Segundo Trabalho de Inteligência Artificial

Matheus Gomes Arante de Souza

Abstract

Trabalho de Inteligência Artificial com o objetivo de comparar experimentalmente um conjunto pré-definido de técnicas de aprendizado e classificação automática aplicadas a problemas de classificação.

Keywords: Inteligência Artificial, Classificadores, ZeroR, OneR, Centróide, Naive Bayes Gaussiano, OneR Probabilístico, OneR Centroide, KNN, Árvore de decisão, Rede Neural, Floresta de árvores, Datasets, Iris, Digits, Wine, Breast Cancer

1. Introdução

Este trabalho consiste em analisar experimentalmente o comportamento das técnicas de aprendizado e classificação automática ZeroR, OneR, OneR Probabilístico, Centróide, OneR Centróide, Naive Bayes Gaussiano, KNN, Árvore de Decisão, Rede Neural e Florestas de Árvores quando submetidas as bases de dados Iris, Digits, Wine e Breast Cancer.

2. Descrição dos Datasets

Nesta seção serão apresentadas, de maneira breve, informações sobre as bases de dados utilizadas neste trabalho: Iris, Digits, Wine e Breast Cancer.

2.1. Iris

A base de dados Iris armazena informações sobre o comprimento e a largura de pétalas e sépalas de três espécies da flor Iris: Setosa, Versicolour e Virginica.

Figura 1: Espécies de Iris: Versicolour, Setosa e Virginica, respectivamente.

Além de utilizar essas medidas para criar um modelo de classificar as espécies de Iris, este dataset é frequentemente usado em exemplos de mineração, classificação e agrupamento de dados e no teste de algoritmos.

Nº DE CLASSES	3
Nº DE EXEMPLOS POR CLASSE	50
Nº TOTAL DE EXEMPLOS	150
Nº DE ATRIBUTOS POR EXEMPLO	4

Tabela 1: Características do dataset Iris.

2.2. Digits

A base de dados Digits é composta de imagens 8x8 de dígitos escritos à mão. Um exemplo destas imagens é apresentado a seguir. Cada imagem é representada por um vetor de tamanho 64, no qual o valor de cada posição é um inteiro no intervalo 0 a 16 e indica a tonalidade de preto daquele pixel.

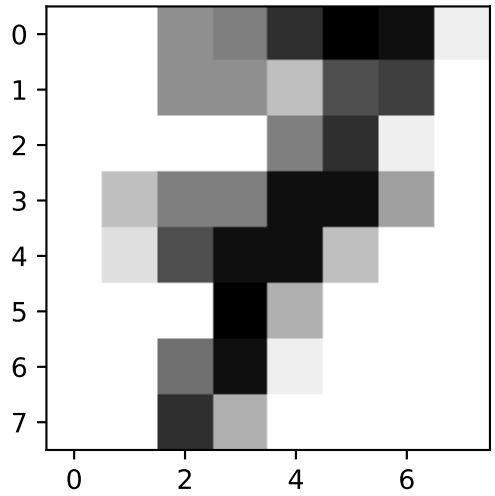

Figura 2: Representação gráfica do dígito 7 da base digits.

Nº DE CLASSES	10
Nº DE EXEMPLOS POR CLASSE	≈ 180
Nº TOTAL DE EXEMPLOS	1797
Nº DE ATRIBUTOS POR EXEMPLO	64

Tabela 2: Características do dataset Digits.

2.3. Wine

A base de dados Wine contém resultados de uma análise química de vinhos cultivados na mesma região na Itália por três agricultores diferentes.

Existem treze medidas diferentes para diferentes constituintes encontradas nos três tipos de vinho:

- Álcool
- Ácido málico
- Cinza
- Alcalinidade das cinzas
- Magnésio
- Fenóis totais
- Flavonóides
- Fenóis não flavonóides
- Proantocianinas
- Intensidade da cor
- Matiz
- OD280 / OD315 de vinhos diluídos
- Proline

Nº DE CLASSES	3
Nº DE EXEMPLOS POR CLASSE	[59, 71, 48]
Nº TOTAL DE EXEMPLOS	178
Nº DE ATRIBUTOS POR EXEMPLO	13

Tabela 3: Características do dataset Wine.

2.4. Breast Cancer

A base de dados Breast Cancer reúne resultados que foram calculados através de imagens digitalizadas de um procedimento médico chamado punção aspirativa por agulha fina (fine needle aspirate (FNA)) de massa mamária. Cada exemplo possui 30 atributos que descrevem características da massa mamária de pacientes. A partir destes atributos, os exemplos são classificados como câncer maligno ou benigno.

A motivação por trás do estudo deste dataset é o desenvolvimento de um algoritmo capaz de prever se um paciente tem um tumor maligno ou benigno, de acordo com os dados calculados a partir de sua massa mamária.

Nº DE CLASSES	2
Nº DE EXEMPLOS POR CLASSE	212 (Malignos) e 357 (Benignos)
Nº TOTAL DE EXEMPLOS	569
Nº DE ATRIBUTOS POR EXEMPLO	30

Tabela 4: Características do dataset Breast Cancer.

3. Descrição dos Métodos Implementados

Nesta seção serão apresentadas breves descrições dos classificadores implementados: ZeroR, OneR, OneR Probabilístico, Centróide e OneR Centróide.

3.1. ZeroR

ZeroR ("Zero Rule") é o classificador mais simples, pois depende apenas das classes e ignora todas as características, de modo que simplesmente prevê a classe majoritária. Embora não haja poder de previsibilidade no ZeroR, ele pode ser utilizado como baseline para determinar o desempenho de outros classificadores.

3.2. OneR

OneR ("One Rule") é um algoritmo de classificação simples, porém de boa precisão, que seleciona o atributo da base de treino com o maior poder preditivo

e cria uma regra para cada valor deste atributo, fazendo uma associação entre o valor do atributo e sua correspondente classe majoritária nos exemplos de treino. Na etapa de testes a predição é feita com base nessas regras, ou seja, os valores da característica com maior poder preditivo (determinada na fase de treino) de cada exemplo são verificados e de acordo com as regras, é feita a classificação dos exemplos de teste.

3.3. OneR Probabilístico

O classificador OneR Probabilístico segue uma ideia semelhante ao OneR: selecionar o atributo dos exemplos de treino que tem o maior poder de predição. A diferença é que o OneR cria uma regra para cada valor deste atributo de acordo com sua respectiva classe majoritária e posteriormente faz a predição baseado nessas regras, enquanto o OneR Probabilístico utiliza um sistema de roleta que mantém a proporção entre as ocorrências de um valor e sua frequência de classificação no treino, mas cria a possibilidade de que um dado exemplo no teste possa ser classificado com uma classe diferente da majoritária.

3.4. Centroide

No classificador Centroide, os exemplos de cada classe são utilizados para determinar seu respectivo centroide. Ao classificar um caso, é feito o cálculo da distância euclidiana do caso para os centroides de cada classe e este é classificado com a classe correspondente ao centroide mais próximo.

3.5. OneR Centroide

No classificador OneR Centroid também se escolhe o atributo tal como no OneR. A partir daí, os exemplos do conjunto de treino são divididos em grupos (cada grupo está relacionado com uma classe e é formado pelos exemplos que tem o mesmo valor do atributo escolhido). De cada grupo calcula-se o centroide e a predição é feita como no classificador Centroide: de cada exemplo é calculada sua distância euclidiana para cada centroide e a classificação é feita de acordo com a classe associada ao centroide mais próximo.

4. Descrição dos Experimentos Realizados

O procedimento experimental foi dividido em duas etapas:

- 1ª etapa:

  Treino e teste com validação cruzada de 10 folds dos classificadores que não possuem hiperparâmetros (ZeroR, OneR, OneR Probabilístico, Centróide, OneR Centróide e Naive Bayes Gaussiano).

- 2ª etapa:

  Treino, validação e teste dos classificadores que precisam de ajuste de hiperparâmetros (KNN, Árvore de Decisão, Redes Neurais e Florestas de Árvores).

  O procedimento de treinamento, validação e teste foi feito por meio 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 levou em consideração os seguintes valores de hiperparâmetros de cada técnica de aprendizado:

  - KNN: [n_neighbors = 1, 3, 5, 7, 10]
  - Árvore de Decisão: [max_depth = None, 3, 5, 10]
  - Rede Neural: {[max_iter = 50, 100, 200], [hidden_layer_sizes=(15,)]}
  - Florestas de Árvores: [n_estimators = 10, 20, 50, 100]

A seguir são apresentados os resultados obtidos pelos classificadores quando aplicados a diferentes bases de dados.

4.1. Iris

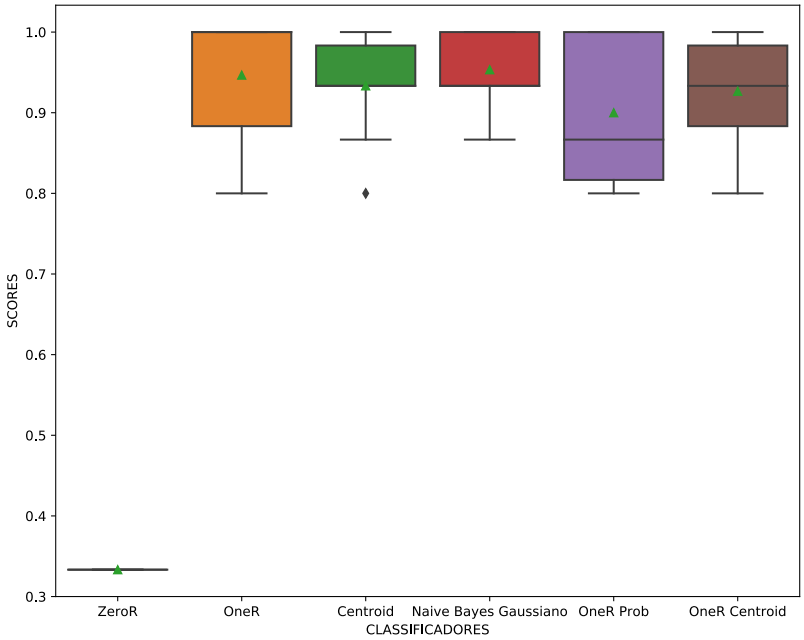

Figura 3: Boxplot dos resultados dos classificadores ZeroR, OneR, Centroid, Naive Bayes Gaussiano, OneR Prob. e OneR Centroid quando aplicados a base Iris.

CLASSIFICADOR	MÉDIA	DESVIO PADRÃO
ZeroR	0.333333	5.55112e-17
OneR	0.946667	0.0718022
Centroid	0.933333	0.0596285
Naive Bayes Gaussiano	0.953333	0.0426875
OneR Prob	0.9	0.0856349
OneR Centroid	0.926667	0.0628932

Tabela 5: Resultados dos classificadores ZeroR, OneR, Centroid, Naive Bayes Gaussiano, OneR Prob. e OneR Centroid quando aplicados a base Iris.

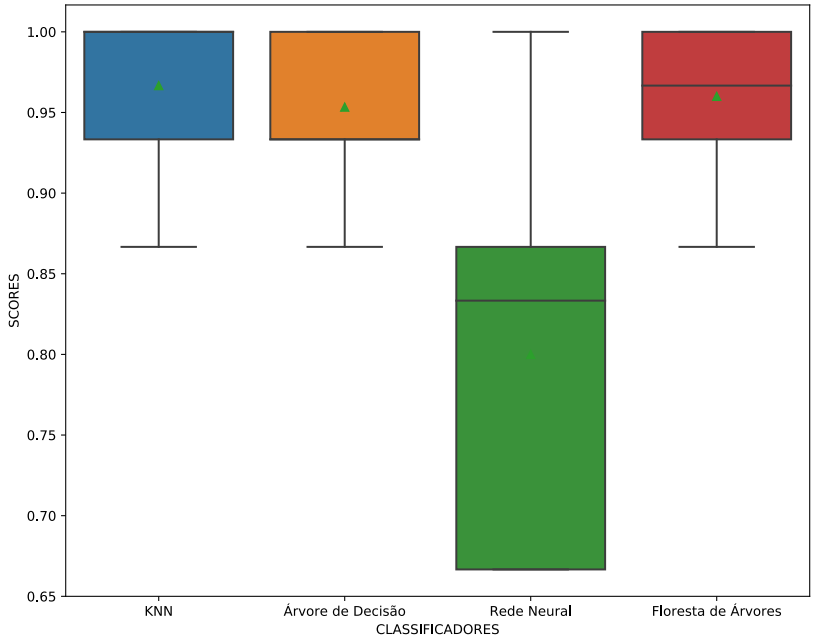

Figura 4: Boxplot dos resultados dos classificadores KNN, Árvore de Decisão, Rede Neural e Floresta de Árvores quando aplicados a base Iris.

CLASSIFICADOR	MÉDIA	DESVIO PADRÃO	MELHOR(ES) PARÂMETRO(S)
KNN	0.966667	0.0447214	{'n_neighbors': 5}
Árvore de Decisão	0.953333	0.0426875	{'max_depth': 5}
Rede Neural	0.8	0.119257	{'hidden_layer_sizes': (15,), 'max_iter': 200}
Floresta de Árvores	0.96	0.0442217	{'n_estimators': 100}

Tabela 6: Resultados dos classificadores KNN, Árvore de Decisão, Rede Neural e Floresta de Árvores quando aplicados a base Iris.

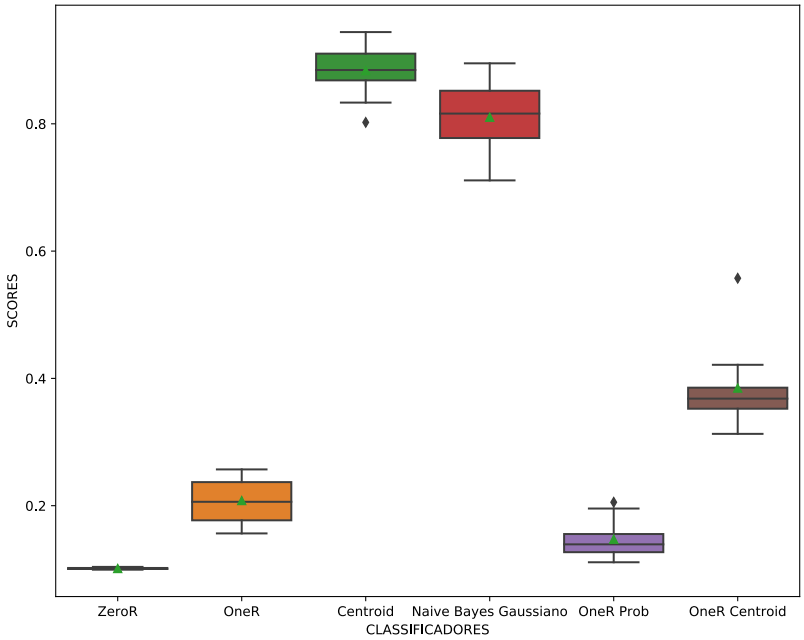

Figura 5: Boxplot dos resultados dos classificadores ZeroR, OneR, Centroid, Naive Bayes Gaussiano, OneR Prob. e OneR Centroid quando aplicados a base Digits.

CLASSIFICADOR	MÉDIA	DESVIO PADRÃO
ZeroR	0.101274	0.0012744
OneR	0.208041	0.0342592
Centroid	0.88361	0.0411268
Naive Bayes Gaussiano	0.810354	0.0566554
OneR Prob	0.147318	0.0298898
OneR Centroid	0.384776	0.0634627

Tabela 7: Resultados dos classificadores ZeroR, OneR, Centroid, Naive Bayes Gaussiano, OneR Prob. e OneR Centroid quando aplicados a base Digits.

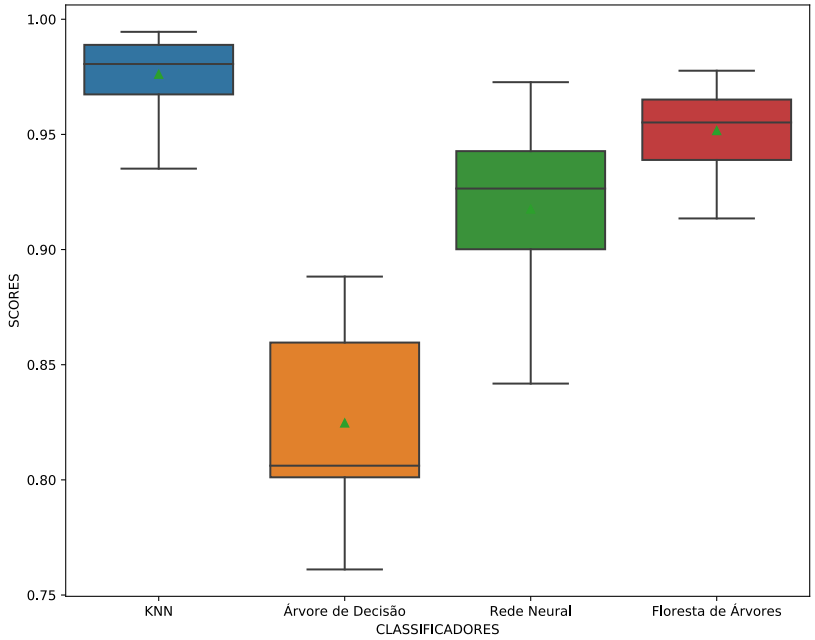

Figura 6: Boxplot dos resultados dos classificadores KNN, Árvore de Decisão, Rede Neural e Floresta de Árvores quando aplicados a base Digits.

CLASSIFICADOR	MÉDIA	DESVIO PADRÃO	MELHOR(ES) PARÂMETRO(S)
KNN	0.976149	0.0179084	{'n_neighbors': 1}
Árvore de Decisão	0.824715	0.0384317	{'max_depth': 10}
Rede Neural	0.917609	0.0370716	{'hidden_layer_sizes': (15,), 'max_iter': 200}
Floresta de Árvores	0.951679	0.0184822	{'n_estimators': 100}

Tabela 8: Resultados dos classificadores KNN, Árvore de Decisão, Rede Neural e Floresta de Árvores quando aplicados a base Digits.

4.3. Wine

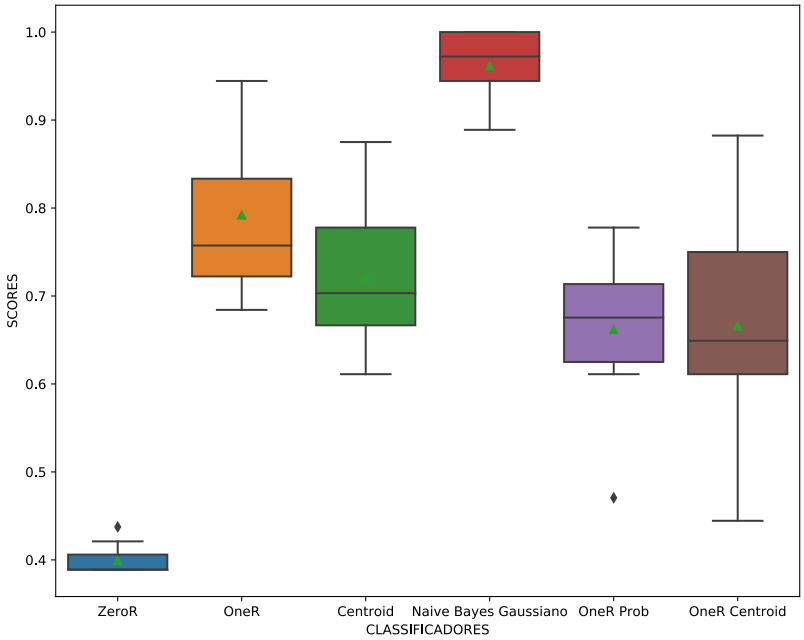

Figura 7: Boxplot dos resultados dos classificadores ZeroR, OneR, Centroid, Naive Bayes Gaussiano, OneR Prob. e OneR Centroid quando aplicados a base Wine.

CLASSIFICADOR	MÉDIA	DESVIO PADRÃO
ZeroR	0.399254	0.0168716
OneR	0.792114	0.0886734
Centroid	0.721607	0.0849013
Naive Bayes Gaussiano	0.961696	0.042442
OneR Prob	0.662008	0.079789
OneR Centroid	0.665977	0.122173

Tabela 9: Resultados dos classificadores ZeroR, OneR, Centroid, Naive Bayes Gaussiano, OneR Prob. e OneR Centroid quando aplicados a base Wine.

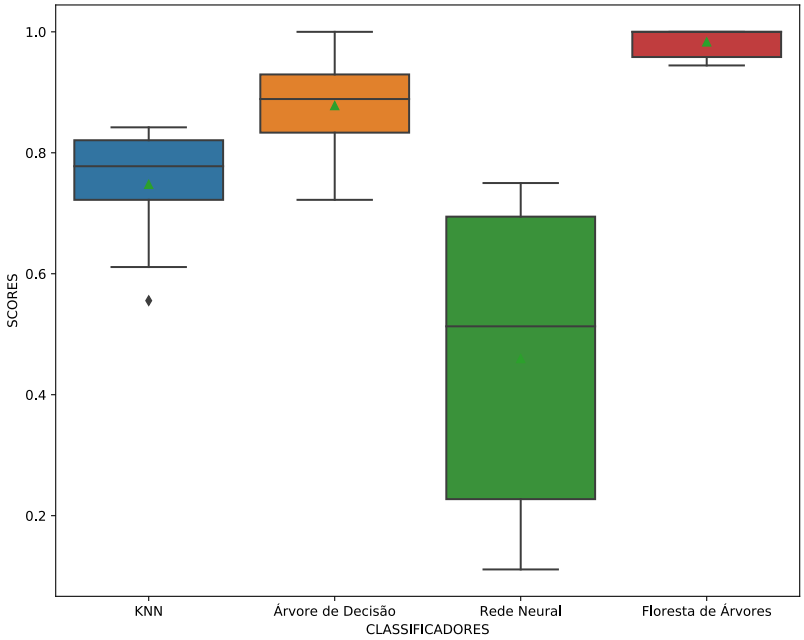

Figura 8: Boxplot dos resultados dos classificadores KNN, Árvore de Decisão, Rede Neural e Floresta de Árvores quando aplicados a base Wine.

CLASSIFICADOR	MÉDIA	DESVIO PADRÃO	MELHOR(ES) PARÂMETRO(S)
KNN	0.747813	0.0921293	{'n_neighbors': 1}
Árvore de Decisão	0.878036	0.0735251	{'max_depth': 3}
Rede Neural	0.459778	0.235989	{'hidden_layer_sizes': (15,), 'max_iter': 200}
Floresta de Árvores	0.983333	0.0254588	{'n_estimators': 50}

Tabela 10: Resultados dos classificadores KNN, Árvore de Decisão, Rede Neural e Floresta de Árvores quando aplicados a base Wine.

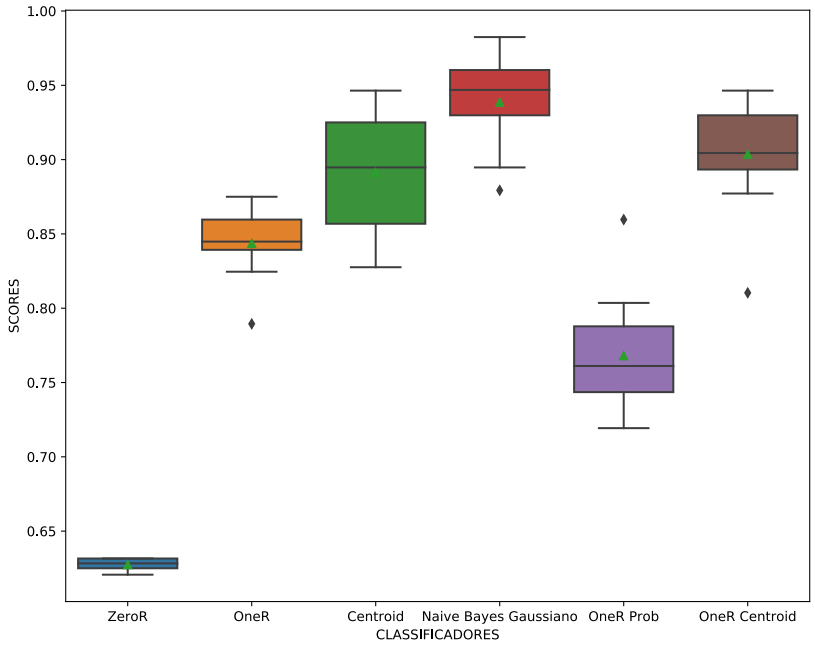

Figura 9: Boxplot dos resultados dos classificadores ZeroR, OneR, Centroid, Naive Bayes Gaussiano, OneR Prob. e OneR Centroid quando aplicados a base Breast Cancer.

CLASSIFICADOR	MÉDIA	DESVIO PADRÃO
ZeroR	0.627427	0.00441189
OneR	0.843621	0.0225633
Centroid	0.891364	0.0387938
Naive Bayes Gaussiano	0.93868	0.0301129
OneR Prob	0.768047	0.0403389
OneR Centroid	0.903647	0.0384147

Tabela 11: Resultados dos classificadores ZeroR, OneR, Centroid, Naive Bayes Gaussiano, OneR Prob. e OneR Centroid quando aplicados a base Breast Cancer.

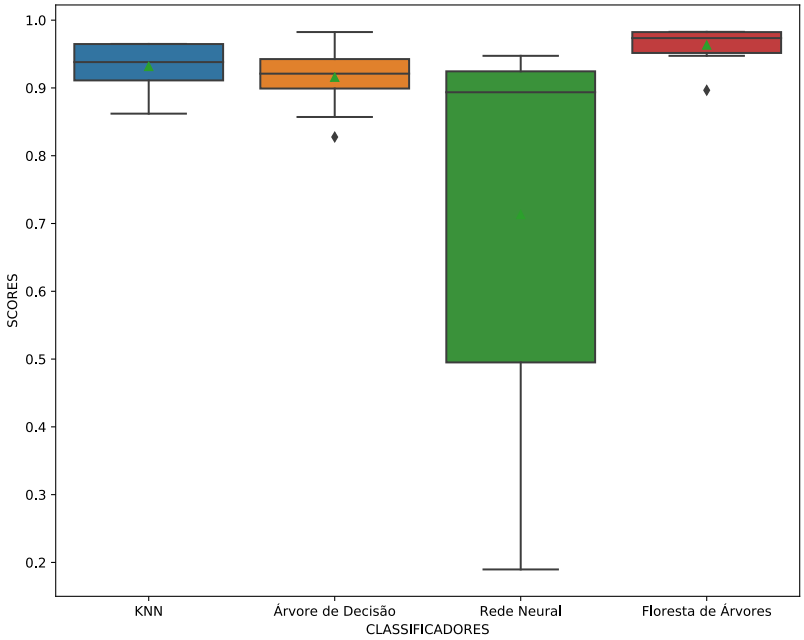

Figura 10: Boxplot dos resultados dos classificadores KNN, Árvore de Decisão, Rede Neural e Floresta de Árvores quando aplicados a base Breast Cancer.

CLASSIFICADOR	MÉDIA	DESVIO PADRÃO	MELHOR(ES) PARÂMETRO(S)
KNN	0.931689	0.0339011	{'n_neighbors': 10}
Árvore de Decisão	0.915806	0.0445551	{'max_depth': 3}
Rede Neural	0.713259	0.307229	{'hidden_layer_sizes': (15,), 'max_iter': 200}
Floresta de Árvores	0.963244	0.026056	{'n_estimators': 50}

Tabela 12: Resultados dos classificadores KNN, Árvore de Decisão, Rede Neural e Floresta de Árvores quando aplicados a base Breast Cancer.

5. Conclusões

Analisando os resultados obtidos, é possível notar que dentre os classificadores, aquele que apresentou o pior desempenho em todas as bases foi o ZeroR, como era esperado, devido a sua incapacidade de previsibilidade.

Na base Iris a maior média ficou por conta do classificador KNN, mas vale ressaltar que dos classificadores implementados, o de maior média foi o OneR e a diferença entre essas médias foi relativamente pequena.

Na base Digits, de maneira geral, as médias dos resultados dos classificadores da parte 1 caíram, se comparadas aos resultados da base anterior. Isso pode ser devido a menor complexidade desses algoritmos (se comparados aos da parte 2) frente a uma base que possui diversas classes e poucos atributos, tornando assim complicada a tarefa de se classificar um dígito de acordo com a tonalidade de preto de apenas 64 pixels. Apesar desta dificuldade, dentre os algoritmos da parte 1, o Centroid foi o que demonstrou melhor resultado, enquanto que todos os classificadores da parte 2 foram superiores aos da parte 1, e mais uma vez com destaque para o KNN.

Na base Wine os destaques vão para os classificadores Naive Bayes Gaussiano e Floresta de Árvores. Note que desta vez, o classificador KNN que vinha apresentando os melhores resultados até então apresentou uma média bem inferior as apresentadas anteriormente. O classificador Floresta de Árvores, além de entregar uma média muito alta, demonstrou pouca variabilidade em cada fold, como é possível ver no boxplot da figura 8.

Na base Breast Cancer mais uma vez os melhores resultados da parte 1 e 2 ficaram por conta, respectivamente, dos classificadores Naive Bayes Gaussiano e Floresta de Árvores. Isso mostra que em bases cujos exemplos possuem muitos atributos, estes tem um bom poder de predição.

Este experimento foi interessante para analisar o comportamento de diversos classificadores quando submetidos a diferentes bases. Entretanto, como os algoritmos KNN, Árvore de Decisão, Rede Neural, Floresta de Árvores e Naive Bayes Gaussiano já estão implementados pelo `sklearn`, provavelmente estes

possuem internamente estratégias que maximizam seus resultados e de maneira geral acabaram sendo superiores aos classificadores que tiveram de ser implementados. Apesar desse fato, como é possível observar nos boxplots e tabelas de resultados, em alguns momentos OneR, Centroid e OneR Centroid se saíram bem. Além disso, como os classificadores da parte 2 necessitam do ajuste de hiperparâmetros, uma variabilidade maior de valores para os atributos da busca em grade seria um trabalho relevante, pois assim, talvez, classificadores que não apresentaram os melhores resultados poderiam obter resultados superiores.

Referências

- Material sobre "Machine Learning" disponibilizado pelo professor Flávio Miguel Varejão.
- https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_iris.html
- https://scikit-learn.org/stable/datasets/index.html#iris-dataset
- http://www.lac.inpe.br/~rafael.santos/Docs/CAP394/WholeStory-Iris.html
- https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_digits.html
- https://scikit-learn.org/stable/auto_examples/datasets/plot_digits_last_image.html
- https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_wine.html
- https://scikit-learn.org/stable/datasets/index.html#wine-dataset
- https://scikit-learn.org/stable/datasets/index.html#breast-cancer-dataset
- https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_breast_cancer.html
- https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)
- https://www.saedsayad.com/classification.htm