

## LISTA 4 — Pré-processamento: tratar os dados (normalização e redução de dimensionalidade)

## Exercício 1 - Cartão de Crédito

Disponibilizei um arquivo chamado CC\_DATASET.csv:

- CUST\_ID
- BALANCE
- BALANCE\_FREQUENCY
- PURCHASES
- ONEOFF\_PURCHASES
- INSTALLMENTS\_PURCHASES
- CASH\_ADVANCE
- PURCHASES\_FREQUENCY
- ONEOFF\_PURCHASES\_FREQUENCY
- PURCHASES\_INSTALLMENTS\_FREQUENCY
- CASH\_ADVANCE\_FREQUENCY
- CASH\_ADVANCE\_TRX
- PURCHASES\_TRX
- CREDIT\_LIMIT
- PAYMENTS
- MINIMUM\_PAYMENTS
- PRC\_FULL\_PAYMENT
- TENURE

Fonte dos dados: kaggle.com/arjunbhasin2013/ccdata

Há propositalmente algumas dificuldades nos dados:

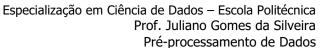
- Aparentemente as ausências já estão preenchidas com zeros, porém, nem todas...
- Há alguns poucos valores extremos espalhados pelas variáveis do dataset.

Antes de mexer nos dados, escolha algumas variáveis e experimente calcular valores médios. Tome nota dos valores que resultam.

## Pré-processamento dos dados

- 1. Importe os dados;
- 2. Tratamento dos outliers:
  - a. Aqui você é quem vai escolher o método de tratamento (excluir registros? sobrescrever com zeros? sobrescrever com a média ou mediana?). Analise os dados, e sob o seu julgamento, escolha a melhor forma de tratar os outliers.
  - b. Ao final, explique porque você tratou usando o método x, y, z...
- 3. Valores ausentes:
  - a. Nesse contexto (cartão de crédito) as ausências devem ser legítimas, ou seja, não existe informação mesmo. Portanto, trate as ausências com zero.
- 4. Normalização:
  - a. O dataset é praticamente todo composto por variáveis numéricas. Aplique a normalização Standard Scaler sobre essas variáveis numéricas.

## Modelagem







- 1. Aplique K-means sobre os dados separando em 2 clusters (k=2) e grave o resultado no df original;
- 2. Aplique K-means novamente, porém, agora k=3 (essa é uma nova clusterização e não deve sobrescrever a clusterização anterior). Grave esse resultado de clusterização no df original também.

#### Análise de dados:

- 1. Escolha duas métricas e plote um gráfico de dispersão, separando por cor os pontos da clusterização k=2.
- 2. Repita o gráfico da questão anterior para k3. Visualmente qual a clusterização ficou melhor?
- 3. Quantos casos ficaram enquadrados em cada cluster em cada uma das clusterizações?
- 4. Qual a média e a mediana de valor de compras, de cada cluster em cada uma das clusterizações?
- 5. Usando a clusterização k=3, qual a mediana de limite de crédito de cada grupo?

# Exercício 2 - Diagnóstico de Câncer

Disponibilizei um dataset clássico: *Breast Cancer Wisconsin*. Nele há informações de diagnósticos de câncer de mama (arquivo **DIAG\_CANCER.csv**).

O arquivo possui um layout adequado para tarefas de classificação, contendo uma variável de desfecho ("diagnosis") que informa se o câncer foi diagnosticado como benigno (B) ou maligno (M). As demais variáveis são medidas discretas/contínuas que explicam o desfecho ("diagnosis"). Abaixo a lista completa de variáveis:

- ic
- diagnosis
- radius\_mean
- texture\_mean
- perimeter\_mean
- area\_mean
- smoothness\_mean
- compactness\_mean
- concavity\_mean
- concave points\_mean
- symmetry\_mean
- fractal\_dimension\_mean
- radius\_se
- texture\_se
- perimeter\_se
- area\_se
- smoothness\_se
- compactness se
- concavity\_se
- concave points\_se
- symmetry\_se
- fractal\_dimension\_se
- radius worst
- texture worst
- perimeter\_worst
- area\_worst
- smoothness\_worst
- compactness\_worst



Especialização em Ciência de Dados – Escola Politécnica Prof. Juliano Gomes da Silveira Pré-processamento de Dados

## Lista de Exercícios 4

- concavity\_worst
- concave points\_worst
- symmetry\_worst
- fractal\_dimension\_worst

• Unnamed: 32

Fonte dos dados: kaggle.com/uciml/breast-cancer-wisconsin-data

## Pré-processamento dos dados

- 1. Importe os dados;
- 2. Tratamento dos outliers:
  - a. Temos outliers apenas na extremidade superior da variável "radius\_mean". Trate esses valores extremos substituindo-os pela mediana da respectiva variável.
- 3. Valores ausentes:
  - a. A última variável do dataset (Unnamed: 32) só possui valores ausentes. Exclua essa variável.
- 4. Normalização:
  - a. Com exceção dos campos *id* e *diagnosis* o dataset só possui variáveis numéricas. Aplique a normalização *Standard Scaler* sobre essas variáveis numéricas.
- 5. Redução de Dimensionalidade:
  - a. Aplique PCA sobre os dados normalizados para obter 5 variáveis (PCA\_1, PCA\_2, PCA\_3, PCA\_4 e PCA\_5).
  - b. Crie um novo df com as variáveis resultantes do PCA (5) e adicione nele a variável de desfecho (*diagnosis*).

### Modelagem

- Crie um modelo de classificação (árvore aleatória, regressão logística, a sua escolha...), aplique sobre os dados originais (dataframe completo, com as 30 variáveis explicativas). Tome nota da acurácia do modelo e os resultados de matriz de confusão.
- 2. Repita o processo (modelagem), só que dessa vez treine e teste o modelo usando o df PCA que você criou anteriormente (5 variáveis explicativas + variável de desfecho). Tome nota da acurácia do modelo e os resultados de matriz de confusão.
- 3. Nesse exercício reduzimos a complexidade computacional na ordem de 6x (de 30 variáveis explicativas para apenas 5). Na sua avaliação observando os resultados de treino/teste, de ambos os modelos, perdemos acurácia? Se sim, quanto?