Comparação de classificadores para dataset Kaggle

Tiago Kenji Umemura

Dataset

- Dataset de classificação de cogumelos
  - Duas classes: Venenoso ou não
  - 22 características

- Necessário converter as características representadas por letras para números.
  - Utilizar if

Kernel Kaggle

- Kernel disponibilizado no Kaggle
  - Verifica quais características possuem maior impacto ao determinar a classe da instância
  - Odor foi a característica de maior influência
  - Random Forest, GBM e RPART
  - Taxa de 100% ao classificar 2000 instâncias

Método

- Utilizar 3 classificadores:
  - Random Forest
  - SVM
  - Nearest Centroid

- Random Forest com 10, 100 e 500 árvores
  - 10 vezes para cada número de árvore
  - calcular a média
- SVM e Nearest Centroid
  - Executado apenas uma vez

Resultado

Random Forest			
Numero de árvores	**Mínimo**	**Máximo**	**Média**
10	0,5316	0,7376	0,5853
100	0,5858	0,8222	0,7746
500	0,6569	0,8182	0,7712

Resultado

SVM	0,8247
Nearest Centroid	0,8182

Conclusão

- SVM apresenta melhor desempenho (0,8247)

- Random Forest apresenta resultados bons resultados mas somente um grande número de árvores, o que aumenta o tempo para executar

- Na Random Forest, as características mais relevantes para determinar as classes foram: odor, gill size e stalk shape