Máster Universitario de Ciencia de Datos

Tipología y ciclo de vida de los datos

Práctica 2: ¿Cómo realizar la limpieza y análisis de datos?

Autores:

Carlos Díez Domínguez
Jacobo Osorio Ríos

1. Descripción del dataset

El conjunto de datos con el que hemos decidido trabajar en esta práctica se llama *Wine Quality*, y ha sido extraído del repositorio de Machine Learning de la universidad UC Irvine.

El conjunto de datos fue donado por *Cortez et. al*., 2009, y se puede encontrar en el siguiente link.

En términos generales, este *dataset* presenta información sobre distintos test psicoquímicos realizados a 6499 *vinhos verdes* portugueses (1600 tintos y 4899 blancos). Las variables son las siguientes:

1. *fixed acidity:* gramos de ácido tartárico por litro.
2. *volatile acidity:* gramos de ácido acético por litro.
3. *citric acid:* gramos de ácido cítrico por litro.
4. *residual sugar:* gramos de azúcar residual por litro.
5. *chlorides:* gramos de cloruro de sodio por litro.
6. *free sulfur dioxide*: miligramos de dióxido de sulfuro libre por mililitro.
7. *total sulfur dioxide:* miligramos de dióxido de sulfuro total por mililitro.
8. *density:* densidad del vino en gramos por mililitro.
9. *pH:* pH del vino.
10. *sulphates:* gramos de sulfato potásico por litro.
11. *alcohol:* graduación alcohólica del vino en %.
12. *quality:* calidad (subjetiva) del vino.

Más allá de nuestro interés personal por el vino (y en concreto, por el *vinho verde*), este conjunto de datos nos parece muy completo tanto por el número de muestras estudiadas como por las variables que se han medido. Así pues, nos parece un conjunto relevante e interesante de estudiar, al menos en el campo de la etnología.

El conjunto de datos se halla dividido en dos, ya que tenemos información tanto para vinos tintos como para vinos blancos. Esta fue otra de las razones por las que nos pareció particularmente interesante el *dataset*, ya que además de medir si hay ciertas **propiedades del vino que puedan estar correlacionadas** entre sí, también podemos **evaluar cuáles de ellas son distintas en función del tipo de vino**, así como tratar de entrenar un clasificador que nos permita **predecir si un vino es tinto o blanco** en base a estas variables.

2. Integración y selección de los datos a analizar

Como hemos comentado anteriormente, el conjunto de datos se divide en dos datasets, por lo que es necesario juntarlos para realizar algunas de las tareas que nos hemos propuesto. Así pues, además de tener los archivos por separado, hemos decidido combinarlos generando una nueva columna llamada *type.* Esta columna, que toma valor 1 si el vino es blanco y valor 0 si el vino es tinto, nos permite distinguir a qué tipo de vino corresponden las muestras observadas.

3. Limpieza de los datos

a. Gestión de ceros y elementos vacíos

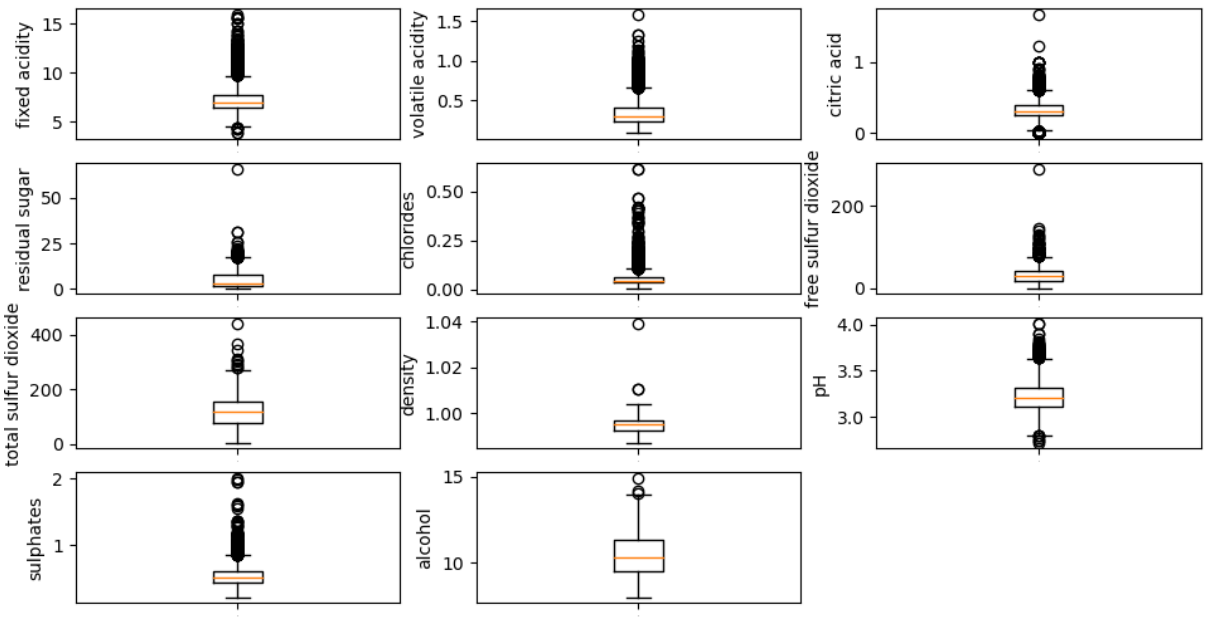
Afortunadamente, el **conjunto de datos está completamente limpio** en relación a ceros y elementos vacíos. Para comprobar esto, hemos hecho un código en Python que nos permite verificar si hay elementos vacíos en algunas de las columnas del conjunto de datos.

En cualquier caso, en el supuesto de que se hubiese presentado algún caso, llegamos a la conclusión de que considerando el tipo de variables que tenemos (todas numéricas), sería razonable imputar los valores utilizando la media de la variable. No obstante, también se podría haber optado por eliminar aquellas muestras que tuviesen algún valor desconocido (por supuesto, si no son excesivas), ya que el dataset es razonablemente grande y el número de muestras de cada tipo de vino está increíblemente desequilibrado.

Finalmente, también hemos decidido **prescindir de la variable *quality*** debido a su aspecto subjetivo, ya que lo que más nos interesa son las variables objetivamente cuantificables del vino.

b. Identificación y gestión de valores extremos

Para estudiar los valores extremos presentes en el conjunto de datos, hemos decidido generar un b*oxplot* para cada una de las columnas. Presentamos el resultado en la imagen inferior:



Como podemos observar, en algunas variables tenemos datos que se sitúan fuera del intervalo [*Q1* - 1.5**RI*, *Q3* + 1.5**RI*], donde *Q1* y *Q3* son los cuartiles 1 y 3, y el *RI* el rango intercuartílico. No obstante, en la mayoría de los casos los valores se encuentran muy próximos a estos intervalos, por lo que incluso desconociendo el significado de las variables podríamos ser bastante escépticos con respecto a considerar estos valores extremos como erróneos.

Afortunadamente, conociendo el significado de las variables (y gracias a cierto conocimiento que tenemos sobre vino), hemos concluido que no hay ningún valor observado que sea increíblemente extraño, incluso en aquellas variables donde se presentan valores más alejados del intervalo. Un ejemplo de esto podría ser la variable *residual sugar*, donde podemos observar un *outlier* bastante alejado del resto de datos. En este caso en particular, sabemos que este valor es muy normal en los denominados vinos dulces, que incluso pueden llegar a tener concentraciones de azúcar muy superiores. Probablemente, la única razón por la que tengamos ese único *outlier* es que de las muestras de vinos que tenemos, este sea el único vino particularmente dulce. En cualquier caso, para salir de dudas, convendría consultar a los productores de los vinos de las muestras (o en su defecto, el creador del conjunto de datos) si alguno de los vinos debería corresponderse con vinos dulces.

4. Análisis de los datos

a. Selección del grupo de datos

Puesto que el número de variables no es exageradamente elevado, nos ha parecido razonable intentar hacer un estudio teniendo en cuenta todas las variables del conjunto de datos. Considerando además que hay dos grupos claros en los que podemos dividir las muestras, hemos planteado un análisis en base a tratar los vinos blancos y tintos por separado.

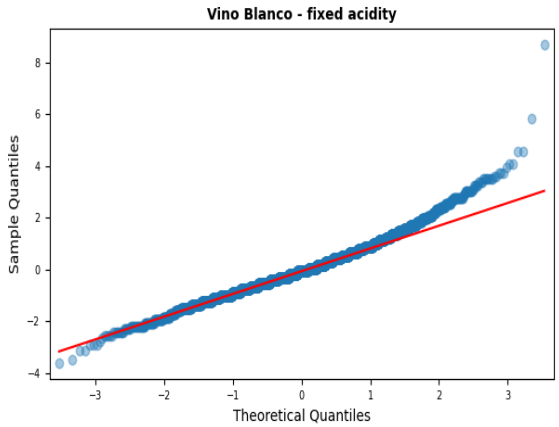
Así pues, además de comprobar la normalidad y homogeneidad de las variables nos ha parecido adecuado realizar las siguientes pruebas:

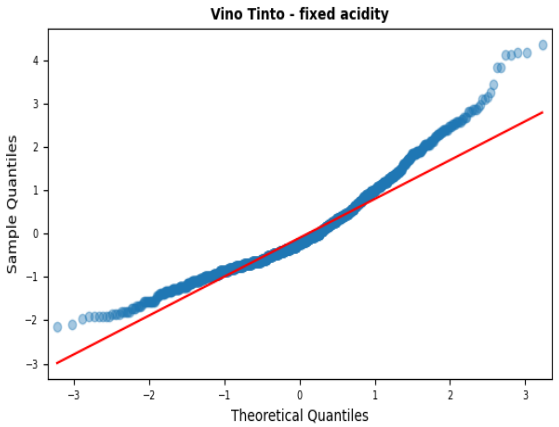
- Matriz de correlación de las variables para los dos tipos de vino.
- t-test para cada una de las variables con la intención de conocer si sus medias para vinos blanco y tinto tienen una diferencia significativa.
- Regresión logística que permita clasificar un vino como tinto o blanco.

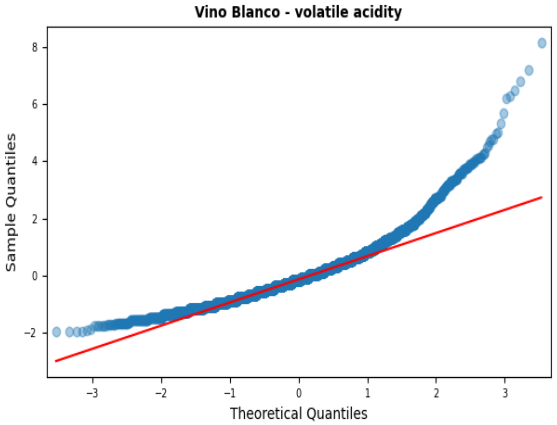
Con estas pruebas / análisis, nos parece que podemos tener una idea bastante clara de aquellas variables que tienen correlación entre sí, determinar cuáles de ellas son distintivas en cada tipo de vinos y finalmente tener una herramienta que nos permita predecir ante qué tipo de vino nos encontramos en caso de recibir una nueva muestra.

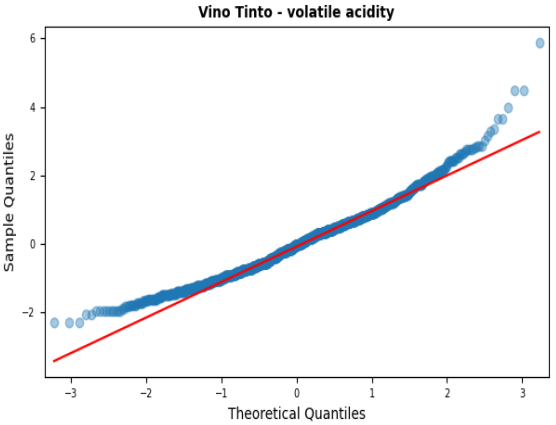
b. Comprobación de normalidad y homogeneidad de la varianza

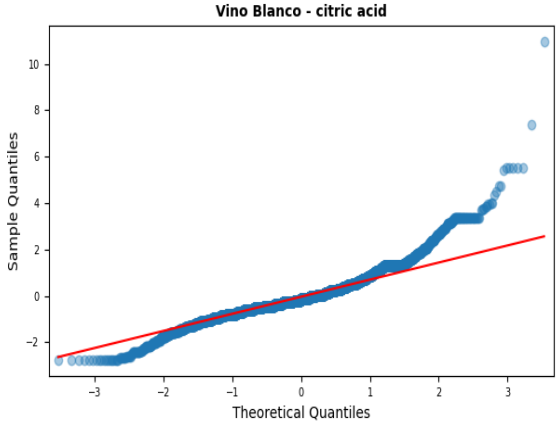
Primero vamos a mostrar las gráficas donde podemos **comprobar la normalidad de las variables**. Las mostramos en pares, a la izquierda vemos los vinos blancos y a la derecha tenemos su homólogo en vino tinto.

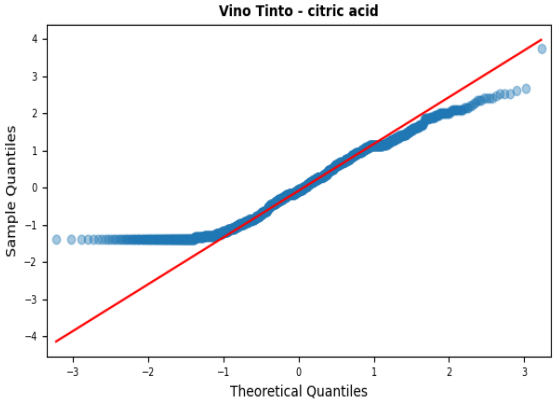


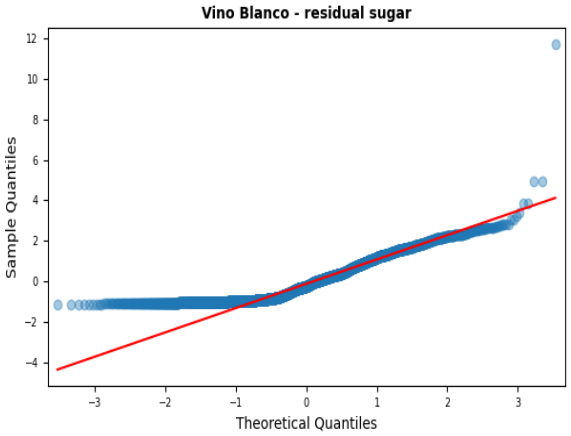


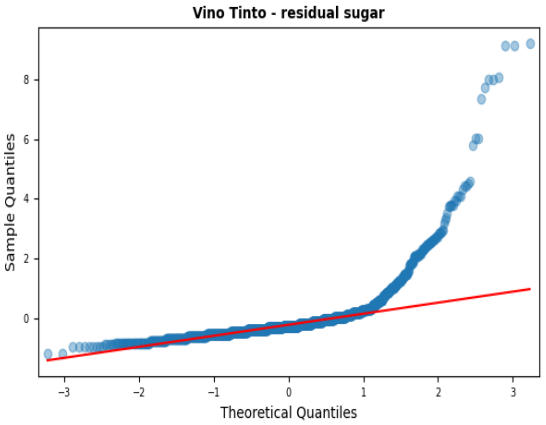

Vino Blanco - volatile acidity
Sample Quantiles
Theoretical Quantiles

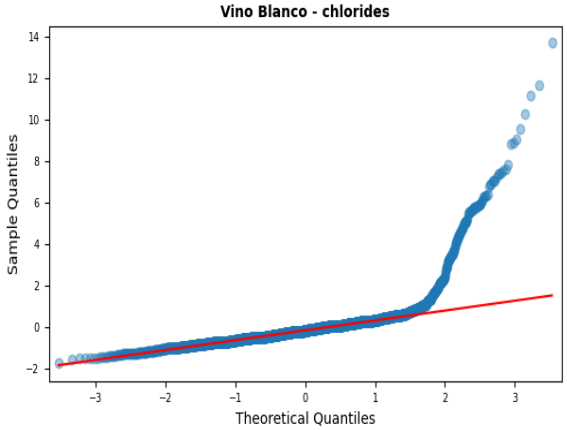

Vino Tinto - volatile acidity
Sample Quantiles
Theoretical Quantiles

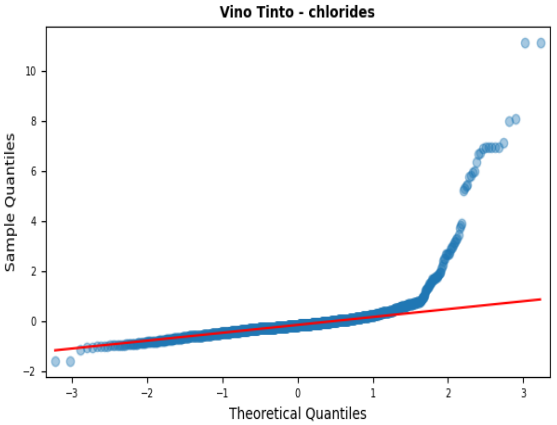

Vino Blanco - citric acid
Sample Quantiles
Theoretical Quantiles

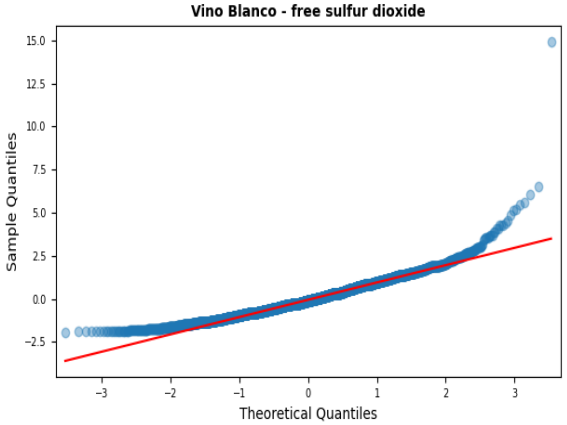

Vino Tinto - citric acid
Sample Quantiles
Theoretical Quantiles

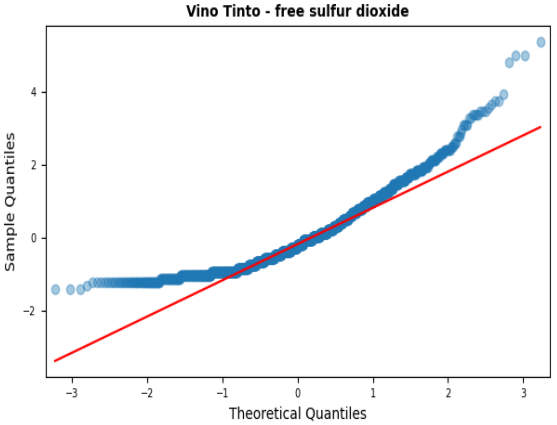

Vino Blanco - residual sugar
Sample Quantiles
Theoretical Quantiles

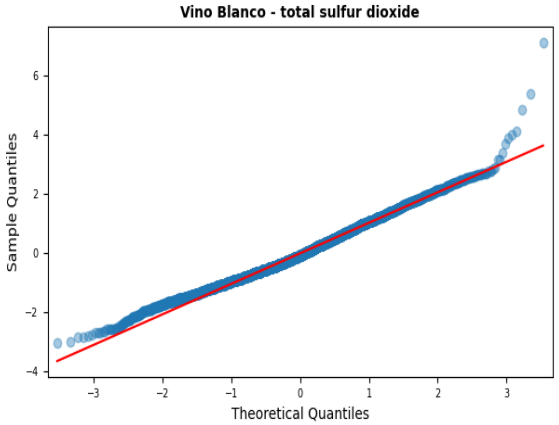

Vino Tinto - residual sugar
Sample Quantiles
Theoretical Quantiles

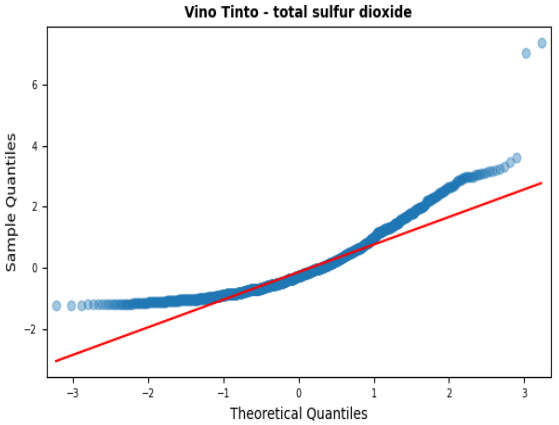

Vino Blanco - chlorides
Sample Quantiles
Theoretical Quantiles

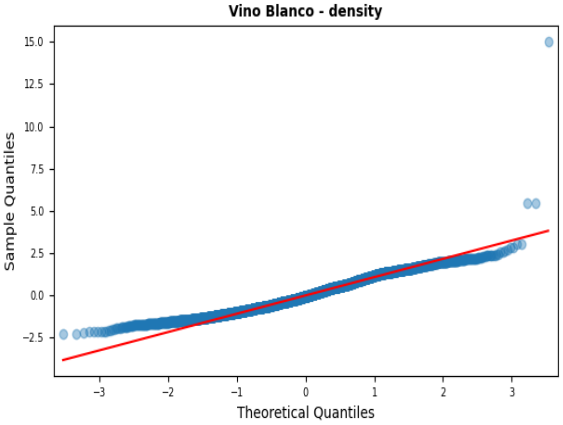

Vino Tinto - chlorides
Sample Quantiles
Theoretical Quantiles

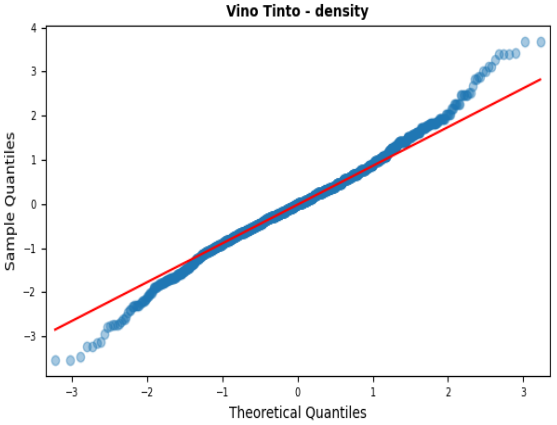

Vino Blanco - free sulfur dioxide
Sample Quantiles
Theoretical Quantiles

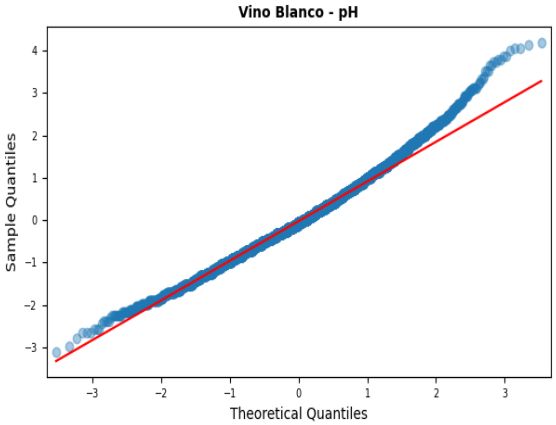

Vino Tinto - free sulfur dioxide
Sample Quantiles
Theoretical Quantiles

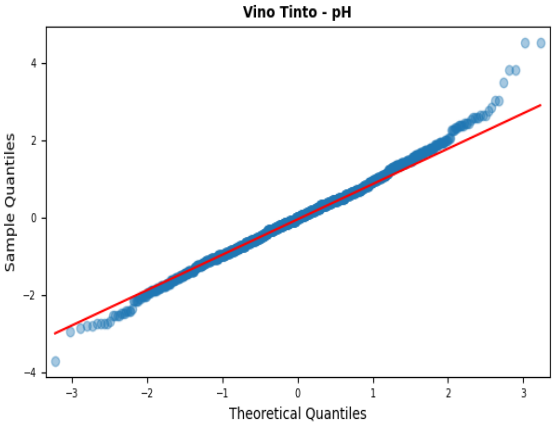

Vino Blanco - total sulfur dioxide
Sample Quantiles
Theoretical Quantiles

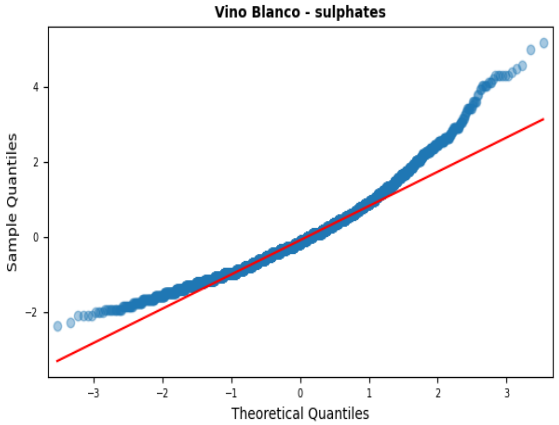

Vino Tinto - total sulfur dioxide
Sample Quantiles
Theoretical Quantiles

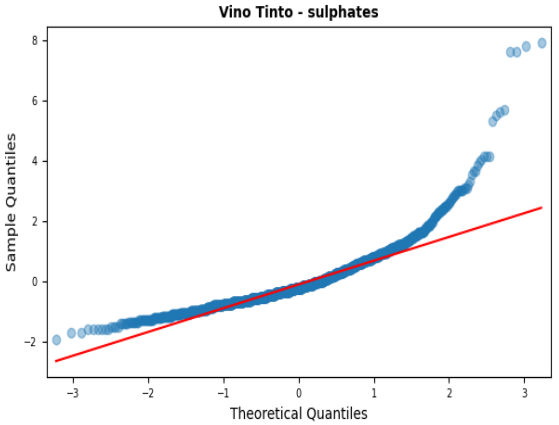

Vino Blanco - density
Sample Quantiles
Theoretical Quantiles

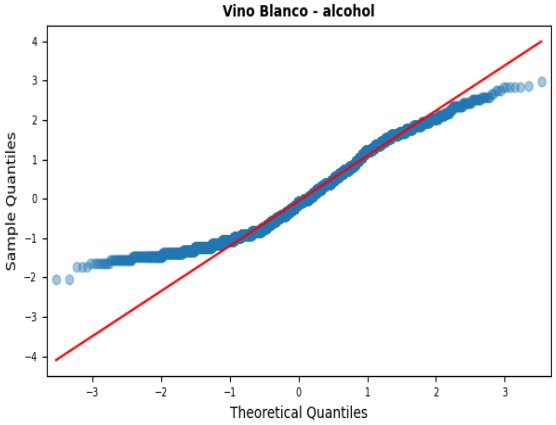

Vino Tinto - density
Sample Quantiles
Theoretical Quantiles

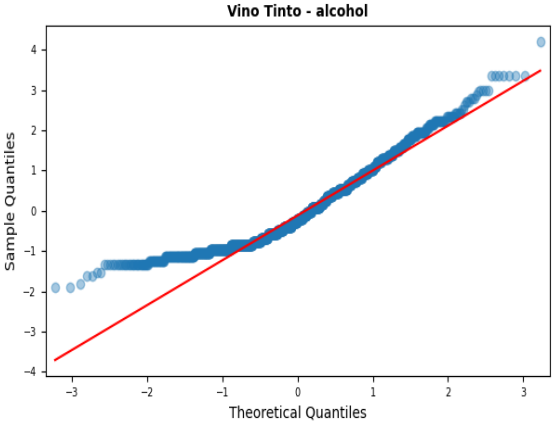

Vino Blanco - pH
Sample Quantiles
Theoretical Quantiles


Vino Tinto - pH
Sample Quantiles
Theoretical Quantiles









Observando las gráficas vemos claramente que ninguna variable es normal, ya que la distribución de los valores no siguen la diagonal roja. Sobre todo en los cuartiles de las colas de la distribución, donde en todas las gráficas tienen una clara desviación respecto a lo que corresponde a una distribución normal.

Realizamos las pruebas de ***kurtosis***, de ***skewness*** y el **Saphiro** a cada una de las variables. Separadas por tipos de vino para reforzar la hipótesis de **normalidad**. El valor de referencia de kurtosis para una distribución normal es de 3, por lo que un valor alejado nos indica una falta de normalidad. La prueba de skewness tiene el valor de referencia cero para distribución normal, así que un valor distinto también nos indicará falta de normalidad en la población que analizamos.

Test para los Vinos Tintos

Variable	Kurtosis	Skewness	p-value Shapiro	Pasa Shapiro
fixed-acidity	1,124856	0,981829	<0.05	No
volatile-acidity	1,217963	0,670962	<0.05	No
citric-acid	-0,790283	0,318039	<0.05	No
residual-sugar	28,524438	4,536395	<0.05	No

chlorides	41,581708	5,675017	<0.05	No
free-sulfur-dioxide	2,01349	1,249394	<0.05	No
total-sulfur-dioxide	3,794172	1,514109	<0.05	No
density	0,927411	0,071221	<0.05	No
pH	0,800671	0,193502	<0.05	No
sulphates	11,679884	2,426393	<0.05	No
alcohol	0,195654	0,860021	<0.05	No

Test para los Vinos Blancos

Variable	Kurtosis	Skewness	p-value Shapiro	Pasa Shapiro
fixed-acidity	5,085205	1,576497	<0.05	No
volatile-acidity	2,168737	0,647553	<0.05	No
citric-acid	6,167374	1,281528	<0.05	No
residual-sugar	3,465054	1,076764	<0.05	No
chlorides	37,525039	5,021792	<0.05	No
free-sulfur-dioxide	11,453416	1,406314	<0.05	No
total-sulfur-dioxide	0,570045	0,39059	<0.05	No
density	9,782587	0,977474	<0.05	No
pH	0,529009	0,457642	<0.05	No
sulphates	1,588081	0,976894	<0.05	No
alcohol	-0,698937	0,487193	<0.05	No

Con estos test de arriba **confirmamos la falta de normalidad** de las variables analizadas. Como podemos observar, de acuerdo al nivel de significación escogido (0.05), la hipótesis nula de normalidad del test de Shapiro-Wilke se rechaza, por lo que podemos confirmar que ninguna de las variables sigue una distribución normal.

Realizamos los test de **homogeneidad** de la varianza para verificar si las variables en los dos conjuntos tienen la misma distribución. Los test de Levene y de Fligner se basan en la mediana de cada variable.

Ahora **mostramos una tabla con los test de homogeneidad**. Aquí vamos que ninguno lo pasa, salvo el pH.

Variable	p-value Levene	Pasa Levene	p-value Bartlett	Pasa Bartlett	p-value Fligner	Pasa Fligner
fixed acidity	0	False	0	False	0	False
volatile acidity	0	False	0	False	0	False
citric acid	0	False	0	False	0	False
residual sugar	0	False	0	False	0	False
chlorides	0	False	0	False	0	False
free sulfur dioxide	0	False	0	False	0	False
total sulfur dioxide	0	False	0	False	0	False
density	0	False	0	False	0	False
pH	0,487705	True	0,274618	True	0,520857	True
sulphates	0	False	0	False	0	False
alcohol	0	False	0	False	0	False

c. Aplicación de pruebas estadísticas

A pesar de haber concluido en el anterior apartado que las variables no parecen seguir una distribución normal, debido al tamaño de la muestra asumimos que se cumple el Teorema del Límite Central. Así pues, todas las pruebas que se hagan a continuación operan bajo el anterior supuesto.

Correlación de variables

En primer lugar, presentamos dos matrices de correlación que nos permitirán estudiar **la correlación de Pearson de todas la variables del conjunto de datos**:

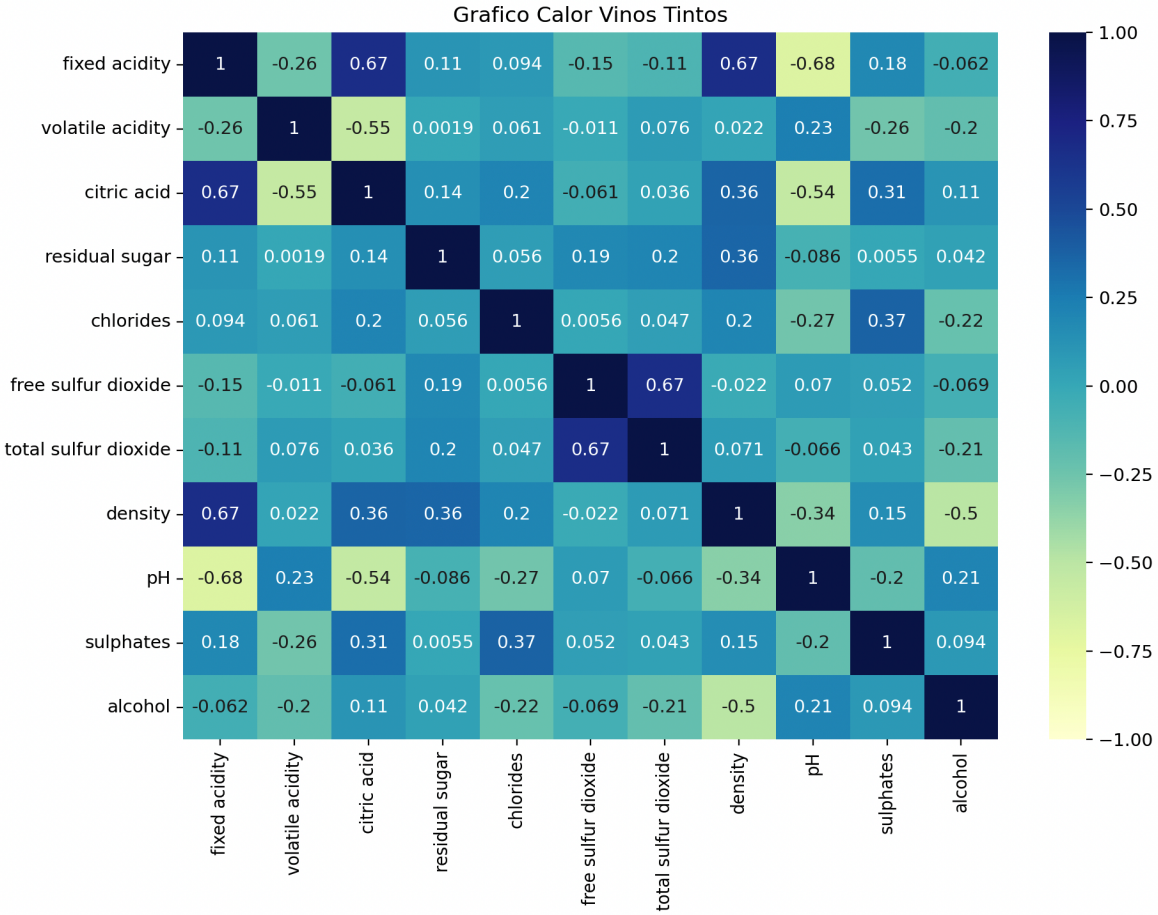
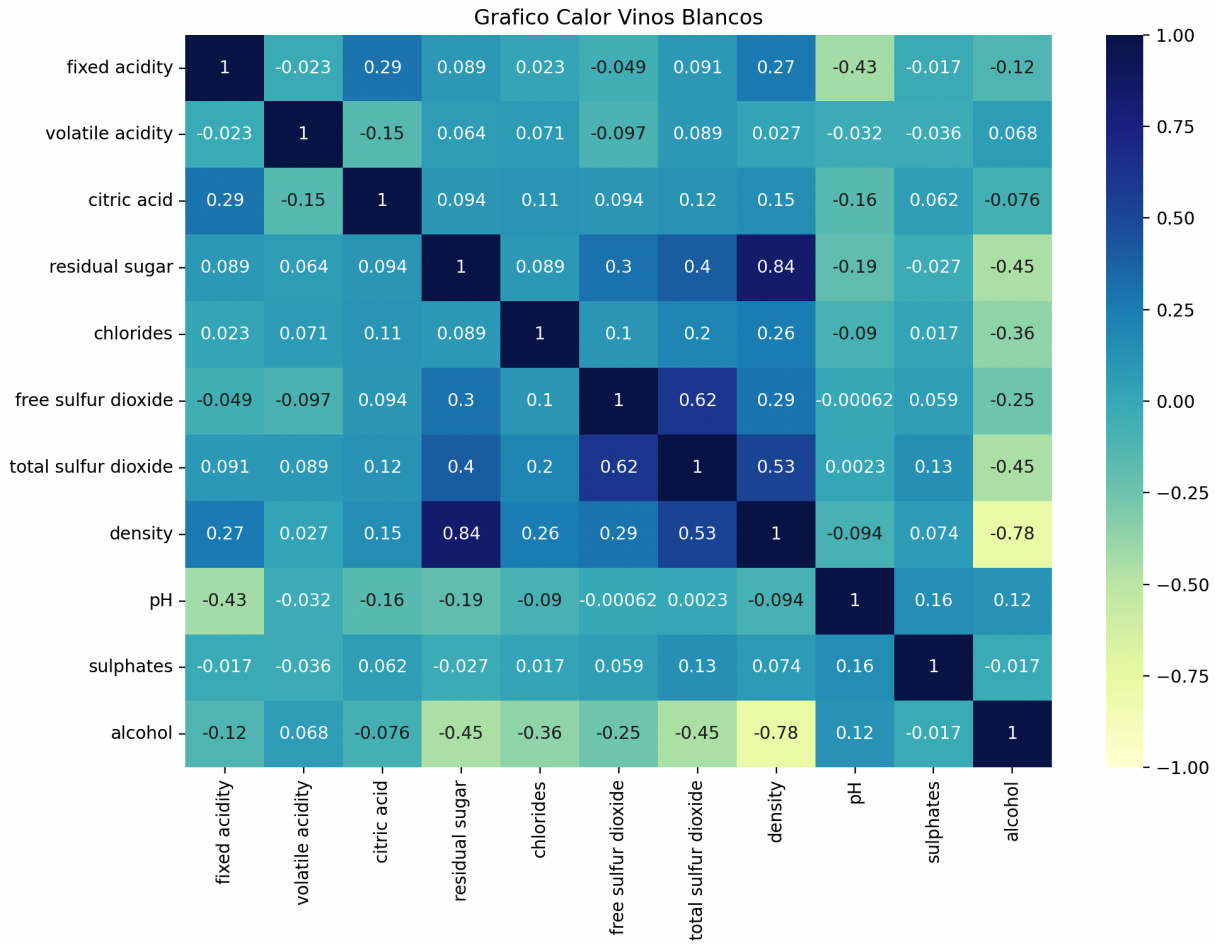


Grafico Calor Vinos Tintos
fixed acidity
volatile acidity
citric acid
residual sugar
chlorides
free sulfur dioxide
total sulfur dioxide
density
pH
sulphates
alcohol
1 -0.26 0.67 0.11 0.094 -0.15 -0.11 0.67 -0.68 0.18 -0.062
-0.26 1 -0.55 0.0019 0.061 -0.011 0.076 0.022 0.23 -0.26 -0.2
0.67 -0.55 1 0.14 0.2 -0.061 0.036 0.36 -0.54 0.31 0.11
0.11 0.0019 0.14 1 0.056 0.19 0.2 0.36 -0.086 0.0055 0.042
0.094 0.061 0.2 0.056 1 0.0056 0.047 0.2 -0.27 0.37 -0.22
-0.15 -0.011 -0.061 0.19 0.0056 1 0.67 -0.022 0.07 0.052 -0.069
-0.11 0.076 0.036 0.2 0.047 0.67 1 0.071 -0.066 0.043 -0.21
0.67 0.022 0.36 0.36 0.2 -0.022 0.071 1 -0.34 0.15 -0.5
-0.68 0.23 -0.54 -0.086 -0.27 0.07 -0.066 -0.34 1 -0.2 0.21
0.18 -0.26 0.31 0.0055 0.37 0.052 0.043 0.15 -0.2 1 0.094
-0.062 -0.2 0.11 0.042 -0.22 -0.069 -0.21 -0.5 0.21 0.094 1
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00



A partir de estas matrices, podemos observar ciertas correlaciones como las siguientes:

1. **Nivel de pH y citric acid:** curiosamente, podemos ver que existe una correlación lineal razonablemente fuerte entre el pH y el ácido cítrico para los vinos tintos, aunque no es así para el caso de los vinos blancos. Lo que nos puede indicar esto es que en el caso de los vinos blancos, el ácido cítrico no tiene un papel relevante a la hora de determinar la acidez de un vino. Esto se puede deber en parte a que los vinos blancos tienen una cantidad muy pequeña de ácido cítrico.
2. **Nivel de pH y fixed acidity (ácido tartárico):** en este caso, podemos observar que sí existe cierta correlación entre el pH y el ácido tartárico en ambos tipos de vinos. Una de las razones principales puede ser que, a diferencia del ácido cítrico, el ácido tartárico sí tiene una presencia significativa en ambos tipos de vinos.
3. **Densidad y múltiples compuestos:** como se puede intuir, la densidad depende de los compuestos que formen al vino, por lo que la concentración de azúcar o distintos tipos de ácidos tienen una obvia correlación lineal con la densidad de cada vino.

T-test

En la siguiente tabla presentamos los resultados obtenidos para los t-test realizados a todas las variables. Con ello, pretendemos **estudiar si las medias de las distintas variables difieren de forma significativa en los vinos tino y blanco**.

El nivel de significación escogido es α = 0.05, un valor bastante habitual y que consideramos más que suficiente para nuestro estudio. Como podemos observar en la tabla inferior, todos los t-test realizados generan p-valores menores a 0.05, lo que indica que la hipótesis nula de que no hay diferencia entre las medias de las variables para los dos grupos ha de ser rechazada (***Anotación: el mayor p-valor que nos indica Python es 0.007868 para el caso del alcohol).**

Así pues, concluimos que para todas las variables existe una diferencia significativa entre las medias de los dos grupos estudiados.

Variable	**p-value**	**Pasa t-test**
fixed acidity	<0.05	No
volatile acidity	<0.05	No
citric acid	<0.05	No
residual sugar	<0.05	No
chlorides	<0.05	No
free sulfur dioxide	<0.05	No
total sulfur dioxide	<0.05	No
density	<0.05	No
pH	<0.05	No
sulphates	<0.05	No
alcohol	<0.05	No

Regresión logística

Puesto que estamos trabajando con un número manejable de variables, hemos considerado que la mejor manera de encontrar un modelo de regresión logística óptimo sería **probar todas las posibles combinaciones de variables**. Para establecer cuál es el mejor modelo

de todos, hemos optado por hacer uso del **criterio de Akaike**, que además de tener en cuenta la bondad del ajuste del modelo también introduce una penalización en base a su complejidad.

Asimismo, para evaluar de forma independiente el modelo hemos decidido dividir el conjunto de datos en un **set de entrenamiento y validación, tomando los porcentajes 80/20**.

Teniendo esto en cuenta, el modelo óptimo que hemos encontrado es el que se muestra la tabla inferior, para el cual obtuvimos un AIC de :

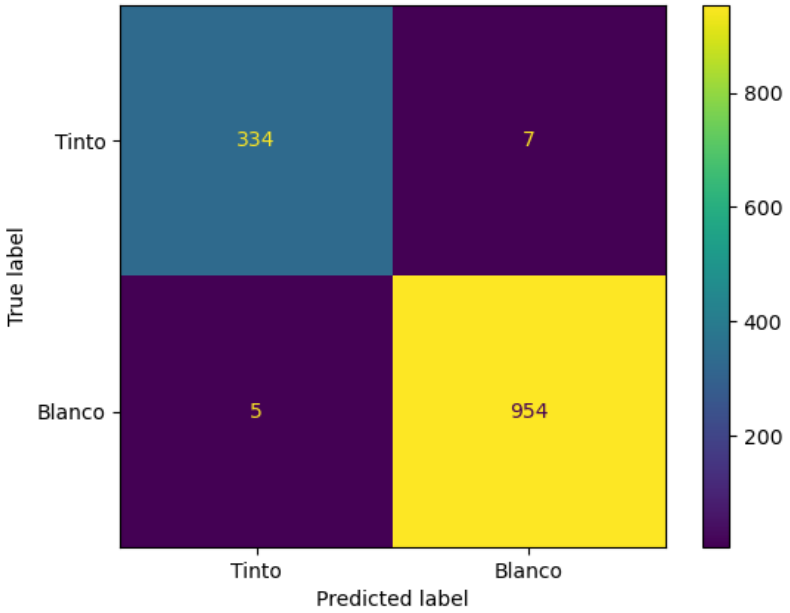
Variable	**Coeficiente**	**Error est.**	**z-score**	**p-value**
residual_sugar	1.0175	0.113	9.020	<0.05
chlorides	-24.6912	5.025	-4.914	<0.05
free sulfur dioxide	-0.0793	0.016	-4.973	<0.05
total sulfur dioxide	0.0597	0.006	9.760	<0.05
density	-1543.6997	134.09	-11.512	<0.05
sulphates	-4.6750	1.4969	-3.182	<0.05
alcohol	-1.4275	0.253	-5.633	<0.05

Como podemos observar de acuerdo al p-value (expresado en Python en todos los casos con el valor 0.000), **todas las variables del modelo son significativa**s.

En cuanto a las métricas que podemos obtener a partir del conjunto de test, encontramos lo siguiente:

- **Accuracy:** 0.99
- **Precision:** 0.99
- **Recall:** 0.99
- **F1 Score:** 0.99
- **ROC AUC:**0.99

Algunas de las métricas más habituales para evaluar la calidad predictiva de los modelos nos indican que nuestro modelo parece ser excelente en cualquier caso, y no parece generar muchos falsos negativos y ni falsos positivos. Por completitud, presentamos la matriz de confusión asociada, donde podemos ver esto con mayor claridad:



En definitiva, podemos afirmar que **nuestro modelo de regresión logística es excelente para diferenciar entre vinos tintos y blancos**

5. Resolución del problema

El estudio propuesto nos ha permitido dar respuesta a aquellas preguntas que nos habíamos realizado en un principio:

- Conocer correlaciones de variables para distintos tipos de vinos.
- Evaluar si las medias de las variables difieren significativamente entre el vino tinto y el blanco
- Saber si es posible generar un modelo predictivo que nos permite conocer si un vino es tinto o blanco en base a ciertas variables.

Tal y como hemos visto, hemos podido extraer información de valor, como que el pH y el ácido cítrico están correlacionados únicamente para un tipo de vino, o que el ácido tartárico juega un papel claro en el pH, independientemente de si el vino es tinto o blanco. Asimismo, hemos comprobado que todas las variables recogidas presentan diferencias significativas en los dos grupos estudiados. De esto último se podría haber intuido que encontrar un modelo predictivo podría ser una tarea realizable, algo que verificamos al obtener un modelo de regresión logística que predice casi a la perfección el tipo de vino con el que se corresponde una muestra.

6. Tabla de contribuciones

Contribuciones	**Firma**
Investigación previa	CDD, JOR
Redacción de las respuestas	CDD, JOR
Desarrollo del código	CDD, JOR
Participación en el vídeo	CDD, JOR