Segundo Trabalho de Inteligência Artificial
2019/2

Lucas Ribeiro Mendes Silva
Universidade Federal do Espírito Santo

Abstract

Este trabalho consiste em realizar uma comparação experimental entre um conjunto pré-definido de técnicas de aprendizado e classificação automática aplicadas a alguns problemas de classificação. As técnicas escolhidas são: ZeroR, OneR, OneR Probabilístico, Centróide, Centróide OneR, Naive Bayes Gaussiano, Knn, Árvore de Decisão, Rede Neural e Florestas de Árvores. As bases de dados a serem utilizadas são iris, digits, wine e breast cancer, todas disponibilizadas pelo pacote do scikit-learn.

Keywords: Inteligência Artificial, `IA`, Machine Learning, Classificadores

1. Introdução

Nesse artigo apresentaremos uma série de resultados provenientes de experimentos com diferentes algoritmos e hiperparâmetros, assim como comparações entre seus resultados com a finalidade de identificarmos aqueles que melhor se destacam para a resolução do problema de Classificação, onde temos como intenção receber uma gama de informações de uma base de dados e classificar seus diferentes objetos em classes, classes essas que simbolizam um determinado grupo (Ex: Informações de animais dadas como entrada podem ser classificadas cada uma como gato, cachorro, papagaio e etc).

2. Descrição dos Datasets

- Iris: Esse talvez seja o banco de dados mais conhecido encontrado na literatura de reconhecimento de padrões. O conjunto de dados contém 3 classes de 50 instâncias cada, em que cada classe se refere a um tipo de planta íris. Atributos: altura de sépala, largura de sépala, altura de pétala, largura de pétala.

- Digits: Este conjunto de dados é composto de 1797 imagens 8x8. Cada imagem é de um dígito escrito à mão. Atributos: números inteiros no intervalo de 0 a 100, código da classe no intervalo de 0 a 9.

- Wine: Esses dados são os resultados de uma análise química de vinhos cultivados na mesma região da Itália, mas derivados de três cultivares diferentes. A análise determinou as quantidades de 13 constituintes encontrados em cada um dos três tipos de vinhos. Atributos: álcool, ácido málico, cinzas, alcalinidade das cinzas, magnésio, fenóis totais, flavonóides, fenóis inflavonóides, proantocianidinas, intensidade da cor, matiz, OD280 ou OD315 de vinhos diluídos, proline.

- Breast Cancer: Os recursos são calculados a partir de uma imagem digitalizada de um aspirado por agulha fina (PAAF) de uma massa mamária. Eles descrevem características dos núcleos celulares presentes na imagem. Atributos: Número de identificação, diagnóstico (M = maligno, B = benigno), raio (média das distâncias do centro aos pontos do perímetro), textura (desvio padrão dos valores da escala de cinza), perímetro, área, suavidade (variação local nos comprimentos do raio), compacidade, concavidade (gravidade das partes côncavas do contorno), pontos côncavos (número de partes côncavas do contorno), simetria, dimensão fractal.

3. Métodos Implementados

Para os experimentos realizados, foram utilizados 8 diferentes algoritmos apresentados em laboratório na disciplina de Inteligência Artificial, entre eles

existem alguns variantes afim de tentar alcançar melhores resultados. A implementação desses algoritmos utilizados encontram-se disponíveis através da caixa de e-mail do professor da disciplina, assim como solicitado para o envio desse trabalho, portanto nessa seção iremos apenas descrever os métodos que foram implementados durante a disciplina, mas de forma não aprofundada.

3.1. ZeroR

ZeroR é o método de classificação mais simples que depende do alvo e ignora todos os preditores. O classificador ZeroR simplesmente prevê a classe majoritária e a atribui para todos as instâncias.

3.2. OneR

O OneR é um algoritmo simples de classificação que gera uma regra para cada preditor nos dados e seleciona a regra com o menor erro total como sua "única regra". Para criar uma regra para um preditor, construímos uma tabela de frequência para cada preditor em relação ao alvo. Foi demonstrado que o OneR produz regras apenas um pouco menos precisas que os algoritmos de classificação de ponta, enquanto produz regras simples para os humanos interpretarem.

O OneR Probabilístico é uma variação do OneR, esse tem como diferencial a utilização do método da Roleta (já dado em sala de aula) para escolher o index da coluna de cada uma das linhas das tabelas de frequência geradas durante a execução do método fit, assim podendo gerar diferentes tabelas de regras. Também nesse variante, no ato da predição, para cada uma das instâncias do problema, geramos diferentes tabelas de regras de predição para a classificação, todas ainda baseadas no atributo escolhido durante o fit.

3.3. Centróide

Para cada uma das classes existentes na base de dados, é criado um ponto médio a partir das médias dos atributos das instâncias que pertencem a essa classe, partindo do mesmo princípio de centro geométrico visto na matemática.

Durante a predição, calculamos as distâncias entre o ponto que queremos estimar até os pontos médios das classes que já foram calculados. Assim que a menor distância for obtida, conseguimos definir a qual classe uma determinada instância supostamente pertence, ou seja, qual a sua categoria (classe).

O Centróide OneR é uma variação do Centróide, tendo como diferencial a utilização de apenas um atributo de cada instância para calcular as médias que caracterizam as classes, esse atributo é escolhido através da execução do OneR.

4. Descrição dos Experimentos Realizados

4.1. Descrição dos valores de hiperparâmetros utilizados na busca em grade de cada algoritmo

KNN: n_neighbors: [1, 3, 5, 7, 10, 20]

Árvore de Decisão: n_neighbors: [1, 3, 5, 7, 10, 20]

Rede Neural: max_iter: [50, 100, 200], hidden_layer_sizes: [(15,)]

Floresta de Árvores: n_estimators: [10, 20, 50, 100]

4.2. Iris

4.2.1. Primeira Etapa

Tabela de médias e desvios-padrão das acurácias na primeira etapa da Iris:

	media	dp
zeror	0.33333333333333337	5.551115123125783e-17
oner	0.9533333333333334	0.06699917080747259
oner-prob	0.9	0.061463629715285906
centroid	0.9333333333333333	0.059628479399994376
centroid-oner	0.9600000000000002	0.05333333333333332
gaussian	0.9533333333333334	0.04268749491621898

Nos algoritmos da primeira etapa de testes, podemos notar que apenas o ZeroR possui uma média de acurácia relativamente baixa, além de ter um dispersão

grande quando comparado aos demais, esse resultado não chega a ser surpreendente, dito que o ZeroR não possui uma taxa de acertividade muito boa.

O OneR probabilístico ficou um pouco abaixo do OneR padrão, mostrando que as regras escolhidas através do método da Roleta não foram tão eficientes para essa base, nessa bateria de testes.

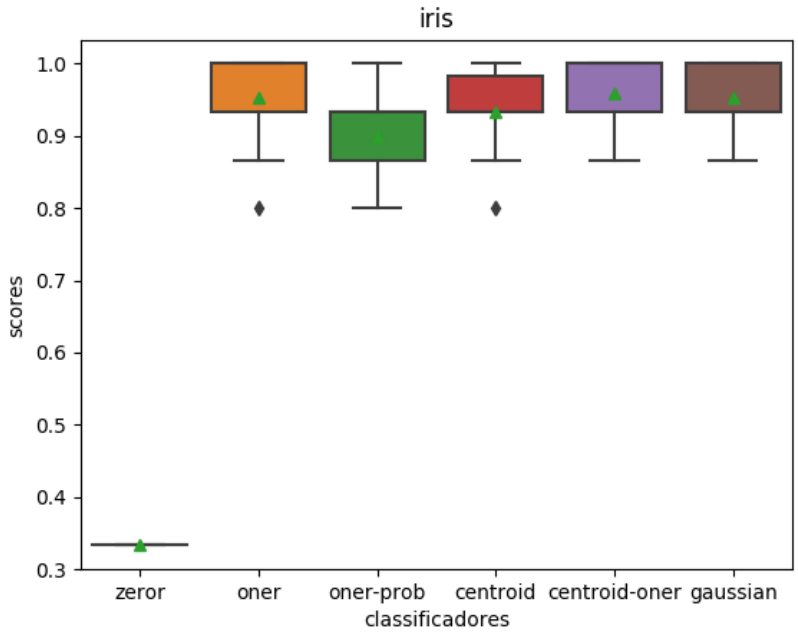

Figure 1: Boxplots das acurácias na primeira etapa da Iris

Podemos notar através dos boxplots do Iris que todos os algoritmos se mantiveram próximos (ou dentro) do intervalo entre 0.9 e 1.0, com exceção do ZeroR, que pode ser descartado, já que sua proposta não se encaixa muito bem para essa base.

No geral, os resultados aqui obtidos foram muito bons para algoritmos que não recebem nenhum hiperparâmetro.

4.2.2. Segunda Etapa

Tabela de médias e desvios-padrão das acurácias na segunda etapa da Iris:

	media	dp
knn	0.9533333333333334	0.059999999999999984
tree	0.9533333333333334	0.04268749491621898
neural	0.8400000000000001	0.11623730516108463
forest	0.96	0.044221663871405324

Analisando as médias e desvios-padrão obtidos na segunda etapa de testes com a base Iris, fica evidente que o algoritmo de Redes Neurais ficou bem atrás dos demais, perdendo até mesmo para os algoritmos da primeira etapa, no boxplot explicaremos melhor o motivo disso ter ocorrido. As dispersões entre as acurácias foram pouca coisa melhor do que na primeira etapa, mas ainda se mantiveram boas, com exceção da neural.

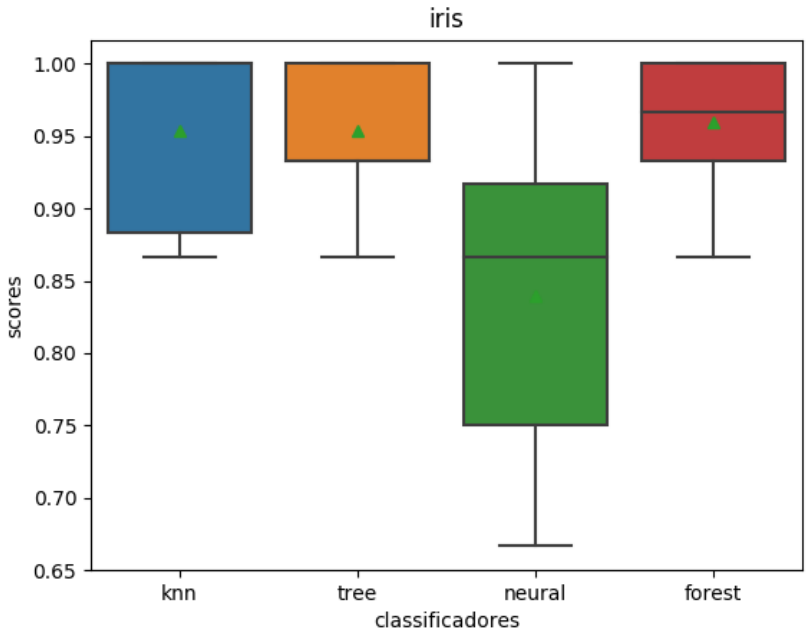

Figure 2: Boxplots das acurácias na segunda etapa da Iris

Perceba que com exceção do algoritmo de redes neurais, os demais encontram-se com o corpo do boxplot quase que totalmente dentro do intervalo entre 0.9 e 1.0. O motivo para o algoritmo de redes neurais não ter apresentado bons resultados se deu pelo fato de que o método não convergiu! As 200 iterações escolhidas entre os hiperparâmetros não foram suficientes para que o método alcançasse uma aproximação satisfatória, perdendo até mesmo para algoritmos simples como o OneR. Em todas as seções a seguir, veremos que o algoritmo de redes neurais não se sai muito bem em comparação aos demais devido ao mesmo problema citado anteriormente.

4.3. Digits

4.3.1. Primeira Etapa

Tabela de médias e desvios-padrão das acurácias na primeira etapa da Digits:

	media	dp
zeror	0.10127425688130931	0.001274401867253491
oner	0.23303354606521257	0.021685445382370397
oner-prob	0.17555141491564957	0.02770063642529908
centroid	0.8836101717889818	0.04112678822301056
centroid-oner	0.23363618614405857	0.024133635663880226
gaussian	0.8103537583567821	0.056655402070708565

Nos algoritmos da primeira etapa, obtivemos médias muito ruins, mas com uma dispersão melhor do que na base Iris por exemplo. Com exceção do Centróide e do Gaussiano, que apesar de não terem resultados tão bons, quando comparado aos demais, esses dois tiveram uma média muito superior.

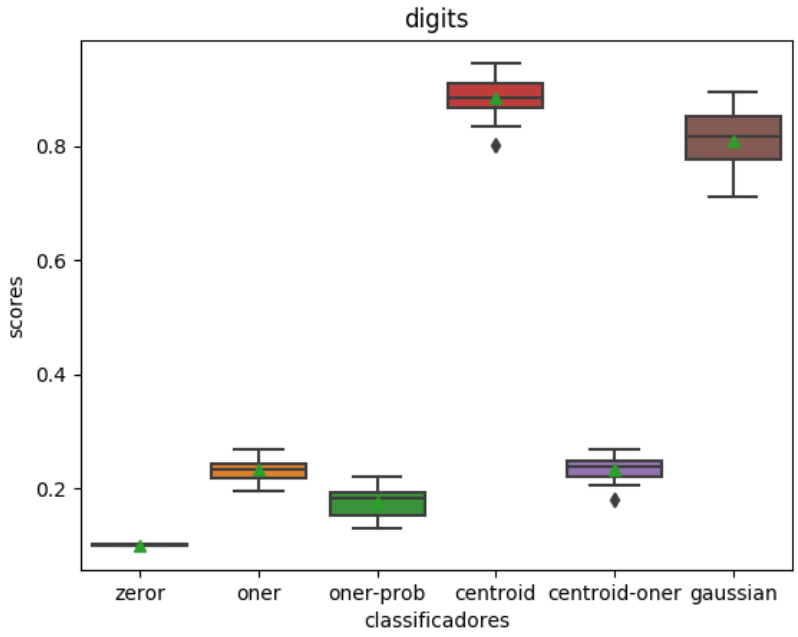

Figure 3: Boxplots das acurácias na primeira etapa da Digits

Com a visão espacial do gráfico acima, podemos perceber como os resultados do ZeroR, OneR, OneR Probabilístico e o Centroide OneR foram muito mal. Esses 4 citados ficaram com o corpo do boxplot por inteiro abaixo de 0.3 de acurácia, sendo inviáveis para essa base. Os algoritmos Centróide e Gaussiano tiveram resultados bem melhores, mas podemos dar destaque ao Centróide que ficou próximo de 0.9, tendo até mesmo um valor muito perto de 1.0.

4.3.2. Segunda Etapa

Tabela de médias e desvios-padrão das acurácias na segunda etapa da Digits:

	media	dp
knn	0.9788938332167104	0.017601243897194148
tree	0.828741699819908	0.03381904734869428
neural	0.9182522897883049	0.033915745203152145
forest	0.9516871108677964	0.022380438784698677

Diferente da primeira etapa, na segunda etapa tivemos bons resultados. Dessa vez o algoritmo de árvore de decisão ficou com uma média bem menor que os demais, tendo desempenho próximo ao Gaussiano na primeira etapa. O algoritmo de redes neurais se mostrou bem melhor nessa base, mesmo com problemas de não convergência. Os algoritmos KNN e Floresta de Árvores tiveram resultados muito bons. Todas as dispersões foram pequenas e muito aceitáveis.

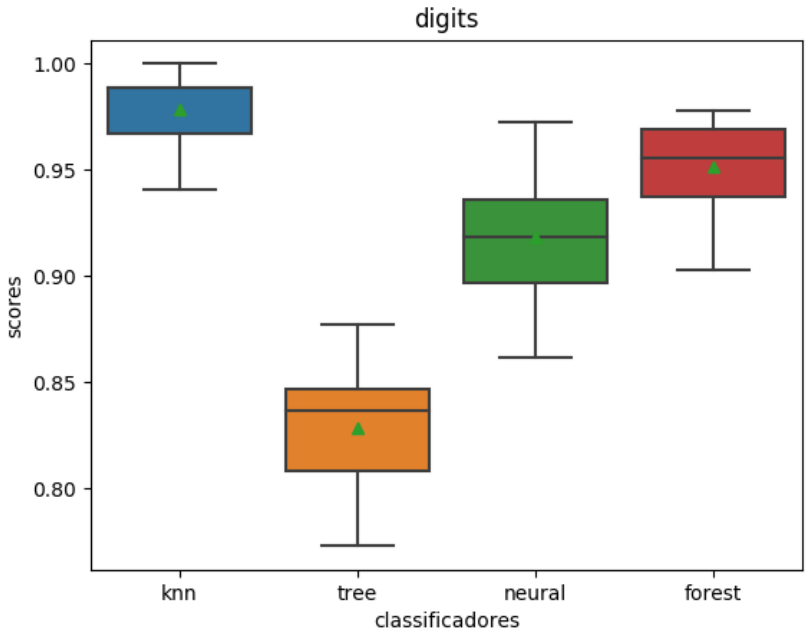

Figure 4: Boxplots das acurácias na segunda etapa da Digits

Com uma visão espacial dessa segunda etapa, podemos de cara dizer que o KNN se saiu muito bem em relação aos demais, tendo até mesmo um valor igual a 1.0, mas todo o corpo do boxplot do KNN se concentra entre 0.95 e 1.0, tendo no geral, ótimos resultados. O algoritmo de Árvore de Decisão ficou bem abaixo dos demais, mas se comparado com a primeira etapa, está ótimo. Os demais foram bem estando acima de 0.9 de acurácia.

4.4. Wine

4.4.1. Primeira Etapa

Tabela de médias e desvios-padrão das acurácias na primeira etapa da Wine:

	media	dp
zeror	0.3992539559683522	0.016871583549832498
oner	0.6235208118335054	0.07712192798491828
oner-prob	0.5584580323357413	0.12608451087413733
centroid	0.7216073271413829	0.08490131985948657
centroid-oner	0.7327184382524939	0.0842522501710735
gaussian	0.9616959064327485	0.042442001415448946

Os algoritmos da primera etapa não tiveram médias boas, mas não chegaram a ser catastróficas como as médias da primeira etapa de testes com a base Digits, porém continuam não sendo nada boas. O algoritmo Gaussiano foi o único que obteve uma média muito boa, assim como uma dispersão pequena.

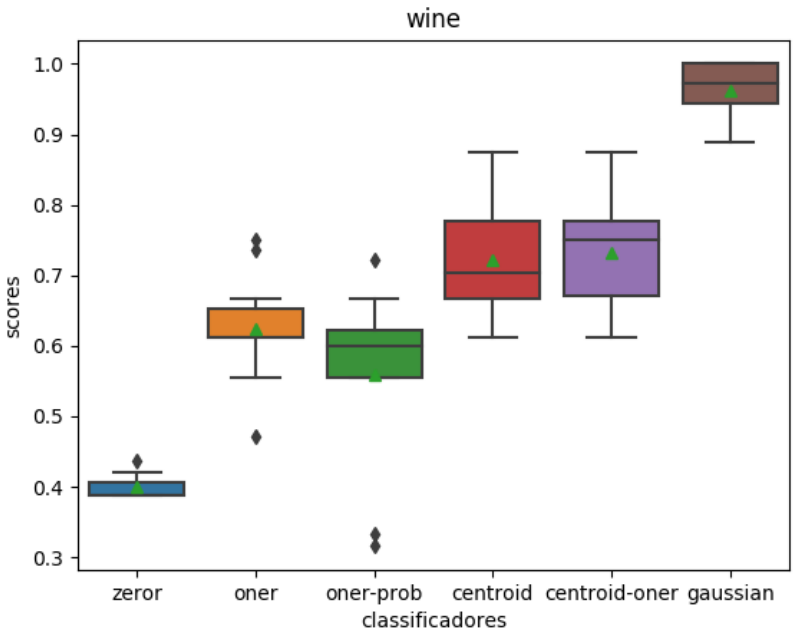

Figure 5: Boxplots das acurácias na primeira etapa da Wine

Como já era de se esperar, o ZeroR teve um resultado bem ruim, estando muito abaixo de todos os boxplots do gráfico. OneR e OneR probabilístico ficaram bem próximos um do outro, tendo resultados ruins. OneR Probabilístico por si só teve outliers bem abaixo dos resultados do ZeroR que não foram nada bons. Os Centróides praticamente tiveram um empate técnico, com resultados não muito bons, mas não chegam a ser estupidamente ruins. Como vista na tabela, a média do Gaussiano foi muito boa, ainda podemos notar como todo o corpo do boxplot dele se encontra entre 0.9 e 1.0.

4.4.2. Segunda Etapa

Tabela de médias e desvios-padrão das acurácias na segunda etapa da Wine:

	media	dp
knn	0.7144801341589267	0.11841049759193698
tree	0.9002579979360166	0.06911723348617782
neural	0.39007782937736496	0.11437130517991893
forest	0.9564327485380117	0.052492506317112934

Através das médias da segunda etapa, nota-se como o KNN que antes teve bastante destaque, acabou perdendo muita precisão nessa base, tendo também uma dispersão bem alta. O algoritmo de redes neurais novamente foi prejudicado pela não convergência do método, assim apresentando um resultado ruim. Os algoritmos de Árvore de Decisão e Floresta de Árvores tiveram resultados bons com baixa dispersão.

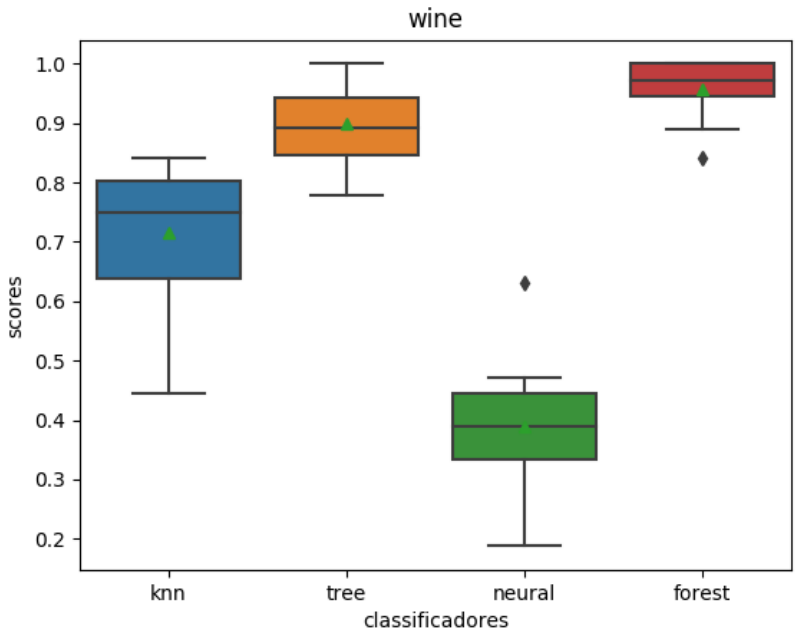

Figure 6: Boxplots das acurácias na segunda etapa da Wine

Com uma visão espacial dos resultados, podemos notar como o algoritmo de redes neurais ficou bem abaixo dos demais, assim como o KNN. O algoritmo de Árvore de Decisão ficou bem fixo em 0.9, variando quase que por igual para cima e para baixo, mostrando bastante estabilidade do método. O algoritmo de Floresta de Árvores apresentou ótimos resultados, tendo todo seu corpo entre 0.9 e 1.0, com um singelo outlier abaixo de 0.9, porém nada gritante.

4.5. Breast Cancer

4.5.1. Primera Etapa

Tabela de médias e desvios-padrão das acurácias na primeira etapa da Breast Cancer:

	media	dp
zeror	0.6274274047186933	0.00441189245975379
oner	0.9141463572724915	0.039822475517287674
oner-prob	0.8735761818338951	0.04937331563721148
centroid	0.8913641863278887	0.03879383921566746
centroid-oner	0.9034309912712816	0.036006020163806926
gaussian	0.9386796733212339	0.030112887483699426

Para essa primeira etapa, como já era de se esperar, o método ZeroR teve um média bem abaixo dos demais algoritmo, mas note que com exceção do ZeroR, todos tiverem resultados aceitáveis. O OneR obteve uma média boa, mas seu variante probabilístico pareceu não conseguir melhorar o resultado, ficando acima apenas do ZeroR, mas ainda assim muito melhor. Ao todo, tiveram resultados bons e uma dispersão pequena.

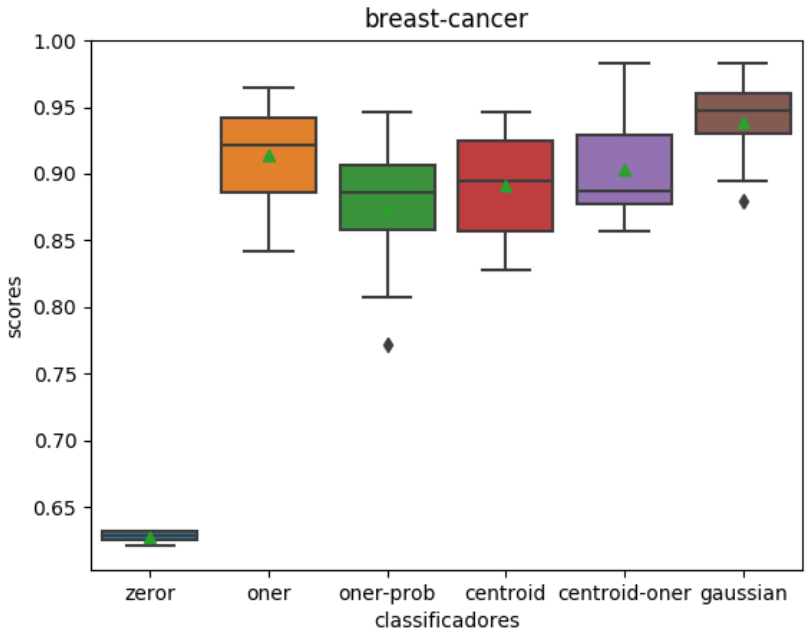

Figure 7: Boxplots das acurácias na primeira etapa da Breast Cancer

Na visão espacial dos resultados, podemos ver claramente que o ZeroR ficou muito abaixo dos demais. Os bons resultados no geral não tiveram nada de muito surpreendente, portanto não focaremos na análise dessa primeira etapa.

4.5.2. Segunda Etapa

Tabela de médias e desvios-padrão das acurácias na segunda etapa da Breast Cancer:

	media	dp
knn	0.9316588886008124	0.031245167972901115
tree	0.9159612393051596	0.04625957411928778
neural	0.8288123325555267	0.12404166891076888
forest	0.9632756460115806	0.0294071662844694

Na segunda etapa tivemos novamente bons resultados, contando também com a volta da boa média do KNN e novamente o algoritmo de Redes Neurais ficou com sua média prejudicada devido a não convergência, tendo também uma dispersão muito alta se comparado aos demais, que por sinal foram boas.

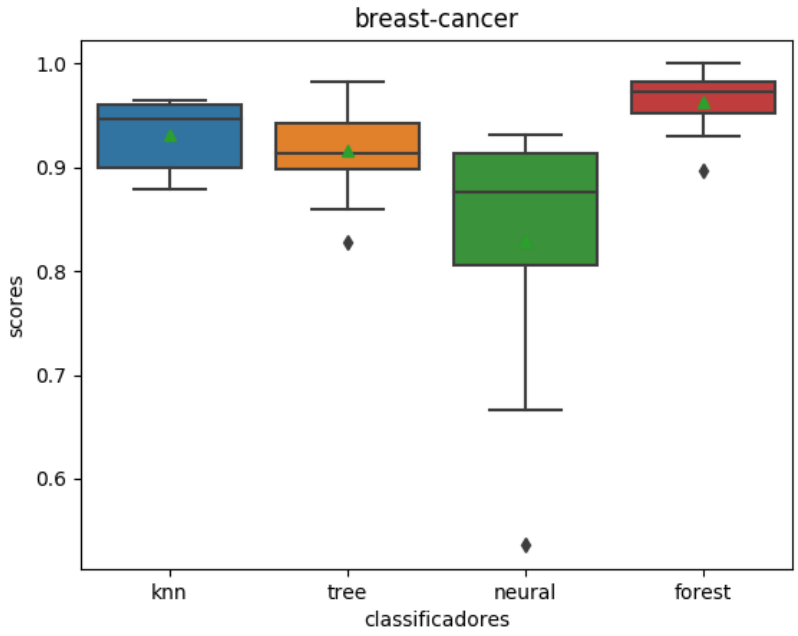

Figure 8: Boxplots das acurácias na segunda etapa da Breast Cancer

Com a visão espacial, nos chama a atenção os resultados muito variados do algoritmo de Redes Neurais, tendo até mesmo outliers na base do gráfico, se monstrando bem instável. Os demais foram bem, tendo seus valores entre 0.9 e 1.0.

5. Conclusões

5.1. Análise Geral dos Resultados

Analisando todos os resultados obtidos entre as bases e os diferentes algoritmos, temos três grandes destaques.

O primeiro deles foi o ZeroR, esse algoritmo se mostrou muito ruim durante todos os testes. A forma como o ZeroR classifica as instâncias é muito ingênua, isso acaba influenciando negativamente seus resultados.

O segundo foi o algoritmo de Redes Neurais, esse algoritmo por sua vez é muito utilizado no ramo de IA, tendo resultados muito satisfatórios, porém em nossos testes, tivemos um máximo de 200 iterações, fazendo com que o método não conseguisse convergir durante a execução, assim nem sempre conseguido alcançar bons resultados.

O terceiro e último foi o algoritmo de Floresta de Árvores, esse algoritmo foi o mais estável de todos, em todos os testes conseguiu manter sua média de resultados em torno de 0.95. Podemos dizer que esse foi o **melhor** algoritmo testado para essas bases de dados.

Os demais algoritmos não tiveram nada muito chamativo, apenas resultados bem variados entre as bases, além da prova de que as variações do OneR e Centróide não tiveram melhorias significativas, até mesmo pioraram seus resultados.

5.2. Contribuições do Trabalho

Esse trabalho impressionou muito em mostrar o poderio dos algoritmos de classificação do scikit-learn, assim como as grandes diferenças existentes entre diferentes técnicas utilizadas por eles. Também foi bem interessante comparar os algoritmos implementados por nós com os algoritmos famosos e já muito utilizados no ramo de IA.

5.3. Melhorias e Trabalhos Futuros

Para melhores resultados, seria interessante utilizar mais hiperparâmetros, principalmente para o algoritmo de Redes Neurais que não conseguiu conver-

gir com os números de iterações fornecidos para ele, assim não conseguido demonstrar seu verdadeiro **poder**. Também seria interessante utilizar alguma(s) outra(s) base(s) que não sejam apenas para estudos, assim poderíamos ver se realmente as técnicas aqui empregadas são efetivas no mundo real.

Referências

- Slides e Notas de aula.
- Aulas ministradas pelo professor da Disciplina.
- Conteúdo de apoio enviado por Email pelo professor da disciplina.
- https://archive.ics.uci.edu/ml/datasets/iris
- https://archive.ics.uci.edu/ml/datasets/Pen-Based+Recognition+of+Handwritten+Digits
- https://scikit-learn.org/stable/auto_examples/datasets/plot_digits_last_image.html
- https://archive.ics.uci.edu/ml/datasets/Wine
- https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)
- https://www.saedsayad.com/oner.htm