Projeto Prático 01

O naufrágio do RMS Titanic é um dos mais famosos da história. Em 15 de Abril de 1912, durante a sua viagem inaugural, o navio naufragou após colidir com um iceberg, matando 1502 dos seus 2224 passageiros e tripulantes. Uma das razões apontadas como causa de tantas mortes foi a falta de botes salva-vidas. Apesar de o fator "'sorte" também existir, alguns grupos tinham mais probabilidade de sobrevivência, tais como mulheres, crianças e pessoas de classes sociais mais altas.

Tarefa

Neste projeto, o objetivo é utilizar diferentes classificadores para identificar quais categorias de pessoas possuem mais chance de sobrevivência. O problema e a base de dados foram extraídos da plataforma *Kaggle* (`https://www.kaggle.com/`). Use como base o arquivo disponibilizado (`titanic_projeto01.py`).

Objetivos

- Inicialmente, você vai precisar manipular os dados de entrada, visando corrigir problemas (por exemplo, falta de dados no campo `Idade`) ou criar características a partir de dados não estruturados (por exemplo, extrair a categoria a partir do campo `Nome`). Para tal, você pode explorar os programas já disponibilizados pelos usuários da plataforma (alguns links estão listados abaixo).
- Após a padronização dos dados, a tarefa consiste em identificar as características potencialmente mais relevantes (ou seja, que conduzem a melhores taxas de classificação). Para tal, uma boa prática consiste em realizar uma análise exploratória, procurando identificar padrões e correlações na base de dados. A categorização de dados também é um passo importante para a seleção do modelo de dados.
- Além disso, é preciso também descobrir qual o melhor classificador, bem como os melhores parâmetros de cada classificador. Neste projeto, utilize (ao menos):
  1. Árvores de decisão
  2. *Random Forests*
  3. *Naive* Bayes
  4. K-*Nearest Neighbors*
- Não se esqueça que não é apenas o resultado final que importa. Salve os parâmetros e os resultados obtidos em cada etapa. Ao final, apresente uma discussão dos aspectos principais de todo o processo (que ações foram importantes para melhoria dos resultados, etc).

Entrega

O trabalho deve ser desenvolvido em equipes de até três pessoas. Cada equipe deve submeter (via moodle) um arquivo .zip com:

- Os arquivos Python com as implementações realizadas. Não se esqueça de comentar adequadamente o código.
- Um arquivo pdf com um breve relatório.

Links

Sugestões de links:

- Geral: https://www.kaggle.com/c/titanic/kernels
- https://www.kaggle.com/ldfreeman3/a-data-science-framework-to-achieve-99-accuracy
- https://www.kaggle.com/startupsci/titanic-data-science-solutions
- https://www.kaggle.com/sachinkulkarni/titanic/an-interactive-data-science-tutorial
- https://www.kaggle.com/creepykoala/titanic/study-of-tree-and-forest-algorithms
- https://www.kaggle.com/dmilla/titanic/introduction-to-decision-trees-titanic-dataset/notebook