Práticas de Aprendizado de Máquina

Gustavo Silva Resende
Departamento de Computação
Universidade Tecnológica
Federal do Paraná
Cornélio Procópio, Brasil
gustavoresende@alunos.utfpr.edu.br

Luis Gustavo de Souza
Departamento de Computação
Universidade Tecnológica
Federal do Paraná
Cornélio Procópio, Brasil
luisouza98@gmail.com

Omar Condori López
Departamento de Computação
Universidade Tecnológica
Federal do Paraná
Cornélio Procópio, Brasil
omarlopez@alunos.utfpr.edu.br

***Resumo*—Neste artigo é apresentado o desenvolvimento de práticas relacionadas aos aprendizados de máquina, envolvendo os seguintes aprendizados: supervisionado, não supervisionado, semi-supervisionado e ativo. Contendo os métodos e discussões destas técnicas sobre diferentes *datasets*.**

***Index Terms*—aprendizado, máquina, supervisionado, não-supervisionado, semi-supervisionado, ativo.**

I. Introdução

Todos os dias, em todos lugares são geradas informações, que podem ser convertidas em dados. A medida que o tempo passa, aumenta a necessidade de manipular estes dados de maneira eficiente. É por este motivo que abordagens como o aprendizado de maquina surgiram, o qual visa "ensinar"o computador a encontrar padrões e classificar de maneira rápida e eficiente. Deste modo, otimizando processos em diversos domínios humanos.

Este artigo tem como objetivo testar o aprendizado de máquina supervisionado, não supervisionado, semi supervisionado e ativo, utilizando *datasets* de imagens públicos, que possuem características distintas, de modo a evidenciar o comportamento das técnicas sobre diferentes domínios.

II. Fundamentação Teórica

Aprendizado de Máquina, ou *Machine Learning*, é uma área da Inteligência Artificial que busca desenvolver algoritmos capazes de tomar decisões com base no conhecimento adquirido de forma automática. Dentre as principais forma de aprendizado estão:

- **Aprendizado Supervisionado:** O agente aprende por meio de um conjunto de dados já classificado e rotulado, desta forma, o algoritmo adquire conhecimento associando as características de uma amostra e a sua classe. Com isso, dada uma entada de dados de teste não rotulada, o algoritmo gera uma saída tentando prever a classe a qual cada amostra pertence.
- **Aprendizado Não supervisionado:** As amostras que são usadas nesse tipo de aprendizado não possuem nenhum rótulo. Desta forma, é por meio de *clusters*, pelas quais as amostras serão agrupadas de acordo com a semelhança de suas características, permitindo realizar análises sobre o conjunto de dados. [1]
- **Aprendizado Semi-Supervisionado:** Funciona de forma semelhante ao aprendizado supervisionado, no entanto, apenas algumas amostras do conjunto de treinamento possuem rótulos de classificação. As amostras não rotuladas formam grupos de acordo com as suas características, com isso é possível associar tais grupos às amostras rotuladas a fim de obter-se uma maior acurácia. [2]
- **Aprendizado Ativo:** Nesta forma de aprendizado, o conjunto de dados inicialmente não possui rótulos, no entanto, ao decorrer da execução do algoritmo, as amostras mais significativas são escolhidas para serem rotuladas, fazendo com que a cada execução sejam inseridas mais amostras rotuladas. Desta forma, existe uma diminuição no esforço humano gasto para a rotulação das amostras. [3]

Em 2017 foi publicado um estudo desenvolvido por [4] em que foi proposto um modelo de classificação de imagens baseado em redes neurais convolucionais e aprendizado ativo. Neste modelo, as amostras mais significativas são selecionadas para a rotulação ativa, e a partir disto, outras amostras são selecionadas e recebem pseudo-rótulos automaticamente. Com isso, obtém-se um maior número de amostras rotuladas, aumentando o desempenho geral da rede neural convolucional sem gerar nenhum esforço humano adicional.

Com o objetivo de realizar experimentos sobre esse modelo de classificação, foram utilizados dois conjuntos de dados, sendo eles:

- **CACD:** Contém 160000 imagens de 2000 celebridades.
- **Caltech-256:** Contém 30607 imagens de 256 categorias de objetos.

Analisando o conjunto de dados CACD, para atingir 91.5% de acurácia, o modelo proposto reduziu em aproximadamente 36% e 18% as anotações realizadas por usuário, quando comparado aos métodos AL_RAND e TCAL, respectivamente.

Para obter-se 73.8% de acurácia no conjunto de dados Caltech-256, o modelo proposto reduziu as anotações realizadas pelo usuário em aproximadamente 19% e 15% quando comparado ás técnicas citadas anteriormente.

III. Metodologia Proposta

Como proposta para selecionar as amostras mais informativas de um *dataset*, pensou-se em construir, primeiramente, um agrupamento sobre os dados, utilizando determinada técnica de aprendizado não supervisionado. Com o agrupamento, as

amostras raízes são extraídas sobre cada *cluster* e servidas como entrada para um classificador de modo a treiná-lo. A partir da distribuição espacial das amostras, é possível construir um vetor que conterá *K* (hiper-parâmetro) amostras de cada *cluster*, localizado nas regiões mais distantes do centróide.

Com a primeira instância do classificador, torna-se possível quantificar as probabilidades de uma amostra ser de determinadas classes (atributo *predict_proba* dos classificadores da biblioteca *scikitlearn* [5]). Com essa informação, é possível verificar o grau de incerteza da amostra, o que infere, por sua vez, o seu grau de informação.

Através de outro hiper-parâmetro α, será determinado se uma amostra deve ser descartada ou mantida. O valor deste hiper-parâmetro representa o limite superior e inferior das probabilidades, na qual se a probabilidade de uma amostra ultrapassar os valores limites, ela é descartada, caso contrário, ela é mantida como uma amostra informativa.

Por uma questão de abstração, faz-se o seguinte exemplo demonstrando o funcionamento da estratégia: Dado um conjunto de dados com duas classes, primeiramente, será realizado o agrupamento sobre as amostras, assim, treinará o classificador *Decision Tree*, pelas amostras raízes. Após isso, é obtido as K=10 amostras mais distantes de cada *cluster*. Por conseguinte, será gerado as probabilidades da amostra pertencer da classe 1 ou classe 2. Em uma análise individual, percebeu-se que uma amostra tem 52% de ser da classe 1 e, consequentemente, 48% da classe 2. Desta forma, com um hiper-parâmetro α=5%, o programa irá verificar se a amostra está dentro do limite, ou seja, entre 45% e 55%. Assim, o código determinará que esta amostra tem alta informatividade.

IV. Avaliação Experimental

Para a realização das práticas, primeiramente, escolheu-se 5 conjunto de dados (Tabela I): Alien vs Predator, Malaria Cell, Fruits, Iris e Rock-Paper-Scissor.

Tabela I
Conjunto de Dados Escolhidos com suas informações.

Dataset	Amostras	Nº Classes	Balanceado
Alien vs Predator	694	2	Sim
Malaria Cell	27560	2	Sim
Fruits	1140	19	Não
Iris	150	3	Sim
Rock-Paper-Scissor	2188	3	Sim

Como observado na tabela I, preocupou-se em escolher *datasets* que continham diferentes características em relação ao: número de classes (problemas de classificação binária e multi-classe), número de amostras e balanceamento/desbalanceamento de classes. Essas diferenças entre os conjunto de dados permite aplicar e entender o comportamento das múltiplas técnicas de Aprendizado de Máquina em diferentes domínios de problemas.

Após a escolha dos *datasets*, selecionou-se os descritores, capazes de extrair informações das imagens e convertê-los em vetores de características, que posteriormente serão utilizadas como entrada para os classificadores. Para isso, utilizou-se 8 descritores distintos, evidenciados na Tabela II.

Tabela II
Características dos Descritores Selecionados.

Descritor	Tipo de Informação	Nº Caract.
Reference Color Similarity	Cor	77
GCH	Cor	255
CEDD	Cor	144
MPOC	Textura	18
LBP (raio 1)	Textura	256
Gabor	Textura	60
FCTH	Cor e Textura	192
JCD	Cor e Textura	336

Nota-se na Tabela II, que atentou-se na escolha de descritores distintos em relação ao seu tipo de informação extraída e número de características geradas. Isto foi feito, a fim de explorar a diversidade de informações produzidas, de modo a encontrar qual melhor desempenha-se nos conjuntos de dados selecionados.

Além dos descritores, uma técnica essencial para a maioria dos classificadores trata-se da normalização. A normalização consiste no processo de alterar os valores das características numéricas a fim de obter uma escala comum [6]. Esta técnica, apesar de sua simplicidade, há grande impacto na performance de classificadores com métodos baseados em distância. Desta forma, escolheu-se 4 exemplos de normalização, presentes na Tabela III.

Tabela III
Tipos de Normalização Considerados. (μ: média; σ: desvio padrão; Q: quartil)

Normalização	Equação
MinMax	$X_{norm} = \frac{X - min(X)}{max(X) - min(X)}$
Standard	$X_{norm} = \frac{X - \mu}{\sigma}$
MaxAbs	$X_{norm} = \frac{X}{max(\lvert X \rvert)}$
Robust	$X_{norm} = \frac{X - Q_1(X)}{Q_3(X) - Q_1(X)}$

A. Aprendizado Supervisionado

Com os *datasets*, descritores e normalizadores definidos, realizou-se um primeiro experimento, que consiste na elaboração completa das combinações (*Grid*) das técnicas (descritores e normalizadores) sobre cada conjunto de dados. Cada combinação gerada é treinada por 8 classificadores distintos (*Naive Bayes, Logistic Regression, Decision Tree, K-NN, LDA, SVM, Random Forest, MLP*), e assim realizado a média das acurácias dos classificadores sobre o *dataset* (conforme evidenciado nas Tabelas da Apêndice A). Com isso, é razoável presumir que a média das acurácias indica potenciais combinações de técnicas.

Nota-se que a métrica tempo é uma medida interessante de ser analisada, em algumas combinações apesar de sua alta acurácia média, tem um tempo de treinamento elevado, o

que pode ser um gargalo em determinadas situações. Nesta análise, como os *datasets* são relativamente pequenos, o tempo enquadra-se na unidade dos segundos, tornando assim, não tão relevante para o estudo.

Deste modo, com o *Grid* e a respectiva média de acurácia concluídas, selecionou-se as 3 combinações com a maior acurácia média para cada *dataset*, na qual os classificadores são re-treinados sobre tais combinações, de modo a analisar cada classificador individualmente. Por conseguinte, através das métricas de acurácia e F1 Score − que combina as métricas *precision* e *recall* − foi analisado e definido o melhor classificador e cenário para cada conjunto de dados.

B. Aprendizado Não-Supervisionado

Com as análises de aprendizado supervisionado concluídas, a fim de entender melhor a constituição dos *datasets* e o agrupamento dos dados, aplicou-se técnicas de aprendizado não-supervisionado, de modo a observar os *clusters* gerados.

Para o agrupamento, considerou-se um cenário, na qual não é conhecido o número de classes de cada *dataset*. Desta forma, utilizou-se a técnica *KMeans*, na qual o número de *clusters* foi determinado sobre um cenário aleatório considerando o método *Elbow* (Imagens dos métodos Elbow disponível na Apêndice C), que auxilia no processo de escolha do número de agrupamentos apropriado através da análise da porcentagem da variância explicada.

Porém, dependendo do cenário utilizado, o agrupamento pode não ser tão eficiente, prejudicando a análise sobre os dados. Desta forma, com o propósito semelhante ao aprendizado supervisionado, faz-se necessário encontrar um extrator e normalizador adequado para os *datasets*. Para isso, através de um *Grid* das combinações e da métrica inércia (soma dos quadrados das distâncias das amostras até o *cluster* mais próximo) do *KMeans*, determinou-se o melhor cenário para cada *dataset*. Após isso, obteve-se o agrupamento, e para alguns *datasets*, exibiu-se a amostra raiz de cada *cluster*, ou seja, a amostra que mais se aproxima do centróide.

Uma observação sobre este processo, trata-se da utilização das PCAs (Principal Component Analysis), que para a prática presente, teve o intuito de diminuir a dimensão e encontrar características mais significantes dos dados, de modo a visualizar grande parte da informação em apenas duas dimensões. Para a agrupação dos dados, o número de PCAs considerado foi 10% do número de características de determinado extrator, caso esse valor seja menor que dois, considerou-se 2 PCAs. Para a geração do gráfico, de modo a situar-se em duas dimensões, utilizou-se apenas a primeira e segunda componentes principais.

C. Aprendizado Semi-Supervisionado

Com o objetivo de analisar melhor o comportamento das técnicas de aprendizado semi-supervisionado sobre os conjuntos de dados descritos − utilizando os melhores cenários e classificadores, que foram determinados na subseção de Aprendizado Supervisionado − escolheu-se quatro taxas de amostras rotuladas, sendo elas 20%, 30%, 40% e 50%. Sendo que, as demais amostras foram rotuladas através da propagação dos rótulos disponíveis. Assim, foi realizado uma comparação sobre o aprendizado supervisionado, de modo a evidenciar a performance da técnica semi-supervisionada.

D. Aprendizado Ativo

Nesta abordagem, foi reunido os aprendizados não supervisionado e supervisionado, de modo a agrupar os dados em *clusters* distintos, na qual serão extraídas as amostras raízes de cada *cluster*, e assim, utilizadas para o treinamento da primeira instância do classificador. Por conseguinte, é recuperado uma amostra aleatória de cada *cluster* por iteração, treinando o classificador por meio delas. Após determinado número de iterações, o aprendizado será interrompido, caso o mesmo tenha estagnado ou alcançado a acurácia obtida no aprendizado supervisionado.

Para esta análise, foram utilizados os melhores cenários já determinados para cada *dataset*, de modo a tornar justo a comparação com as demais técnicas. Além disso, o aprendizado é realizado sobre os 8 classificadores já mencionados, de forma a evidenciar o comportamento de cada.

Para este aprendizado, também, foi apresentado uma ideia, discutida na seção III, que poderia ser implementada, a fim de escolher amostras mais significativas, de modo a aumentar a performance de treinamento dos classificadores.

V. Resultados

A primeira situação que deve-se analisar, trata-se da determinação do melhor cenário (descritor e normalizador) para cada conjunto de dados. Desta forma, através da heurística apresentada na seção anterior, obteve-se as tabelas apresentadas na Apêndice A.

A partir das tabelas referenciadas, foram selecionados os três melhores cenários, de acordo com a acurácia média, de cada *dataset* para a realização de uma análise específica de modo a determinar o melhor classificador e cenário para cada conjunto de dados.

Com isso, foi possível obter a Tabela IV, na qual é mostrado a melhor combinação de extrator, normalização e classificador para cada *dataset*, de acordo com a acurácia e F1 score. Além disso, na Tabela V e Figuras 1, 2, 3, 4, 5, é possível evidenciar as métricas para tais cenários e classificadores.

Tabela IV
Classificadores e cenários escolhidos para cada dataset.

Dataset	Extrator	Normalização	Classificador
Alien vs Predator	FCTH	Nenhuma	MLP
Malaria Cell	GCH	Robust	RF
Fruits	JCD	Nenhuma	KNN
Iris	GCH	Standard	MLP
Rock-Paper-Scissor	LBP	Standard	MLP

Como pode ser evidenciado, apesar do tempo não ser um obstáculo para os *datasets* utilizados, percebe-se um tempo de treinamento relativamente maior nos que utilizam o classificador MLP em relação a média de todos os classificadores para o mesmo cenário.

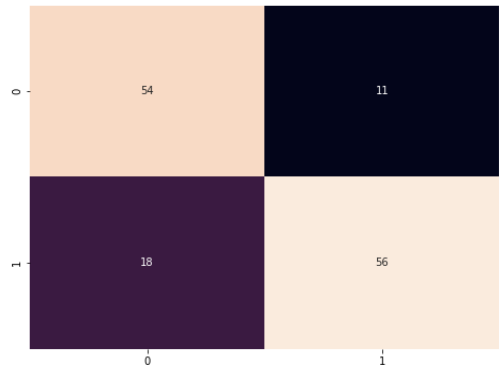

Figura 1. Matriz de Confusão sobre o *Dataset* Alien vs Predator.

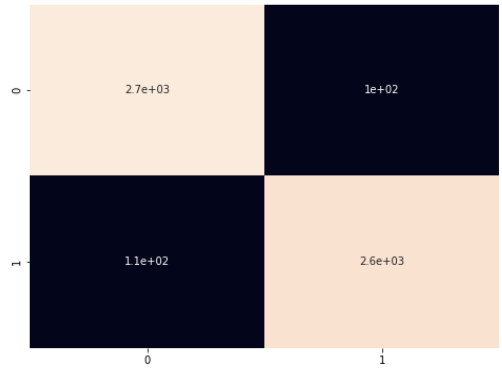

Figura 2. Matriz de Confusão sobre o *Dataset* Malaria Cell.

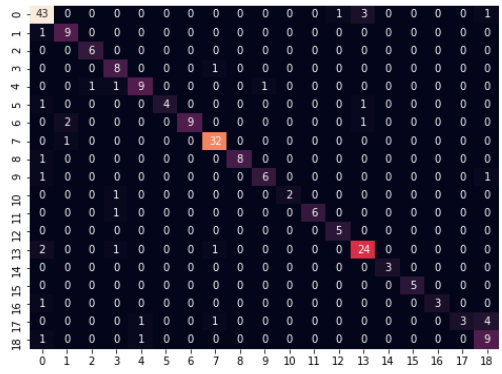

Figura 3. Matriz de Confusão sobre o *Dataset* Fruits.

Tabela V
MÉTRICAS DOS MELHORES CLASSIFICADORES DE CADA DATASET.

Dataset	Acurácia	F1	Precisão	Recall	Tempo(s)
Alien vs Predator	0.791	0.791	0.792	0.793	0.486
Malaria Cell	0.960	0.960	0.961	0.960	0.509
Fruits	0.850	0.843	0.893	0.828	0.002
Iris	0.966	0.961	0.952	0.974	0.152
Rock-Paper-Scissor	0.956	0.956	0.956	0.957	1.977

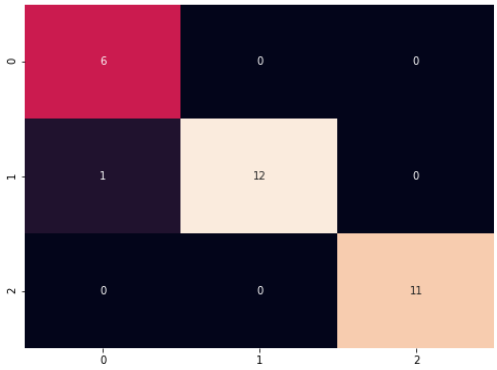

Figura 4. Matriz de Confusão sobre o *Dataset* Iris.

Através das Matrizes de Confusão, é possível evidenciar o comportamento dos falsos positivos e negativos para cada classe. Desta forma, é notável que para determinados *datasets* algumas classes possuem taxas de falsos positivos relativamente superiores as demais, indicando um problema individual na classe que deve ser explorado, como nos casos:

- *Dataset* Alien vs Predator: A classe 0 - *Alien* possui 25% de falsos positivos enquanto a classe 1 - *Predator* possui 16.4%. Supostamente, um motivo para esta diferença, remete-se a premissa que as imagens da classe *Alien*

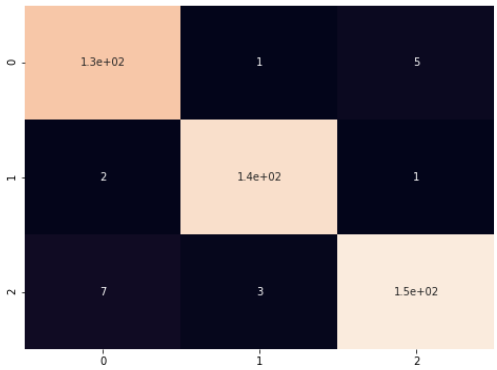

Figura 5. Matriz de Confusão sobre o *Dataset* Rock-Paper-Scissor.

vem de diversas distribuições distintas (conforme Figura 6), mesmo possuindo apenas 347 exemplos, consequentemente, não possuirá um padrão tão consistente quanto aquela que possui uma distribuição mais uniforme, como aparenta a classe *Predator*, desta forma, sendo mais difícil para o classificar identificá-lo.

- *Dataset* Fruits: A classe 18 - *Satsuma* (variedade da Tangerina) tem 40% de falsos positivos, sendo 26% das vezes classificada como a classe 17 - *Red-grapefruit* (Toranja). A razão disto pode ser atribuída a complexidade da tarefa em separar essas duas classes, que pode ser desafiadora até para humanos, como evidenciado na Figura 7.
- *Dataset* Iris: Apesar da classe Setosa apresentar 14.3% de falsos negativos, e as outras duas classes 0%. Este número não é tão informativo, pois foi necessário apenas uma classificação errada para apresentar esse valor, devido a pouca quantidade de exemplos no conjunto de teste.

Figura 6. Exemplos da classe *Alien* de distribuições distintas.

Figura 7. *Red-grapefruit* vs *Satsuma*.

A. Agrupamento dos Dados

Na etapa de clusterização, através dos passos descritos na seção IV, encontrou-se os cenários e número de *clusters* mais adequados para o agrupamento (Tabelas de combinações disponível na Apêndica B), que pode ser evidenciado pela Tabela VI, que contém informações sobre o número de *clusters*, descritor, normalização, quantidade de componentes principais utilizadas.

Tabela VI
NÚMERO DE CLUSTERS E CENÁRIO ESCOLHIDOS NA CLUSTERIZAÇÃO.

Dataset	Clusters	Descritor	Norm.	PCAs
Alien vs Predator	4	MPOC	MinMax	2
Malaria Cell	4	C. Similarity	Nenhuma	7
Fruits	3	MPOC	MaxAbs	2
Iris	3	C. Similarity	Nenhuma	7
Rock-Paper-Scissor	5	C. Similarity	Nenhuma	7

Analisando a Tabela VI, é possível perceber, que de acordo com as premissas utilizadas, os extratores *MPOC* e *Reference Color Similarity* tem uma boa relação com a clusterização, já que foram os únicos descritores selecionados.

Os agrupamentos dos dados realizados pela técnica *KMeans* com o seu respectivo cenário, que foi descrito na tabela citada, são destacados nas Figuras 8, 9, 10, 11, 12.

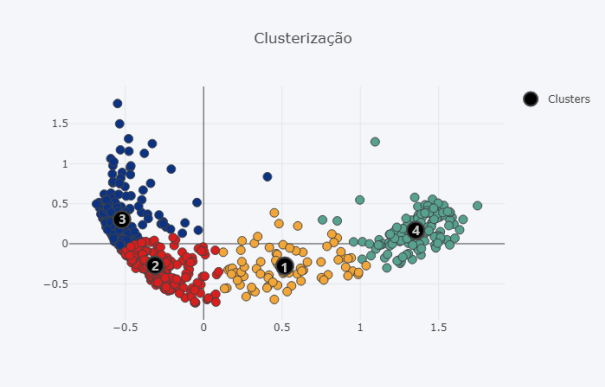

Figura 8. Clusterização do *dataset* Alien vs Predator.

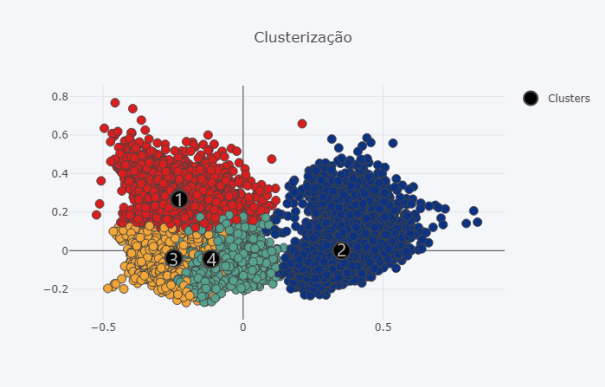

Figura 9. Clusterização do *dataset* Malaria Cell.

Nota-se que para determinados *datasets* não há uma boa clusterização, como no caso do *Rock-Paper-Scissor*. Um motivo para este fenômeno advém da natureza do dataset – que

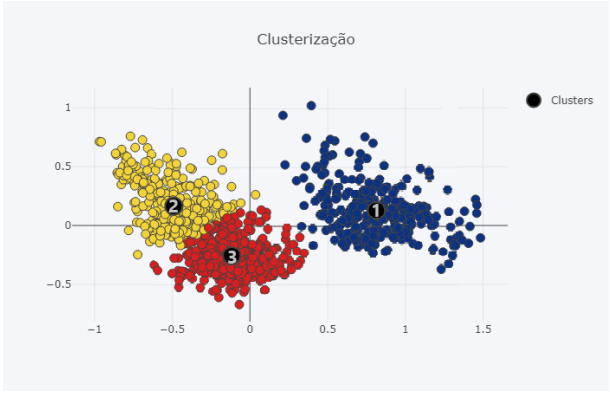

Figura 10. Clusterização do *dataset* Fruits.

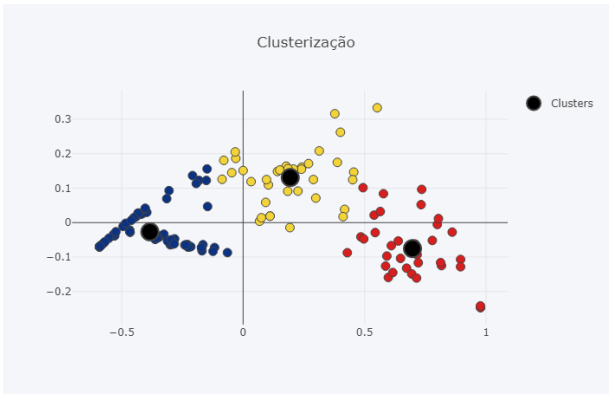

Figura 11. Clusterização do *dataset* Iris.

tem características baseadas, principalmente, em formas – e dos descritores utilizados, que extraem apenas informações de cor e/ou textura.

Outra análise que pode ser efetuada sobre os *datasets*, a fim de compreender melhor a estrutura dos dados, refere-se a investigação das amostras raízes dos *clusters*, na qual é possível induzir o tipo de informação aglomerado em cada *cluster*, como demonstrado nas Figuras 13, 14, 15. Observa-se que não foram mostradas as amostras dos *datasets* Iris, que possui imagens ininteligíveis, e *Rock-Paper-Scissor* que possui informação pouco relevante, como citado previamente.

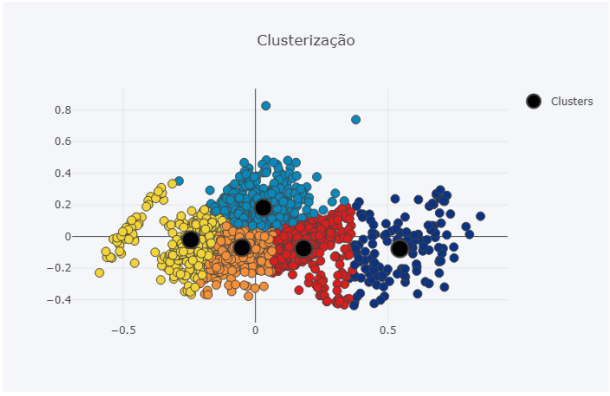

Figura 12. Clusterização do *dataset* Rock-Paper-Scissor.

Figura 13. Amostras raízes dos clusters do *dataset* Alien vs Predator.

Figura 14. Amostras raízes dos clusters do *dataset* Fruits.

Nas imagens das clusterizações e amostras mais representativas é possível observar que existem números ordinais, que tem o objetivo de identificar quais amostras representam quais *clusters*.

B. Aprendizado Semi-Supervisionado

Aplicando as técnicas de Aprendizado Semi-Supervisionado descritas na seção IV, obteve-se as acurácias mostradas na

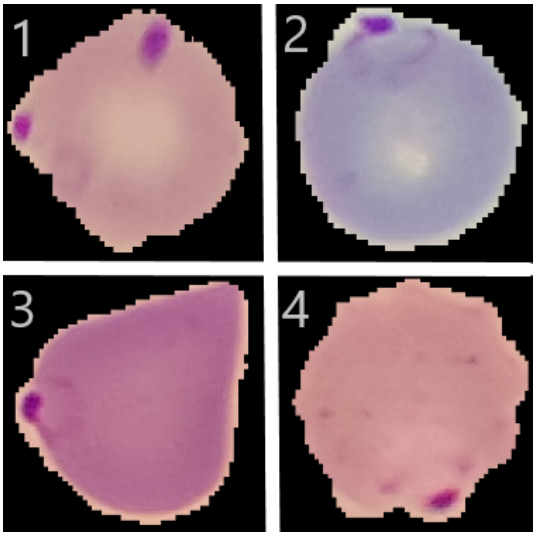

Figura 15. Amostras raízes dos clusters do *dataset* Malaria Cell.

Tabela VII. Com isso, é possível perceber que a acurácia está diretamente ligada a porcentagem de amostras rotuladas, como já era esperado, sendo que quanto mais amostras rotuladas, consequentemente, maior é a acurácia obtida.

Tabela VII
ACURÁCIAS DOS CLASSIFICADORES EM FUNÇÃO DA PORCENTAGEM DE AMOSTRAS ROTULADAS.

Dataset	**Acurácias**				
	20%	30%	40%	50%	100%
Alien vs Predator	0.474	0.496	0.539	0.575	0.791
Malaria Cell	0.513	0.533	0.577	0.666	0.960
Fruits	0.263	0.302	0.407	0.495	0.850
Iris	0.200	0.366	0.466	0.533	0.966
Rock-Paper-Scissor	0.385	0.413	0.504	0.648	0.956

Outra análise interessante, remete-se a comparação da taxa de 50% com a de 100%, na qual é possível perceber uma diferença considerável, mostrando que para este tipo de aprendizado, é de extrema importância conter grandes quantidades de amostras.

C. Aprendizado Ativo

Na análise de aprendizado ativo, de acordo com as etapas descritas pela seção IV, foi possível obter os gráficos (Figuras 16, 17, 18, 19, 20) de iterações $\times$ acurácia de cada *dataset*, utilizando os 8 classificadores já mencionados.

Uma observação sobre os agrupamentos que foram utilizados, trata-se da escolha do número de *clusters*, necessário para o método *KMeans*. Para isso, foram utilizados 3 vezes o número de classes existentes, de forma a conter diferentes distribuições de dados, com exceção do *dataset* Fruits, que tem muitas classes e poucas amostras.

Como pode-se notar na Tabela VIII, as acurácias obtidas pelo melhor classificador, são próximas ou idênticas as obtidas pelo aprendizado supervisionado, como mostrado na Tabela V, mesmo utilizando um número de amostras bem inferior. Para

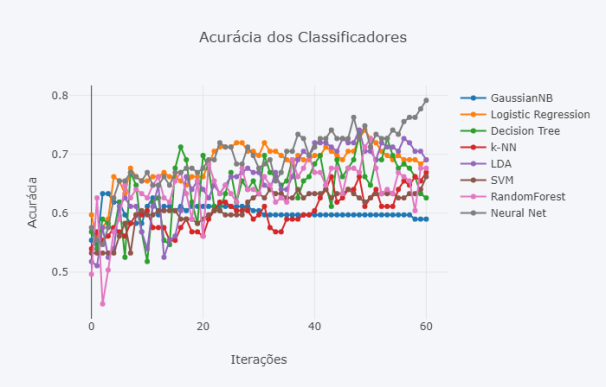

Figura 16. Acurácia dos classificadores utilizando a técnica de aprendizado ativo sobre o *dataset* Alien vs Predator.

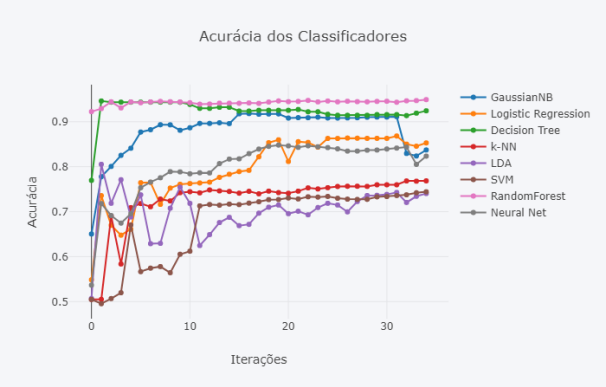

Figura 17. Acurácia dos classificadores utilizando a técnica de aprendizado ativo sobre o *dataset* Malaria Cell.

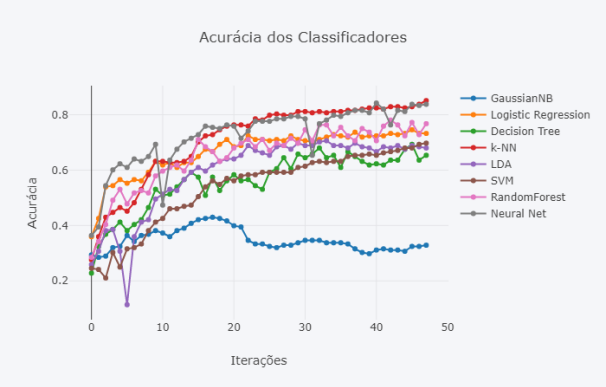

Figura 18. Acurácia dos classificadores utilizando a técnica de aprendizado ativo sobre o *dataset* Fruits.

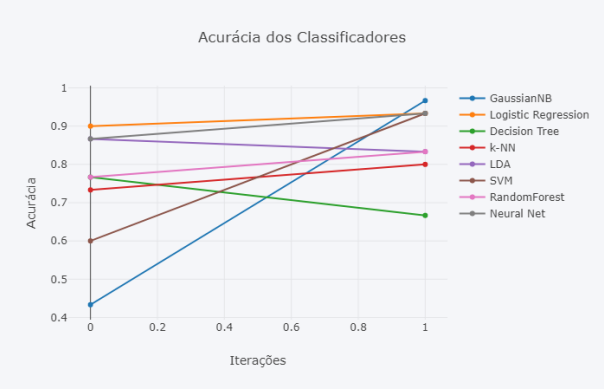

Figura 19. Acurácia dos classificadores utilizando a técnica de aprendizado ativo sobre o *dataset* Iris.

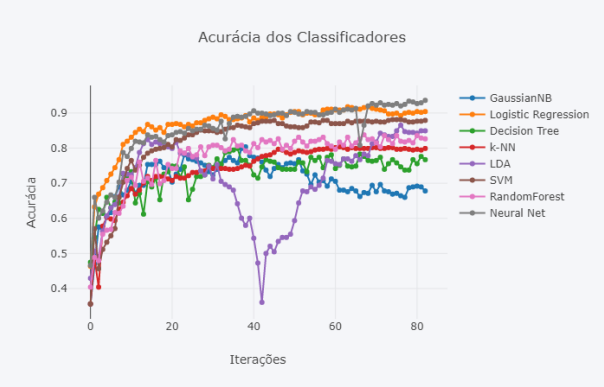

Figura 20. Acurácia dos classificadores utilizando a técnica de aprendizado ativo sobre o *dataset* Rock-Paper-Scissor.

o conjunto Malaria Cell, esta diferença torna-se ainda maior, pois utilizando apenas 0.7% dos dados rotulados, foi possível obter uma acurácia de 0.951, contra 0.960 utilizando todos os dados.

Se comparado com a Tabela VII pertencente ao Aprendizado Semi-Supervisionado, é notável a eficiência do Aprendizado Ativo sobre esta, pois para a maioria dos *datasets*, mesmo utilizando um número de amostras consideravelmente inferior, obtém uma acurácia demasiadamente superior.

Tabela VIII
CENÁRIO DO APRENDIZADO ATIVO COM A MAIOR ACURÁCIA OBTIDA.

Dataset	Clusters	Amostras Usadas	Acurácia
Alien vs Predator	6	51.8%	0.791
Malaria Cell	6	0.7%	0.951
Fruits	19	25.2%	0.850
Iris	6	4%	0.966
Rock-Paper-Scissor	6	22.4%	0.936

VI. CONSIDERAÇÕES FINAIS

Com base neste estudo, foi possível analisar e entender as diferentes técnicas de aprendizado de máquina, bem como compará-los, de forma a entender os seus diferentes comportamentos frente aos mais diversos cenários e *datasets*.

Além disso, para este estudo, apresentou-se a ideia de uma solução viável para recolher as amostras mais significativas, de modo a aumentar a performance do aprendizado ativo. Sendo que, para trabalhos futuros, pretende-se implementar tal solução e analisar as suas consequências.

REFERÊNCIAS

[1] S. Russell and P. Norvig, "Artificial intelligence: A modern approach. third edit," 2010.

[2] X. Zhu and A. B. Goldberg, "Introduction to semi-supervised learning," *Synthesis lectures on artificial intelligence and machine learning*, vol. 3, no. 1, pp. 1–130, 2009.

[3] S. Tong and E. Chang, "Support vector machine active learning for image retrieval," in *Proceedings of the ninth ACM international conference on Multimedia*. ACM, 2001, pp. 107–118.

[4] K. Wang, D. Zhang, Y. Li, R. Zhang, and L. Lin, "Cost-effective active learning for deep image classification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 12, pp. 2591–2600, 2016.

[5] 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, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[6] S. Kotsiantis, D. Kanellopoulos, and P. Pintelas, "Data preprocessing for supervised leaning," *International Journal of Computer Science*, vol. 1, no. 2, pp. 111–117, 2006.

APÊNDICE

A. Tabelas das combinações das técnicas de descritores e normalizadores para cada dataset com base na acurácia

Tabela IX
ALIEN-VS-PREDATOR DATASET

Descritor	**Normalização**	Acurácia	Tempo (s)
CEDD	Nenhuma	0.677 ± 0.037	0.06 ± 0.13
	MinMax	0.659 ± 0.046	0.06 ± 0.13
	Standard	0.667 ± 0.058	0.07 ± 0.15
	MaxAbs	0.659 ± 0.046	0.06 ± 0.13
	Robust	0.668 ± 0.045	0.06 ± 0.14
FCTH	**Nenhuma**	**0.704 ± 0.060**	**0.07 ± 0.16**
	MinMax	0.661 ± 0.071	0.07 ± 0.15
	Standard	0.683 ± 0.063	0.08 ± 0.17
	MaxAbs	0.661 ± 0.071	0.07 ± 0.15
	Robust	0.687 ± 0.058	0.07 ± 0.16
Gabor	Nenhuma	0.562 ± 0.025	0.03 ± 0.06
	MinMax	0.576 ± 0.016	0.05 ± 0.11
	Standard	0.579 ± 0.019	0.05 ± 0.11
	MaxAbs	0.574 ± 0.016	0.05 ± 0.10
	Robust	0.577 ± 0.017	0.05 ± 0.11
GCH	Nenhuma	0.577 ± 0.053	0.04 ± 0.07
	MinMax	0.580 ± 0.034	0.06 ± 0.12
	Standard	0.602 ± 0.038	0.06 ± 0.13
	MaxAbs	0.580 ± 0.034	0.05 ± 0.11
	Robust	0.594 ± 0.034	0.06 ± 0.13
JCD	Nenhuma	0.691 ± 0.052	0.07 ± 0.15
	MinMax	0.669 ± 0.040	0.07 ± 0.15
	Standard	0.665 ± 0.052	0.07 ± 0.15
	MaxAbs	0.669 ± 0.040	0.07 ± 0.15
	Robust	0.685 ± 0.050	0.07 ± 0.17
LBP	Nenhuma	0.647 ± 0.076	0.13 ± 0.17
	MinMax	0.684 ± 0.045	0.10 ± 0.19
	Standard	**0.707 ± 0.051**	**0.10 ± 0.19**
	MaxAbs	0.683 ± 0.044	0.09 ± 0.18
	Robust	**0.704 ± 0.053**	**0.10 ± 0.20**
MPOC	Nenhuma	0.578 ± 0.051	0.02 ± 0.02
	MinMax	0.600 ± 0.030	0.04 ± 0.09
	Standard	0.607 ± 0.032	0.04 ± 0.10
	MaxAbs	0.603 ± 0.034	0.04 ± 0.09
	Robust	0.609 ± 0.036	0.05 ± 0.10
Color Similarity	Nenhuma	0.611 ± 0.060	0.05 ± 0.11
	MinMax	0.642 ± 0.041	0.06 ± 0.12
	Standard	0.658 ± 0.039	0.06 ± 0.13
	MaxAbs	0.616 ± 0.057	0.06 ± 0.11
	Robust	0.662 ± 0.044	0.06 ± 0.14

Tabela X
MALARIA CELL DATASET

Descritor	**Normalização**	Acurácia	Tempo (s)
CEDD	Nenhuma	0.801 ± 0.081	7.52 ± 17.48
	MinMax	0.782 ± 0.080	11.03 ± 27.30
	Standard	0.800 ± 0.088	8.43 ± 17.20
	MaxAbs	0.782 ± 0.080	11.10 ± 27.46
	Robust	**0.802 ± 0.083**	**7.94 ± 18.59**
FCTH	Nenhuma	0.699 ± 0.054	13.90 ± 32.77
	MinMax	0.679 ± 0.058	16.66 ± 40.68
	Standard	0.689 ± 0.066	15.16 ± 35.10
	MaxAbs	0.678 ± 0.058	17.80 ± 42.48
	Robust	0.692 ± 0.057	16.83 ± 40.42
Gabor	Nenhuma	0.541 ± 0.015	10.68 ± 21.60
	MinMax	0.548 ± 0.010	9.04 ± 20.88
	Standard	0.552 ± 0.018	9.13 ± 20.77
	MaxAbs	0.547 ± 0.010	9.40 ± 21.68
	Robust	0.551 ± 0.016	8.89 ± 20.56
GCH	Nenhuma	0.796 ± 0.150	29.37 ± 71.07
	MinMax	0.780 ± 0.107	8.80 ± 21.07
	Standard	**0.862 ± 0.086**	**5.74 ± 10.06**
	MaxAbs	0.780 ± 0.107	9.04 ± 21.56
	Robust	**0.862 ± 0.088**	**19.30 ± 45.42**
JCD	Nenhuma	0.794 ± 0.079	13.13 ± 29.27
	MinMax	0.775 ± 0.081	14.86 ± 37.10
	Standard	0.798 ± 0.088	10.30 ± 22.90
	MaxAbs	0.776 ± 0.081	14.14 ± 34.13
	Robust	0.792 ± 0.080	10.71 ± 25.08
LBP	Nenhuma	0.658 ± 0.078	75.86 ± 181.93
	MinMax	0.684 ± 0.050	23.98 ± 56.07
	Standard	0.702 ± 0.062	25.60 ± 55.42
	MaxAbs	0.684 ± 0.050	23.71 ± 55.96
	Robust	0.702 ± 0.063	20.40 ± 43.21
MPOC	Nenhuma	0.685 ± 0.083	14.03 ± 35.53
	MinMax	0.739 ± 0.033	3.03 ± 6.07
	Standard	0.774 ± 0.052	2.79 ± 4.72
	MaxAbs	0.735 ± 0.034	3.33 ± 6.45
	Robust	0.770 ± 0.048	2.60 ± 4.66
Color Similarity	Nenhuma	0.674 ± 0.034	10.59 ± 23.85
	MinMax	0.679 ± 0.032	10.64 ± 22.61
	Standard	0.717 ± 0.054	9.01 ± 18.46
	MaxAbs	0.673 ± 0.034	9.11 ± 20.14
	Robust	0.711 ± 0.048	8.42 ± 16.51

Tabela XI
FRUITS DATASET

Descritor	**Normalização**	Acurácia	Tempo (s)
CEDD	**Nenhuma**	**0.689 ± 0.002**	**3.50 ± 0.11**
	MinMax	0.607 ± 0.002	3.47 ± 0.13
	Standard	0.674 ± 0.001	4.38 ± 0.34
	MaxAbs	0.601 ± 0.005	3.18 ± 0.09
	Robust	0.672 ± 0.004	3.32 ± 0.12
FCTH	Nenhuma	0.644 ± 0.002	3.48 ± 0.06
	MinMax	0.557 ± 0.000	3.71 ± 0.09
	Standard	0.630 ± 0.001	4.22 ± 0.06
	MaxAbs	0.555 ± 0.001	3.54 ± 0.11
	Robust	0.607 ± 0.000	3.62 ± 0.11
Gabor	Nenhuma	0.217 ± 0.001	1.28 ± 0.09
	MinMax	0.230 ± 0.001	1.98 ± 0.03
	Standard	0.243 ± 0.002	2.10 ± 0.00
	MaxAbs	0.223 ± 0.003	2.01 ± 0.00
	Robust	0.242 ± 0.002	2.15 ± 0.03
GCH	Nenhuma	0.536 ± 0.002	5.98 ± 0.26
	MinMax	0.520 ± 0.007	2.13 ± 0.01
	Standard	0.619 ± 0.002	2.48 ± 0.06
	MaxAbs	0.521 ± 0.003	2.11 ± 0.08
	Robust	0.604 ± 0.002	2.57 ± 0.15
JCD	**Nenhuma**	**0.696 ± 0.003**	**3.49 ± 0.16**
	MinMax	0.606 ± 0.002	3.37 ± 0.03
	Standard	**0.675 ± 0.001**	**4.89 ± 0.28**
	MaxAbs	0.604 ± 0.003	3.43 ± 0.07
	Robust	0.672 ± 0.005	3.46 ± 0.13
LBP	Nenhuma	0.416 ± 0.009	29.34 ± 0.35
	MinMax	0.417 ± 0.003	6.42 ± 0.13
	Standard	0.468 ± 0.004	10.52 ± 0.41
	MaxAbs	0.407 ± 0.003	6.55 ± 0.26
	Robust	0.456 ± 0.003	8.99 ± 0.10
MPOC	Nenhuma	0.406 ± 0.007	1.17 ± 0.09
	MinMax	0.450 ± 0.002	1.61 ± 0.02
	Standard	0.516 ± 0.002	1.68 ± 0.04
	MaxAbs	0.447 ± 0.002	1.47 ± 0.03
	Robust	0.500 ± 0.000	1.58 ± 0.03
Color Similarity	Nenhuma	0.452 ± 0.002	2.54 ± 0.15
	MinMax	0.498 ± 0.003	2.96 ± 0.08
	Standard	0.567 ± 0.005	2.90 ± 0.11
	MaxAbs	0.450 ± 0.003	2.48 ± 0.11
	Robust	0.551 ± 0.004	2.05 ± 0.27

Tabela XII
IRIS DATASET

Descritor	**Normalização**	Acurácia	Tempo (s)
CEDD	Nenhuma	0.871 ± 0.054	0.02 ± 0.06
	MinMax	0.808 ± 0.177	0.02 ± 0.05
	Standard	0.867 ± 0.062	0.03 ± 0.06
	MaxAbs	0.808 ± 0.177	0.02 ± 0.05
	Robust	0.850 ± 0.058	0.02 ± 0.05
FCTH	Nenhuma	0.742 ± 0.105	0.02 ± 0.05
	MinMax	0.662 ± 0.203	0.02 ± 0.05
	Standard	0.708 ± 0.123	0.02 ± 0.05
	MaxAbs	0.662 ± 0.203	0.02 ± 0.06
	Robust	0.679 ± 0.177	0.02 ± 0.04
Gabor	Nenhuma	0.754 ± 0.088	0.01 ± 0.03
	MinMax	0.812 ± 0.058	0.02 ± 0.04
	Standard	0.829 ± 0.051	0.02 ± 0.04
	MaxAbs	0.812 ± 0.058	0.02 ± 0.04
	Robust	0.838 ± 0.056	0.02 ± 0.04
GCH	Nenhuma	0.808 ± 0.236	0.01 ± 0.01
	MinMax	0.913 ± 0.029	0.02 ± 0.05
	Standard	**0.929 ± 0.026**	**0.02 ± 0.04**
	MaxAbs	**0.913 ± 0.029**	**0.02 ± 0.04**
	Robust	**0.921 ± 0.023**	**0.02 ± 0.05**
JCD	Nenhuma	0.879 ± 0.053	0.02 ± 0.05
	MinMax	0.783 ± 0.227	0.02 ± 0.05
	Standard	0.879 ± 0.071	0.02 ± 0.05
	MaxAbs	0.779 ± 0.226	0.02 ± 0.05
	Robust	0.850 ± 0.062	0.02 ± 0.06
LBP	Nenhuma	0.700 ± 0.239	0.01 ± 0.01
	MinMax	0.733 ± 0.153	0.03 ± 0.06
	Standard	0.783 ± 0.139	0.02 ± 0.05
	MaxAbs	0.738 ± 0.158	0.03 ± 0.07
	Robust	0.762 ± 0.124	0.02 ± 0.05
MPOC	Nenhuma	0.721 ± 0.237	0.01 ± 0.01
	MinMax	0.875 ± 0.036	0.02 ± 0.04
	Standard	0.875 ± 0.032	0.02 ± 0.04
	MaxAbs	0.879 ± 0.037	0.02 ± 0.04
	Robust	0.883 ± 0.041	0.02 ± 0.04
Color Similarity	Nenhuma	0.725 ± 0.199	0.02 ± 0.05
	MinMax	0.838 ± 0.059	0.02 ± 0.04
	Standard	0.863 ± 0.048	0.02 ± 0.05
	MaxAbs	0.746 ± 0.155	0.02 ± 0.04
	Robust	0.858 ± 0.049	0.02 ± 0.04

Tabela XIII
ROCK-PAPER-SCISSOR DATASET

Descritor	**Normalização**	Acurácia	Tempo (s)
CEDD	Nenhuma	0.856 ± 0.180	0.24 ± 0.52
	MinMax	0.839 ± 0.176	0.17 ± 0.30
	Standard	0.844 ± 0.182	0.26 ± 0.56
	MaxAbs	0.838 ± 0.176	0.20 ± 0.36
	Robust	0.855 ± 0.179	0.23 ± 0.51
FCTH	Nenhuma	0.677 ± 0.052	0.25 ± 0.44
	MinMax	0.658 ± 0.074	0.25 ± 0.41
	Standard	0.686 ± 0.056	0.33 ± 0.63
	MaxAbs	0.658 ± 0.074	0.25 ± 0.42
	Robust	0.680 ± 0.054	0.31 ± 0.58
Gabor	Nenhuma	0.520 ± 0.021	0.12 ± 0.18
	MinMax	0.530 ± 0.015	0.18 ± 0.34
	Standard	0.536 ± 0.022	0.20 ± 0.39
	MaxAbs	0.525 ± 0.016	0.18 ± 0.32
	Robust	0.537 ± 0.021	0.17 ± 0.32
GCH	Nenhuma	0.719 ± 0.159	0.29 ± 0.36
	MinMax	0.747 ± 0.094	0.19 ± 0.34
	Standard	0.799 ± 0.092	0.20 ± 0.40
	MaxAbs	0.747 ± 0.095	0.20 ± 0.38
	Robust	0.781 ± 0.090	0.18 ± 0.23
JCD	Nenhuma	0.862 ± 0.181	0.25 ± 0.52
	MinMax	0.846 ± 0.178	0.24 ± 0.44
	Standard	0.856 ± 0.185	0.26 ± 0.53
	MaxAbs	0.844 ± 0.178	0.25 ± 0.46
	Robust	**0.863 ± 0.180**	**0.25 ± 0.55**
LBP	Nenhuma	0.775 ± 0.172	0.56 ± 0.70
	MinMax	0.859 ± 0.074	0.43 ± 0.62
	Standard	**0.877 ± 0.077**	**0.45 ± 0.76**
	MaxAbs	0.851 ± 0.075	0.56 ± 0.87
	Robust	**0.878 ± 0.076**	**0.44 ± 0.72**
MPOC	Nenhuma	0.613 ± 0.133	0.09 ± 0.12
	MinMax	0.701 ± 0.076	0.15 ± 0.32
	Standard	0.749 ± 0.077	0.17 ± 0.38
	MaxAbs	0.678 ± 0.086	0.15 ± 0.33
	Robust	0.749 ± 0.077	0.17 ± 0.39
Color Similarity	Nenhuma	0.619 ± 0.088	0.21 ± 0.34
	MinMax	0.628 ± 0.080	0.24 ± 0.43
	Standard	0.677 ± 0.068	0.24 ± 0.46
	MaxAbs	0.625 ± 0.082	0.24 ± 0.43
	Robust	0.674 ± 0.067	0.21 ± 0.37

B. Tabelas das combinações das técnicas de descritores e normalizadores para cada dataset com base na Inércia do KMeans

Tabela XIV
ALIEN-VS-PREDATOR DATASET

Descritor	Normalização	Inércia
CEDD	Nenhuma	27139.240
	MinMax	977.132
	Standard	47217.457
	MaxAbs	977.132
	Robust	14437.489
FCTH	Nenhuma	19879.249
	MinMax	985.824
	Standard	52692.439
	MaxAbs	985.965
	Robust	11409.728
Gabor	Nenhuma	24567.238
	MinMax	272.765
	Standard	6345.689
	MaxAbs	248.265
	Robust	3045.552
GCH	Nenhuma	176679779866.468
	MinMax	170.535
	Standard	14640.695
	MaxAbs	170.526
	Robust	90918.524
JCD	Nenhuma	23983.877
	MinMax	1032.577
	Standard	59101.563
	MaxAbs	1032.579
	Robust	21081.677
LBP	Nenhuma	6475650883.949
	MinMax	1609.680
	Standard	72499.695
	MaxAbs	1493.530
	Robust	45280.885
MPOC	Nenhuma	2139530734.070
	MinMax	**57.372**
	Standard	1453.332
	MaxAbs	76.916
	Robust	4640.483
Color Similarity	Nenhuma	115.979
	MinMax	541.172
	Standard	19607.521
	MaxAbs	204.453
	Robust	12681.953

Tabela XV
MALARIA-CELL DATASET

Descritor	Normalização	Inércia
CEDD	Nenhuma	316976.931
	MinMax	11894.495
	Standard	780287.500
	MaxAbs	11685.368
	Robust	154289.470
FCTH	Nenhuma	440653.701
	MinMax	15253.908
	Standard	701246.334
	MaxAbs	14833.894
	Robust	134426.918
Gabor	Nenhuma	173539.866
	MinMax	8011.978
	Standard	463649.528
	MaxAbs	3747.912
	Robust	224745.414
GCH	Nenhuma	1838046121537.312
	MinMax	2886.648
	Standard	398971.108
	MaxAbs	2886.531
	Robust	2944653159.750
JCD	Nenhuma	340843.128
	MinMax	14041.279
	Standard	879563.237
	MaxAbs	13775.723
	Robust	178701.456
LBP	Nenhuma	45080725828.717
	MinMax	21247.911
	Standard	2089397.049
	MaxAbs	21046.077
	Robust	1307081.432
MPOC	Nenhuma	23772954058.967
	MinMax	1507.578
	Standard	116450.801
	MaxAbs	1166.299
	Robust	78011.852
Color Similarity	**Nenhuma**	**545.647**
	MinMax	10847.058
	Standard	536811.661
	MaxAbs	1135.114
	Robust	277218.604

Tabela XVI
FRUITS DATASET

Descritor	Normalização	Inércia
CEDD	Nenhuma	46556.244
	MinMax	1837.544
	Standard	65145.676
	MaxAbs	1837.544
	Robust	23329.405
FCTH	Nenhuma	27695.804
	MinMax	1375.083
	Standard	56564.243
	MaxAbs	1375.087
	Robust	11376.694
Gabor	Nenhuma	16815.665
	MinMax	633.818
	Standard	27318.613
	MaxAbs	252.391
	Robust	13753.328
GCH	Nenhuma	486401340032.634
	MinMax	589.944
	Standard	33607.827
	MaxAbs	566.699
	Robust	72715.689
JCD	Nenhuma	35188.416
	MinMax	2014.539
	Standard	78949.803
	MaxAbs	2014.539
	Robust	26460.552
LBP	Nenhuma	27086893476.486
	MinMax	2707.994
	Standard	139332.439
	MaxAbs	2172.430
	Robust	103675.470
MPOC	Nenhuma	6932755824.245
	MinMax	116.078
	Standard	4638.399
	MaxAbs	**82.365**
	Robust	2859.327
Color Similarity	Nenhuma	103.070
	MinMax	1208.939
	Standard	38468.760
	MaxAbs	248.534
	Robust	18032.100

Tabela XVII
IRIS DATASET

Descritor	Normalização	Inércia
CEDD	Nenhuma	2863.353
	MinMax	98.972
	Standard	1062.129
	MaxAbs	98.826
	Robust	292.583
FCTH	Nenhuma	909.973
	MinMax	18.571
	Standard	178.215
	MaxAbs	18.571
	Robust	122.301
Gabor	Nenhuma	15060.114
	MinMax	170.907
	Standard	2026.479
	MaxAbs	170.907
	Robust	732.296
GCH	Nenhuma	27827139630.802
	MinMax	113.648
	Standard	4099.327
	MaxAbs	113.648
	Robust	9573.848
JCD	Nenhuma	1613.753
	MinMax	93.613
	Standard	1062.799
	MaxAbs	93.613
	Robust	331.429
LBP	Nenhuma	3363724752.278
	MinMax	519.365
	Standard	17256.548
	MaxAbs	516.129
	Robust	1278798.018
MPOC	Nenhuma	562001377.099
	MinMax	44.580
	Standard	493.494
	MaxAbs	54.038
	Robust	158.924
Color Similarity	**Nenhuma**	**4.242**
	MinMax	143.507
	Standard	2292.526
	MaxAbs	8.117
	Robust	859.193

Tabela XVIII
ROCK-PAPER-SCISSOR DATASET

Descritor	Normalização	Inércia
CEDD	Nenhuma	17933.761
	MinMax	1897.121
	Standard	76102.491
	MaxAbs	1889.188
	Robust	16565.633
FCTH	Nenhuma	9154.529
	MinMax	559.460
	Standard	44807.527
	MaxAbs	529.714
	Robust	8140.552
Gabor	Nenhuma	7513.576
	MinMax	424.116
	Standard	30509.687
	MaxAbs	201.605
	Robust	23702.061
GCH	Nenhuma	478693964211.992
	MinMax	669.692
	Standard	36543.725
	MaxAbs	659.967
	Robust	213609224.616
JCD	Nenhuma	12617.438
	MinMax	1991.993
	Standard	89944.791
	MaxAbs	1971.584
	Robust	21237.183
LBP	Nenhuma	2911400204.739
	MinMax	3281.317
	Standard	145041.298
	MaxAbs	2356.580
	Robust	91313.684
MPOC	Nenhuma	1508579640.260
	MinMax	131.568
	Standard	6852.409
	MaxAbs	85.442
	Robust	5315.025
Color Similarity	**Nenhuma**	**49.754**
	MinMax	960.360
	Standard	61014.708
	MaxAbs	122.128
	Robust	51231.844

C. Imagens dos Métodos Elbow aplicados sobre cada Dataset

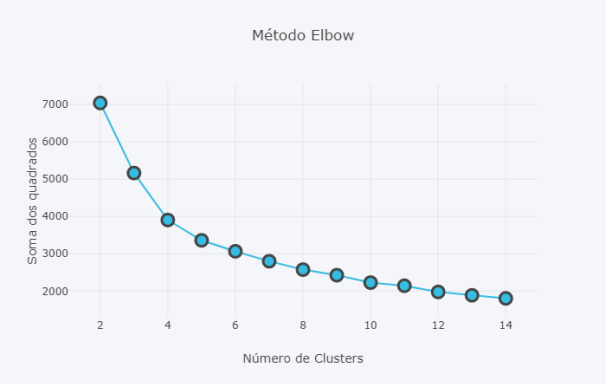

Figura 21. Método Elbow do *dataset* Alien vs Predator.

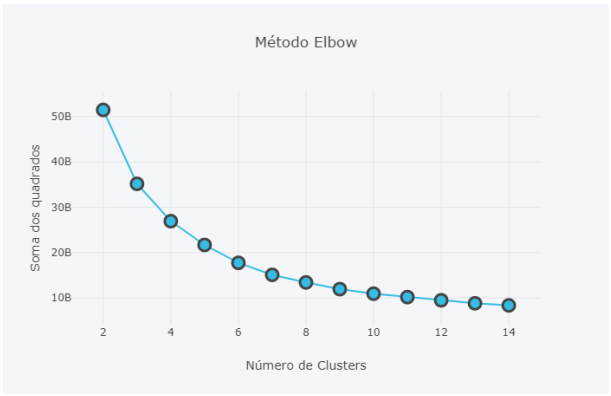

Figura 22. Método Elbow do *dataset* Malaria Cell.

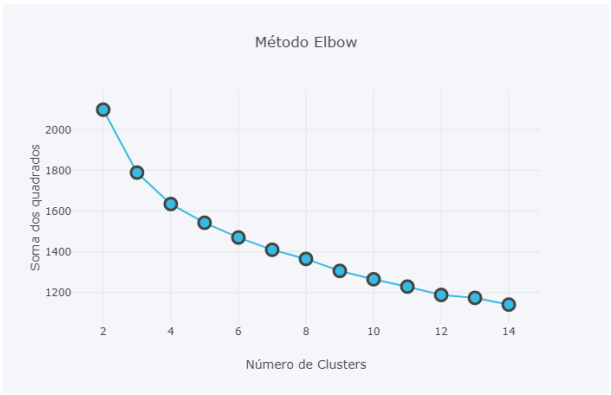

Figura 23. Método Elbow do *dataset* Fruits.

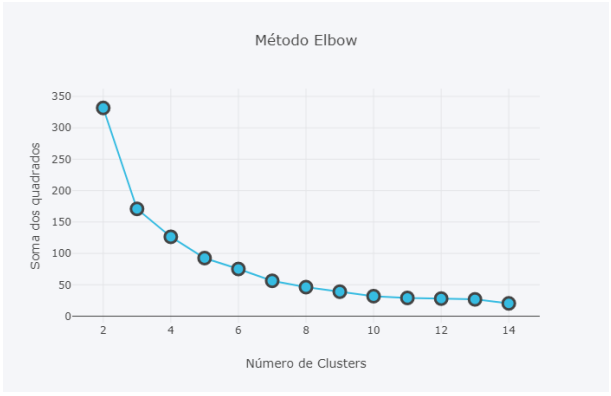

Figura 24. Método Elbow do *dataset* Iris.

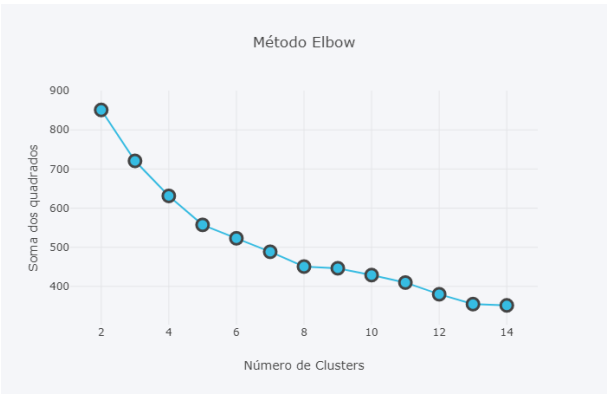

Figura 25. Método Elbow do *dataset* Rock-Paper-Scissor.