# UFPA PPGCC: Aprendizado de Máquina

Lista de Exercício Final - Valor 10 pts

**1) [2.0 pts]** Use os dados Breast Cancer Wisconsin (Diagnostic) Data Set do UCI Machine Learning Repository. Use validação cruzada para avaliar qual dos algoritmos tem maior acurácia nos dados:

- SVM Linear
- SVM RBF

Decida que tipo de padronização (normalização) dos dados você usará para cada algoritmo (ou nenhuma). Justifique.

Em principio, fiz um escalado dos dados, normalizando eles entre o intervalo (0,1), dado que assim o *accuracy* do modelo é incrementado. A seguinte fórmula deixa ver como foi a normalização, dado que **x** é um vector de caracteristicas:

$$x = \frac{x - min(x)}{max(x) - min(x)}$$

| | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | ... |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 17.99 | 10.38 | 122.80 | 1001.0 | 0.11840 | 0.27760 | 0.3001 | 0.14710 | 0.2419 | 0.07871 | ... |
| **1** | 20.57 | 17.77 | 132.90 | 1326.0 | 0.08474 | 0.07864 | 0.0869 | 0.07017 | 0.1812 | 0.05667 | ... |
| **2** | 19.69 | 21.25 | 130.00 | 1203.0 | 0.10960 | 0.15990 | 0.1974 | 0.12790 | 0.2069 | 0.05999 | ... |
| **3** | 11.42 | 20.38 | 77.58 | 386.1 | 0.14250 | 0.28390 | 0.2414 | 0.10520 | 0.2597 | 0.09744 | ... |
| **4** | 20.29 | 14.34 | 135.10 | 1297.0 | 0.10030 | 0.13280 | 0.1980 | 0.10430 | 0.1809 | 0.05883 | ... |

5 rows × 30 columns

| | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | ... |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 0.521037 | 0.022658 | 0.545989 | 0.363733 | 0.593753 | 0.792037 | 0.703140 | 0.731113 | 0.686364 | 0.605518 | ... |
| **1** | 0.643144 | 0.272574 | 0.615783 | 0.501591 | 0.289880 | 0.181768 | 0.203608 | 0.348757 | 0.379798 | 0.141323 | ... |
| **2** | 0.601496 | 0.390260 | 0.595743 | 0.449417 | 0.514309 | 0.431017 | 0.462512 | 0.635686 | 0.509596 | 0.211247 | ... |
| **3** | 0.210090 | 0.360839 | 0.233501 | 0.102906 | 0.811321 | 0.811361 | 0.565604 | 0.522863 | 0.776263 | 1.000000 | ... |
| **4** | 0.629893 | 0.156578 | 0.630986 | 0.489290 | 0.430351 | 0.347893 | 0.463918 | 0.518390 | 0.378283 | 0.186816 | ... |

5 rows × 30 columns

A continuação, fiz uma validação cruzada, com 10 folds para o *SVM Linear* e o *SVM RBF*, as figuras 1 e 2 mostram os resultados por separado e logo a figura 3 mostra a comparação deles.

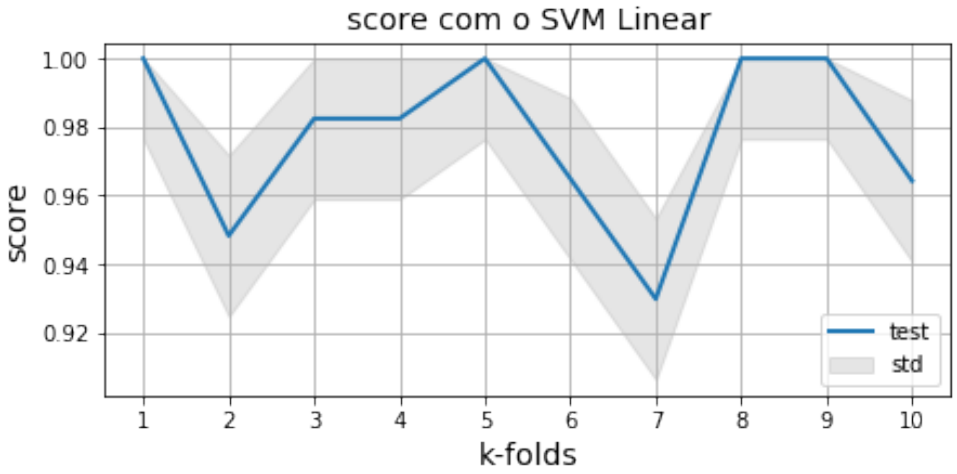


Figura 1: *Accuracy* do modelo *SVM Linear* com sua desviação estándar para cada k-fold.