



# Atividade 2

## Exploração e Análise de Dados em Big Data

Prof<sup>a</sup>. Dr<sup>a</sup>. Emilia Colonese Carrard

Esta atividade deve ser realizada em grupo.

O contexto a seguir define o requisito de análise a ser implementado com uma tarefa análise de dados em ambiente de Big Data.

O ambiente de análise a ser utilizado para este trabalho são o RSpark ou PySpark.

### Contexto:

Uma Seguradora possui o histórico de seus clientes e o comportamento destes em relação a compras de seguros imobiliários contraídos previamente. O gerente da área de vendas precisa de informação sobre potenciais clientes da empresa para adquirir seguro imobiliário. Para obter tal informação, é necessário saber o perfil de clientes que têm alta probabilidade de comprar seguro imobiliário. O resultado será utilizado pela equipe de vendas para oferecer seguros a novos clientes.

Os dados dos clientes da Seguradora está no arquivo customer.csv.

### O aluno deverá realizar os seguintes passos para a exploração de dados:

#### 1. Carregamento dos Dados

- Carga de dados do arquivo para o ambiente de trabalho.

#### 2. Exploração dos Dados

- Entender o conjunto de dado fornecido, executando:
  - Testes estatísticos de correlação das variáveis em relação à variável alvo.
  - Testes de hipóteses (t-value, e chi-squared).
  - Sumarizações; e
  - Visualizações gráficas de distribuição de cada variável.

#### 3. Transformar dos Dados

- Limpar, transformar e selecionar os dados relevantes para o contexto da análise.

#### 3. Preparar o conjunto de dados para a tarefa de análise

- De acordo com os dados e o requisito de análise escolher o tipo de tarefa a ser executada na análise.
- Preparar o *dataset* para se adequar ao algoritmo de análise selecionado para atender ao requisito especificado.
- Durante a construção do conjunto de dados, novas informações podem ser criadas.

4. Aplicar a tarefa de mineração

- Execução do algoritmo de análise escolhido para entregar a resposta esperada.
- Geração de resultados (métricas, plotagens ou visualizações) próprios do algoritmo. Mostre a figura do resultado se houver.
- Análise dos resultados.
- Explicação dos resultados.

5. Apresentar (descrever) o perfil de clientes com alta probabilidade de compra de seguro imobiliário.

Os grupos devem postar o seu trabalho (arquivo .pdf ou .doc) na pasta da Atividade 2 do *classroom*, contendo os tópicos de 1 a 5 descritos acima.