Projeto Aprendizado de Máquina Preditivo
CMC- 15 Inteligência Artificial
Profa. Ana Carolina Lorena
Trabalho em Grupo de Três ou Quatro Alunos

1. Objetivo

Exercitar e fixar conhecimentos adquiridos sobre Aprendizado de Máquina preditivo, por meio do uso e comparação de técnicas do paradigma de aprendizado supervisionado em um problema prático.

2. Descrição do Trabalho

O trabalho envolverá analisar dados para prever a presença de meteoros em imagens coletadas de um ponto de observação na cidade de São José dos Campos. Uma descrição do problema pode ser encontrada em: https://www.facom.ufu.br/~kddbrcompetition/

Os dados estão em:
https://www.kaggle.com/competitions/can-i-make-a-wish-detecting-shooting-stars/overview

O que deve ser feito:

a) Cada imagem foi descrita por um conjunto de características extraídas por métodos distintos de processamento de imagens. São 21 conjuntos de características. Realize uma exploração de **cinco** desses subconjuntos de treinamento (*escolhidos por vocês*), usando estatísticas descritivas e visualizações pertinentes. Analise e discuta seus resultados.

b) Avalie o desempenho preditivo de classificadores do tipo k-vizinhos mais próximos (kNN) e de árvores de decisão nestes cinco conjuntos de treinamento, usando validação cruzada com 5 pastas. Dentro de cada treinamento, realize a busca pelo hiperparâmetro k do kNN entre os valores 1, 3 e 5 usando uma validação cruzada interna com 3 pastas. O desempenho deve ser reportado pela média e desvio-padrão da medida desempenho log-loss na validação cruzada externa. Os resultados devem ser reportados em uma tabela, como segue:

Conjunto	**kNN**	**AD**
AutoColorCorrelogram	0.3 ± 0.2	0.4 ± 0.1
CEDD	…	…
…		

c) Realize agora o teste dos modelos treinados anteriormente nos conjuntos de teste independentes, usando a plataforma Kaggle. Reporte os resultados no formato de uma tabela, semelhante à anterior. Neste caso, não é necessário fazer a validação cruzada, basta treinar no conjunto train inteiro e testar no test correspondente.

d) Compare, analise e discuta os resultados alcançados.

Atenção ao *data leakage*, tentem usar a estrutura de pipelines do sklean para fazer seus experimentos. Fiquem atentos às eventuais necessidades de pré-processamento, justifiquem suas escolhas no material entregue.

3. Material a ser entregue e prazo

Material: Notebook com as implementações, resultados e discussões
Prazo de Entrega: 20/setembro/2024

Estrutura sugerida:

Nomes dos Membros da Equipe
1. Análise Descritiva dos Dados
2. Resultados da Validação Cruzada
3. Resultados em Conjunto de Teste Independente
4. Discussões
5. Conclusões: Comentários e sugestões sobre o trabalho
(complexidade/facilidade, sugestões, etc.).

Bom Trabalho!
Profa. Ana Carolina Lorena
aclorena@ita.br