Estudo de Técnicas de Aprendizado Aplicadas a Problemas de Classificação

Fernando Amaral Musso

Universidade Federal do Espírito Santo (UFES)

Novembro, 2019

Resumo

Este artigo tem como objetivo realizar uma comparação experimental entre as técnicas de aprendizado para problemas de classificação ZeroR, OneR, OneR Probabilístico, Centroide, Centroide OneR, Naive-Bayes Gaussiano, KNN, Árvore de Decisão, Rede Neural e Florestas de Árvores. Para isso, foram realizadas etapas de treino e teste com validação cruzada para as bases de dados iris, digits, wine e breast cancer.

Palavras-chave: inteligência artificial, aprendizado de máquina, problemas de classificação

1. Introdução

Problemas de classificação consistem em identificar categorias em que cada instância de uma base de dados deve ser classificada. Para resolvê-los, são utilizadas técnicas de reconhecimento de padrões, que realizam a análise dos dados, com procedimentos de treino e teste. Como essas técnicas não são exatas por conta da complexidade dos problemas, busca-se maximizar a acurácia alcançada.

Este artigo utiliza uma metodologia a fim de comparar a acurácia de dez técnicas de aprendizado para problemas de classificação, também chamadas de classificadores. São eles: ZeroR, OneR, OneR Probabilístico, Centroide, Centroide OneR, Naive-Bayes Gaussiano, KNN, Árvore de Decisão, Rede Neural e Florestas de Árvores. Para isso, cada classificador foi treinado, testado e validado com as bases de dados iris, digits, wine e breast cancer. Adicionalmente,

para as técnicas que exigem ajuste de hiperparâmetros foi realizada também uma busca em grade com validação cruzada.

A estrutura do artigo é como segue: na seção 2 são apresentadas as bases de dados utilizadas nas etapas de treino e teste, na seção 3 são descritos cada um dos classificadores implementados, na seção 4 são detalhados os procedimentos experimentais realizados e seus resultados, e na seção 5 são apresentadas as principais conclusões deste trabalho.

2. Descrição das Bases de Dados Utilizadas

2.1. Iris

A base de dados iris é uma das mais famosas contidas na literatura, compreendendo dados relativos a plantas do gênero íris. Possui 150 instâncias no total, cada uma com 4 atributos contínuos referentes a medições das plantas:

1. Comprimento da sépala em centímetros.
2. Largura da sépala em centímetros.
3. Comprimento da pétala em centímetros.
4. Largura da pétala em centímetros.

Cada instância pode ser classificada em uma dentre três classes, que representam tipos de planta íris:

1. Íris Setosa.
2. Íris Versicolor.
3. Íris Virgínica.

2.2. Digits

A base de dados digits é composta por imagens de tamanho $8x8$ de dígitos escritos à mão, representados por um conjunto de *pixels*. Possui 5620 instâncias no total, cada uma com 64 atributos inteiros, que indicam o valor de cada *pixel* dentro do intervalo fechado $[0, 16]$.

Cada instância pode ser classificada em uma dentre dez classes, referentes aos dígitos de 0 a 9.

2.3. Wine

A base de dados wine é composta pelos resultados de uma análise química de vinhos cultivados em uma mesma região da Itália por três produtores distintos. Possui 178 instâncias no total, cada uma com 13 atributos contínuos, que compreendem as medições de diferentes constituintes encontrados nos três tipos de vinho:

1. Teor alcoólico.
2. Ácido málico.
3. Cinzas.
4. Alcalinidade das cinzas.
5. Magnésio.
6. Fenois totais.
7. Flavonoides.
8. Fenois não flavonoides.
9. Proantocianidinas.
10. Intensidade da cor.
11. Matiz.
12. OD280/OD315 de vinhos diluídos.
13. Prolina.

Cada instância pode ser classificada em uma dentre três classes, que representam os três tipos de vinho analisados.

2.4. Breast Cancer

A base de dados breast cancer é composta por dados extraídos de biópsias de amostras celulares de mamas, usadas para determinar a presença de câncer. Possui 569 instâncias no total, cada uma com 30 atributos contínuos, que compreendem as características analisadas do núcleo celular das amostras.

Cada instância pode ser classificada em uma dentre duas classes, que determinam o diagnóstico da amostra analisada (maligno ou benigno).

3. Descrição dos Métodos Implementados

Para este trabalho, foram implementados os classificadores ZeroR, OneR, OneR Probabilístico, Centroide e Centroide OneR. Os demais classificadores estão disponibilizados para uso dentro da biblioteca *scikit-learn (Python 3.7)*.

3.1. ZeroR

O classificador ZeroR apresenta uma abordagem bem ingênua, que não estabelece nenhuma regra de classificação. A classe mais frequente no conjunto de dados de treino é utilizada para classificar o conjunto de dados de teste.

- ***fit():*** armazena a classe mais frequente do conjunto de dados de treino.
- ***predict():*** classifica todas as instâncias do conjunto de teste dentro da classe armazenada em *fit()*.

3.2. OneR

O classificador OneR estabele uma única regra de classificação baseada em um dos atributos do conjunto de dados de treino. O atributo escolhido é aquele que fornece a maior acurácia de resultados.

- ***fit():*** constrói tabelas de frequência para cada atributo do conjunto de dados de treino, associando para cada valor do atributo o número de ocorrências de cada classe. Para cada valor do atributo na tabela, a classe mais frequente é escolhida como regra. Com isso, calcula a acurácia de cada atributo, utilizando a sua regra sobre o conjunto de dados de treino, e armazena aquele que apresentou a maior acurácia, junto com sua regra.
- ***predict():*** classifica as instâncias do conjunto de teste segundo a regra armazenada em *fit()*, levando em consideração apenas o atributo associado a ela.

3.3. OneR Probabilístico

O classificador OneR Probabilístico apresenta um comportamento similar ao OneR, com a única diferença de que as classes escolhidas na regra não são aquelas de maior frequência. Ao invés disso, são atribuídas probabilidades de escolha para cada classe e utiliza-se o método da roleta para escolhê-las.

- ***fit():*** constrói tabelas de frequência para cada atributo do conjunto de dados de treino, associando para cada valor do atributo o número de ocorrências de cada classe. Para cada valor do atributo na tabela, são atribuídas as probabilidades de cada classe a serem escolhidas como regra. Com isso, calcula a acurácia de cada atributo, utilizando a sua regra sobre o conjunto de dados de treino, juntamente com o método da roleta, e armazena aquele que apresentou a maior acurácia, junto com sua regra.

- ***predict():*** classifica as instâncias do conjunto de teste segundo a regra armazenada em *fit()*, levando em consideração apenas o atributo associado a ela. Note que também é utilizado o método da roleta para determinar a classe a ser escolhida.

3.4. Centroide

O classificador Centroide utiliza noções de distância euclidiana para decidir a classe utilizada para a classificação. Para isso, centroides são definidos a partir da análise do conjunto de dados de treino. Instâncias do conjunto de teste são classificadas de acordo com o centroide mais próximo a elas.

- ***fit():*** divide o conjunto de treino segundo cada classe associada. Com isso, armazena os centroides de cada conjunto, calculados como sendo o centro geométrico das instâncias.

- ***predict():*** classifica as instâncias do conjunto de teste segundo a distância euclidiana para cada centroide armazenado em *fit()*. A classe escolhida é aquela associada ao centroide mais próximo da instância.

3.5. Centroide OneR

O classificador Centroide OneR combina os classificadores Centroide e OneR, de forma a fornecer uma forma de classificação mais robusta. O atributo associado a regra escolhida pela OneR é utilizado para a determinação dos centroides. Ao invés de se calcular o centro geométrico utilizando todos os atributos do conjunto de treino, apenas os valores associados ao atributo escolhido pelo OneR são considerados. Devido a isso, uma particularidade deste método é que os centroides são todos unidimensionais.

- ***fit():*** constrói tabelas de frequência para cada atributo do conjunto de dados de treino, associando para cada valor do atributo o número de ocorrências de cada classe. Para cada valor do atributo na tabela, a classe mais frequente é escolhida como regra. Com isso, calcula a acurácia de cada atributo, utilizando a sua regra sobre o conjunto de dados de treino, e armazena aquele que apresentou a maior acurácia, junto com sua regra. Por fim, divide o conjunto de treino segundo cada classe associada e armazena os centroides de cada conjunto, calculados como sendo o centro geométrico das instâncias, levando em consideração apenas o atributo escolhido pelo OneR.

- ***predict():*** classifica as instâncias do conjunto de teste segundo a distância euclidiana para cada centroide armazenado em *fit()* (como os centroides são unidimensionais, pode ser calculado apenas o módulo de uma diferença simples). A classe escolhida é aquela associada ao centroide mais próximo da instância. Note que apenas o atributo armazenado em *fit()* será considerado para o cálculo das distâncias.

4. Descrição dos Experimentos Realizados

A metodologia empregada neste trabalho consiste em realizar uma comparação experimental de todos os classificadores. Os experimentos foram divididos em duas etapas: a primeira compreendendo os classificadores que não

precisam de ajuste de hiperparâmetros, e a segunda compreendendo os classificadores que precisam.

A primeira etapa consiste no treino e teste com validação cruzada de 10 *folds* dos classificadores ZeroR, OneR, OneR Probabilístico, Centroide, Centroide OneR e Naive-Bayes Gaussiano.

A segunda etapa consiste na validação, treino e teste dos classificadores KNN, Árvore de Decisão, Redes Neurais e Florestas de Árvores. O procedimento será realizado através de ciclos aninhados de validação e teste, com o ciclo interno de validação contendo 4 *folds* e o externo de teste com 10 *folds*. A busca em grade do ciclo interno considera os seguintes valores de hiperparâmetros:

1. **KNN:** $n\_neighbors = [1, 3, 5, 7, 10]$.
2. **Árvore de Decisão:** $max\_depth = [None, 3, 5, 10]$.
3. **Rede Neural:** $max\_iter = [50, 100, 200]$, $hidden\_layer\_sizes = [(15,)]$.
4. **Florestas de Árvores:** $n\_estimators = [10, 20, 50, 100]$.

Os resultados das duas etapas serão apresentados em forma de tabela, com a média de desvio padrão obtidos pelos classificadores em cada *fold* do ciclo mais externo. Além disso, as acurácias obtidas em cada *fold* também serão apresentadas em forma de *boxplots*.

4.1. Iris

A média e o desvio padrão das acurácias obtidas pelos classificadores da primeira etapa com a base de dados iris são apresentados na Tabela 1. O *boxplot* das acurácias obtidas em cada *fold* é apresentado na Figura 1.

Tanto por meio da tabela, quanto pelo *boxplot*, nota-se que a acurácia do ZeroR foi muito inferior à acurácia dos outros classificadores, devido à sua ingenuidade. O Centroide OneR apresentou a melhor acurácia, por ser bem mais criterioso. Também vale notar que o OneR Probabilístico apresentou o maior desvio padrão, por conta de sua natureza aleatória.

Classificadores	Média	Desvio Padrão
ZeroR	0,33333333	0,00000000
OneR	0,95333333	0,06699917
OneR Probabilístico	0,91333333	0,07333333
Centroide	0,93333333	0,05962848
Centroide OneR	0,96000000	0,05333333
Naive-Bayes Gaussiano	0,95333333	0,04268749

Tabela 1: Média e desvio padrão das acurácias obtidas pelos classificadores da primeira etapa com a base de dados iris.

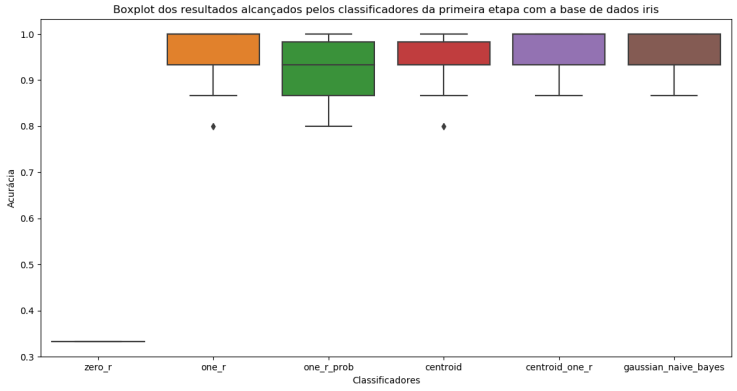

Figura 1: Boxplot dos resultados alcançados pelos classificadores da primeira etapa com a base de dados iris.

A média e o desvio padrão das acurácias obtidas pelos classificadores da segunda etapa com a base de dados iris são apresentados na Tabela 2. O *boxplot* das acurácias obtidas em cada *fold* é apresentado na Figura 2.

Nota-se de imediato que o método Rede Neural mostrou-se muito variável, apresentando uma média de acurácias significativamente inferior aos demais métodos. Observa-se que houve um empate entre os métodos Árvore de Decisão e Florestas de Árvores, que inclusive apresentaram o mesmo valor de desvio padrão.

Classificadores	Média	Desvio Padrão
KNN	0,95333333	0,06000000
Árvore de Decisão	0,96000000	0,04422166
Rede Neural	0,74000000	0,18487233
Florestas de Árvores	0,96000000	0,04422166

Tabela 2: Média e desvio padrão das acurácias obtidas pelos classificadores da segunda etapa com a base de dados iris.

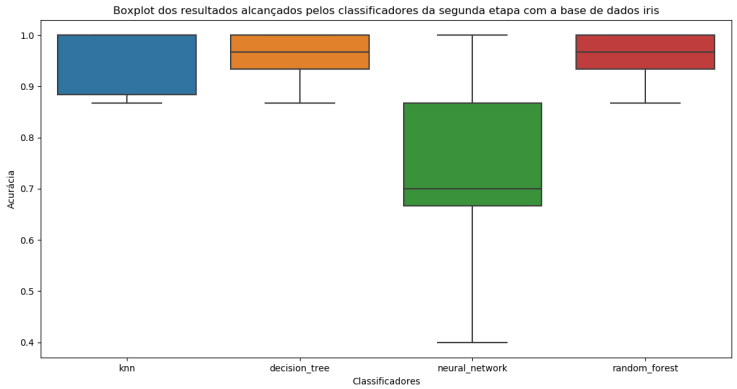

Figura 2: Boxplot dos resultados alcançados pelos classificadores da segunda etapa com a base de dados iris.

Comparando os resultados das duas etapas, observa-se um empate triplo entre os métodos Centroide OneR, Rede Neural e Florestas de Árvores em termos de média das acurácias. No entanto, o método Centroide OneR apresenta maior variabilidade.

4.2. Digits

A média e o desvio padrão das acurácias obtidas pelos classificadores da primeira etapa com a base de dados digits são apresentados na Tabela 3. O *boxplot* das acurácias obtidas em cada *fold* é apresentado na Figura 3.

Observa-se que o Centroide e o Naive-Bayes Gaussiano foram os únicos classificadores que apresentaram bons resultados. A base de dados digits possui atributos inteiros, e como os métodos que utilizam o OneR realizam uma discretização do conjunto de treino, pode ter ocorrido alguma discrepância indesejada.

Classificadores	**Média**	**Desvio Padrão**
ZeroR	0,10127426	0,00127440
OneR	0,23807089	0,02539605
OneR Probabilístico	0,17695275	0,02400862
Centroide	0,88361017	0,04112679
Centroide OneR	0,23028109	0,02332096
Naive-Bayes Gaussiano	0,81035376	0,05665540

Tabela 3: Média e desvio padrão das acurácias obtidas pelos classificadores da primeira etapa com a base de dados digits.

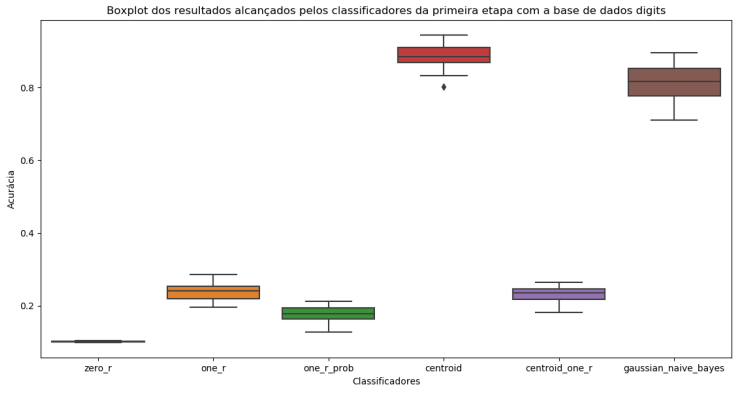

Figura 3: Boxplot dos resultados alcançados pelos classificadores da primeira etapa com a base de dados digits.

A média e o desvio padrão das acurácias obtidas pelos classificadores da segunda etapa com a base de dados digits são apresentados na Tabela 4. O *boxplot* das acurácias obtidas em cada *fold* é apresentado na Figura 4.

Por meio dos resultados obtidos, nota-se que o classificador KNN apresentou os melhores resultados para esta base de dados, tanto em média quanto em desvio padrão das acurácias. Além disso, quando comparado aos resultados com a base de dados iris, o método Rede Neural mostrou-se bem mais estável, apresentando resultados de média também muito melhores.

Classificadores	**Média**	**Desvio Padrão**
KNN	0,97889383	0,01760124
Árvore de Decisão	0,83542787	0,03475215
Rede Neural	0,92318879	0,03193339
Florestas de Árvores	0,94945413	0,03050055

Tabela 4: Média e desvio padrão das acurácias obtidas pelos classificadores da segunda etapa com a base de dados digits.

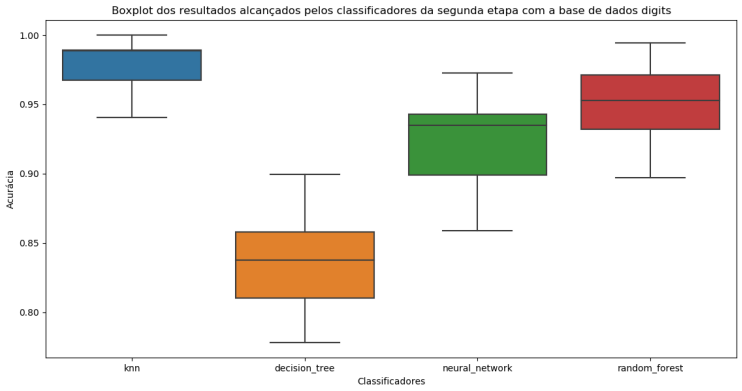

Figura 4: Boxplot dos resultados alcançados pelos classificadores da segunda etapa com a base de dados digits.

Comparando os resultados das duas etapas, nota-se que os métodos da segunda etapa foram, em geral, bem melhores para esta base de dados. O único método da primeira etapa que chegou perto das médias obtidas na segunda etapa foi o classificador Centroide.

4.3. Wine

A média e o desvio padrão das acurácias obtidas pelos classificadores da primeira etapa com a base de dados wine são apresentados na Tabela 5. O *boxplot* das acurácias obtidas em cada *fold* é apresentado na Figura 5.

Classificadores	Média	Desvio Padrão
ZeroR	0,33333333	0,00000000
OneR	0,95333333	0,06699917
OneR Probabilístico	0,90000000	0,09067647
Centroide	0,93333333	0,05962848
Centroide OneR	0,96000000	0,05333333
Naive-Bayes Gaussiano	0,95333333	0,04268749

Tabela 5: Média e desvio padrão das acurácias obtidas pelos classificadores da primeira etapa com a base de dados wine.

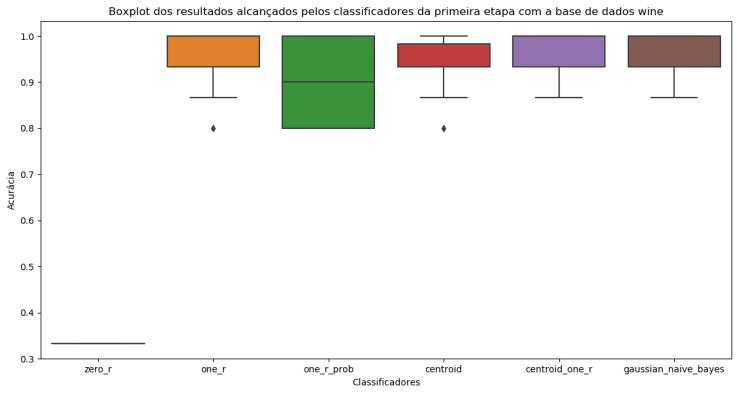

Figura 5: Boxplot dos resultados alcançados pelos classificadores da primeira etapa com a base de dados wine.

Com exceção do ZeroR, todos os classificadores apresentaram excelentes médias das acurácias. O Centroide OneR destaca-se mais uma vez devido à sua robustez. Além disso, o classificador Naive-Bayes Gaussiano obteve o me-

nor desvio padrão, enquanto que o OneR Probabilístico apresentou a maior variabilidade.

A média e o desvio padrão das acurácias obtidas pelos classificadores da segunda etapa com a base de dados wine são apresentados na Tabela 6. O *boxplot* das acurácias obtidas em cada *fold* é apresentado na Figura 6.

Classificadores	**Média**	**Desvio Padrão**
KNN	0,95333333	0,06000000
Árvore de Decisão	0,95333333	0,04268749
Rede Neural	0,79333333	0,14126413
Florestas de Árvores	0,95333333	0,06000000

Tabela 6: Média e desvio padrão das acurácias obtidas pelos classificadores da segunda etapa com a base de dados wine.

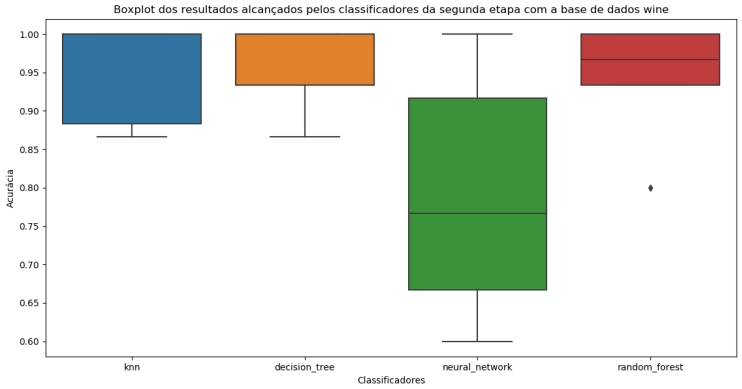

Figura 6: Boxplot dos resultados alcançados pelos classificadores da segunda etapa com a base de dados wine.

Por meio dos resultados obtidos, nota-se que todos os métodos obtiveram a mesma média, com exceção do Rede Neural, que novamente mostrou-se muito instável. Os métodos KNN e Florestas de Árvores empataram inclusive no quesito desvio padrão, mas por meio do *boxplot* observam-se melhores acurácias para o Florestas de Árvores, que ficou comprometido devido a um *outlier*.

Comparando os resultados das duas etapas, nota-se que o classificador Centroide OneR apresentou a maior média das acurácias, muito próximo do empate triplo da segunda etapa entre o KNN, Árvore de Decisão e Florestas de Árvores.

4.4. Breast Cancer

A média e o desvio padrão das acurácias obtidas pelos classificadores da primeira etapa com a base de dados breast cancer são apresentados na Tabela 7. O *boxplot* das acurácias obtidas em cada *fold* é apresentado na Figura 7.

Classificadores	**Média**	**Desvio Padrão**
ZeroR	0,62742740	0,00441189
OneR	0,91414636	0,03982248
OneR Probabilístico	0,86320003	0,04135683
Centroide	0,89136419	0,03879384
Centroide OneR	0,90343099	0,03600602
Naive-Bayes Gaussiano	0,93867967	0,03011289

Tabela 7: Média e desvio padrão das acurácias obtidas pelos classificadores da primeira etapa com a base de dados breast cancer.

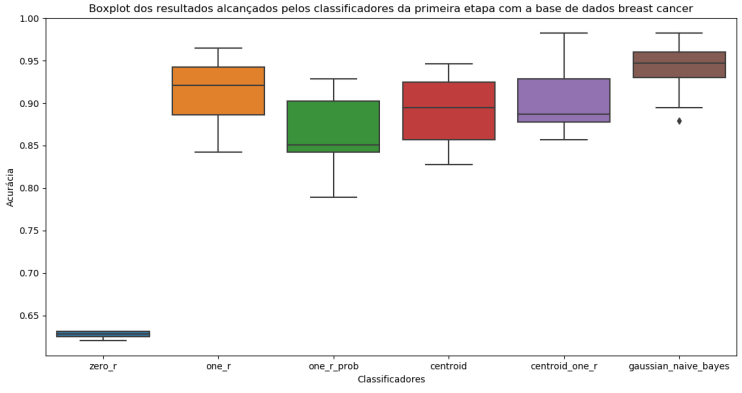

Figura 7: Boxplot dos resultados alcançados pelos classificadores da primeira etapa com a base de dados breast cancer.

Observa-se que o classificador Naive-Bayes Gaussiano apresentou os melhores resultados, relativamente próximo ao OneR. É interessante notar que o classificador OneR mostrou-se melhor que o Centroide OneR. Também vale destacar que, surpreendentemente, o classificador ZeroR apresentou uma melhora, mesmo que pequena, para esta base de dados.

A média e o desvio padrão das acurácias obtidas pelos classificadores da segunda etapa com a base de dados breast cancer são apresentados na Tabela 8. O *boxplot* das acurácias obtidas em cada *fold* é apresentado na Figura 8.

Classificadores	**Média**	**Desvio Padrão**
KNN	0,93165889	0,03124517
Árvore de Decisão	0,90885079	0,04859750
Rede Neural	0,76296128	0,24621764
Florestas de Árvores	0,96139703	0,02162908

Tabela 8: Média e desvio padrão das acurácias obtidas pelos classificadores da segunda etapa com a base de dados breast cancer.

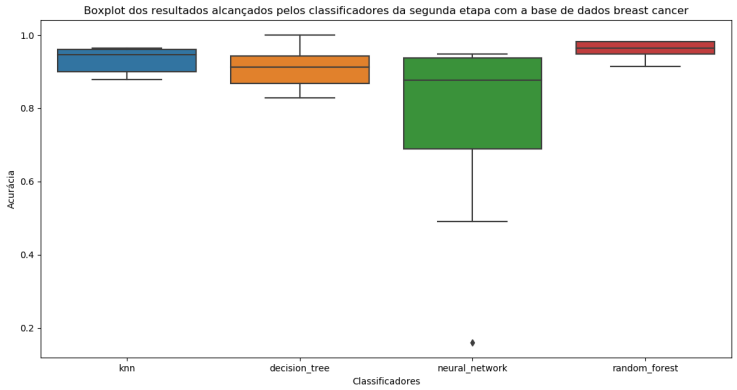

Figura 8: Boxplot dos resultados alcançados pelos classificadores da segunda etapa com a base de dados breast cancer.

Por meio dos resultados obtidos, nota-se que o método Florestas de Árvores apresentou os melhores resultados. Pelo *boxplot* observa-se pouca variabilidade

de suas acurácias. Novamente, o método Rede Neural demonstrou instabilidade, com grande variabilidade e presença de um *outlier*.

Comparando os resultados das duas etapas, observa-se que o método Florestas de Árvores apresentou os melhores resultados, inclusive de desvio padrão, desconsiderando o classificador ZeroR. No geral, os métodos da segunda etapa mostraram-se melhores que os da primeira etapa.

5. Conclusões

Este trabalho foi importante para estabelecer um estudo comparativo dos principais métodos de aprendizado para problemas de classificação apresentados na disciplina de Inteligência Artificial, entre eles: ZeroR, OneR, OneR Probabilístico, Centroide, Centroide OneR, Naive-Bayes Gaussiano, KNN, Árvore de Decisão, Rede Neural e Florestas de Árvores.

A etapa de implementação dos classificadores foi essencial para o aprendizado da utilização da biblioteca *scikit-learn* e da criação de classificadores customizados.

A partir dos resultados obtidos, verifica-se que o classificador OneR Centroide mostrou-se o mais robusto dentre os classificadores implementados. Por outro lado, o ZeroR apresentou péssimos resultados para todos os testes realizados. O OneR Probabilístico mostrou-se muito bom na maioria dos casos, mas sua natureza aleatória comprometeu um possível excelente desempenho.

Entre os classificadores utilizados da biblioteca *scikit-learn*, o Naive-Bayes Gaussiano foi bem estável, apresentando resultados bem satisfatórios. O método Rede Neural mostrou-se muito variável e pouco eficiente em termos de acurácia na maioria dos casos de teste. Particularmente, o método Florestas de Árvores apresentou os resultados mais satisfatórios, sendo melhor que os demais em vários casos. Além disso, esse método mostrou-se melhor que os métodos implementados.

Em resumo, conclui-se que os métodos Centroide OneR e Florestas de Árvores foram bem adequados para as bases de dados utilizadas, com exceção da

base de dados digits, com a qual o Naive-Bayes Gaussiano, KNN e Florestas de Árvores mostraram-se mais estáveis.

Por fim, este trabalho mostrou-se bastante relevante para o estudo de métodos de aprendizado aplicados a problemas de classificação. Uma possível extensão interessante englobaria outros classificadores existentes na literatura, além de outros classificadores customizados que poderiam ser idealizados. Além disso, poderiam ser consideradas bases de dados mais heterogêneas e com maiores volumes de dados.

Referências Bibliográficas

- https://en.wikipedia.org/wiki/Statistical_classification
- https://www.saedsayad.com/oner.htm
- https://scikit-learn.org/stable/datasets/index.html
- https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.GaussianNB.html
- https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html
- https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html
- https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html
- https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html
- Informações do professor disponibilizadas por e-mail