# O IMPACTO DE ATRIBUTOS IRRELEVANTES NA PRECISÃO DOS DIFERENTES MÉTODOS DE APRENDIZADO

Henrique de S. **Oliveira**[1]; Jacques **Wainer**[2]

**INSTITUTO DE COMPUTAÇÃO – IC, UNICAMP**

Agência: CNPq

Machine-Learning-Artificial-Intelligence-Irrelevant-Random-Attributes

## INTRODUÇÃO

Um algoritmo de aprendizado é tipicamente representado por um conjunto de dados em N-dimensões. O algoritmo deve, a partir de pontos previamente demarcados, classificar um novo ponto a ser inserido no conjunto. Na prática, um conjunto de dados inicial deve ser cuidadosamente selecionado para aprimorar o aprendizado e assim aumentar a precisão do algoritmo. Em alguns casos o conjunto de atributos de entrada são bastante grandes contendo uma fração deles relevante para a função de classificação e outra não, quanto maior a quantidade de atributos irrelevantes maior a degradação da performance do algoritmo de aprendizado.

Existem algumas técnicas para seleção de variáveis e remoção de atributos, porém este não é o foco deste trabalho e sim a medição de quão impactante é a adição de diversas quantidades de atributos irrelevantes ao conjunto de dados inicial. A motivação para a pesquisa nessa área foi a falta de bibliografia e informações a respeito de como e quanto esses atributos afetam os diferentes métodos de classificação já desenvolvidos.

## METODOLOGIA

No experimento realizado, foram utilizados quatro algoritmos de aprendizado de máquina: Naive Bayes, Decision Trees, Support Vector Machines (SVM) e K-Nearest-Neighbors (k-NN), sendo os dois primeiros realizados com a versão estritamente categórica dos cinquenta conjuntos de dados de pesquisa e os outros dois algoritmos testados com a versão estritamente numérica.

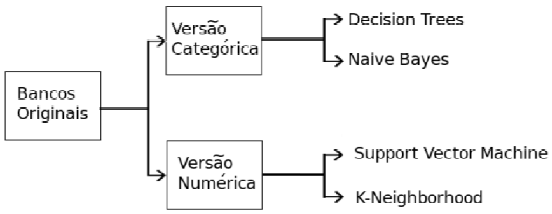


Figura 1 – Fluxo da formatação dos conjuntos de dados para os testes

Posteriormente foram adicionados atributos com valores gerados aleatoriamente em uma quantidade diretamente proporcional aos atributos presentes no conjunto de dados original, para se obter frações definidas de conteúdo inicial e conteúdo "artificial" que é o foco do estudo nos resultados dos testes. Finalmente todos os conjuntos de dados (originais e manipulados) foram submetidos aos quatro algoritmos implementados pelo Weka utilizando o método 5-Fold.

## RESULTADOS E DISCUSSÃO

Para todos os quatro algoritmos foram plotados gráficos com o valor relativo de acerto e a quantidade de atributos aleatórios inseridos no conjunto. A análise da degradação do desempenho foi explicitada com regressão linear aplicada a todos os pontos dos gráficos.

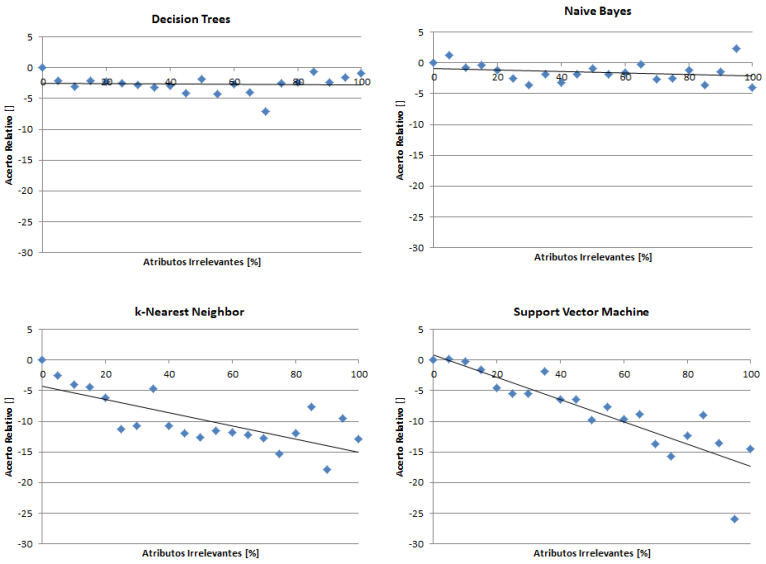


Figura 2 – Gráficos com regressão linear para os quatro algoritmos

Analisando os gráficos e suas respectivas regressões notamos que os algoritmos que utilizam conjuntos de dados categóricos tendem a ter uma diminuição de performance quase imperceptível, enquanto nos algoritmos com dados numéricos há uma degradação bastante acentuada principalmente no caso de Support Vector Machines.

## CONCLUSÃO

Ao final desta pesquisa concluímos que os dados aleatórios interferem significativamente nos resultados gerados por algoritmos classificadores de conjuntos numéricos, enfatizando a necessidade de um pré-processamento dos dados para sua utilização.

Surpreendentemente o algoritmo k-NN apresentou uma maior resistência à esses atributos com relação a SVM que acredita-se ser mais eficiente nos casos em que há números elevados de dimensões para análise, o que pode ser estudado mais profundamente no futuro.

1. Bolsista CNPq: Graduação em Engenharia de Computação, UNICAMP, Campinas-SP. henrique.oliveira@gmx.com
2. Orientador: Pesquisador e Professor, IC-UNICAMP, Campinas-SP.