Riesling

Esta hipótesis la determinamos con el fin de intentar buscar qué variables eran las más influyentes a la hora de determinar la calidad de un vino Riesling que, como comprobamos antes, tiende a tener mejores calificaciones. Está bien explicado en el transcurso de este informe que es muy difícil o poco probable poder determinar la calidad de un vino, ya que este depende de muchas características distintas, y ninguna es tan significativa como para realizar alguna aproximación estimada. Pero aun así, con todo en contra, planteamos utilizar regresión logística con el fin de encontrar esas variables que necesitamos para determinar la calidad. Como este tipo de regresión necesita utilizar una variable objetivo binaria, decidimos dividir la calidad de los vinos en dos grupos, mediano/bajo y mediano/alto (o binaria 0, 1).

La primera regresión que realizamos (sin utilizar regularizador), la probamos con una validación cruzada. Esta resultó bastante positiva, con un resultado de exactitud promedio de 0.72 y un desvío de la exactitud menor a 0.04, dando como variables más importantes la densidad, azúcar residual, acidez volátil, pH y alcohol. Es interesante notar que dentro de la lista se encuentra la densidad, la acidez volátil y el alcohol, y otras variables que coinciden con la visualización de los boxplots que realizamos en el material de trabajo.



La segunda regresión que realizamos (utilizando regularizador), también la probamos con una validación cruzada. Esta resultó bastante positiva (por no decir casi igual a la primera), con un resultado de exactitud promedio de 0.72, también dando como variables más importantes la densidad, azúcar residual, acidez volátil, pH, y alcohol.



Para probar el rendimiento de los modelos, probamos la predicción con el conjunto de datos de test y graficamos las matrices de confusión para cada regresión. Ambas fueron iguales, y lo que indicaron es que el modelo tiene una cierta efectividad a la hora de clasificar a un vino como medio/bueno pero no como medio/malo (casi un 50% de aciertos y fallos lo que no es deseable).

Finalmente, probamos como alternativa un árbol de decisión, el cual también fue entrenado con el 70% de los datos. El resultado en este caso fue ligeramente mejor, obteniendo una mejora del 2%. Cabe destacar que la variables más significativas de este modelo fueron el alcohol, la acidez volátil, los sulfitos libres, los cloruros y los sulfitos totales, otra vez, hay tres variables que trabajamos en las hipótesis.

La matriz de confusión en este caso, dió mejores resultados, distribuyendo mejor los aciertos y los fallos para cada clase.

En relación a lo que planteamos en la hipótesis, vemos que es mucho más complicado predecir la calidad de un vino de lo que parece. Cada modelo, interpretó de distinta forma lo que debía ser importante para la clasificación, obteniendo resultados similares. A modo general vemos que hay variables como el alcohol y la acidez volátil que parecen repetirse y que cabe destacar, son independientes del tipo uva. Lo más probable es que haya varias combinaciones que permiten que los vinos sean buenos y que no haya una certeza de lo que está bien y mal. Lo que sí se podría hacer, es analizar tendencias generales que tienen los vinos de calidad e intentar replicarlas.

Aclaración: podría haber una relación entre la efectividad de aciertos de cada clase y la proporción de vinos de mediana/alta y mediana/baja calidad en Riesling.

Conclusiones

No se puede determinar la calidad de un vino teniendo en cuenta un margen acotado de componentes químicas, pero tampoco todas ellas tienen la misma importancia a la hora de determinar la calificación. Existen distintas proporciones de atributos químicos que resultan en buenos vinos, lo que implica que no hay un criterio que determine con una alta precisión si un vino es de alta calidad solo observando las mediciones. Esto no quiere decir que no haya tendencias, por ejemplo, explorando el dataset pudimos descubrir que los vinos de buena calidad suelen tener más porcentaje de alcohol, menos densidad, ácido volátil y cloruros, esto sin importar el tipo de uva. Además, si se dividen los vinos por tipo, es posible divisar otras tendencias que distinguen a Riesling de Garnacha. Sin embargo, estos no son axiomas, sino que son, más bien, parámetros generales que se pueden seguir con el fin de obtener buenos resultados pero que no garantizan el éxito. Pueden haber múltiples combinaciones que sean válidas.

Referencias

Fixed Acidity

<https://waterhouse.ucdavis.edu/whats-in-wine/fixed-acidity>

Volatile Acidity

<https://www.bodegastrespiedras.com/acidez-del-vino/#:~:text=La%20acidez%20vol%C3%A1til%20calcula%20el,al%20vinagre%20y%20al%20barniz.>

Citric Acid

<https://www.cdrfoodlab.es/cdrwinelab/analisis/acido-citrico-vino>
<https://www.valtea.es/que-es-la-acidez-en-un-vino-y-que-importancia-tiene/>

Residual Sugar

[https://winedecoded.com.au/wine-words/residual-sugar/#:~:text=in%20new%20window\)-,Residual%20Sugar%20refers%20to%20the%20amount%20of%20sugar%20left%20in,Typically%202g%2FL%20or%20less.](https://winedecoded.com.au/wine-words/residual-sugar/#:~:text=in%20new%20window)-,Residual%20Sugar%20refers%20to%20the%20amount%20of%20sugar%20left%20in,Typically%202g%2FL%20or%20less.)
<https://catatu.es/blog/el-azucar-en-el-vino/#:~:text=Los%20vinos%20tranquilos%20se%20rigen%20por%20la%20cantidad,45%20g%2FL%204%20Dulce%3A%20m%C3%A1s%20de%2045%20g%2FL>

Chlorides

<https://foro.e-nologia.com/thread-37415-page-1.html>
<https://www.neuralword.com/es/cocina/hierbas-y-especias/origen-del-cloruro-en-el-vino-de-donde-proviene-este-elemento>

Free Sulfur Dioxide

[https://www.hannacolombia.com/blog/post/801/dioxido-azufre-libre-y-total-en-vinos-mediante-minitulador-hi-84500](https://www.hannacolombia.com/blog/post/801/dioxido-azufre-libre-y-total-en-vinos-mediante-minititulador-hi-84500)
<https://www.cdrfoodlab.es/cdrwinelab/analisis/dioxido-azufre-total-vino>

Total Sulfur Dioxide

<https://www.extension.iastate.edu/wine/total-sulfur-dioxide-why-it-matters-too/#:~:text=Simply%20put%2C%20Total%20Sulfur%20Dioxide%20%28TSO2%29%20is%20the,the%20wine%20such%20as%20aldehydes%2C%20pigments%2C%20or%20sugars>

Density

<https://pacolola.com/vendimia-2015-6-fermentacion-alcoholica/>

pH

<https://delariberavinos.com/blog/noticias/que-es-ph-vino#>
<https://vinotecavirtual.com/como-influye-el-ph-en-el-vino/>

Sulphates

https://www.scielo.cl/scielo.php?script=sci_arttext&pid=S0718-07642021000100017&lng=en&nrm=i&so&tlng=en

Alcohol

<https://www.debuenavid.es/blog/elaboracion-del-vino/graduacion-alcoholica-vino>

Hipótesis: Los vinos de uva Garnacha, tienen mejor calidad que los vinos de uva Riesling

<https://consensus.app/results/?q=Can%20I%20use%20Kruskal%20to%20test%20ordinal%20variables%3F&pro=on>