Descripción:

Se trabajará con un conjunto de datos que contas sobre muestras de dos tipos de vinos, vino tinto y vino blanco, dichos datos fueron recolectados en Portugal durante los años 2009 y 2010, los datos cuentan con 12 atributos y 4898 muestras:

1 - acidez fija

2 - acidez volátil

3 - ácido cítrico

4 - azúcar residual

5 - cloruros

6 - anhídrido sulfuroso libre

7 - anhídrido sulfuroso total

8 - densidad

9 - pH

10 - sulfatos

11 - alcohol

Variable de salida (basada en datos sensoriales):

12 - calidad (puntuación entre 0 y 10)

Como se puede observar la variable numero 12 se puede usar como etiqueta tanto para el modelo de regresión como para el modelo de clasificación ya que nos ayudará a hacer grupos de acuerdo con las propiedades de los dos tipos de vinos.

Objetivo general:

* Realizar la mayor inferencia estadística sobre el conjunto de datos

Objetivos específicos:

* Realizar un modelo de regresión multivariada para la predicción de datos
* Obtener las componentes principales de los datos para posteriormente realizar un modelo de regresión multivariado más sencillo.
* Obtener las componentes principales de los datos para saber que propiedades son mas relevantes en cada tipo de vino
* Usar algoritmos de clasificación como k-means o k-medoids para identificar propiedades que influyen en la calidad del vino

Referencias:

* P. Cortez, A. Cerdeira, F. Almeida, T. Matos y J. Reis. Modelado de preferencias de vino mediante minería de datos a partir de propiedades fisicoquímicas. En Decision Support Systems, Elsevier, 47(4):547-553, 2009