Segundo trabalho de inteligencia artificial

Wanderson Ralph Silva Vita[1]

Abstract

Este artigo tem o objetivo de comparar diversos classificadores de dados, tais como OneR Probabilístico, KmeansCentroides e KGACentroides que serão implementados, e os já implementados na biblioteca Sklearn ZeroR, Aleatório, Aleatório Estratificado, Naive Bayes Gaussiano, Knn, DistKnn, Árvore de Decisão e Florestas de Árvores. As bases de dados utilizadas foram iris, digits, wine e breast cancer, todas disponíveis na Sklearn.

Keywords: Classificadores, Validação Cruzada, Inteligência Artificial, Classificação Automática

1. Introdução

Neste trabalho serão comparados diferentes técnas de classificação automática de dados. Os algoritimos analisados foram ZeroR, Aleatório, Aleatório Estratificado, OneR Probabilístico, Naive Bayes Gaussiano, KmeansCentroides, KGACentroides, Knn, DistKnn, Árvore de Decisão e Florestas de Árvores.

Estes foram treinados e testados com quatro diferentes bases de dados, a Iris Flower, Digits, Wine, e Breast Cancer. Para cada base, cada classificador foi submetido a três validações cruzadas de 10 folds e a as acuráricas como métrica de avaliação.

Com base na acurácia foi aplicado diferentes testes estatísticos a fim de classifica-los e compará-los. O teste t pareado e o teste de wilcoxon, foram utilizado utilizados a fim de havaliar a hipotese nula $H_0 : \mu_k = \mu_r$, para um nivel de confiança $(1 - \alpha) = 95\%$, onde se deseja saber se a média da acuracia

[1]Aluno de Engenharia de Computação da Universidade Federal do Espirito Santo

de um classificador k, é igual a média de outro classificador r. A hipótese nula será rejeitada para $p - valor < \alpha = 0.05$, e destacada em negrito nas tabelas pareadas.

2. Descrição dos Métodos Implementados

Neste trabalho foram implementados dois algorimos classificadores, descritos a seguir.

2.1. OneR Probabilístico

Essa técnica consiste em escolher a caracteristica que mais se destaca na base de treino. Onde pra cada caracteristica é gerada uma matriz pivoteada pelo valor da característica e a classe, retornando a quantidade do valor da característica por classe.

Com essas quantidades, é calculado a soma das maiores quantidades por classe em cada matriz, e escolhida a matriz da característica em que a soma foi maior. Essa matriz é utilizada como uma matriz de probabilidades na hora fazer a classificação. Onde se pega o valor da característica destaque, no dado de entrada, e com base na distribuição em cada classe é feita o sorteio de uma classe como resposta.

2.2. KCentróides

Essa técnica consistem em dividir cada classe em k grupos com o auxilio de um algoritimo de agrupamento.

Foi utilizado dois algoritimos diferentes, o Kmeans já implementado no Sklearn, e o Algoritimo Genético implementando no trabalho 1.

3. Descrição dos Experimentos Realizados

3.1. Iris

Método	Média	Desvio Padrão	Limite Inferior	Limite Superior
ZeroR	0.33	0.00	0.33	0.33
Aleatório	0.33	0.14	0.28	0.38
Aleatório Estratificado	0.33	0.11	0.29	0.37
OneR Probabilístico	0.78	0.15	0.72	0.83
Naive Bayes Gaussiano	0.95	0.05	0.93	0.97
KmeansCentroides	0.96	0.05	0.94	0.98
KGACentroides	0.94	0.06	0.92	0.96
Knn	0.95	0.06	0.92	0.97
DistKnn	0.95	0.06	0.93	0.97
Árvore de Decisão	0.96	0.05	0.94	0.97
Florestas de Árvores	0.96	0.05	0.94	0.98

Table 1: Média da acurácia por classificador, para base Iris

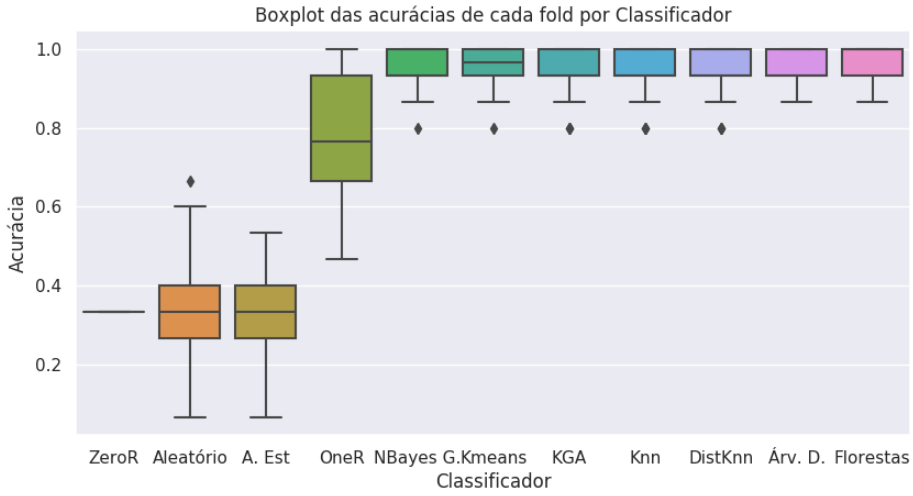

Figure 1: Boxplot das acurácias de cada fold por Classificador

Para a base Iris Flower, os algoritimos Naive Bayes Gaussiano, KmeansCentroides, KGACentroides, Knn, DistKnn, Árvore de Decisão e Florestas de

Árvores tiveram ótimos resultados, tendo uma média de acurácia entre 92% e 98%, como visto na Tabela 1.

E todos esses citados, tem as médias das acurárias iguais, o que pode pode ser comprovado pela Tabela 2, onde somente a hipótese nula dos classificadores Algoritimo Genético e Florestas foi rejeitada.

ZeroR	1.00	0.75	**5.11e-16**	**6.61e-33**	**4.27e-34**	**2.53e-30**	**9.10e-31**	**1.14e-30**	**3.13e-34**	**4.27e-34**
0.70	**Aleatório**	0.85	**1.54e-11**	**5.30e-19**	**4.30e-20**	**5.92e-19**	**2.31e-18**	**1.63e-18**	**4.97e-20**	**1.50e-19**
0.59	0.86	**A. Est**	**7.22e-14**	**1.44e-22**	**3.15e-22**	**5.46e-22**	**1.19e-22**	**2.69e-22**	**1.66e-22**	**7.83e-23**
1.57e-06	**2.53e-06**	**1.70e-06**	**OneR**	**4.39e-08**	**2.87e-07**	**7.48e-07**	**1.21e-07**	**2.68e-07**	**2.14e-08**	**3.50e-08**
1.07e-06	**1.66e-06**	**1.62e-06**	**8.63e-06**	**NBayes G.**	0.35	0.16	0.65	0.79	0.57	0.10
9.42e-07	**1.63e-06**	**1.64e-06**	**2.35e-05**	0.36	**Kmeans**	**2.26e-02**	0.16	0.20	0.63	1.00
1.18e-06	**1.68e-06**	**1.63e-06**	**2.94e-05**	0.15	**2.55e-02**	**KGA**	0.25	0.17	0.07	**2.26e-02**
1.07e-06	**1.68e-06**	**1.62e-06**	**1.59e-05**	0.64	0.15	0.25	**Knn**	0.66	0.33	0.08
1.07e-06	**1.66e-06**	**1.65e-06**	**2.26e-05**	0.78	0.19	0.17	0.65	**DistKnn**	0.48	0.13
1.06e-06	**1.65e-06**	**1.64e-06**	**1.11e-05**	0.56	0.62	0.07	0.32	0.47	**Árv. D.**	0.42
9.88e-07	**1.67e-06**	**1.61e-06**	**7.90e-06**	0.10	1.00	**2.54e-02**	0.08	0.13	0.41	**Florestas**

Table 2: Tabela paredada da base Iris Flower

3.2. Digits

Método	**Média**	**Desvio Padrão**	**Limite Inferior**	**Limite Superior**
ZeroR	0.10	0.00	0.10	0.10
Aleatório	0.09	0.03	0.08	0.10
Aleatório Estratificado	0.11	0.02	0.10	0.11
OneR Probabilístico	0.10	0.03	0.09	0.11
Naive Bayes Gaussiano	0.78	0.03	0.77	0.80
KmeansCentroides	0.95	0.02	0.94	0.96
KGACentroides	0.91	0.02	0.90	0.92
Knn	0.97	0.01	0.97	0.98
DistKnn	0.98	0.01	0.97	0.98
Árvore de Decisão	0.85	0.02	0.84	0.86
Florestas de Árvores	0.98	0.01	0.97	0.98

Table 3: Média da acurácia por classificador, para base Digits

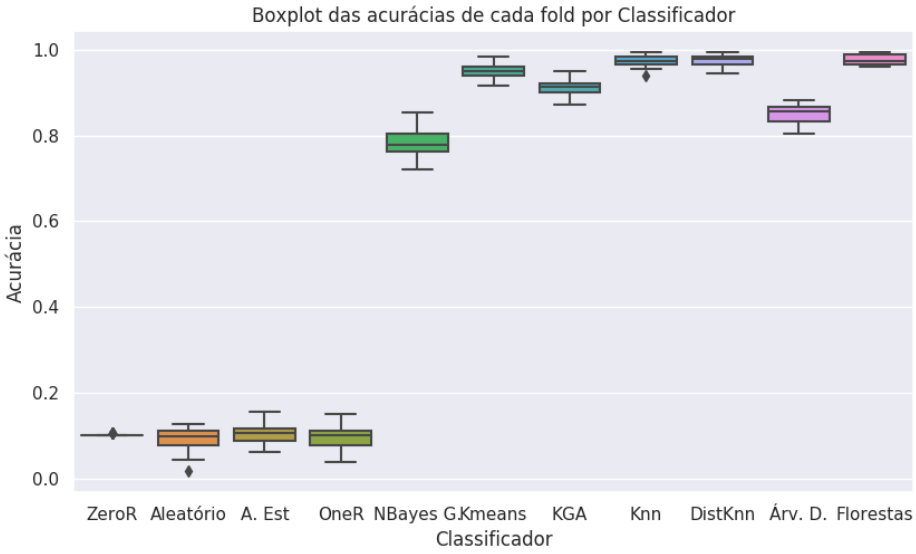

Figure 2: Boxplot das acurácias de cada fold por Classificador

Já para a base digits, o Kmeans, Knn, DistKnn e Floresta de Árvores, tiveram exelentes resultados, ficando entre 95% e 98% (Tabela 3).

Na Tabela 4 podemos observar que desses, o Kmeans rejeia a hipótese nula, indicando não ter a mesma média de acurácia. E os algoritimos que tiveram um resultado ruim, a baixo de 20% de acurácia, estão todos com a mesma média de acurácia.

ZeroR	0.06	0.40	0.44	**4.23e-41**	**4.62e-51**	**6.11e-49**	**5.02e-56**	**3.15e-56**	**3.73e-45**	**6.66e-56**
0.09	**Aleatório**	**3.85e-02**	0.41	**2.42e-37**	**4.63e-42**	**1.74e-42**	**1.60e-44**	**1.95e-44**	**1.06e-39**	**1.05e-44**
0.57	**4.78e-02**	**A. Est**	0.26	**1.34e-40**	**1.00e-43**	**1.28e-42**	**7.98e-46**	**5.86e-45**	**2.00e-41**	**1.17e-45**
0.29	0.55	0.33	**OneR**	**1.67e-36**	**1.11e-42**	**7.58e-42**	**6.43e-44**	**3.60e-44**	**9.35e-40**	**2.74e-44**
1.71e-06	**1.72e-06**	**1.72e-06**	**1.73e-06**	**NBayes G.**	**1.05e-22**	**1.42e-19**	**8.46e-26**	**2.67e-25**	**1.67e-10**	**1.06e-24**
1.70e-06	**1.73e-06**	**1.73e-06**	**1.72e-06**	**1.72e-06**	**Kmeans**	**9.56e-12**	**8.41e-10**	**6.42e-11**	**6.99e-18**	**1.26e-09**
1.69e-06	**1.70e-06**	**1.72e-06**	**1.73e-06**	**1.72e-06**	**2.55e-06**	**KGA**	**5.18e-19**	**4.57e-19**	**5.86e-11**	**3.93e-17**
1.69e-06	**1.72e-06**	**1.72e-06**	**1.73e-06**	**1.72e-06**	**3.82e-06**	**1.72e-06**	**Knn**	0.11	**5.85e-21**	0.41
1.68e-06	**1.71e-06**	**1.72e-06**	**1.73e-06**	**1.72e-06**	**2.56e-06**	**1.72e-06**	0.13	**DistKnn**	**1.07e-20**	0.87
1.70e-06	**1.71e-06**	**1.72e-06**	**1.72e-06**	**2.60e-06**	**1.72e-06**	**2.01e-06**	**1.73e-06**	**1.72e-06**	**Árv. D.**	**2.32e-22**
1.64e-06	**1.72e-06**	**1.72e-06**	**1.73e-06**	**1.72e-06**	**7.28e-06**	**1.71e-06**	0.51	0.95	**1.71e-06**	**Florestas**

Table 4: Tabela paredada da base Digits

3.3. *Wine*

Método	**Média**	**Desvio Padrão**	**Limite Inferior**	**Limite Superior**
ZeroR	0.40	0.02	0.39	0.41
Aleatório	0.37	0.09	0.33	0.40
Aleatório Estratificado	0.31	0.09	0.28	0.35
OneR Probabilístico	0.60	0.11	0.57	0.64
Naive Bayes Gaussiano	0.97	0.05	0.96	0.99
KmeansCentroides	0.97	0.05	0.95	0.98
KGACentroides	0.97	0.04	0.96	0.99
Knn	0.96	0.05	0.94	0.98
DistKnn	0.96	0.05	0.94	0.98
Árvore de Decisão	0.89	0.08	0.86	0.92
Florestas de Árvores	0.98	0.03	0.97	0.99

Table 5: Média da acurácia por classificador, para base Wine

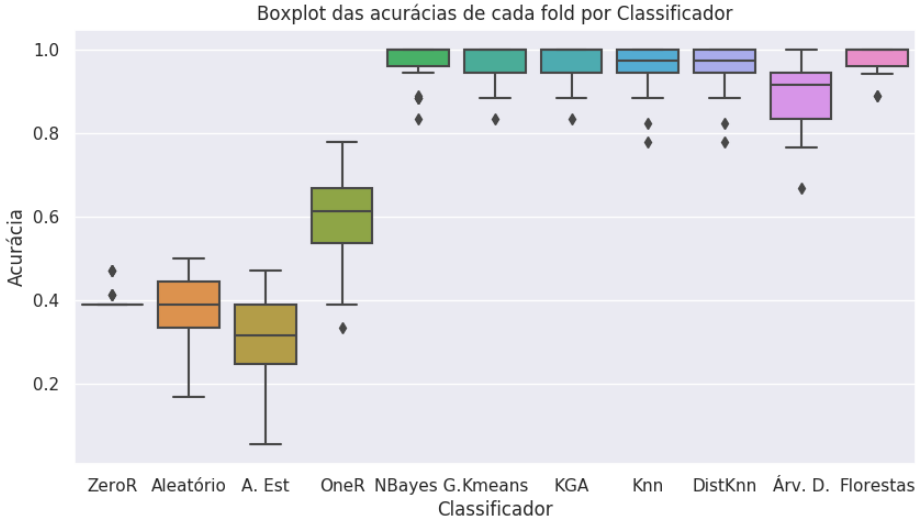

Figure 3: Boxplot das acurácias de cada fold por Classificador

Nessa base vemos que os classificadores Naive Bayes Gaussiano, KmeansCentroides, KGACentroides, Knn, DistKnn e Florestas de Árvores tiveram exelentes resultados, com acuária entre 96% e 99%. Sendo a Foresta de Árvores a melhor, com acuária média entre 98% e 99% (Tabela 5).

ZeroR	**4.84e-02**	**4.55e-05**	**2.48e-10**	**1.14e-29**	**2.43e-30**	**9.53e-32**	**1.13e-28**	**2.68e-28**	**2.00e-24**	**5.39e-34**
2.26e-02	**Aleatório**	**4.09e-02**	**1.07e-08**	**6.85e-23**	**1.55e-23**	**4.38e-24**	**7.37e-23**	**9.44e-23**	**4.71e-21**	**2.22e-24**
8.63e-05	**4.03e-02**	**A. Est**	**2.08e-11**	**6.43e-25**	**5.58e-25**	**3.97e-25**	**2.64e-24**	**2.46e-24**	**9.27e-21**	**2.37e-25**
4.36e-06	**4.76e-06**	**2.82e-06**	**OneR**	**4.62e-18**	**5.95e-18**	**6.57e-18**	**4.19e-17**	**7.21e-17**	**5.85e-13**	**4.65e-18**
8.06e-07	**1.68e-06**	**1.61e-06**	**1.62e-06**	**NBayes G.**	0.39	0.97	0.05	**1.79e-02**	**5.40e-08**	0.20
1.16e-06	**1.69e-06**	**1.64e-06**	**1.68e-06**	0.56	**Kmeans**	0.18	0.21	0.13	**9.75e-07**	**4.77e-02**
1.07e-06	**1.68e-06**	**1.63e-06**	**1.69e-06**	0.56	0.18	**KGA**	**3.12e-02**	**1.73e-02**	**4.68e-08**	0.15
1.25e-06	**1.64e-06**	**1.61e-06**	**1.68e-06**	0.11	0.17	**2.93e-02**	**Knn**	0.33	**2.06e-07**	**8.91e-03**
1.25e-06	**1.64e-06**	**1.61e-06**	**1.68e-06**	**2.01e-02**	0.09	**1.68e-02**	0.32	**DistKnn**	**9.69e-07**	**8.07e-03**
1.52e-06	**1.58e-06**	**1.71e-06**	**1.71e-06**	**2.81e-05**	**4.36e-05**	**1.52e-05**	**3.48e-05**	**9.37e-05**	**Árv. D.**	**8.13e-09**
9.03e-07	**1.67e-06**	**1.65e-06**	**1.67e-06**	0.17	**4.35e-02**	0.15	**9.94e-03**	**9.94e-03**	**7.18e-06**	**Florestas**

Table 6: Tabela paredada da base Wine

3.4. Breast Cancer

Método	**Média**	**Desvio Padrão**	**Limite Inferior**	**Limite Superior**
ZeroR	0.63	0.01	0.62	0.63
Aleatório	0.50	0.06	0.48	0.52
Aleatório Estratificado	0.53	0.05	0.51	0.55
OneR Probabilístico	0.62	0.10	0.58	0.65
Naive Bayes Gaussiano	0.93	0.03	0.92	0.94
KmeansCentroides	0.95	0.03	0.94	0.96
KGACentroides	0.93	0.03	0.92	0.95
Knn	0.97	0.02	0.96	0.97
DistKnn	0.97	0.02	0.96	0.97
Árvore de Decisão	0.93	0.03	0.92	0.94
Florestas de Árvores	0.96	0.03	0.95	0.97

Table 7: Média da acurácia por classificador, da base Breast Cancer

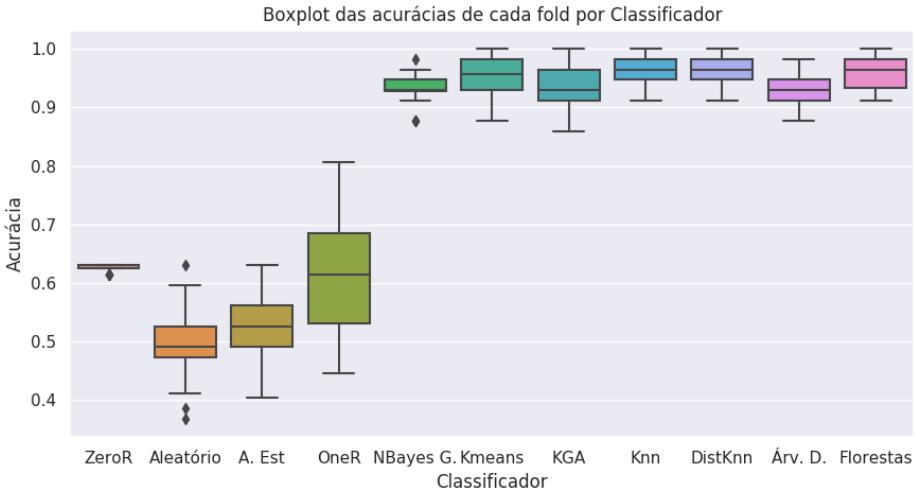

Figure 4: Boxplot das acurácias de cada fold por Classificador

Para base Breast Cancer os classificadores aive Bayes Gaussiano, KmeansCentroides, KGACentroides, Knn, DistKnn, Árvore de Decisão e Florestas de Árvores, possuem uma ótima média de acurácia entra 93% e 97%, como visto na Tabela 7

Os melhores algoritimos em geral tem a mesma média, como visto na Tabela 8. A Árvore de Decisão e o Algoritimo Genético que fogem da média.

ZeroR	**3.18e-12**	**6.53e-12**	0.61	**1.35e-31**	**2.99e-29**	**3.56e-29**	**2.09e-34**	**1.36e-34**	**2.21e-30**	**3.91e-33**
2.49e-06	**Aleatório**	0.05	**1.09e-05**	**2.92e-27**	**1.92e-26**	**6.84e-26**	**4.61e-28**	**5.69e-28**	**1.26e-25**	**1.54e-28**
3.64e-06	0.07	**A. Est**	**1.61e-04**	**4.72e-26**	**1.10e-26**	**2.26e-26**	**1.39e-29**	**1.12e-29**	**3.67e-26**	**1.17e-27**
0.54	**9.79e-05**	**1.03e-03**	**OneR**	**5.77e-16**	**1.10e-18**	**6.21e-16**	**1.09e-17**	**7.60e-18**	**5.59e-17**	**1.13e-17**
1.60e-06	**1.72e-06**	**1.71e-06**	**1.73e-06**	**NBayes G.**	**1.22e-02**	0.90	**1.72e-06**	**9.55e-07**	0.40	**1.04e-05**
1.68e-06	**1.72e-06**	**1.71e-06**	**1.72e-06**	**3.02e-02**	**Kmeans**	**1.77e-02**	0.05	**4.19e-02**	**1.83e-03**	0.34
1.64e-06	**1.70e-06**	**1.71e-06**	**1.71e-06**	0.77	**2.16e-02**	**KGA**	**2.42e-06**	**1.37e-06**	0.37	**1.57e-04**
1.62e-06	**1.69e-06**	**1.71e-06**	**1.72e-06**	**6.96e-05**	0.07	**7.29e-05**	**Knn**	0.33	**2.20e-07**	0.09
1.63e-06	**1.70e-06**	**1.70e-06**	**1.72e-06**	**5.44e-05**	0.06	**5.73e-05**	0.32	**DistKnn**	**7.08e-08**	0.06
1.65e-06	**1.71e-06**	**1.72e-06**	**1.71e-06**	0.62	**3.43e-03**	0.34	**2.77e-05**	**1.58e-05**	**Árv. D.**	**9.08e-06**
1.59e-06	**1.72e-06**	**1.71e-06**	**1.71e-06**	**1.03e-04**	0.32	**6.59e-04**	0.11	0.10	**1.54e-04**	**Florestas**

Table 8: Tabela paredada da base Breast Cancer

4. Conclusões

4.1. Análise geral dos resultados

Neste trabalho vimos que pra cada base, classificadores diferentes se destacaram. Ou seja, para cada problema há algoritimos diferentes que o resolve com mais precisão. Em geral, para essas bases analisadas, o classificado Florestas de Árvores se destacou, com a menor faixa de acurácia sendo de 94% para a Iris e a maior, com 99% para a Wine.

4.2. Contribuições do Trabalho

Aprendenmos como automátizar a comparação dos algoritimos, para escolha do melhor para resolver um determinado problema.

4.3. Melhorias e trabalhos futuros

References