Segundo Trabalho de Inteligência Artificial: Estudo de Métodos para Problemas de Classificação

Gabriel Ferrari Cipriano [1]

Ufes — Departamento de Informática [2]

Abstract

Este artigo busca realizar uma comparação experimental entre diferentes técnicas de aprendizado e classificação aplicadas a diferentes bases de dados. As técnicas que serão comparadas são *ZeroR, Aleatório, Aleatório Estratificado, Naive Bayes Gaussiano, Knn, DistKnn, Árvore de Decisão* e *Florestas de Árvores*, todas disponibilizados pelo Scikit-learn[1], a maior *biblioteca* de *Machine Learning* para *Python*, juntamente dos classificadores implementados para este trabalho: *OneR Probabilístico, KmeansCentroides* e *KGACentroides*.

Após a análise estatística das etapas de *treino* e *teste* com validação cruzada, foi constatado que o método que melhor performou em todas as bases de Dados foi o *Floresta de Árvores*.

Keywords: metaheuristics, Classification, Supervised Learning

1. Introdução

O problema de classificação é um tipo de *Reconhecimento de Padrão*, uma parte importante e extensa da área de *Machine Learning*, que consiste em identificar a que classe uma determinada Observação pertence. Um exemplo de problema de classificação seria tentar descobrir a Nacionalidade de uma pessoa com base em informações desta pessoa, como Sexo, Idade, Salário Anual e Grau de Educação.

[1] gabriel.cipriano 'at' edu.ufes.br

[2] Ufes - Dep. de Informática https://informatica.ufes.br/

Um algoritmo desenvolvido para resolver um problema de classificação é também conhecido como *Classificador*. O objetivo deste estudo é fazer uma análise de desempenho de diferentes classificadores, são eles *ZeroR, Aleatório, Aleatório Estratificado, Naive Bayes Gaussiano, Knn, DistKnn, Árvore de Decisão*, *Florestas de Árvores*, *OneR Probabilístico, KmeansCentroides* e *KGACentroides*.

Os *Classificadores* objetos deste estudo serão todos treinados por meio de Aprendizado Supervisionado, onde os dados disponibilizados para a etapa de Treino já possuem sua classe conhecida, e o classificador deve, portanto, aprender a classificar uma Observação utilizando a estratégia desenvolvida com base nas informações obtidas no treino. Para treinarmos e mensuramos o desempenho dos Métodos de Classificação, cada Classificador será submetido a diferentes bases de dados: Iris, Wine, Digits e Breast Cancer.

1.1. Classificadores fornecidos pelo Scikit-learn

Os métodos *ZeroR, Aleatório* e *Aleatório Estratificados* são diferentes instâncias do classificador *Dummy Classifier*[3], tendo o valor do parâmetro *strategy* como 'most_frequent', 'uniform' e 'stratified', respectivamente. O *DummyClassifier* tenta classificar utilizando regras muito simples, servindo apenas como uma referência de comparação para outros classificadores.

Os métodos KNN e DistKnn são instâncias do classificador KneighborsClassifier[4], estabelecendo o valor do parâmetro *weights* como 'uniform' e 'distance', respectivamente. O *KNeighborsClassifier* classifica uma *Observação* levando em conta os K vizinhos mais próximos, atribuindo a classe mais popular à *Observação* de entrada.

O método Naive Bayes Gaussiano corresponde ao classificador GaussianNB[5], que classifica fazendo estimativas da Distribuição Normal Gaussiana dos dados.

[3] sklearn.dummy.DummyClassifier https://scikit-learn.org/stable/modules/generated/sklearn.dummy.DummyClassifier.html

[4] sklearn.neighbors.KNeighborsClassifier https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html

[5] sklearn.naive_bayes.GaussianNB https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.GaussianNB.html

O método Árvore de decisão corresponde ao classificador DecisionTreeClassifier[6], que classifica as Observações percorrendo uma árvore de descisão onde as classes são os nós-folhas dessa árvore. Por fim, o método Floresta de Árvores, que corresponde ao classificador RandomForestClassifier[7], que classifica utilizando várias Árvores de Decisão e aleatoriedade para tentar evitar *Overfitting*.

1.2. Bases de dados utilizadas neste estudo[2]

Iris. Base de dados com 150 *Observações* de plantas do gênero *Iris*, contendo 50 observações de cada uma das 3 três espécies presentes no *Dataset*:

- Íris Setosa
- Íris Versicolor
- Íris Virgínica

A classificação será feita com base em quatro atributos contínuos referentes a medições realizadas nas pétalas e nas sépalas das plantas.

Digits. Base de dados com 1797 imagens de tamanho 8x8, sendo cada imagem um dígito escrito a mão. Cada imagem (*Observação*) possui 64 atributos que representam a matriz de *pixels* dessa imagem, e pode ser classificada como um dígito de 0 a 9.

Wine. Base de dados com 178 *Observações* de uma análise química feita em três tipos de vinho vindos da mesma região. Sua classificação pode ser feita levando em conta os 13 atributos que as observações possuem, como *Teor Alcóolico, Intensidade de Cor* e *Alcalinidade das Cinzas*.

Breast Cancer. Base de dados com 569 instâncias de diagnósticos de câncer de mama possuindo 32 atributos cada. Sua classificação é binária, sendo classificados como *Tumor Benigno* ou *Tumor Maligno*.

[6] sklearn.tree.DecisionTreeClassifier https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html

[7] sklearn.ensemble.RandomForestClassifier https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

2. Métodos Implementados

Para este estudo foram implementados os classificadores *OneR Probabilístico* e *KCentróides*. Os métodos *KmeansCentroides* e *KGACentróides* são variações do classificador *KCentróides*, mudando apenas o algoritmo de agrupamento: *Kmeans* e *GeneticAlgorithm*, respectivamente.

Todos os métodos foram desenvolvidos seguindo os padrões de desenvolvimento[3] sugeridos pela biblioteca *Scikit-learn*, sendo então totalmente compatíveis com o ecossistema *Scikit-learn*. Assim, por convenção as duas principais etapas dos algoritmos são as de *Ajustar* os dados de treino (***fit()***), e de Predizer (***predict()***), onde é realizada a classificação das observações de entrada.

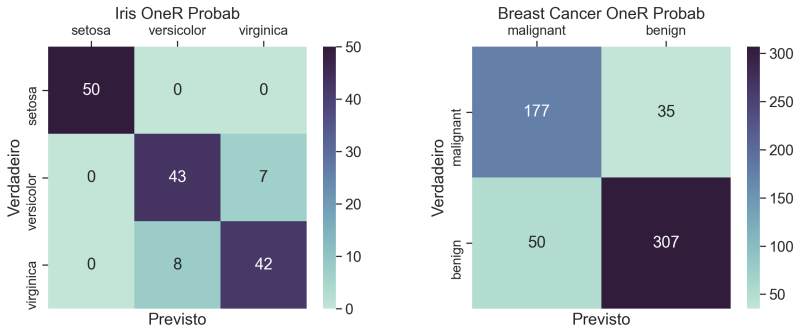
2.1. OneR Probabilístico

O método *OneR Probabilístico* se chama assim pois utiliza apenas **uma regra** (*One Rule*) como critério de classificação. No caso do classificador implementado, a regra de classificação é escolher o atributo que tem o maior poder de predição de classe dentre todos, e assim, predizer considerando a probabilidade de aquele valor de atributo pertencer a uma determinada classe. As etapas do algoritmo são como segue:

- ***fit()***: Constrói uma tabela de contingência para cada atributo dos dados de treino, para isso os dados precisam estar *discretizados*; Escolhe a tabela do melhor atributo para diferenciação; Obtém a distribuição de probabilidades de classes para cada valor que o atributo escolhido possa assumir.

- ***predict()***: Verifica qual valor o *atributo-regra* da *observação* assumiu; Faz um sorteio ponderado para predizer a classe que a *observação* pertence.

Na figura 1 é possível ver dois exemplos de classificações realizadas pelo *OneR Probabilístico* nas bases de dados *Iris* e *Breast Cancer*, respectivamente.



(a) Base de dados Iris (b) Base de dados Breast Cancer

Figura 1: Matrizes de dispersão representando classificações feitas pelo OneR após um Treino de validação cruzada com 10 folds.

2.2. KCentroides

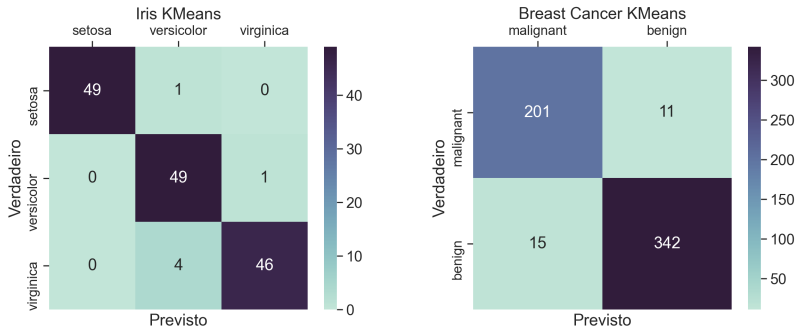
O método KCentroides utiliza os conceitos de *clusterização* para realizar a classificação dos dados. Para isso ele precisa receber um método de agrupamento e o número de K grupos que serão feitos para cada *classe* dos dados de treino. As etapas do algoritmo são como segue:

- ***fit()***: Faz uma lista de classes mapeando as classes em índices; Gera os K centróides de cada classe utilizando método passado como parâmetro; Guarda a lista de centróides de cada classe.
- ***predict()***: Calcula a distância euclidiana ao quadrado da *observação* para todos os centróides; Identifica qual foi o centróide mais próximo da observação; Atribui a classe deste centróide à *observação*.

Dentre as inúmeras abordagens de clusterização para o problema de agrupamento, os algoritmos de agrupamento propostos para serem usados no KCentróides foram o ***KMeans***[4] e o ***Genetic Algorithm***[5].

2.2.1. *KMeansCentroides*

O *KMeans* passado como parâmetro do *KCentroides* foi uma implementação do fornecido pela biblioteca Scikit-learn[8]. Na figura 2 é possível ver dois exemplos de classificações realizadas pelo *KMeansCentroides* nas bases de dados *Iris* e *Breast Cancer*, respectivamente.



(a) Iris (b) Breast Cancer

Figura 2: Matrizes de dispersão representando classificações feitas pelo *KmeansCentroides* com $k = 7$ após um Treino de validação cruzada com 10 folds.

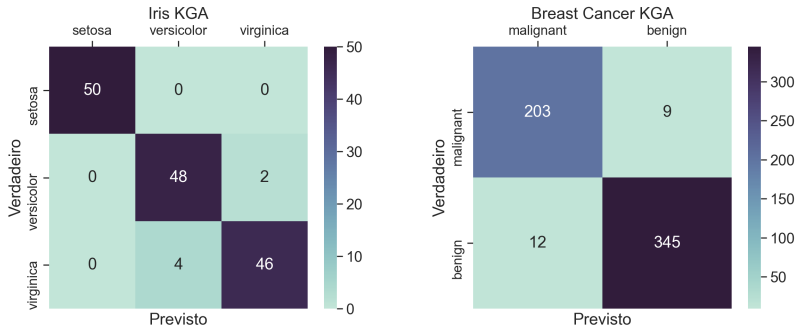
2.2.2. *KGACentroides*

O *Algoritmo Genético* passado como parâmetro do *KCentroides* foi de implementação própria, com os seguintes atributos:

- Tamanho da Populacao: **10**
- Taxa de Crossover: **0.95**
- Taxa de Mutacao: **0.2**

Na figura 3 é possível ver dois exemplos de classificações realizadas pelo *KGACentroides* nas bases de dados *Iris* e *Breast Cancer*, respectivamente.

[8] sklearn.cluster.KMeans https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html



(a) Iris (b) Breast Cancer

Figura 3: Matrizes de dispersão representando classificações feitas pelo *KGACentroides* com $k = 7$ após um Treino de validação cruzada com 10 folds.

3. Descrição dos Experimentos Realizados

O objetivo desse trabalho é realizar uma comparação experimental de todos os classificadores observando como eles desempenharam em diferentes bases de dados, sendo a ***Acurácia*** o principal critério de avaliação. Para isso, os experimentos foram divididos em duas etapas.

Primeira Etapa. Compreende os classificadores que não possuem hiperparâmetros, são eles: *ZeroR, Aleatório, Aleatório Estratificado, Naive Bayes Gaussiano* e *OneR Probabilístico*. Nesta etapa é realizado o *Treino* e *Teste* com três rodadas de validação cruzada estratificada de 10 *folds*.

Segunda Etapa. Compreende os classificadores que possuem hiperparâmetros, são eles: *KmeansCentroides, KGACentroides, Knn, DistKnn, Árvore de Decisão* e *Florestas de Árvores*. Nesta etapa, é realizado o *Treino, Validação,* e *Teste* através de três rodadas de ciclos aninhados de validação e teste, sendo o ciclo externo de teste com 10 *folds*, e o ciclo interno de validação de 4 *folds* realizando uma busca em grade (*grid search*[9]) para identificar o valor de hiperparâmetro com maior potencial.

[9]Tuning the hyper-parameters of an estimator https://scikit-learn.org/stable/modules/grid_search.html

Hiperparâmetros da Etapa Dois:.

- *KMeansCentroides*: $[k = 1, 3, 5, 7]$
- *KGACentroides*: $[k = 1, 3, 5, 7]$
- *Knn*: $[n\_neighbors = 1, 3, 5, 7]$
- *DistKnn*: $[n\_neighbors = 1, 3, 5, 7]$
- *Árvore de Decisão*: $[max\_depth = None, 3, 5, 10]$
- *Florestas de Árvores*: $[n\_estimators = 10, 20, 50, 100]$

Os experimentos foram conduzidos num computador *Dual-Core Intel® Core™* i5-7360U CPU @ 2.30GHz, 16GB de memória RAM 2133MHz LPDDR3, 260GB APPLE® SSD AP0256J, *Intel Iris Plus Graphics* 640 1,5GB e Sistema Operacional MacOS. Foi utilizada a *random seed* 36851234 para que os resultados possam ser reproduzidos. É importante citar que o *KGACentroides* não pode ter seus resultados reproduzidos pois seu principal critério de parada é o tempo de execução do algoritmo, que pode variar bastante a depender da CPU.

A tabela 1 mostra os tempos de execução dos experimentos, com tempo total de ***30,3*** **minutos**. Vale destacar que a base de dados que mais demorou para executar foi a *Digits*, com $20,8$ minutos de *running time*, o que era esperando devido ao seu tamanho. O *KGACentroides* foi o método com maior *running time*, com $18,5$ minutos de execução, pois foi implementado sem foco em desempenho quando comparado com os métodos fornecidos pelo *Scikit-learn*.

3.1. Iris

A tabela 2 apresenta a *Média, Desvio padrão, Limite Superior* e *Limite inferior* das acurácias de cada classificador na base de dados *Iris*. A figura 4 mostra o *boxplot* de acurácias dos *folds* para cada método na base de dados *Iris*.

Como esperado, o *Aleatorio* juntamente dos outros *DummyClassifiers* obtiveram os piores resultados. O *OneR Probabilístico* teve o maior desvio-padrão graças à sua natureza probabilística, mesmo assim performou relativamente bem

	Iris	Digits	Wine	Breast Cancer	Total por método
ZeroR	0.07	0.09	0.05	0.05	0.25
Aleatorio	0.05	0.09	0.05	0.05	0.24
Estratificado	0.06	0.10	0.05	0.06	0.27
OneR Probab	1.44	266.99	4.21	19.44	292.08
Naive Bayes	0.08	0.15	0.06	0.07	0.36
KMeans	24.53	122.79	24.92	38.32	210.56
KGA	61.78	765.84	68.33	215.33	1111.28
KNN	1.68	18.10	1.66	5.84	27.28
DistKNN	1.24	9.45	1.22	2.96	14.86
Árvore Desc	1.21	5.96	1.17	3.01	11.35
Floresta	26.16	61.30	27.05	35.53	150.04
Total por *Dataset*	118.31	1250.86	128.76	320.66	**1818.58**

Tabela 1: Tempos de execução do experimento por *dataset* e *método* (em segundos). Tempo total de execução em destaque.

para um algoritmo simples, com acurácia média de 90%. Os outros classificadores obtiveram resultados muito próximos, sendo o ***KMeansCentroides*** o método que obteve o melhor resultado nesta primeira análise, com acurácia média de 95,3%, menor desvio-padrão dentre os algoritmos sérios, e sequer apresentando *outliers*.

	Média	Desvio Padrão	Lim. Inferior	Lim. Superior
ZeroR	0.333	0.000	0.333	0.333
Aleatorio	0.267	0.000	0.267	0.267
Estratificado	0.467	0.000	0.467	0.467
OneR Probab	0.909	0.076	0.882	0.936
Naive Bayes	0.951	0.051	0.933	0.970
KMeans	0.953	0.043	0.938	0.969
KGA	0.949	0.064	0.926	0.972
KNN	0.942	0.070	0.917	0.967
DistKNN	0.947	0.058	0.926	0.967
Árvore Desc	0.951	0.045	0.935	0.967
Floresta	0.949	0.051	0.931	0.967

Tabela 2: Iris - Informações estatísticas do desempenho de cada classificador

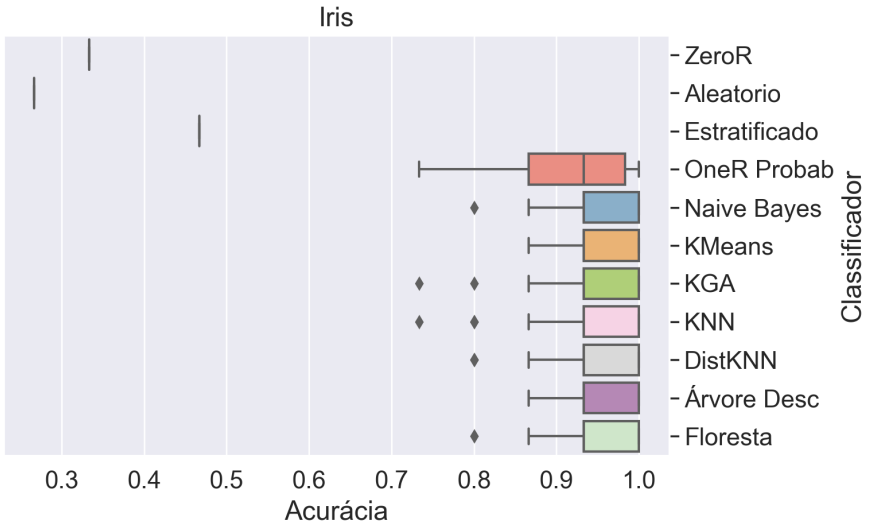


Figura 4: Boxplot do desempenho em cada *fold* dos diferentes métodos no *dataset* Iris.

Para fazermos uma análise estatística mais aprofundada recorremos aos testes pareados dos métodos de classificação. A tabela 3 mostra os resultados (p-values) do Teste t de *Student* na matriz triangular superior, e os resultados do teste de *Wilcoxon* na matriz triangular inferior. Os valores que apresentam fortes evidências para a rejeição da hipótese nula para um nível de significância de 95% ($p - value \leq 0.05$) estão em destaque.

Uma análise das diferenças estatísticas entre os classificadores será realizada na seção *4.1 - Análise Geral dos Resultados*. Vale citar que os métodos *ZeroR, Aleatório, Aleatório Estratificado* e *OneR Probabilistico* apresentaram rejeição da hipótese nula para todos os seus pares, em contrapartida, os métodos *Naive Bayles Gaussiano, KMeansCentroides, KGACentroides, KNN, DistKNN, Árvore* e *Floresta de Árvores* não rejeitaram a hipótese nula entre si.

ZeroR	**0.00000**	**0.00000**	**0.00000**	**0.00000**	**0.00000**	**0.00000**	**0.00000**	**0.00000**	**0.00000**	**0.00000**
0.00000	Aleat	**0.00000**	**0.00000**	**0.00000**	**0.00000**	**0.00000**	**0.00000**	**0.00000**	**0.00000**	**0.00000**
0.00000	**0.00000**	Estrat	**0.00000**	**0.00000**	**0.00000**	**0.00000**	**0.00000**	**0.00000**	**0.00000**	**0.00000**
0.00000	**0.00000**	**0.00000**	OneR P	**0.00222**	**0.00536**	**0.02170**	**0.04091**	**0.00877**	**0.00368**	**0.00143**
0.00000	**0.00000**	**0.00000**	**0.00354**	Naive	0.84538	0.87266	0.42350	0.60148	1.00000	0.74501
0.00000	**0.00000**	**0.00000**	**0.00791**	0.86843	KMeans	0.64526	0.28266	0.44793	0.78676	0.60148
0.00000	**0.00000**	**0.00000**	**0.02869**	0.97626	0.78151	KGA	0.41462	0.81302	0.83888	1.00000
0.00000	**0.00000**	**0.00000**	**0.04388**	0.47550	0.30494	0.40538	KNN	0.32558	0.38007	0.41462
0.00000	**0.00000**	**0.00000**	**0.01094**	0.59298	0.43858	1.00000	0.25684	DistK	0.60148	0.71223
0.00000	**0.00000**	**0.00000**	**0.00461**	1.00000	0.78151	0.96652	0.43858	0.59298	Árvore	0.74501
0.00000	**0.00000**	**0.00000**	**0.00242**	0.73888	0.59298	0.85105	0.47950	0.70546	0.73888	Florest

Tabela 3: Iris - p-values dos Testes Pareados. Teste t de Student na matriz triangular superior e Teste de Wilcoxon na matriz triangular inferior. Valores da tabela que rejeitaram a hipótese nula para um nível de significância de 95% estão escritos em negrito. Valores arredondados para cinco casas decimais.

3.2. Digits

A tabela 4 apresenta as informações estatísticas de cada classificador na base de dados *Digits*. A figura 5 mostra o *boxplot* de acurácias dos *folds* para cada método na base de dados *Digits*.

Como esperado, o *Zero R* juntamente dos outros *DummyClassifiers* obtiveram os piores resultados. O *OneR Probabilístico* performou mal também, graças a dimensionalidade da base de dados *Digits*, com 64 atributos, fazendo com que nenhum atributo seja unicamente decisivo ao classificar. Entre os Algoritmos implementados para o estudo, o *KGACentroides* foi o que obteve a melhor média, com 95,6% de taxa de acerto.

Os classificadores que obtiveram as melhores médias foram *KmeansCentroides, KGACentroides, KNN, DistKNN* e *Floresta de Árvores*. Desses, ***KNN, DistKNN*** e ***Floresta de Árvores*** foram os métodos que obtiveram os melhores resultados nesta análise inicial, todos os três com acurácia média de 97,6%, menores desvios-padrão dentre os algoritmos sérios, e sequer apresentando *outliers*.

	Média	Desvio Padrão	Lim. Inferior	Lim. Superior
ZeroR	0.101	0.002	0.100	0.102
Aleatorio	0.101	0.019	0.094	0.108
Estratificado	0.116	0.024	0.107	0.124
OneR Probab	0.178	0.023	0.170	0.187
Naive Bayes	0.784	0.030	0.773	0.795
KMeans	0.951	0.017	0.945	0.957
KGA	0.956	0.014	0.952	0.961
KNN	0.976	0.011	0.972	0.980
DistKNN	0.976	0.010	0.973	0.980
Árvore Desc	0.853	0.022	0.845	0.861
Floresta	0.976	0.012	0.972	0.980

Tabela 4: Digits - Informações estatísticas do desempenho de cada classificador

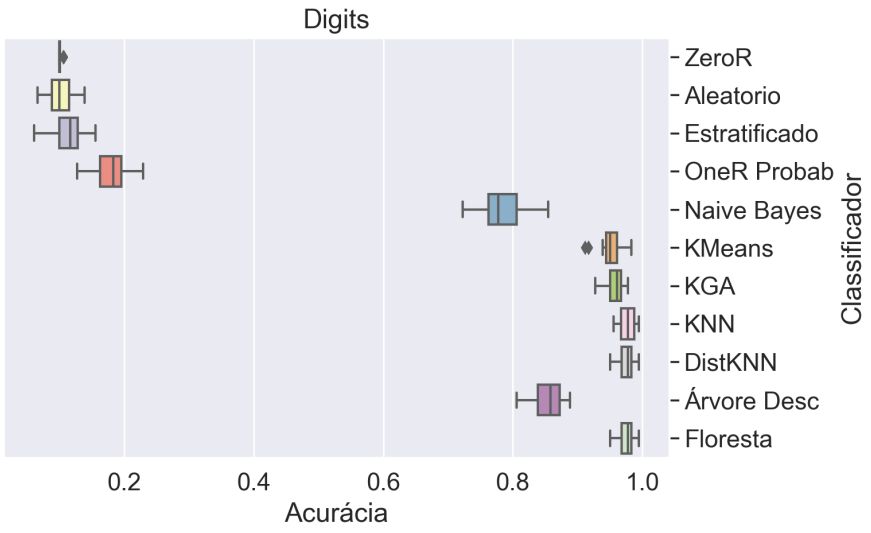


Figura 5: Boxplot do desempenho em cada *fold* dos diferentes métodos no *dataset* Digits.

A tabela 5 mostra os resultados (p-values) do Teste t de *Student* na matriz triangular superior, e os resultados do teste de *Wilcoxon* na matriz triangular inferior. Os valores que apresentam fortes evidências para a rejeição da hipótese nula para um nível de significância de 95% ($p-value \leq 0.05$) estão em destaque.

Vale destacar que os métodos que melhor performaram, *KNN, DistKNN* e *Floresta de Árvores* não rejeitaram a hipótese nula entre si. Uma análise das diferenças estatísticas entre os classificadores será realizada na seção *4.1 - Análise Geral dos Resultados.*

ZeroR	0.91462	**0.00409**	**0.00000**	**0.00000**	**0.00000**	**0.00000**	**0.00000**	**0.00000**	**0.00000**	**0.00000**
0.82013	Aleat	**0.02447**	**0.00000**	**0.00000**	**0.00000**	**0.00000**	**0.00000**	**0.00000**	**0.00000**	**0.00000**
0.00879	**0.01976**	Estrat	**0.00000**	**0.00000**	**0.00000**	**0.00000**	**0.00000**	**0.00000**	**0.00000**	**0.00000**
0.00000	**0.00000**	**0.00000**	OneR P	**0.00000**	**0.00000**	**0.00000**	**0.00000**	**0.00000**	**0.00000**	**0.00000**
0.00000	**0.00000**	**0.00000**	**0.00000**	Naive	**0.00000**	**0.00000**	**0.00000**	**0.00000**	**0.00000**	**0.00000**
0.00000	**0.00000**	**0.00000**	**0.00000**	**0.00000**	Kmeans	0.05900	**0.00000**	**0.00000**	**0.00000**	**0.00000**
0.00000	**0.00000**	**0.00000**	**0.00000**	**0.00000**	**0.03373**	KGA	**0.00000**	**0.00000**	**0.00000**	**0.00000**
0.00000	**0.00000**	**0.00000**	**0.00000**	**0.00000**	**0.00000**	**0.00000**	KNN	0.40420	**0.00000**	0.79806
0.00000	**0.00000**	**0.00000**	**0.00000**	**0.00000**	**0.00000**	**0.00000**	0.72351	DistK	**0.00000**	0.91286
0.00000	**0.00000**	**0.00000**	**0.00000**	**0.00000**	**0.00000**	**0.00000**	**0.00000**	**0.00000**	Arvore	**0.00000**
0.00000	**0.00000**	**0.00000**	**0.00000**	**0.00000**	**0.00002**	**0.00003**	0.94604	0.96083	**0.00000**	Florest

Tabela 5: Digits - p-values dos Testes Pareados. Teste t de Student na matriz triangular superior e Teste de Wilcoxon na matriz triangular inferior. Valores da tabela que rejeitaram a hipótese nula para um nível de significância de 95% estão escritos em negrito. Valores arredondados para cinco casas decimais.

3.3. Wine

A tabela 6 apresenta as informações estatísticas de cada classificador na base de dados *Wine*. A figura 6 mostra o *boxplot* de acurácias dos *folds* para cada método na base de dados *Wine*.

Novamente, o *Aleatorio Estratificado* juntamente dos outros *DummyClassifiers* obtiveram os piores resultados. Entre os Algoritmos implementados para o estudo, o *KGACentroides* foi o que obteve a melhor média, com 97% de acurácia.

Os classificadores que obtiveram as melhores médias foram, com resultados muito próximos, *Naive Bayes*, *KmeansCentroides, KGACentroides, KNN, DistKNN* e *Floresta de Árvores*, sendo a ***Floresta de Árvores*** o método que obteve os melhores resultados nesta análise inicial, com acurácia média de 98,3%, menor desvio-padrão entre os algoritmos sérios, sem apresentar *outliers*, e alcançando interessantes 99,2% no teto da Acurácia.

	Média	Desvio Padrão	Lim. Inferior	Lim. Superior
ZeroR	0.399	0.025	0.391	0.408
Aleatorio	0.320	0.030	0.309	0.331
Estratificado	0.281	0.007	0.279	0.283
OneR Probab	0.685	0.079	0.657	0.713
Naive Bayes	0.973	0.048	0.956	0.991
KMeans	0.966	0.050	0.948	0.984
KGA	0.970	0.043	0.954	0.985
KNN	0.951	0.054	0.932	0.971
DistKNN	0.949	0.053	0.930	0.968
Árvore Desc	0.901	0.067	0.877	0.925
Floresta	0.983	0.026	0.974	0.992

Tabela 6: Wine - Informações estatísticas do desempenho de cada classificador

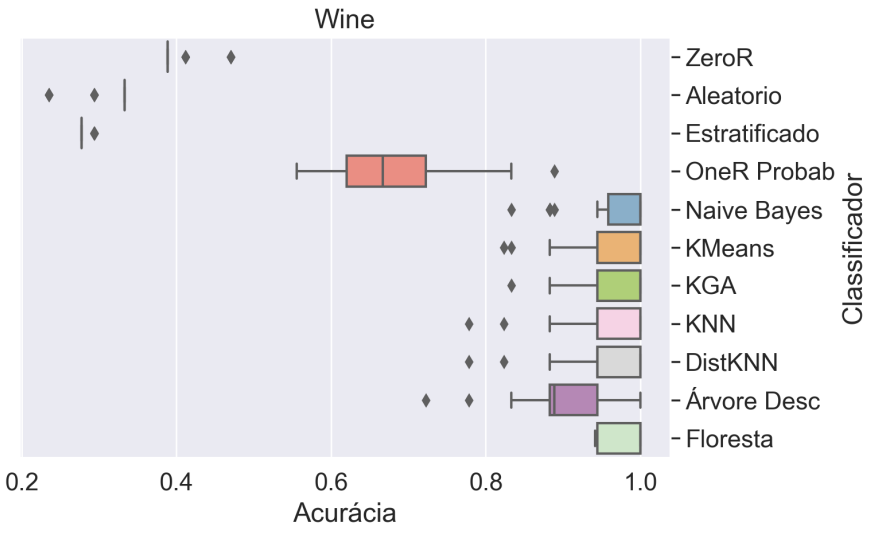


Figura 6: Boxplot do desempenho em cada *fold* dos diferentes métodos no *dataset* Wine.

A tabela 7 mostra os resultados (p-values) do Teste t de *Student* na matriz triangular superior, e os resultados do teste de *Wilcoxon* na matriz triangular inferior. Os valores que apresentam fortes evidências para a rejeição da hipótese nula para um nível de significância de 95% ($p-value \leq 0.05$) estão em destaque.

Os métodos que obtiveram as melhores médias de acurácia, *Naive Bayles Gaussiano e Floresta de Árvores*, não rejeitaram a hipótese nula entre si, com *p-value* de 0,161 no Teste t de *Student* e 0,088 no Teste de *Wilcoxon*. Uma análise das diferenças estatísticas entre os classificadores será realizada na seção *4.1 - Análise Geral dos Resultados*.

ZeroR	**0.00000**	**0.00000**	**0.00000**	**0.00000**	**0.00000**	**0.00000**	**0.00000**	**0.00000**	**0.00000**	**0.00000**
0.00000	Aleat	**0.00000**	**0.00000**	**0.00000**	**0.00000**	**0.00000**	**0.00000**	**0.00000**	**0.00000**	**0.00000**
0.00000	**0.00349**	Estrat	**0.00000**	**0.00000**	**0.00000**	**0.00000**	**0.00000**	**0.00000**	**0.00000**	**0.00000**
0.00000	**0.00000**	**0.00000**	OneR	**0.00000**	**0.00000**	**0.00000**	**0.00000**	**0.00000**	**0.00000**	**0.00000**
0.00000	**0.00000**	**0.00000**	**0.00000**	Naive	0.15773	0.50504	**0.00317**	**0.00167**	**0.00000**	0.16114
0.00000	**0.00000**	**0.00000**	**0.00000**	0.13167	Kmeans	0.32558	**0.03367**	**0.01978**	**0.00001**	**0.03693**
0.00000	**0.00000**	**0.00000**	**0.00000**	0.76302	0.31731	KGA	**0.00574**	**0.00302**	**0.00000**	**0.03070**
0.00000	**0.00000**	**0.00000**	**0.00000**	**0.01569**	0.17244	**0.02577**	KNN	0.32558	**0.00002**	**0.00172**
0.00000	**0.00000**	**0.00000**	**0.00000**	**0.01029**	0.12326	**0.01697**	0.31731	DistK	**0.00006**	**0.00098**
0.00000	**0.00000**	**0.00000**	**0.00000**	**0.00003**	**0.00026**	**0.00008**	**0.00042**	**0.00068**	Árvore	**0.00000**
0.00000	**0.00000**	**0.00000**	**0.00000**	0.08808	**0.02779**	**0.02720**	**0.00278**	**0.00179**	**0.00002**	Florest

Tabela 7: Wine - p-values dos Testes Pareados. Teste t de Student na matriz triangular superior e Teste de Wilcoxon na matriz triangular inferior. Valores da tabela que rejeitaram a hipótese nula para um nível de significância de 95% estão escritos em negrito. Valores arredondados para cinco casas decimais.

3.4. Breast Cancer

A tabela 8 apresenta as informações estatísticas de cada classificador na base de dados *Breast Cancer*. A figura 7 mostra o *boxplot* de acurácias dos *folds* para cada método na base de dados *Breast Cancer*.

Mais uma vez, o *Aleatorio* juntamente dos outros *DummyClassifiers* obtiveram os piores resultados. Entre os Algoritmos implementados para o estudo, novamente o *KGACentroides* foi o que obteve a melhor média, com 95,6% de taxa de acerto.

Os classificadores que obtiveram as melhores médias foram *KmeansCentroides, KGACentroides, KNN* e *Floresta de Árvores*. Desses, ***KNN*** **e** ***DistKNN*** foram os métodos que obtiveram os melhores resultados nesta análise inicial, ambos com resultados idênticos: acurácia média de 96,4%, desvio-padrão de *0,024*, e não apresentando *outliers*.

	Média	Desvio Padrão	Lim. Inferior	Lim. Superior
ZeroR	0.627	0.007	0.625	0.630
Aleatorio	0.468	0.062	0.446	0.490
Estratificado	0.529	0.057	0.509	0.549
OneR Probab	0.834	0.038	0.821	0.848
Naive Bayes	0.934	0.026	0.924	0.943
KMeans	0.954	0.028	0.944	0.964
KGA	0.956	0.029	0.946	0.967
KNN	0.964	0.024	0.956	0.973
DistKNN	0.964	0.024	0.956	0.973
Árvore Desc	0.924	0.033	0.913	0.936
Floresta	0.957	0.024	0.949	0.966

Tabela 8: Breast Cancer - Informações estatísticas do desempenho de cada classificador

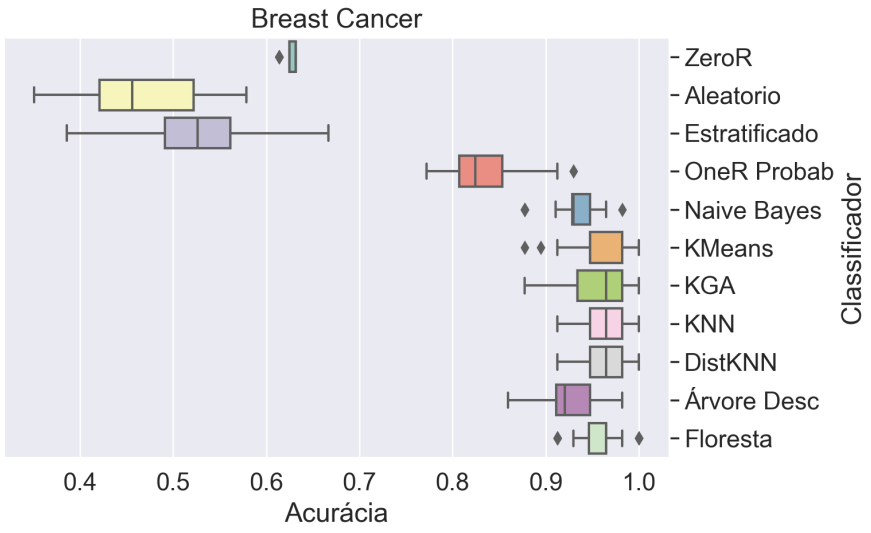


Figura 7: Boxplot do desempenho em cada *fold* dos diferentes métodos no *dataset* Breast Cancer.

A tabela 9 mostra os resultados (p-values) do Teste t de *Student* na matriz triangular superior, e os resultados do teste de *Wilcoxon* na matriz triangular inferior. Os valores que apresentam fortes evidências para a rejeição da hipótese nula para um nível de significância de 95% ($p-value \leq 0.05$) estão em destaque.

Vale citar que os métodos *ZeroR, Aleatório, Aleatório Estratificado* e *OneR Probabilistico* apresentaram rejeição da hipótese nula para todos os seus pares, em contrapartida, os métodos *Naive Bayles Gaussiano, KMeansCentroides, KGACentroides, KNN, DistKNN, Árvore* e *Floresta de Árvores* não rejeitaram a hipótese nula entre si. Uma análise das diferenças estatísticas entre os classificadores será realizada na seção *4.1 - Análise Geral dos Resultados.*

ZeroR	**0.00000**	**0.00000**	**0.00000**	**0.00000**	**0.00000**	**0.00000**	**0.00000**	**0.00000**	**0.00000**	**0.00000**
0.00000	Aleat	**0.00000**	**0.00000**	**0.00000**	**0.00000**	**0.00000**	**0.00000**	**0.00000**	**0.00000**	**0.00000**
0.00000	**0.00000**	Estrat	**0.00000**	**0.00000**	**0.00000**	**0.00000**	**0.00000**	**0.00000**	**0.00000**	**0.00000**
0.00000	**0.00000**	**0.00000**	OneR P	**0.00222**	**0.00536**	**0.02170**	**0.04091**	**0.00877**	**0.00368**	**0.00143**
0.00000	**0.00000**	**0.00000**	**0.00354**	Naive	0.84538	0.87266	0.42350	0.60148	1.00000	0.74501
0.00000	**0.00000**	**0.00000**	**0.00791**	0.86843	KMeans	0.64526	0.28266	0.44793	0.78676	0.60148
0.00000	**0.00000**	**0.00000**	**0.02869**	0.97626	0.78151	KGA	0.41462	0.81302	0.83888	1.00000
0.00000	**0.00000**	**0.00000**	**0.04388**	0.47550	0.30494	0.40538	KNN	0.32558	0.38007	0.41462
0.00000	**0.00000**	**0.00000**	**0.01094**	0.59298	0.43858	1.00000	0.25684	DistK	0.60148	0.71223
0.00000	**0.00000**	**0.00000**	**0.00461**	1.00000	0.78151	0.96652	0.43858	0.59298	Árvore	0.74501
0.00000	**0.00000**	**0.00000**	**0.00242**	0.73888	0.59298	0.85105	0.47950	0.70546	0.73888	Florest

Tabela 9: Breast Cancer - p-values dos Testes Pareados. Teste Boxplot do desempenho em cada *fold* dos diferentes métodos no *dataset* de Student na matriz triangular superior e Teste de Wilcoxon na matriz triangular inferior. Valores da tabela que rejeitaram a hipótese nula para um nível de significância de 95% estão escritos em negrito. Valores arredondados para cinco casas decimais.

4. Conclusões

4.1. Análise Geral dos Resultados

O problema de seleção de método. O objetivo deste estudo é comparar os diferentes métodos propostos e escolher o que melhor performou, que no problema de classificação significa obter o melhor desempenho ao classificar dados desconhecidos levando em conta informações estatísticas como média e desvio-padrão das acurácias.

Por se tratar de dados estatísticos, surge o questionamento: *A diferença entre dois métodos é real ou é devido a uma chance estatística?* É para isso que recorremos aos *Testes Pareados* para que possamos rejeitar ou não a hipótese de haver diferença apenas por um acaso estatístico. Quando não rejeitamos a Hipótese nula, assumimos a possibilidade de não haver diferença entre os grupos experimentais dos métodos.

Isso nos leva a resultados interesses em cada base de dados:.

Iris. O método *KmeansCentroides* foi o que obteve os melhores resultados na análise inicial (tabela 2), mas não é possível rejeitar a hipótese nula dele com os métodos *Naive Bayles Gaussiano, KGACentroides, KNN, DistKNN, Árvore* e *Floresta de Árvores* (tabela 3). Assim, não podemos afirmar que KMeansCentroides foi o melhor classificador, mas sim que ***KmeansCentroides, Bayles Gaussiano, KGACentroides, KNN, DistKNN, Árvore* e *Floresta de Árvores* foram os que melhor performaram na Base de Dados *Iris.***

Digits. Os métodos *KNN, DistKNN* e *Floresta de Árvore* foram os que obtiveram os melhores resultados na análise inicial (tabela 4). Nos testes pareados (tabela 5) rejeitamos as hipóteses nulas deles para todos outros pares. Portanto, podemos afirmar que ***KNN, DistKNN* e *Floresta de Árvores* foram os métodos que melhor performaram na Base de Dados *Digits.***

Wine. Embora a *Floresta de Árvores* tenha obtido a melhor média de acurácia (tabela 6), não é possível rejeitar a hipótese nula dele com o método *Naive Bayles Gaussiano* (tabela 6). Assim, podemos afirmar que **_Naive Bayles Gaussiano_ e _Floresta de Árvores_ foram os que melhor performaram na Base de Dados _Wine_.**

Breast Cancer. Os métodos *KNN* e *DistKNN* foram os que obtiveram os melhores resultados na análise inicial (tabela 8), mas não é possível rejeitar a hipótese nula deles com os métodos *Naive Bayles Gaussiano, KmeansCentroides, KGACentroides, Árvore* e *Floresta de Árvores* (tabela 9). Assim, não podemos afirmar que *KNN* e *DistKNN* foram os melhores classificadores, mas sim que **_Bayles Gaussiano, KmeansCentroides, KGACentroides, KNN, DistKNN, Árvore_ e _Floresta de Árvores_ foram os que melhor performaram na Base de Dados _Breast Cancer_.**

O melhor método. Diante dos resultados expostos para cada base de dados deste estudo, podemos concluir que **o classificador _Floresta de Árvores_ foi o único método que performou melhor em todas as Bases de Dados.**

4.2. Contribuições do Trabalho

Dos métodos avaliados neste trabalho, Todos os classificadores da etapa dois tiverem um limite inferior de 91% de acurácia em todas bases de dados, com exceção da *Árvore de Decisão* que teve um limite inferior abaixo de 90% para alguns casos. Mesmo assim, pudemos ver que todos os algoritmos da Etapa 2, que são os que possuem ajuste de hiperparâmetro, se mostram opções melhores que os métodos da Etapa 1, com exceção do *Naive Bayes Gaussiano* que também performou bem para alguns casos.

Dos métodos implementados, o *KmeansCentroides* e *KGACentroides* performaram muito bem quando comparado com métodos já consolidados da biblioteca *Scikit-learn*, sendo considerados uns dos melhores métodos para algumas bases de dados.

4.3. Melhorias e Trabalhos Futuros

Um melhoria interessante seria buscar outros testes estatísticos como uma alternativa para podermos diferenciar melhor o desempenho dos classificadores, pois em alguns casos como o da Base de Dados *Breast Cancer* tivemos 7 melhores métodos ao considerar a Hipótese Nula desses pares.

A implementação do KCentroides com outros métodos de agrupamento, como as meta-heurísticas *GRASP* e *Simulated Annealing* seria um possível trabalho futuro, realizando uma análise de performance dos diferentes métodos de clusterização usados no classificador KCentróides em diferentes Bases de Dados.

5. Referências

[1] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830.

[2] D. Dua, C. Graff, UCI machine learning repository (2017).
URL http://archive.ics.uci.edu/ml

[3] L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler, R. Layton, J. VanderPlas, A. Joly, B. Holt, G. Varoquaux, API design for machine learning software: experiences from the scikit-learn project, in: ECML PKDD Workshop: Languages for Data Mining and Machine Learning, 2013, pp. 108–122.

[4] J. B. MacQueen, Some methods for classification and analysis of multivariate observations, in: L. M. L. Cam, J. Neyman (Eds.), Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1, University of California Press, 1967, pp. 281–297.

[5] H.-J. Lin, F.-W. Yang, Y.-T. Kao, An efficient ga-based clustering technique 8 (06 2005).