Saulo Pinedo, Bruna Santos e Caue Oliveira
Silvia Silva
8 de outubro de 2018

Relatório de Atividade

Datasets, classificadores e árvores de decisão

A motivação do experimento

Nosso grupo procurou aplicar os conhecimentos de algoritmos inteligentes em um tema bastante real: a distinção entre tumores benignos e malignos. O câncer é uma doença que acomete pessoas de todo o mundo, e como sabemos, existem diversos fatores que podem contribuir para o seu surgimento. Não é de se esperar o pior sem ao menos realizar uma consulta médica - o tumor possui características notáveis dentro da medicina que permitem distingui-lo dentre os dois quadros possíveis. Por se tratar de algo bastante ocorrente na sociedade, e por haver somente dois resultados possíveis, optamos por aplicar os algoritmos aprendidos em aula para construir um classificador usando um dataset de tumores.

Os desafios do experimento

Fazer um experimento dessa proporção nos trouxe dificuldades medianas, porém nada que não pudéssemos resolver. Tivemos certas incoerências nos comandos do terminal do Linux, cuja situação nos fez perder boa parcela do tempo disponível. Outros pequenos detalhes relacionados às bibliotecas também foram um pouco controversos no início, mas conseguimos resolver logo. Os algoritmos não foram complicados de implementar, afinal, todo o material da disciplina estava à nossa disposição. Seguimos todos os slides cuidadosamente, e, no final, obtivemos os resultados que esperávamos.

Os objetivos do experimento

O objetivo geral dessa atividade é testar a eficácia do classificador com valores reais baseado na sua "intensidade" de treino. Quanto mais treinamento o classificador recebe, mais eficaz ele se torna. Separamos alguns valores para treinamento e teste final, conforme descritos mais adiante.

A composição de todo o experimento se deve aos objetivos secundários:

1. Escolher um dataset;
2. Fazer uma análise de histograma;
3. Desenvolver um classificador usando os métodos Árvore de Decisão e KNN;
4. Analisar a sensibilidade do método (acurácia x split).

A metodologia aplicada

Como visto anteriormente, a escolha do dataset não foi aleatória: utilizamos o critério de simplicidade dos resultados para tornar o nosso trabalho mais notável.

Utilizamos do material disponibilizado na plataforma do Moodle alguns algoritmos para seguir com a implementação da Árvore de Decisão e do KNN.

Finalmente, fizemos cinco testes para as situações onde o classificador recebia pouco treinamento ou grande treinamento. Como já esperávamos, os resultados de cada situação não possuíram grandes semelhanças.

Os resultados do experimento

Simulamos três situações com amostragens diferentes. Para cada situação, realizamos cinco testes consecutivos. Segue abaixo as acurácias procuradas:

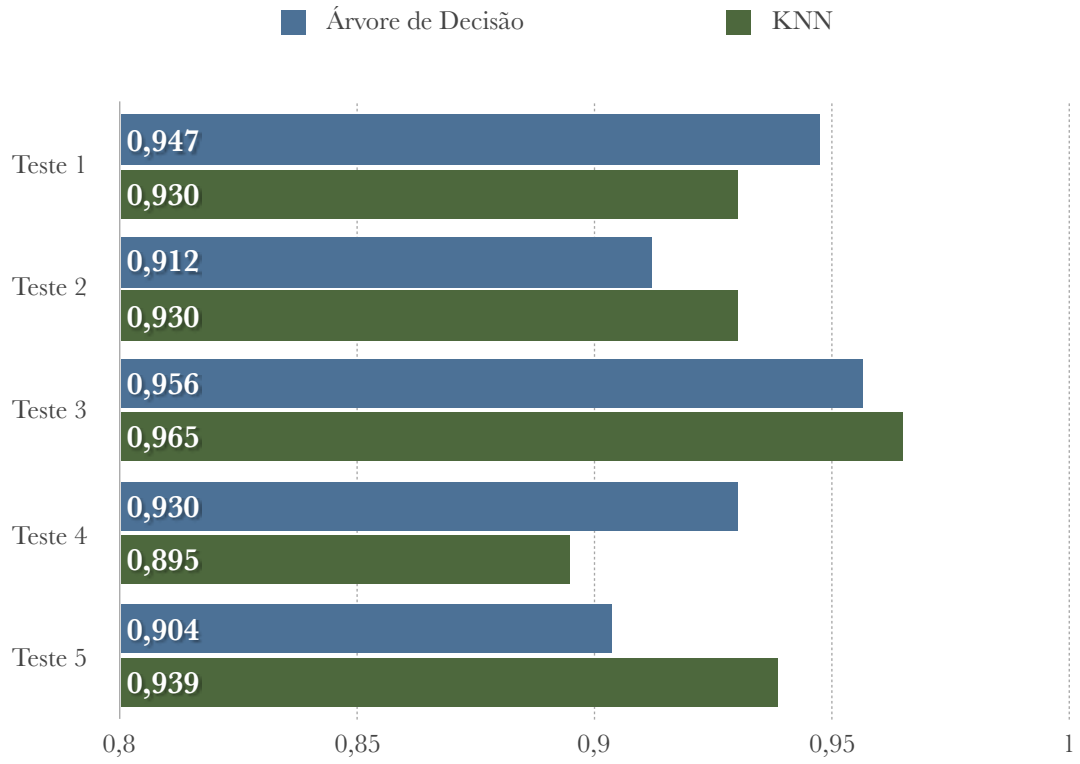

Gráfico 1 - Situação com 80% das amostras voltadas para o treinamento.

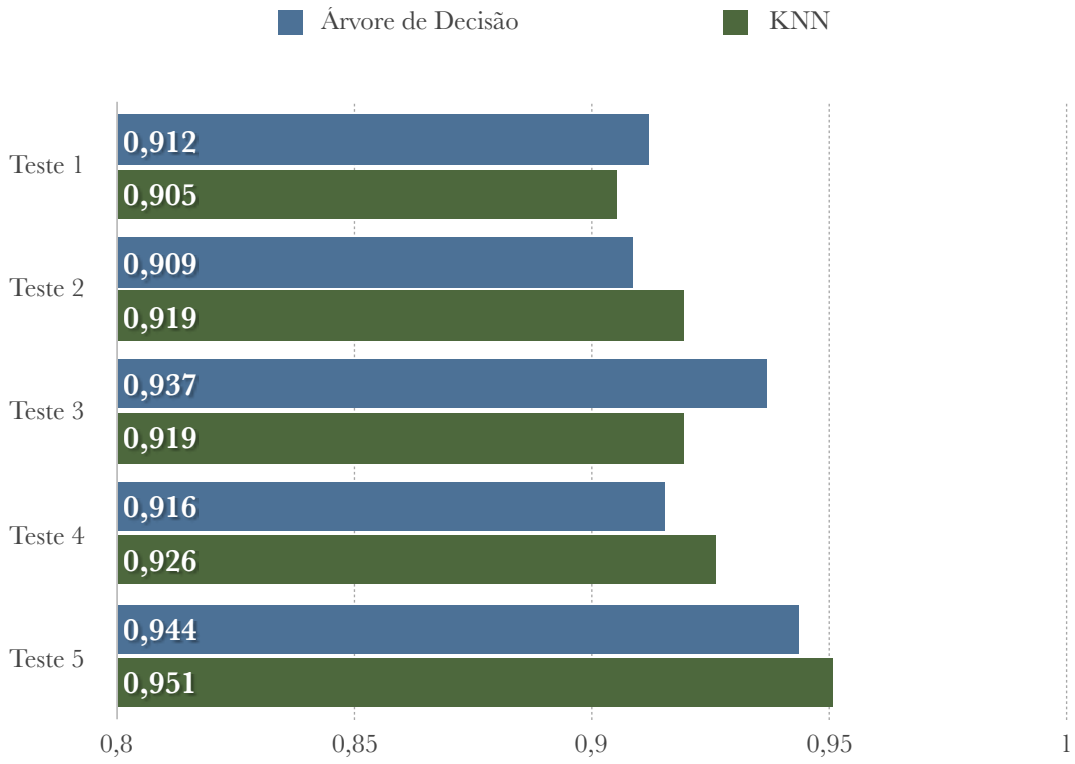

Gráfico 2 - Situação com 50% das amostras voltadas para o treinamento.

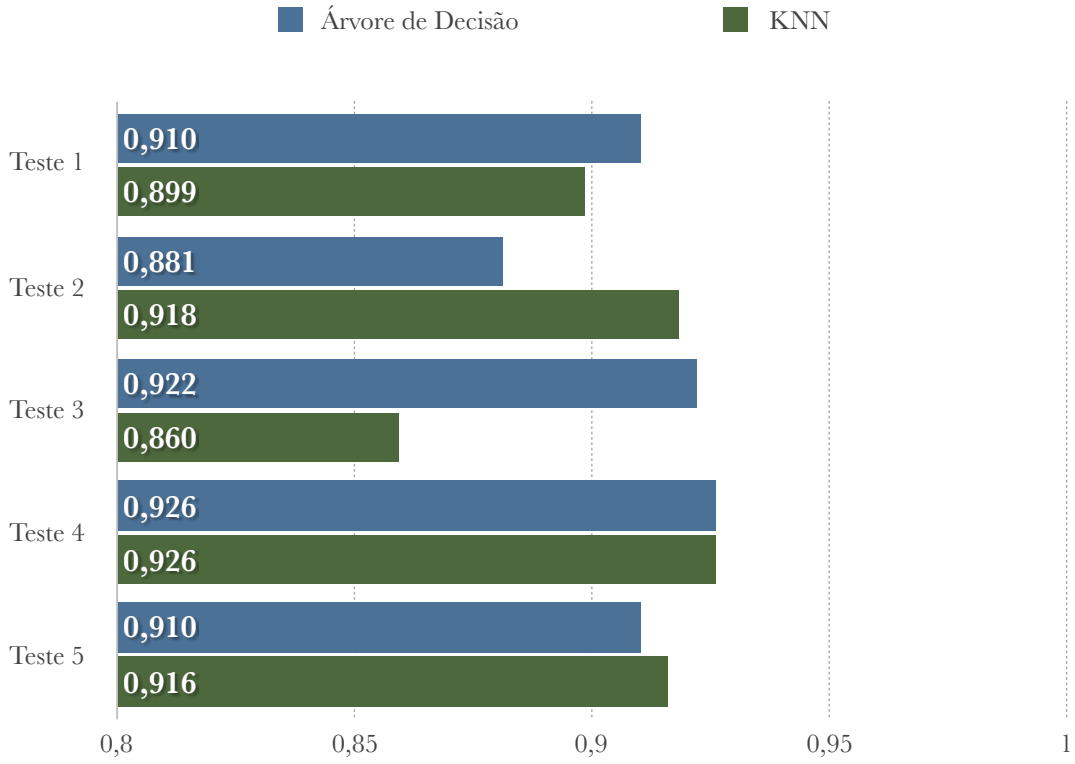

Gráfico 3 - Situação com 10% das amostras voltadas para o treinamento.

Conclusão

Concluímos que ao aplicar os classificadores nesse assunto, obtemos resultados melhores com taxas de erros menores, levando assim a resultados ágeis, e consequentemente, salvando mais vidas.