# 1 Analisando Vinhos

**Objetivo:** Familiarização com a linguagem Python e com as bibliotecas numpy, sklearn, pandas, matplotlib, etc.

Neste projeto, iremos empregar uma base de dados reais contendo informações de cultivo de vinho branco e tinto em plantações de Portugal. Serão empregados conceitos de análise exploratória dos dados, pré-processamento dos dados, técnicas clássicas de filtragem e balanceamento de classes no contexto de análise de variáveis psico-químicas de cultivos de vinho.

## 1.1 DESENVOLVIMENTO E QUESTÕES

1. Baixar a base de dados disponível no repositório da *UCI machine learning*: `ttps://archive.ics.uci.edu/ml/datasets/wine+quality`. Gere uma única planilha para os dados de vinho branco e tinto.
2. Visualize a dimensão da matriz de dados e do vetor de rótulos.
3. Realize uma análise criteriosa dos dados da base de dados. Pontos que devem ser considerados.
   - **i.** A base de dados é consistente?
   - **ii.** Há dados faltantes?
   - **iii.** Há dados não numéricos?
   - **iv.** A base de dados é balanceada (considere as classes vinho tinto e branco).

   Sugestão: leia o artigo *Modeling wine preferences by data mining from physicochemical properties. Cortez et al., 2009*.
4. Apresente uma análise estatística dos dados que embase quais variáveis de entrada são mais relevantes se quisermos classificar o vinho pelo tipo (tinto vs. branco). Justifique.
5. Escolha duas das variáveis de entrada e faça um gráfico de dispersão para visualizar a distribuição dos dados de cada classe de vinho. Justifique sua escolha.
6. Considerando agora a variável de qualidade do vinho, avalie como é a prevalência nas duas classes para os vinhos mais bem avaliados ($nota > 7$) e para os avaliados com $nota < 3$.