MACHINE LEARNING

Curso de especialização em Ciência de Dados e Big Data

Prof. Hugo de Paula

ATIVIDADE EM LABORATÓRIO 3

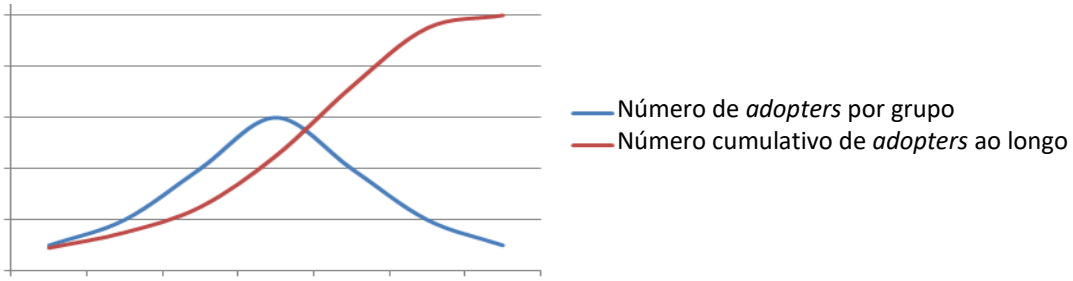
ÁRVORES DE DECISÃO

Adaptado de North, Matthew A. Data Mining for the Masses. 2012.

OBJETIVOS

1- O objetivo dessa atividade é classificar pessoas pela sua propensão a comprar um novo leitor digital (eReader), de modo a melhorar a efetividade de campanhas de marketing.

2- Para alcançar este objetivo, iremos utilizar o modelo de árvores de decisão para minerar hábitos de consumo de usuários de um site de e-commerce. O sociólogo Everett Rogers publicou, na década de 60, um trabalho que mostrava a adoção de novas tecnologias por consumidores. Ele identificou os seguintes grupos de consumidores: *inovators, early adopters, early majority e late majority*. Os dois primeiros grupos compreendem os usuários mais propensos a comprar uma nova tecnologia, enquanto os dois últimos eventualmente poderão comprar a tecnologia, se achar conveniente.



3- Resumo dos dados:

a. ***ID_usuário***: um identificador numérico associado a cada pessoa que possui uma conte no site.

b. ***Gênero***: M (masculino) ou F (feminino).

c. ***Idade***: numérico.

d. ***Estado_civil***: C – casado, S – não casados (solteiros, viúvos, divorciados, etc.).

e. ***Atividade_no_site***: Esporádico, Intermitente, Frequente.

f. ***Pesquisou_eletronicos_12m***: Sim/Não indicando se é o usuário andou pesquisando eletrônicos no site nos últimos 12 meses.

g. ***Comprou_eletronicos_12m***: Sim/Não indicando se é o usuário comprou eletrônicos no site nos últimos 12 meses.

h. ***Pesquisou_mídia_digital_18m***: Sim/Não indicando se é o usuário andou pesquisando mídia digital (como mp3) no site nos últimos 18 meses.

i. ***Comprou_mídia_digital_18m***: Sim/Não indicando se é o usuário comprou mídia digital (como mp3) no site nos últimos 18 meses.

j. ***Forma_pagamento***: Transferência; Website; Cartão e Boleto bancário.

- ***Adoção_eReader***: Esse atributo será o rótulo classificador (***label***)
    - ***Inovator***: quem comprou eReaders de geração anteriores até 1 semana após o lançamento.
    - ***Early Adopter***: quem comprou eReaders de geração anteriores entre 1 e 3 semanas após o lançamento.
    - ***Early Majority***: quem comprou eReaders de geração anteriores entre 3 semanas e 2 meses após o lançamento.
    - ***Late Majority***: quem comprou eReaders de geração anteriores após 2 meses do lançamento.

PREPARAÇÃO DOS DADOS

1- Serão usadas duas bases de dados disponíveis em duas planilhas do arquivo ***Atividade 3 - Bases.xlsx***. A primeira planilha se chama ***TREINAMENTO*** e contém os registros já classificados. A segunda planilha se chama ***VALIDAÇÃO***, e contém novos registros sem classes definidas para podermos avaliar o modelo.

2- Analise os dados usando a descrição estatística. Você perceberá que, aparentemente, não há valores omissos nem inconsistentes, mas ainda temos que preparar os dados.

3- O campo ***ID_usuário*** não tem relação com os consumidores, propriamente dito e não deverá ser utilizado na modelagem.

4- A classe destino deve ser o campo ***Adoção_eReader***.

MODELAGEM

1- Vamos construir a árvore de decisão para os dados de treinamento.

2- Lembrando-se que o processo de *Data Science* sugere que **o aprendizado de máquina é um processo iterativo**, volte ao modelo e **altere os valores dos diversos parâmetros da árvore**.

   a. Altere o critério de seleção de atributos para e veja o resultado.

   b. Para evitar o *overfitting*, habilite as opções de ***prunning***. Altere valores e simplifique a árvore de decisão para uma complexidade que julgar aceitável.

   c. Baseado nesses valores, tente interpretar os resultados. Por exemplo, tente entender qual o potencial de alcance daquela regra na população. Quantos registros ela está representando. Analise o nó folha e identifique se as "confusões" geradas por cada regra podem ajudar a diferenciar uma regra boa de uma regra ruim.

A atividade deve ser entregue na forma de um Notebook desenvolvido no Google Colab (https://colab.research.google.com/), contendo os nomes dos integrantes da dupla, e compartilhado com o e-mail hugodepaula@gmail.com, até 7 dias após a realização na aula.