Práctica 2:

Componentes: **LAURA BELENGUER QUEROL - LEIRE ALEGRIA MURILLO**

Contribuciones	Firma
Investigación previa	Laura Belenguer Querol
Redacción de las respuestas	Laura Belenguer Querol
Desarrollo código	Laura Belenguer Querol (2,3, 4,5) Leire Alegria Murillo (1,2,4)

1. Descripción del dataset. ¿Por qué es importante y qué pregunta/problema pretende responder?

El conjunto de datos seleccionado comprende las variantes de tinto (red-wine) del "Vinho" Verde Portugués.

Tenemos 1599 observaciones y 12 variables

Se presentan datos referentes a los *parámetros fisicoquímicos (entradas) y sensoriales (salida/target)*

Variable entrada - parámetros físico-químicos:

- Fixed acidity: ácidos (no-volátiles) relacionados con el vino
- Volatile acidity: nivel de ácido acético, niveles altos significan sabor avinagrado
- Citric acid: aporta frescura y sabor a los vinos
- Residual sugar: cantidad remanente de azúcar cuando termina la fermentación. Extraño encontrar vinos con menos de 1
- Chlorides: cantidad de sal
- Free sulfur dioxide: sulfitos que previenen el crecimiento de microbial y la oxidación del vino
- Total sulfur dioxide: niveles altos de dióxido de sulfuro 50 ppm resultan detectable al olor y sabor
- Density: quan similar es la densidad del vino dependiendo del contenido de alcohol y azúcar a la densidad del agua
- pH: cómo de ácido o básico es un vino en una escala de 0 (muy ácido) a 14 (muy básico). La mayoría de vinos se encuentran entre 3-4.
- Sulphates: aditivo del vino que puede contribuir a los niveles de SO2 (sulfitos)
- Alcohol: porcentaje de alcohol en el vino

Target - datos sensoriales

- quality: basado en datos sensoriales, valores entre 0 y 10

La pregunta que el dataset pretende responder son los factores fisicoquímicos que contribuyen a la calidad del vino. Que combinación o variables influyen más en la calidad de un vino.

En este apartado mostramos las estadísticas e histogramas de cada variable y del target quality.

2. Integración y selección de los datos de interés a analizar

La integración o fusión de los datos consiste en la combinación de datos procedentes de múltiples fuentes, con el fin de crear una estructura de datos coherente y única que contenga mayor cantidad de información.

En este caso los datos se encuentran todos en el mismo dataset, por lo tanto no hay que integrar datos de otro dataset.

Una de las primeras etapas en el preprocesado de los datos es el filtrado o selección de datos de interés. En esta fase también es habitual realizar una exploración de los datos.

El preprocesado de los datos también puede incluir la creación de nuevas variables a partir de la extracción de características de los datos originales. Como posible creación de nuevas variables, nos planteamos la reducción de la variable 'quality', para etapas posteriores.

Si un vino es excelente tiene una nota (‘quality’) de 7 o superior, por lo que más adelante creamos una variable llamada 'rating' donde clasificamos los vinos entre 'superior', 'inferior' o 'correcto'. Si su nota es superior a 7 será superior, de lo contrario será inferior.

3. Limpieza de los datos.

3.1. ¿Los datos contienen ceros o elementos vacíos?

Para comprobar la presencia de ceros o elementos vacíos, aplicamos los algoritmos correspondientes en python, tal como se muestra en el código adjunto.

Gracias a ellos, observamos que no hay valores nulos ni vacíos.

¿Cómo gestionarías cada uno de estos casos?

ESi hubiesen datos NaN en alguna de las columnas, utilizariamos: wine2.dropna() para eliminar aquellas filas que contienen dichos valores NaN.

También podríamos asignar a los registros perdidos la media o la mediana de ese atributo, dependiendo de la distribución de los datos. Esta media se puede calcular para toda la muestra o para cada una de las clases o categorías que la describan.

Otras aproximaciones se basan en la implementación de métodos probabilistas para predecir (o imputar) los valores perdidos. Algunos de estos métodos son las regresiones, las inferencias basadas en modelos bayesianos o los árboles de decisión.

Ejemplos: método kNN o missForest para completar los valores pérdidos (NaN) o vacíos.

3.2. Identificación y tratamiento de valores extremos.

Los valores extremos (extreme scores o outliers) son aquellos datos que se encuentran muy alejados de la distribución normal de una variable o población.

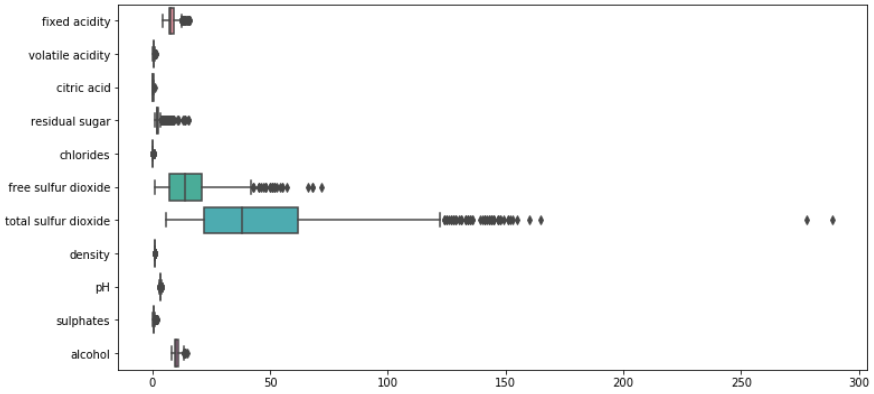
Son observaciones que se desvían tanto del resto que levantan sospechas sobre si fueron generadas mediante el mismo mecanismo.

Estos valores pueden afectar de forma adversa los resultados de los análisis posteriores, al incrementar el error en la varianza de los datos y sesgar significativamente los cálculos y estimaciones.

En el apartado 1.2 hemos visto que algunas variables/atributos tienen una desviación standard (std) bastante elevada si observamos el valor de la media.

- citric acid
- volatile acidity
- residual sugar
- chlorides
- free sulfur dioxide
- total sulfur dioxide

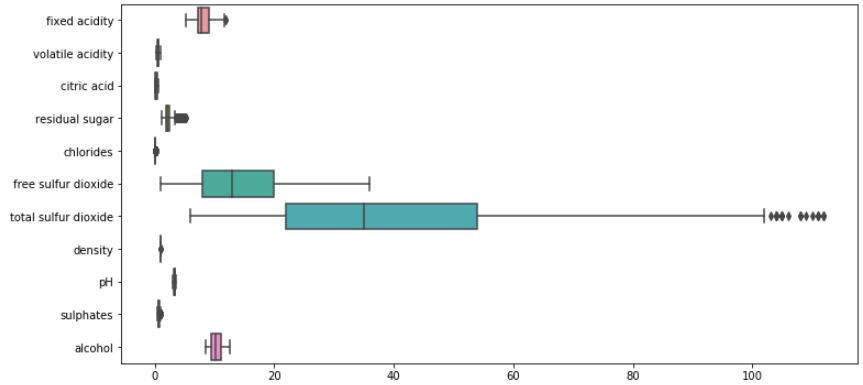
Con los gráficos boxplot analizamos las variables de manera visual.



También basados en los valores de calidad, visualizamos como dichos parámetros afectan a la calidad del vino.

Observamos que casi todas las variables presentan valores extremos. Intentamos identificar como outliers aquellos puntos que se encuentran, unidimensionalmente a más de 2 desviaciones estándar de la media (Distancia Mahalanobis)

Aplicando el criterio de Mahalanobis para eliminar los valores extremos hemos reducido el tamaño de nuestro dataset de 1599 muestras a 1124.



Con los boxplots vemos que se han reducido considerablemente los outliers para la mayoría de variables. Aunque ciertas variables como 'residual sugar' y 'chlorides' presentan un alto número de valores por encima del tercer cuartil (75% de los datos).

De todos modos eliminar dichos valores extremos puede dar lugar a eliminar aquellos vinos que marcan la diferencia por atributos excepcionales, por ello nos quedaremos con los atributos originales sin eliminar ningún valor extremo.

4. Análisis de los datos

4.1. Selección de los grupos de datos que se quieren analizar/comparar (planificación de los análisis a aplicar).

En los apartados anteriores hemos realizado un análisis descriptivo de los datos (describe(), boxplot) donde hemos encontrado los valores de la media, mediana, desviación estándar, cuartiles, valores mínimo y máximo antes y después de eliminar los outliers.

En los siguientes apartados realizaremos el análisis inferencial que tiene por objetivo modelar los datos a través de una distribución conocida. Partiendo de la premisa que el conjunto de datos estudiado representa una fracción de la totalidad de una población, su objetivo es inferir cómo es esa población, asumiendo un grado de error en las estimaciones por el hecho de disponer de una muestra reducida de los datos.

4.2. Comprobación de la normalidad y homogeneidad de la varianza.

- Normalidad

Para la comprobación de que los valores que toman nuestras variables cuantitativas provienen de una población distribuida normalmente, utilizaremos la prueba de normalidad de Shapiro.

Se comprueba que para que cada prueba se obtiene un p-valor superior al nivel de significación prefijado $\alpha = 0.05$. Si esto se cumple, entonces se considera que dicha variable sigue una distribución normal.

H_0 = Tiene una distribución normal
H_1 = No tiene una distribución normal

p <= alpha: rechazamos H_0, no hay normalidad.
p > alpha: no rechazamos H_0, hay normalidad.

Todas las variables tienen el p valor inferior a 0.05, por lo tanto, ninguna variable sigue una distribución normal.

- Homogeneidad

Cuando los datos no cumplen con la condición de normalidad, cómo hemos visto en el apartado anterior, utilizamos el test Fligner-Killeen

La hipótesis nula asume igualdad de varianzas en los diferentes grupos de datos, por lo que p-valores inferiores al nivel de significancia (0.05) indicán heterocedasticidad.

En este caso lo aplicaremos a todas las variables, comparando los atributos (data_wine) con la variable target (quality).

Como todos los valores de p-valor < 0.05 los datos presentan heterocedasticidad.

4.3. Aplicación de pruebas estadísticas para comparar los grupos de datos. En función de los datos y el objetivo del estudio, aplicar pruebas de contraste de hipótesis, correlaciones, regresiones, etc. Aplicar al menos tres métodos de análisis diferentes.

- **Contraste de hipótesis**

  - **Test Mann Whitney**

  Aplicamos el test comparando dos grupos los vinos de calidades inferiores (‘quality’<=5) con aquellos de calidades superiores (‘quality’>=6)

  Se comprueba que para que cada prueba se obtiene un p-valor superior al nivel de significación prefijado $\alpha = 0.05$. Si esto se cumple, entonces se considera que dicha variable sigue una distribución normal. Por lo tanto aquellas variables con un p-valor mayor al nivel de significación con este test son: *Residual sugar, Free sulfur dioxide, pH*

  - **Test Kruskal-Wallis**

  Podemos comparar más de dos grupos con este test. En este caso comparamos los vinos inferiores (<5), con los correctos (5,6) y los superiores(>6).

  Con este test sólo obtenemos una variable con un nivel de significación > 0.05: *Residual sugar*

- **Correlación**

  Procedemos a realizar un análisis de correlación entre las distintas variables para determinar cuáles de ellas ejercen una mayor influencia sobre la calidad del vino. Se utiliza el coeficiente de correlación de Person.

  A partir del diagrama de matriz (heatmap) de correlación se puede observar que:

  - 'fixed acidity':
    - altamente correlacionada con el 'citric acid' y 'density'
    - alta correlación negativa con el pH
    - correlación con 'quality' es muy baja
  - 'volatile acidity':
    - correlaciona negativamente con 'citric acid' y 'quality'.
  - 'citric acid':
    - alta correlación positiva con la 'fixed acidity'
    - negativamente correlada con 'volatile acidity' y pH
  - 'density':

    - altamente correlada con la 'fixed acidity'
    - negativamente correlada con el 'alcohol'
- 'pH':
    - negativamente correlada con el 'fixed acidity' y 'citric acid')
- 'alcohol':
    - negativamente correlada con la 'density'
- 'quality':
    - positivamente correlada con 'alcohol' y en menor medida con los 'sulphates'
- 'free sulfur dioxide' y 'total sulfur dioxide' están bastante correlacionadas entre ellas pero no con el resto de variables.

Ninguna variable está muy correlacionada con la calidad del vino. La correlación mas alta la encontramos con alcohol 0.48, no es un valor muy alto.

- Regresión lineal

Por último aplicamos regresión lineal. Para ello una vez identificadas las variables X (atributos) y y (target) dividimos el dataset en entrenamiento (train) y test (test) y aplicamos LinearRegression.

Calculamos el Root Mean Squared Error que es una medida que se utiliza para identificar las diferencias entre los valores. Si hemos construido un buen módelo el RMSE debe ser muy similar. Si el RMSE es mayor para el test que para en train, seguramente tengamos un mal ajuste de los datos.

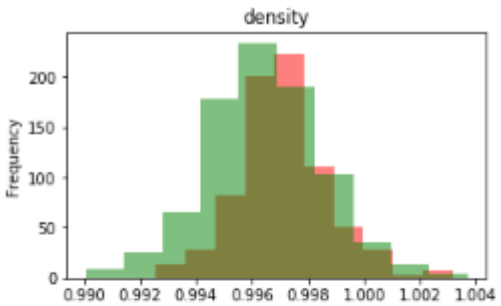
Los resultados para RMSE del train y test son bastante similares, por lo tanto hemos construido un buen modelo.

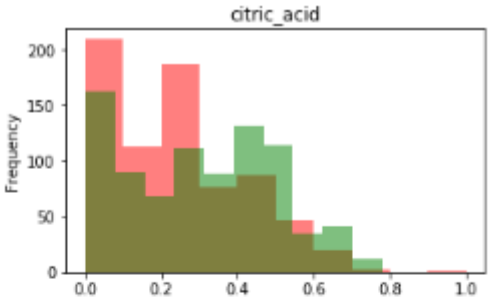
La precisión del módelo lineal no es muy elevado (37%). Si miramos los coeficientes de RMSE de cada atributo vemos que los que más impacto tienen (positivo o negativo): *volatile acidity, chlorides, pH, sulphates.*

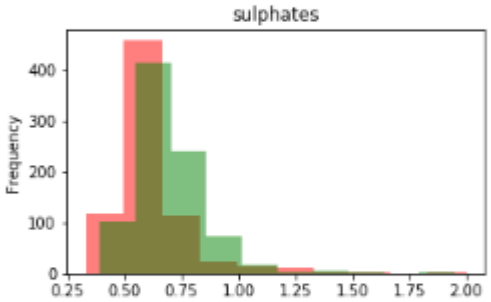
Algunas variables tienen una fuerte correlación con la variable 'quality' pero no significa que dicho atributo cause una mejora en 'quality'. Ello se debe a que la calidad del vino es una variable cualitativa y esta es la razón por la que el algoritmo no ajusta bien. El modelo lineal funciona mejor con variables cuantitativas.

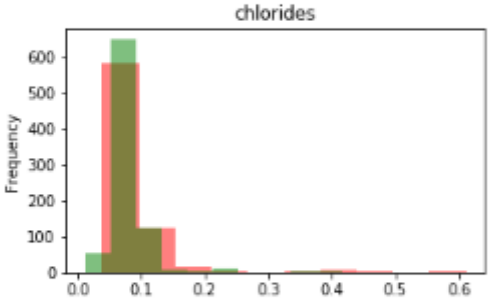
5. Representación de los resultados a partir de tablas y gráficas

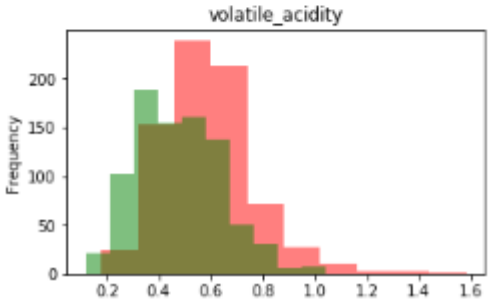
HISTOGRAMAS

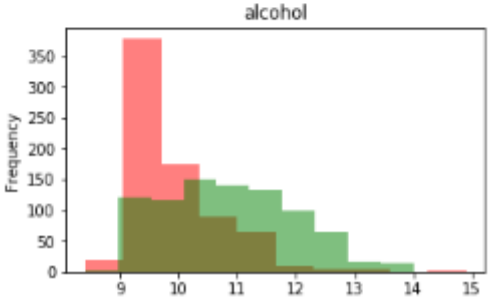




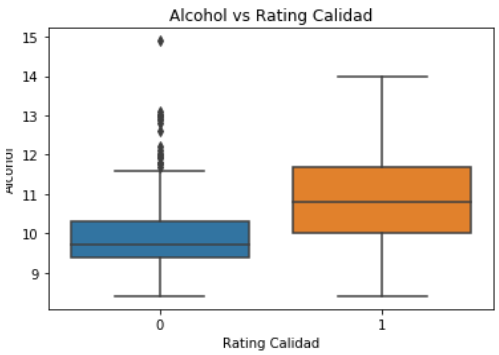


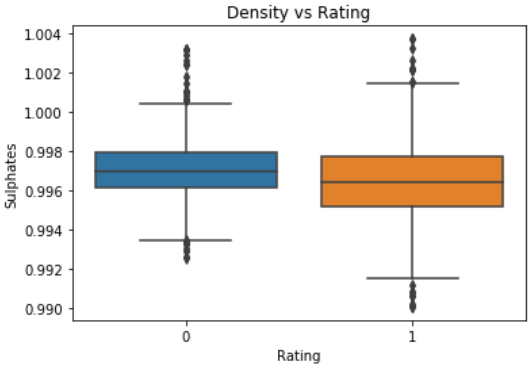


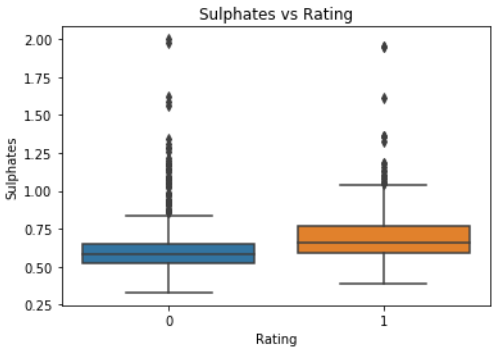


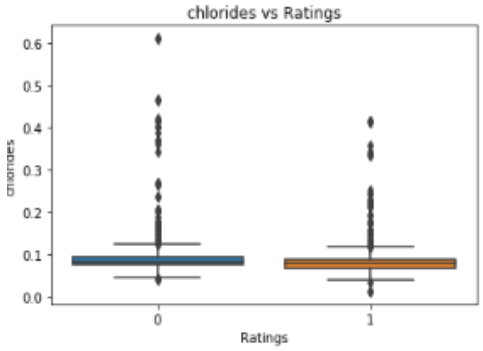


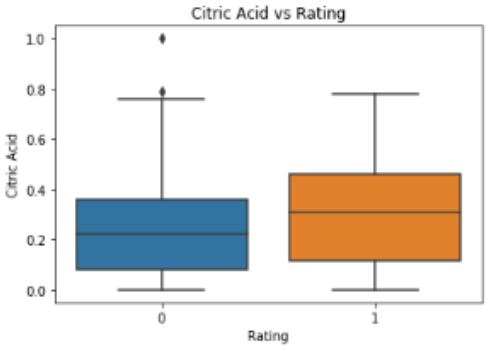
BOXPLOTS

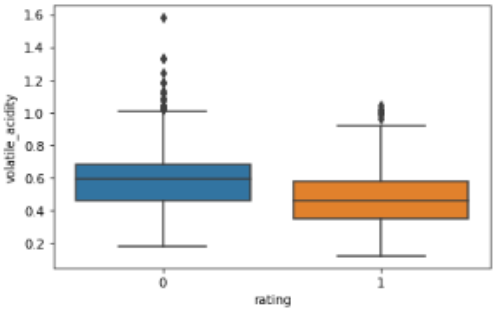




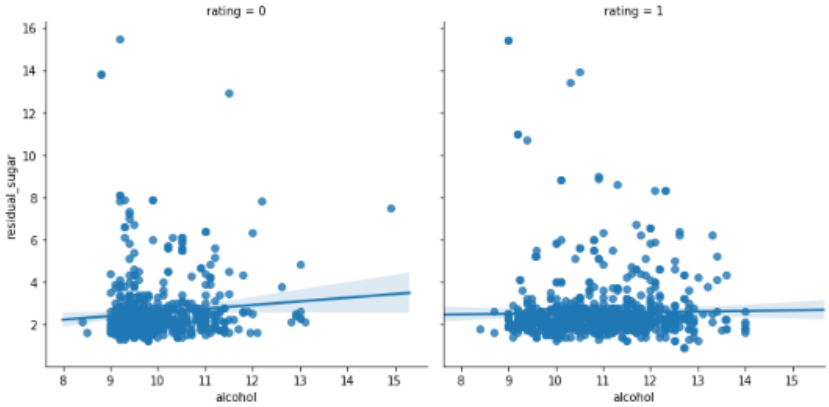


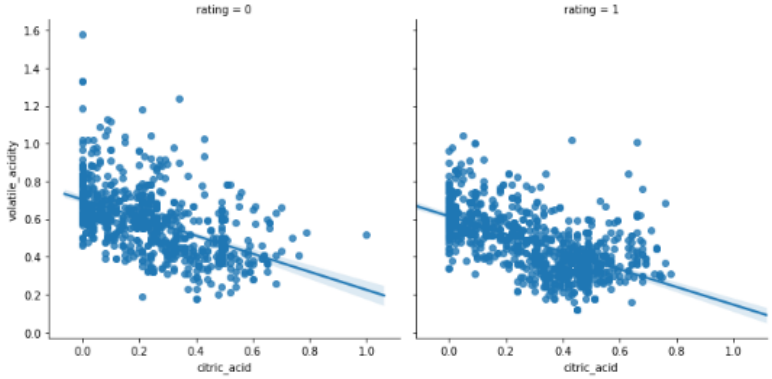






LINEAR REGRESSION





6. Resolución del problema. A partir de los resultados obtenidos, ¿cuáles son las conclusiones?

En el conjunto de datos dado para el vino tinto, los vinos de calidad 5 y calidad 6 son los más disponibles. Los vinos tintos de calidad 1, calidad 2, calidad 9 y calidad 10 no están disponibles.

Ninguna variable está muy correlacionada con la calidad del vino, por lo tanto, es previsible que ningún algoritmo nos dé una predicción muy buena.
Por los test de contraste de hipótesis vemos que los datos no cumplen la condición de normalidad en la varianza ni homogeneidad.

En el análisis de los datos vía contraste de hipótesis, correlación y regresión lineal encontramos las variables más significativas que influyen en la calidad del vino.

- Hipótesis: encontramos significativas: residual sugar, pH

- Correlación: no encontramos fuertes correlaciones, las más altas: volatile acidity, citric acid, sulphates y alcohol

- Regresión: La precisión del módelo lineal no es muy elevado (~30%). Si miramos los coeficientes de RMSE de cada atributo vemos que los que más impacto tienen (positivo o negativo): density, volatile acidity, chlorides, pH, sulphates, alcohol

- Árbol de clasificación: nos da una mejor predicción del modelo, pero no aclara mucho más que atributos contribuyen más a la calidad.

De las gráficas encontramos primero los histogramas con cada uno de los atributos en función de calidad '1' y '0'.

- Density vs Rating: la media de la densidad de los vinos de mayor calidad es más baja.
- Sulphates vs Rating: valores de la media de sulfatos ligeramente mayores para los vinos de calidad alta
- Alcohol vs Rating: la media de contenido alcohol es más alta en los vinos con calificación alta.
- Volatile acidity vs Rating: la media de volatile acidity es menor para los vinos de alta calidad
- Citric acid vs Rating: las medias son similares para vinos de alta y baja calidad
- Chlorides vs Rating: las medias son similares para vinos de alta y baja calidad

Las gráficas boxplot confirman los resultados del histograma con las siguientes precisiones:

- Citric acid vs Rating: las medias son mayores para los vinos de alta calidad

Gráficas de linear regresión

- citric acid y volatile acidity: tanto en los vinos de calidad alta y baja, el citric acid está inversamente proporcional al volatile acidity
- residual sugar vs alcohol: observamos que para los vinos de calidad más alta la relación entre estas dos variables permanece constante. En los vinos de calidad baja, el residual sugar incrementa gradualmente con el contenido de alcohol.

Este análisis puede ayudar a producir el vino de alta calidad manteniendo estos parámetros en la relación indicada para obtener un buen vino tinto.

¿Los resultados permiten responder al problema?

Nos permiten saber que variables están correlacionadas y como tenemos que modificarlas para obtener los vinos de alta calidad.

Aunque con los datos obtenidos donde las medias son bastante similares, no hay un parámetro o parámetros que influyan significativamente, sino que es una combinación de ellos.

Biografía:

P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553, 2009

https://github.com/ranjitkumarpatnaik/winequality-redwine/blob/master/winequality-red.ipynb

https://www.kaggle.com/vishalyo990/prediction-of-quality-of-wine

http://rpubs.com/jeswin_george/explore_red_quality_wines

https://github.com/paulgx/tipologiapractica2/blob/master/practica2Tipologia.pdf

https://www.kaggle.com/datacog314/tutorial-machine-learning-interpretability

https://machinelearningmastery.com/a-gentle-introduction-to-normality-tests-in-python/

Donde se pueden descargar los datos:

https://archive.ics.uci.edu/ml/datasets/wine+quality