



# IA 1

UT 02 - TA 02

Participantes:  
Federico Beconha  
Bruno Cattaneo  
Gerardo Fernandez  
Felipe Mestre

# UT02 TA 02

## El problema del dataset wine

El dataset contiene datos provenientes del análisis químico de vinos hechos en la misma región de Italia pero que derivan de distintos cultivadores de uvas. El análisis determinó la cantidad de los 13 componentes encontrados en los tres tipos de vinos.

El problema de esta dataset se basa en construir varios modelos de clasificación para la clase del vino.


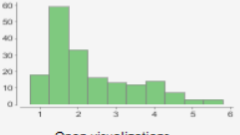
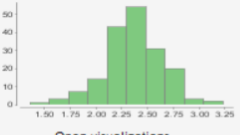

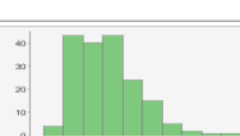
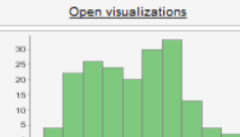
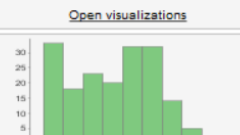
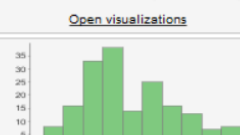
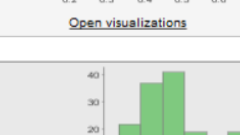
**Variable objetivo:** consiste en la clase de vino que puede tener valores 1, 2 o 3.

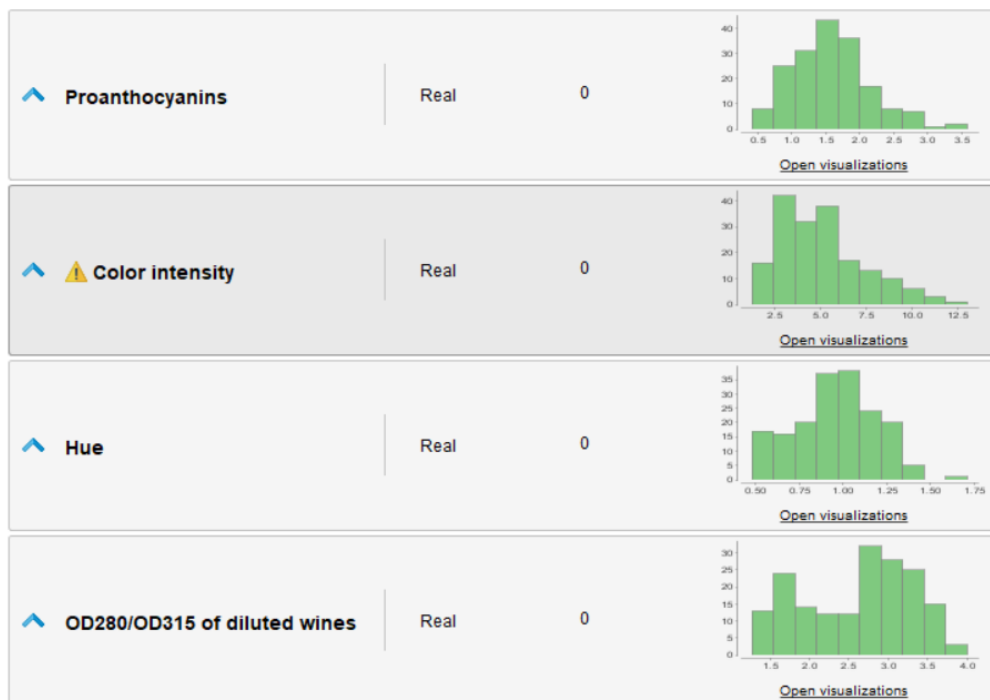
### Atributos:

- 1) Alcohol: Porcentaje de alcohol presente en vino
- 2) Malic acid: Concentración de ácido málico en el vino. Medida (g/l)
- 3) Ash: Se define como toda la materia inorgánica que queda tras la ignición del residuo que queda de la evaporación del mosto o del vino. Medida (g/l).
- 4) Alkalinity of ash: La alcalinidad del ash se define como la suma de cationes, distintos del ion amonio, combinados con los ácidos orgánicos en el vino. Medida: (g/l)
- 5) Magnesium: La cantidad de magnesio que contiene una botella de vino. Medida (mg)
- 6) Total phenols: Es la concentración de fenoles en el vino. Medida (mg/l)
- 7) Flavanoids: Es la concentración de flavanoides en el vino. Medida (mg/l)
- 8) Nonflavanoid phenols: Es el porcentaje de Fenoles no flavonoides presentes en el vino. Medida (%).
- 9) Proanthocyanins: Es la concentración de Proantocianidinas en el vino. Medida (mg/l)
- 10) Color intensity: La intensidad de color del vino.
- 11) Hue: El tono del vino
- 12) OD280/OD315 of diluted wines: Es la concentración de proteínas en el vino. Medida (mg/l).
- 13) Proline: Es un aminoácido que regula el sabor del vino, se mide su concentración. Medida: (mg/l).

Todos los atributos del dataset son ordinales (reales y enteros).

## Distribuciones de los atributos

Alcohol	Real	0	
Malic acid	Real	0	
Ash	Real	0	
Alcalinity of ash	Real	0	
Magnesium	Integer	0	
Total phenols	Real	0	
Flavanoids	Real	0	
Nonflavanoid phenols	Real	0	
Proline	Integer	0	



## Análisis de Normalización/Estandarización

Para este dataset tomamos en cuenta sus atributos y llegamos a las siguientes conclusiones:

- En los algoritmos de clasificación es importante aplicar estandarización o normalización a datos que están en diferentes unidades de medidas a efectos de compararlos o combinarlos.
- Dentro del dataset Wine se encuentran casos como este por ejemplo con el atributo alcohol que es un porcentaje y el ácido málico que se mide en g/l. Hay varios atributos que corresponden a distintos tipos de medidas: concentraciones, porcentajes e incluso colores que no sabemos cómo se miden.
- Como se puede visualizar en las imágenes dispuestas anteriormente, la mayoría de los atributos en el dataset tienen una distribución que se asemeja a una gaussiana y por lo tanto sería conveniente normalizarlos.
- Es prudente estandarizar atributos que tienen una varianza muy alta, en este caso el atributo que tiene mayor varianza es Proline.
- Además las medidas de los atributos que aparecen arriba son inferidas en base a información obtenida fuera de la descripción del dataset. Esto es un buen motivo para estandarizar todos los atributos numéricos dado que es una medida prudente y los hace comparables.