2019

PRAC 2 – Regresión Lineal

ALEJANDRO TORRES BRITO / CRISTINA RODRIGUEZ MARTINEZ - 10/06/2019

PRAC 2 Análisis de Datos

ANÁLISIS DE LA CALIDAD DEL VINO EN FUNCIÓN A DISTINTOS PARÁMETROS MEDIDOS

Esta memoria es un breve resumen de las preguntas clave de la práctica. Para seguir mejor los razonamientos y conclusiones, por favor consultar el código R.

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

El conjunto de datos objeto de análisis se ha obtenido a partir de este enlace en Kaggle y está constituido por 12 características (columnas) que presentan 1599 muestras de vino portigués (filas o registros). Las variables que se disponen son fisicoquímicas (entradas) y sensoriales (la salida) (por ejemplo, no hay datos sobre tipos de uva, marca de vino, precio de venta del vino, etc.).

Entre los campos de este conjunto de datos, encontramos los siguientes:

fixed.acidity: Acidez fija, la mayoría de los ácidos relacionados con el vino. Fijos o no volátiles (no se evaporan fácilmente).

volatile.acidity: Acidez volátil, la cantidad de ácido acético en el vino, que en niveles demasiado altos puede provocar un sabor desagradable a vinagre.

citric.acid: El ácido cítrico se encuentra en pequeñas cantidades, el ácido cítrico puede agregar 'frescura' y sabor a los vinos.

residual.sugar: azúcar residual, la cantidad de azúcar restante después de que se detiene la fermentación, es raro encontrar vinos con menos de 1 gramo / litro y vinos con más de 45 gramos / litro se consideran dulces.

chlorides: Cloruro, la cantidad de sal en el vino.

free.sulfur.dioxide: sin dioxido de azufre, la forma sin SO2 existe en equilibrio entre el SO2 molecular (como un gas disuelto) y el ión bisulfito; Previene el crecimiento microbiano y la oxidación del vino.

total.sulfur.dioxide: Cantidad total de dióxido de azufre de formas libres y unidas de SO2; en bajas concentraciones, el SO2 es mayormente indetectable en el vino, pero a concentraciones de SO2 libres de más de 50 ppm, el SO2 se hace evidente en la nariz y el sabor del vino.

density: La densidad del agua es cercana a la del agua dependiendo del porcentaje de alcohol y de contenido de azúcar.

pH: El pH describe qué tan ácido o básico es un vino en una escala de 0 (muy ácido) a 14 (muy básico); La mayoría de los vinos están entre 3-4 en la escala de pH.

sulphates: Aditivo de vino sulphatesa que puede contribuir a los niveles de dióxido de azufre (S02), que actúa como un antimicrobiano y antioxidante.

alcohol: alcohol el porcentaje de alcohol del vino .

quality: Variable de calidad de salida (basada en datos sensoriales, puntuación entre 0 y 10).

En esta práctica unificaremos los datos, los limpiaremos y trataremos de predecir cuáles serán los de mejor calidad. Se compararán las características de los vinos calificados como de calidad media/baja y de calidad alta, para tratar de establecer las diferencias en cuanto a cualidades de los mismos.

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

fixed.acidity: Si a la acidez total le quitamos la acidez volátil debida al acético, la diferencia se llama acidez fija. Con lo que si hacemos la operación a la inversa podemos obtener la acidez total.

volatile.acidity: La acidez volátil puede oscilar entre 0,2 - 1 gr/L hasta un gramo por litro, una larga crianza en roble pueden situarse alrededor de 0,8 gr/L de acidez volátil sin manifestar sensaciones desagradables.

citric.acid: Este ácido desaparece lentamente al ser fermentado por las bacterias. No es muy abundante en la uva.

residual.sugar: Se considera vino seco o sin azúcar, cuando ese contenido es inferior a cinco gramos por litro. En función de esta cantidad de azúcares residuales, el vino puede ser: seco, semiseco, semidulce y dulce. Variando de menor a mayor cantidad de azúcar, pudiendo ir desde 1 gramo hasta 200 gramos por litro de azúcar. En los vinos secos en muchas ocasiones no suele sobrepasarse los 2 gramos.

chlorides: 1 a 50 mg/L: potasio (agente diurético, efecto que se potencia en los vinos espumosos debido al alto contenido en dióxido de carbono), sodio, hierro, manganeso, boro, nitratos, cloruros y silicio

free.sulfur.dioxide :

total.sulfur.dioxide: En Europa su valor máximo no suele sobrepasar los 50 miligramos por litro pero en E.E.U.U. puede alcanzar los 350 miligramos por litro con lo que se puede saber si este vino es de buena calidad y se puede exportar.

density:

- Vino blanco seco: 0,9880-0,9930 g/mL.

- Vinos tinto seco: 0,9910-0,9950 g/mL.

- Vino espumoso: 0,9890-1,0080 g/mL.

- Vino de licor (moscatel): 1,0500-1,0700 g/mL.

- Mosto: 1,0590-1,1150 g/mL.

pH: El pH de la mayoría de los vinos se encuentra en el intervalo de 2,5 a 4,5 , lo que lógicamente recae en el lado ácido de la escala. Un vino con un pH de 2,8 es extremadamente ácido mientras que uno con un pH en torno a 4 es plano, carente de acidez. Los vinos blancos suelen estar entre 3 y 3,3 y la mayoría de los tintos entre 3,3 y 3,6

sulphates: Cualquier vino que contiene más de 10 mg por litro de dióxido de azufre deberá indicar en su etiqueta la expresión "Contiene sulfitos"

alcohol: La graduación alcohólica que oscila entre los 3,5 y los 15 grados

quality: Puntuación que se da al vino de 1 a 10 en función de su calidad

3. Limpieza de los datos.

3.1. ¿Los datos contienen ceros o elementos vacíos? ¿Cómo gestionarías cada uno de estos casos?

3.2. Identificación y tratamiento de valores extremos.

Identificación de valores NA y nulos, búsqueda de outliers: NAs corregidos y outliers no relevantes en este caso. Consultar código R para más detalles.

4. Análisis de los datos.

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

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

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.

Se seleccionan los grupos dentro del conjunto de datos que pueden resultar interesantes para analizar y/o comparar. Se intenta describir qué variables de las que describen el vino influyen sobre la calidad del mismo, por lo que se irán haciendo operaciones para comparar la influencia de cada una de ellas sobre la calificación dada a cada uno de los vinos del dataset.

Consultar código R para ver mayor detalle de las estadísticas y variables que no siguen una distribución normal.

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

Consultar código R.

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

A partir del análisis de correlaciones entre las distintas características del vino y su calidad, se ha visto que para los vinos etiquetados como de calidad media baja las variables más correlacionadas son volatile.acidity, free.sulfur, total.sulfur. Para los de calidad alta, las variables más correlacionadas son: volatile.acidity, sulphates y alcohol.

Al aplicar el análisis de regresión lineal, se confirma mediante el coeficiente R^2 que efectivamente las tres variables con alta correlación generan un buen modelo de regresión para los vinos de calidad baja. Sin embargo, para los de calidad alta el modelo de regresión más ajustado se obtiene teniendo en cuenta las variables residual.sugar, total.sulfur.dioxide y density.

Se puede decir que los vinos de calidad alta tienen una acidez volátil más baja que los de menor calidad.Su graduación alcohólica es más alta que los de menor calidad y la cantidad de sulfitos presentes en los mismos es menor.La cantidad de sulfitos libres con respecto a los totales es mayor también para los vinos de calidad alta.

7. Código: Hay que adjuntar el código, preferiblemente en R, con el que se ha realizado la limpieza, análisis y representación de los datos. Si lo preferís, también podéis trabajar en Python.

Contribuciones:

Análisis y código inicial: Alejandro Torres Brito

Replanteamiento, correcciones sobre el código inicial y memoria: Cristina Rodríguez Martínez.