Projeto 3

Ciência dos Dados 2017

Este documento apresenta as premissas do projeto final de Ciência dos Dados.

Objetivos

O principal objetivo do Projeto 3 é conduzir uma análise de dados com grau elevado de autonomia e liberdade de escolha de tema e de técnica.

Para que este fim possa ser alcançado, os estudantes deverão se aprofundar na técnica escolhida enquanto realizam o projeto.

É importante que o trabalho produza uma conclusão analítica e vá além da análise exploratória.

Grupos

O projeto pode ser realizado em grupos de no máximo 3 alunos.

Datas

Data	Entregável
26/10	Definição de grupo e até 3 propostas de tema (técnica e dataset) por grupo
31/10	Até 18 horas ter escolhido dataset e tema
1/11	Devolutiva dos professores
7/11	Check intermediário - dataset lido e exemplo de aplicação da técnica
14/11	Análises concluídas - algoritmo gera resposta
16/11	FIM: Relatório com conclusões e fundamentação teórica

Sugestões de temas a utilizar

1. Teste A/B

Compare o desempenho entre duas alternativas A e B.

Exemplos:

- Qual *user interface* é melhor para uma dada tarefa?
- Este tratamento é mais efetivo que outro?

Para seguir esta vertente de projeto, recomendamos uma de duas alternativas:

- Conduzir um experimento. Veja a pasta TesteAB sobre como conduzir um experimento online
- Encontrar um dataset em que um experimento foi conduzido

2. Regressão (linear ou logística)

Prever o valor de uma coluna de um dataset em função das outras. Pode ser uma regressão linear (se a variável de saída for quantitativa) ou regressão logística (se a variável de saída for qualitativa)

Exemplos de datasets:

Predição de preços de casas em King County, Seattle

Predição de por quanto uma casa vai ser vendida

Predição de se funcionário vai deixar empresa ou não

Predição de qual *rating* alguém vai dar para um filme no Netflix

3. Classificadores - extensão do Naive Bayes

Baseado em todos os dados existentes, classificar em categorias

Exemplos de datasets:

Porto Seguro - cliente vai acionar o seguro?

Deteção de fraude no cartão de crédito

Deteção de fraude financeira

Predição de se funcionário vai deixar empresa ou não

Predição de sucesso de um filme

4. Clusterização

Agrupe os dados de um conjunto baseado em similaridade. Neste problema em geral pode-se escolher o número de *clusters* e o algoritmo precisa fazer o agrupamento.

Datasets interessantes para esta técnica

Pokémon

Fifa 18

Datasets interessantes

Ainda não há pergunta definida, mas são datasets interessantes

Futebol Europeu

Reviews de smartphones na Amazon

Filtro Anti Spam

Dataset da Enron. Mensagens classificadas em relação a assunto e sentimento

Predição se um produto entrou em falta - *backorder* - ou não

Lista de todos os datasets do Kaggle

Alguns datasets disponíveis publicamente