Primera entrega - Proyecto final

Análisis estadístico de datos

Manuela Acosta Fajardo, Carlos Sebastián Matrínez Vidal, Jessenia Piza Londoño

Universidad del Rosario, Escuela de Ingeniería, Ciencia y Tecnología

Introducción

Los viñedos son terrenos de plantado de vides, que están vinculados a la producción de uvas y, por tanto, a la elaboración del vino. El vino se describe con características como color, cuerpo, aroma, sabor, cosecha y variedad. Hay muchos tipos de vino, entre ellos consideraremos el vino blanco y el vino tinto.

El vino blanco se caracteriza porque en su elaboración se prescinde de la piel de la uva, tiene un sabor ácido y un bajo nivel de alcohol. Por el contrario, el vino tinto contiene mas calorías, posee un nivel alto de alcohol y usa la piel de la uva para su producción.

Descripción del problema

Un viñedo quiere hacer un estudio de los vinos en la región norte de Portugal. Teniendo en cuenta algunas de las características de los vinos, busca identificar la calidad de los mismos. Esto con el objetivo de ver cuál de los vinos de la región tiene mayor calidad. Así mismo, quiere conocer cuáles son las características que más influyen para determinar dicha calidad, de forma que puedan enfocarse en ellas para producir un excelente vino.

Objetivos

1. Identificar el tipo de vino con mejor calidad en el norte de Portugal.

2. Analizar detalladamente el dataset seleccionado para determinar cuáles son las características de los vinos que más influyen en la calidad de los mismos.

3. Determinar qué tipo de vino, entre blanco y tinto, suele tener mayor calidad.

Dataset seleccionado

Para el desarrollo del proyecto, seleccionamos el dataset `Wine Quality Data Set`, de `UCI Repository`. Este cuenta con dos datasets, uno para vinos blancos y otro para vinos tintos. Cada uno de ellos contiene doce variables, o características de los vinos. Estas características, en conjunto, indican la calidad de cada uno de los vinos, que está dada en una escala del 0 al 10.

Análisis descriptivo inicial de los datos

Realizamos un primer acercamiento a los datos seleccionados, calculando algunas medidas de tendencia central y dispersión, como se muestra en la siguiente tabla:

	fixed.acidity <dbl>	volatile.acidity <dbl>	citric.acid <dbl>	residual.sugar <dbl>	chlorides <dbl>
dat_red	8.319637	0.5278205	0.2709756	2.538806	0.08746654
dat_white	6.854788	0.2782411	0.3341915	6.391415	0.04577236

2 rows | 1-6 of 12 columns

free.sulfur.dioxide <dbl>	total.sulfur.dioxide <dbl>	density <dbl>	pH <dbl>	sulphates <dbl>	alcohol <dbl>	quality <dbl>
15.87492	46.46779	0.9967467	3.311113	0.6581488	10.42298	5.636023
35.30808	138.36066	0.9940274	3.188267	0.4898469	10.51427	5.877909

2 rows | 7-13 of 12 columns

Imagen 1: Media de los atributos de cada tipo de vino.

	fixed.acidity <dbl>	volatile.acidity <dbl>	citric.acid <dbl>	residual.sugar <dbl>	chlorides <dbl>
dat_red	3.0314164	0.03206238	0.03794748	1.987897	0.0022151427
dat_white	0.7121136	0.01015954	0.01464579	25.725770	0.0004773337

2 rows | 1-6 of 12 columns

free.sulfur.dioxide <dbl>	total.sulfur.dioxide <dbl>	density <dbl>	pH <dbl>	sulphates <dbl>	alcohol <dbl>	quality <dbl>
109.4149	1082.102	3.562029e-06	0.02383518	0.02873262	1.135647	0.6521684
289.2427	1806.085	8.945524e-06	0.02280118	0.01302471	1.514427	0.7843557

2 rows | 7-13 of 12 columns

Imagen 2: Varianza de los atributos de cada tipo de vino.

	fixed.acidity <dbl>	volatile.acidity <dbl>	citric.acid <dbl>	residual.sugar <dbl>	chlorides <dbl>
dat_red	1.7410963	0.1790597	0.1948011	1.409928	0.04706530
dat_white	0.8438682	0.1007945	0.1210198	5.072058	0.02184797

2 rows | 1-6 of 12 columns

free.sulfur.dioxide <dbl>	total.sulfur.dioxide <dbl>	density <dbl>	pH <dbl>	sulphates <dbl>	alcohol <dbl>	quality <dbl>
10.46016	32.89532	0.001887334	0.1543865	0.1695070	1.065668	0.8075694
17.00714	42.49806	0.002990907	0.1510006	0.1141258	1.230621	0.8856386

2 rows | 7-13 of 12 columns

Imagen 1: Desviación estándar de los atributos de cada tipo de vino.