

# Introdução à Ciência de Dados

## Projeto final

O projeto final deve ser um notebook do Jupyter com todos os passos comentados e os respectivos códigos em Python. Use os notebooks das aulas como exemplo.  
O projeto deverá ser enviado até 04 de Abril de 2020.

Considere a base de dados sobre vinhos (wine).

<https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009>

### 1. Pré-processamento

Realize todas as fases de pré-processamento que vimos no curso, incluindo:

- \* Remoção de atributos faltantes,
- \* Normalização ou padronização.
- \* Transformação para valores numéricos (one-hot-encoding) se precisar.

### 1 – Classificação:

Classifique os vinhos de acordo com a qualidade. Compare os métodos:

- \* k-vizinhos (encontre o melhor k usando validação cruzada)
- \* Árvore de decisão
- \* Naive Bayes
- \* SVM (encontre o melhor C usando validação cruzada)
- \* Random Forest (encontre o melhor número de estimadores usando validação cruzada)

### 3. Ordenação dos atributos

Para o algoritmo random forest, mostre a importância de cada atributo.

### 4. Regressão

- \* Usando regressão linear, tente prever a porcentagem de álcool.
- \* Compare os métodos Lasso, Ridge Regression, calculando o erro quadrático médio em função dos seus parâmetros (alpha).

### 5. Conclusão

Escreva uma conclusão (um ou dois parágrafos) discutindo os resultados.