Um estudo sobre métodos de Aprendizagem de Máquina

1st Ana Lívia Franco
DACOM
UTFPR - CP
Cornélio Procópio, PR
analivia@live.com

2nd Bruna Almeida Osti
DACOM
UTFPR - CP
Cornélio Procópio, PR
brunaosti@alunos.utfpr.edu.br

3rd Katharina Akemi Ikeda Rosa
DACOM
UTFPR - CP
Cornélio Procópio, PR
kath.akemi@gmail.com

***Resumo*—Este artigo tem como objetivo demonstrar as atividades realizadas utilizando-se abordagens de Aprendizado de Máquina, bem como Aprendizado Supervisionado, Aprendizado Não Supervisionado, Aprendizado Semi-supervisionado e Aprendizado Ativo.**

***Index Terms*—Aprendizado de Máquina, Aprendizado Supervisionado, Aprendizado Não Supervisionado, Aprendizado Semi-Supervisionado, Aprendizado Ativo.**

I. Introdução

O Aprendizado de Máquina é uma subárea de estudo pertencente à inteligência artificial. Em outras palavras, este campo científico dedica-se à construção de algoritmos e técnicas que permitam aprender novas tarefas e aperfeiçoar seu desempenho em atividades rotineiras [1].

Ademais, os estudos relacionados à esta área foram iniciados devido a uma necessidade em programas que padrões e comportamentos por meio da observação de exemplos [1].

Enfim, este artigo tratará de um estudo prático em relação ao Aprendizado de Máquina, aprofundando-se no Aprendizado Supervisionado, Não-Supervisionado, Semi-Supervisionado e Ativo. Sendo assim, serão utilizados três diferentes datasets de imagens e quatro descritores tradicionais [1].

A. Motivação

Buscamos com este trabalho o aprendizado e o aprofundamento da teoria vista em sala, e que ao final deste trabalho sejamos capazes de compreender os diferentes tipos de aprendizado e aplicá-los em problemas reais.

B. Objetivos

O objetivo geral deste trabalho é compreender os diferentes tipos de aprendizagem, e aplicá-los em problemas reais de classificação de imagens.

1) Objetivos Específicos:

- Compreender os Descritores;
- Obter a extração de característica através de vários descritores;

II. Fundamentação Teórica

Nesta seção abordaremos os conceitos fundamentais para o entendimento do trabalho, sendo este os tipos de aprendizado existentes e suas características de funcionamento.

A. Aprendizado Supervisionado

Modelos de Aprendizagem de Máquina Supervisionados são caracterizados pela utilização de dados de treinamentos rotulados, isto é, contém a variável dependente resultante das variáveis independentes observadas. A partir disso, o objetivo é encontrar parâmetros ótimos que ajustem um determinado modelo, para então predizer rótulos desconhecidos em objetos do conjunto de teste. Tal técnica pode ser divida em dois modelos distintos, sendo esses a regressão e a classificação. A regressão visa encontrar um eixo de linearidade entre os dados. A classificação, por sua vez, prevê variáveis categóricas. Os principais algoritmos para a implementação do Aprendizado Supervisionado são: *Gaussian Naive Bayes (GNB), Logistic Regression (LR), Decision Tree (DT), K-Nearest Neighbors (KNN), Linear Discriminant Analysis (LDA), Support Vector Machine (SVM), Random Forest (RF) e Multi-layer Perceptron (MLP)* [2].

B. Aprendizado Não Supervisionado

Na Aprendizagem Não Supervisionado o conjunto de dados de treinamento não é rotulado. Desse modo, os algoritmos verificam similaridades entre os objetos e os inclui em grupos apropriados, criando os *clusters*. Objetos que diferem largamente de todos os grupos criados podem ser considerados anomalias. Pode-se citar como exemplo de algoritmos o *k-Means, Hierarchical Cluster Analysis (HCA), Expectation Maximization, Principal Component Analysis (PCA), Kernel PCA e Locally-Linear Embedding (LLE)* [2].

C. Aprendizado Semi-Supervisionado

Algoritmos de Aprendizado Semi-Supervisionados são capazes de trabalhar com dados com rotulação incompleta, ou seja, em que apenas parte dos dados são rotulados. Grande parte desses algoritmos são resultado da combinação de técnicas supervisionadas e não supervisionadas. Como exemplo, têm-se o algoritmo *Restricted Boltzmann Machines (RBMs)*, comumente utilizado no treinamento não supervisionado, contudo, posteriormente, são aplicadas técnicas supervisionadas afim de otimizar o sistema e melhorar sua acurácia [3].

D. Aprendizado Ativo

Um dos subcampos do Aprendizado de Máquina é o Aprendizado Ativo, onde os classificadores são treinados a partir de amostras representativas. Desse modo, os classificadores demonstram melhorias em sua performance, o que faz com que uma menor quantidade de dados seja necessária para o treinamento. O Aprendizado Ativo é uma abordagem recomendada quando a rotulação dos dados é dificultosa, uma vez que são selecionadas as amostras consideradas mais importantes para que sejam rotuladas por um especialista, otimizando o processo [4].

III. Metodologia Proposta

Nesta seção será descrito como será cada etapa do trabalho, desde os datasets utilizados, a utilização dos descritores e como será feita as métricas para comparação dos métodos.

A. Datasets

Em primeira instância, foi realizado a escolha de quais os datasets de imagens e os descritores tradicionais à serem utilizados para a extração de características das imagens propostas. Portanto, os *datasets* escolhidos foram o Monkeys, Malaria-Cell e um conjunto de imagens de retinopatia diabética, disponível no [5] .

1) Dataset Monkeys: O conjunto de dados se trata de um conjunto de imagens para classificação de macacos entre 10 espécies distintas. O *dataset* contém 1098 imagens e 10 classes, de modo que cada classe representa uma espécie como algumas estão sendo mostradas na Figura 1, sendo elas: Alouatta Palliata, Erythrocebus Patas, Cacajao Calvus, Macaca Fuscata, Cebuella Pygmea, Cebus Capucinus, Mico Argentatus, Saimiri Sciureus, Aotus Nigriceps, Trachypithecus Johnii.

Figura 1. Dataset Monkeys

2) Dataset Malaria-Cell: A malária é uma doença infecciosa febril aguda, causada por protozoários transmitidos pela fêmea infectada do mosquito Anopheles. Indivíduos que tiveram vários episódios de malária podem atingir um estado de imunidade parcial, apresentando poucos ou mesmo nenhum sintoma no caso de uma nova infecção [6]. A base de dados fornecida previamente pela professora contém 27.558 imagens divida em duas classes como mostrado na Figura 2.

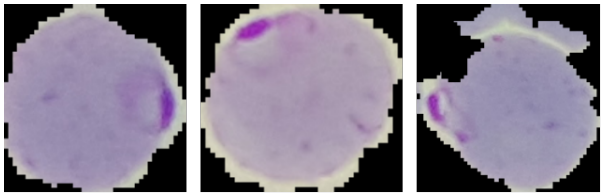

Figura 2. Dataset Células de Malária

3) Dataset Retinopatia Diábetica: A Retinopatia Diabética (DR) é uma complicação do diabetes que pode levar à cegueira, se não for descoberta em tempo hábil, a base de dados [5] como mostrada na figura 3 conta com dois módulos: *DR1 e DR2*, além disso, é separada em 7 classes de imagens: Cotton-wool Spots, Deep Hemorrhages, Drusen, Hard Exudates, Normal Images, Red Lesions, Superficial Hemorrhages. Utilizou-se o dataset DR1, aplicou-se normalização das imagens e também data-augmentation, totalizando 7.083 imagens, divididas entre treinamento e teste como mostrado na Figura 3.

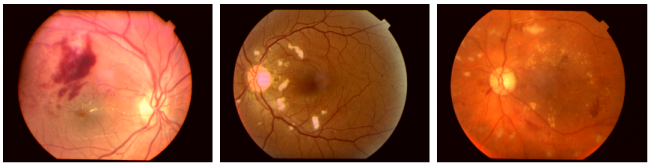

Figura 3. Dataset Retinopatia Diabética

B. Descritores

Devido à grande quantidade de características presentes nos *datasets* escolhidos, foram utilizados extratores de características, também conhecidos como descritores. Para as práticas realizadas no presente trabalho, os descritores selecionados foram:

1) **BIC:** Extrator de características baseado em cor, descreve 128 características [7].
2) **LBP:** Extrator de características baseado em textura, descreve 256 características [8].
3) **FCHT:** Extrator de características baseado em cor e textura, descreve 192 características [9].
4) **GCH:** Extrator de características baseado em cor, descreve 255 características [10].

Por fim, utilizou-se dos quatro algoritmos citados anteriormente para extrair características dos conjuntos de imagens e coletar o conjunto de dados elaborado por cada execução.

Neste trabalho dividiu-se os métodos em cinco práticas no qual avaliaríamos a acurácia de cada tipo de aprendizagem levando-se em conta os quatro extratores (*BIC, LBP, FCHT e GCH*), e além disso, se havia algum tipo de normalização (MinMax, StandardScaler, MaxAbsScaler ou RobustusScaler) empregada.

IV. Avaliação Experimental

Nesta seção avaliaremos a metologia de cada abordagem sendo elas: Aprendizado supervisionado, aprendizado não su-

pervisionado, aprendizagem semi supervisionado e o aprendizado ativo.

A. Aprendizado Supervisionado

Dando continuidade ao experimento, utilizou-se dos conjuntos de dados extraídos anteriormente para aplicá-los em oito classificadores diferentes, sendo esses, Gaussian Naive Bayes, Logistic Regression, Decision Tree, K-Nearest Neighbors, Linear Discriminant Analysis, Support Vector Machine, RandomForest e Neural Net. Contudo, todos as bases foram classificadas de maneira crua, ou seja, sem nenhum tipo de normalização, assim como, foram tratadas anteriormente à este processo por meio de 4 técnicas de normalização: MinMax Scaler, Standard Scaler, MaxAbs Scaler e Robustus Scaler.

Portanto, a escolha do extrator preferível para cada um dos três *datasets* realizou-se por meio da seleção do melhor cruzamento das técnicas, normalização e classificação. Ademais, deve-se considerar esta predileção embasada na acurácia, precisão, *recall*, *f1 score* e tempo de execução superiores.

Por fim, justificasse o usufruto desta técnica para o aprendizado supervisionado, pois, apesar do alto custo computacional devido a operação das 32 possibilidades para cada *dataset*, torna-se precisa e pouco excludente a busca pelo extrator mais aconselhável para cada categoria.

B. Aprendizado Não Supervisionado

O Aprendizado Não Supervisionado e suas técnicas são evidenciados na etapa presente do experimento. Portanto, novamente submeteu-se os *datasets* ao seu melhor extrator e normalizador caracterizados dentro da tripla encontrada na etapa anterior. Por conseguinte, o método de agrupamento de dados *k-means*, em conjunto ao método *Elbow*, foram utilizados para o encontro do melhor *k* para cada um dos *datasets*.

Ademais, aplicou-se o método PCA à cada um dos descritores de todos os *datasets*, permitindo, contemplar o agrupamento *k-means*. Por consequência, a variância foi essencialmente utilizada para a extração das duas características mais singulares, permitindo, a geração de gráficos dos centroides, estabelecidos pelo conceito de clusterização.

Enfim, as técnicas de extração e normalização responsáveis por treinar o *k-means* foram novamente aplicadas, a partir da escolha da melhor configuração para cada *dataset*. Posteriormente, realizou-se a eleição do melhor *k* para cada *dataset*, por meio do método *Elbow*. Em suma, avaliou-se a qualidade do agrupamento, utilizando de comparações, aos quais se dão pelos gráficos, das *clusters* fundadas pelo *k-means* com as classes reais de cada conjunto de imagens.

C. Aprendizado Semi Supervisionado

Nesta abordagem, foram resgatados os mesmos classificadores utilizados no treinamento supervisionado. Enfim, dentre as amostras rotuladas e não rotuladas, foram selecionadas diferentes razões de quantidade: 10%, 20%, 30%, 40%, 50% e 60%. Portanto, todos os classificadores, quantidades de rótulos e extratores foram aplicados aos três *datasets*, para que ao fim, fosse selecionado o melhor conjunto de técnicas para cada conjunto.

D. Aprendizado Ativo

Para a execução do Aprendizado Ativo, foram utilizados os mesmos descritores, normalizadores e classificadores abortados no Aprendizado Supervisionado. Visto isso, o método *k-means* foi utilizado com a finalidade de realizar a separação de amostras significativas dos conjuntos, de modo que, o valor de k utilizado para cada um dos *datasets* foi encontrado com a implementação do Aprendizado Não Supervisionado. O Aprendizado Ativo foi iterado até que se encontrasse acurácia próxima às alcançadas no Aprendizado Supervisionado ou até que não houvesse melhorias após 10 incrementos consecutivos.

V. Resultados e Discussões

Nesta seção apresentaremos os resultados encontrados em cada abordagem, além da análise e comparação desses dados, bem como as melhores triplas considerando descritor, normalizador e classificador.

A. Aprendizado Supervisionado

Foram realizados os testes descritos na seção anterior, para com os classificadores, normalizadores e extratores. Desse modo, pôde-se reunir na Tabela (I) as melhores combinações identificadas. Acerca disso, as métricas das melhores triplas são apresentadas na Tabela (II). [1]

Tabela I
Melhores Combinações

Dataset	Descritor	Normalizador	Classificador
Monkeys	BIC	StandardScaler	RandomForest
Malaria-Cell	BIC	StandardScaler	RandomForest
Retinopatia Diabética	BIC	StandardScaler	RandomForest

Tabela II
Melhores resultados obtidos em cada dataset

Dataset	Acurácia	Precisão	Revocação	F1 Score	Tempo (segundos)
Monkeys	0.56	0.55980	0.55217	0.54241	0.01478
Malaria-Cell	0.95	0.95000	0.94988	0.94992	0.09713
Retinopatia Diabética	0.51	0.27984	0.28138	0.27787	0.04575

Além do mais, observando os resultados finais obtidos por meio da extração de características do *dataset* Malaria-Cell a tripla com melhor acurácia é composta pelos mesmos componentes do conjunto de imagens citado anteriormente. Contudo, a Tabela II trás dados que permitem concluir que o descritor BIC com os dados normalizados pelo StandardScaler e o classificados pelo RandomForest, possuem acurácia muito alta de 95% e um tempo de execução equivalente a 0,09713 segundos, ao qual pode ser considerado muito baixo. Abaixo está apresentado a matriz de confusão no qual representa em sua diagonal a quantidade de informações corretamente

[1]Os resultados completos estão disponíveis em: $< \ https \ : \ //github.com/analiviafr/Praticas\_Inteligencia\_Artificial >$

classificadas pelo modelo.

Matriz de confusão:

2658	125
151	2578

No que tange o *dataset* Retinopatia Diabética, o conjunto gerador com melhor acurácia é composto pelo extrator *Border Interior Classification* BIC, que utiliza a correlação entre *pixels* de borda e *pixels* internos que formam a imagem. Faz parte da tripla também o normalizador StandartScaler, caracterizado pela produção de bons resultados para recursos com diferentes magnitudes. O melhor classificador neste caso foi o RandomForest, um algoritmo com estrutura em árvore que não necessita de dados normalizados para um bom desempenho. De mesmo modo abaixo está descrito quantidade de informações corretamente classificadas pelo modelo.

Matriz de confusão:

173	4	0	0	0	0	0	0	0	0	0	4	2	4	11
1	85	0	0	0	0	0	0	0	0	0	19	1	36	47
1	0	1	5	4	2	0	3	4	7	0	0	2	0	2
1	0	4	0	0	2	3	2	2	2	1	1	1	1	1
1	1	0	1	4	1	0	0	0	2	2	1	4	0	1
0	1	3	2	3	1	3	6	1	1	2	3	0	0	0
0	0	0	2	2	1	2	0	0	0	2	3	1	0	0
1	1	1	4	2	1	0	0	2	2	3	1	0	0	2
0	0	5	1	1	1	1	2	2	1	1	1	3	0	1
0	0	3	1	2	1	4	0	2	2	2	1	2	0	0
0	0	1	1	3	1	3	3	0	3	1	0	2	0	1
12	32	0	0	0	0	0	0	0	1	0	98	7	37	11
15	10	0	1	0	0	0	1	0	2	0	14	149	7	2
11	59	1	0	0	0	1	0	0	0	1	44	2	51	59
10	44	0	0	0	0	0	0	3	0	0	11	1	25	120

Por fim, no caso do *dataset* Monkeys, a tripla com melhor acurácia é composta pelo extrator de características *Border Interior Classification* BIC, juntamente com o normalizador StandartScaler, utilizando-se como classificador o RandomForest. Apesar de o melhor resultado obtido ser 56% de acurácia ele ainda fica longe de ser o valor desejado, pois apresenta uma baixa precisão em conjunto com uma baixa revocação, além disso, o tempo também é um diferencial, com apenas 0,01478 o que pode ser considerado muito baixo.

Matriz de confusão:

11	2	**0**	0	0	**1**	1	0	0	**4**
3	15	0	1	1	0	0	2	0	1
0	2	19	0	0	0	0	0	0	0
0	1	2	19	1	0	0	0	2	0
0	4	0	2	8	1	1	1	0	0
6	1	1	1	0	11	1	1	1	4
0	3	1	2	1	0	18	0	0	0
4	1	0	2	3	1	1	14	0	0
2	3	0	2	4	0	1	1	4	2
4	1	0	1	0	4	1	2	1	5

B. Aprendizado Não Supervisionado

Para o *dataset* Retinopatia Diabética, a melhora acurácia foi de 51%, vide à Tabela II, resultado obtido através da tripla formada pelo extrator BIC, normalizador StandardScaler e classificador RandomForest. Inicialmente, utilizou-se uma quantidade de *clusters* igual a 7, após a analise do Método *Elbow*, demonstrado pela Figura 4, a mesma foi reduzida para 2. Desse modo, o *k-means* pôde gerar os agrupamentos representados na Figura 5.

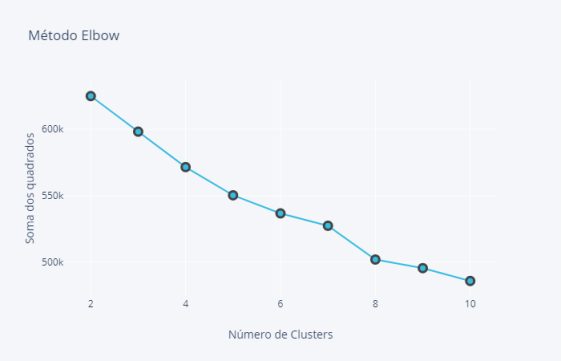

Figura 4. Método *Elbow* aplicado ao *dataset* Retinopatia Diabética

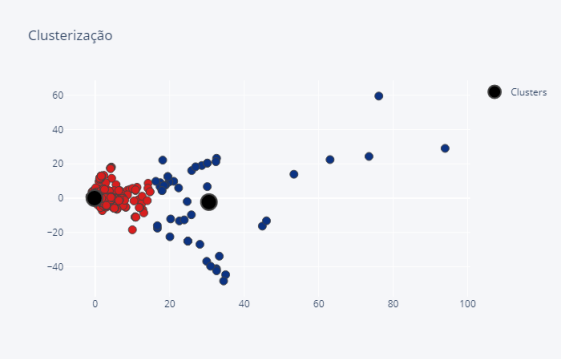

Figura 5. Clusterização do *dataset* Retinopatia Diabética

Em consequência, acompanhando a tendência, o *dataset* Malária, com melhor acurácia de 95%, como visto na Tabela II, teve sua tripla de sucesso formada pelo extrator BIC e normalizador StandardScaler. Portanto, utilizou-se destes para o encontro da melhor quantidade de *clusters*, segunda análise do Método *Elbow*, sendo assim, o valor total de dois, como pode-se observar na Figura 6. Sendo assim, este é o mesmo valor, previamente testado e correspondenteo número de classes do conjunto de imagens. Por fim, o *k-means* gerou os agrupamentos vistos na Figura 7.

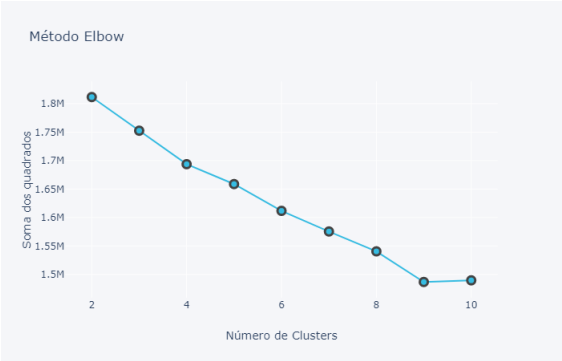

Figura 6. Método *Elbow* aplicado ao *dataset* Malária

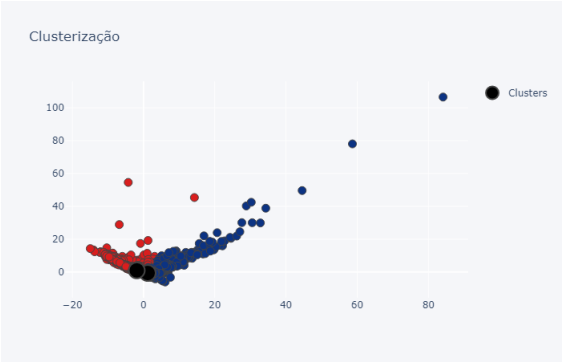

Figura 7. Clusterização do *dataset* Malária

Por outro lado, o *dataset Monkeys*, a melhor acurácia foi de 56%, vide Tabela II, resultado obtido através da tripla formada pelo extrator BIC, normalizador StandardScaler e classificador RandomForest. Primeiramente, utilizou-se uma quantidade de *clusters* igual a 10 (mesma quantidade de classes do *dataset*), mas após o método *Elbow*, demostrado na figura 9, a mesma foi reduzida para 2. Desse modo, o *k-means* pôde gerar os agrupamentos representados na figura 9.

Figura 8. Método *Elbow* aplicado ao *dataset Monkeys*

C. Aprendizado Semi Supervisionado

Foi realizada a aplicação do Aprendizado Semi Supervisionado sobre o *dataset* Retinopatia Diabética, utilizando o

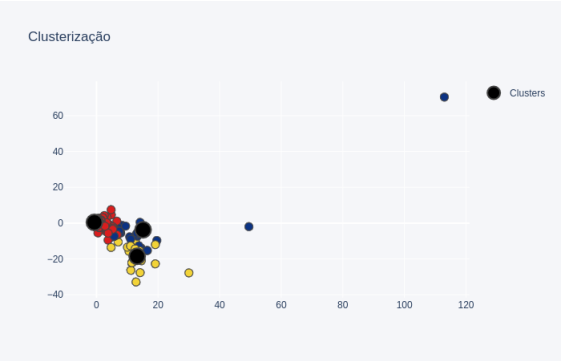

Figura 9. Clusterização do *dataset Monkeys*

descritor BIC e a normalização *StandardScaler*, de modo que as acurácias dos classificadores utilizados estão representados na Figura 10 e suas respectivas métricas na Figura 11. Após análise, verificou-se que o melhor resultado foi obtido através do classificador *Decision Tree*, o qual obteve 24% de acurácia de teste com 60% de amostras rotuladas. Se comparado com os resultados obtidos pelo aprendizado supervisionado, que alcançou 51% de acurácia no teste para o *dataset* em questão, têm-se que o aprendizado semi supervisionado retornou resultados inferiores, não sendo o ideal para este caso.

Figura 10. Acurácia dos Classificadores para o *dataset* Retinopatia Diabética

Figura 11. Métricas de Avaliação para o *dataset* Retinopatia Diabética

Ademais, a aplicação do Aprendizado Semi Supervisionado sobre o *dataset* Malária, usufruindo do descritor BIC e o normalizador *StandardScaler* , de tal modo que as acurácias dos classificadores utilizados estão representador na Figura 12 e suas respectivas métricas na Figura 13. Enfim, ao analisar os resultados, identificou-se que o melhor resultado foi obtido por meio do classificador *RandomForest* e é o classificador que atinge o melhor resultado em ambas, mas que infelizmente está muito inferior ao resultado obtido pelo método supervisionado, portanto, o outro método é mais preciso e eficaz.

Figura 12. Acurácia dos Classificadores para o *dataset* Malária

Figura 13. Métricas de Avaliação para o *dataset* Malária

Por outro lado a aplicação do aprendizado semi supervisionado sobre o *dataset* Monkeys, utilizando a melhor tripla obtida anteriormente, fazem parte dela o descritor BIC e o normalizador *StandardScaler* em conjunto com o classificador *RandomForest*, de modo que as acurácias de todos os classificadores utilizados estão representadas na Figura 14 e suas respectivas métricas de avaliação na Figura 15. Observou-se de acordo com as avaliações que o classificador que mais aproximou-se dos resultados obtidos na abordagem semi supervisionado foi o *Gaussian Naive Bayes*, utilizando 60% dos dados rotulados e obtendo apenas 25% de acurácia. No qual é possível observar que o método não é adequado para este caso pois não se aproximou dos dados obtidos anteriormente.

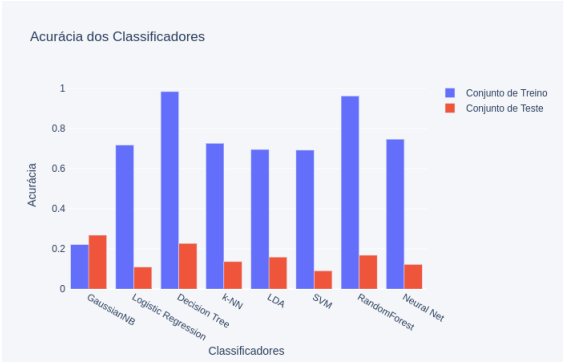

Figura 14. Acurácia dos Classificadores para o *dataset* Monkeys

Figura 15. Métricas de Avaliação para o *dataset* Monkeys

D. *Aprendizado Ativo*

Trabalhando com o *dataset* Retinopatia Diabética, utilizando o descritor BIC e a normalização *StandardScaler*, o melhor resultado foi obtido com o classificador *RandomForest*, obtendo uma acurácia de 29% no teste, o qual foi atingido utilizando 342 amostras e 170 iterações. Contudo, a acurácia obtida foi distante da acurácia alcançada pela mesma tripla no Aprendizado Supervisionado, que foi de 51% no teste. A Figura 16 demonstra as alterações de acurácia ao longo das iterações e Figura 17 demostra as métricas de avaliação utilizadas.

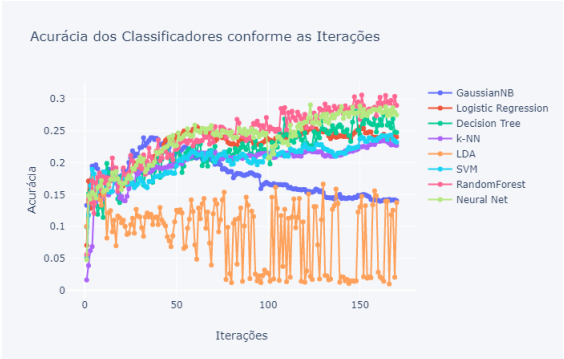

Figura 16. Acurácia dos classificadores para o *dataset* Retinopatia Diabética

Figura 17. Métricas de Avaliação para o *dataset* Retinopatia Diabética

Ademais, ao implementar com o *dataset* Malária, utilizou-se o descritor BIC e normalizador *StandardScaler*, assim como, no experimento anterior, logo o melhor resultado obtido foi por meio do classificador *RandomForest*. Em suma, a acurácia do experimento foi de 93% no teste, após, a utilização de 167 amostras e 80 interações. Portanto, este resultado foi muito próximo do alcançado pela mesma tripla durante o experimento do Aprendizado Supervisionado, o qual foi de 95%. A figura 18 demonstra as alterações de acurácia a partir das iterações e a Figura 19 demonstra as métricas de avaliação utilizadas.

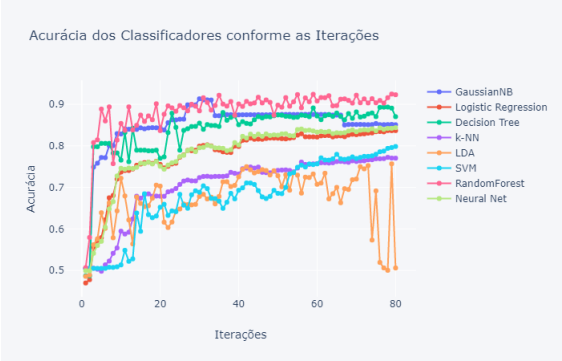

Figura 18. Acurácia dos classificadores para o *dataset* Malária

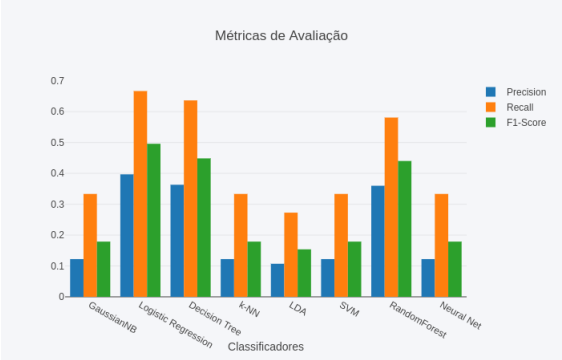

Figura 19. Métricas de Avaliação para o *dataset* Malária

Por fim, ao implementar o método ativo utilizando o *dataset Monkeys*, utilizou-se a melhor tripla encontrada para o aprendizado supervisionado sendo ele o descritor BIC em conjunto com o normalizador *StandardScaler*, de mesmo modo, o melhor classificador utilizado foi o *RandomForest*. Alcançando de modo bem próximo 54% a acurácia obtida pelo método supervisionado em apenas 591 amostras e 196 iterações. Portanto, este resultado foi muito próximo do alcançado pela mesma tripla durante o experimento supervisionado, no qual obteve-se 56% de acurácia. A Figura 20 mostra a acurácia obtida pelos classificadores e a Figura 21 demonstra as métricas de avaliação utilizadas.

Figura 20. Acurácia dos Classificadores para o *dataset* Monkeys

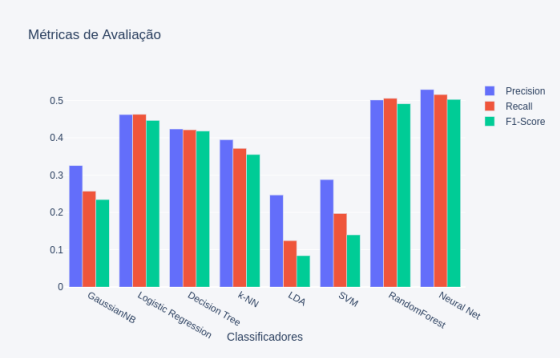

Figura 21. Métricas de Avaliação para o *dataset* Monkeys

VI. Metodologia Proposta

Como nova abordagem para a seleção de amostras para o aprendizado ativo, tem-se a utilização de uma rede neural profunda para extração de características melhores como mostrado na Figura 22.

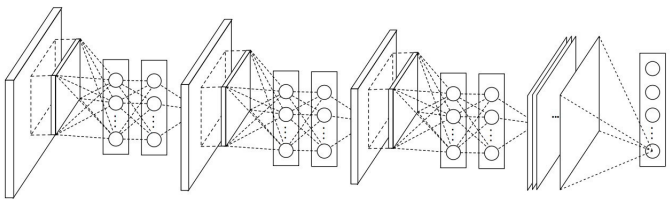

Figura 22. Extração de características

Para aumentar a possibilidade de selecionar amostras de classes distintas e de fato as mais significativas mais rapidamente é realizado um agrupamento das amostras do conjunto de dados e os pares de amostras de fronteira são pré-organizados, calculando-se as distâncias entre as amostras de cada par e organizá-las em ordem crescente de distância. Ao final, tem-se uma lista de arestas de fronteira pré-organizadas em ordem crescente de distância. A ideia é priorizar amostras mais similares e que sejam de classes distintas, as quais seriam as mais difíceis de serem classificadas.

Portanto, podemos utilizar a distância euclideana para encontrar a distância entre esses pontos, e assim encontrar as amostras que seriam mais simples de serem classificadas, como mostra a equação abaixo:

$$Distância = \sqrt{(x_2^2 - x_1^2) + (y_2^2 - y_1^2)} \quad (1)$$

E pode ser aplicada seguindo os passos da Figura 23, quando os clusters já estão separados, calcula-se a distância euclidiana para cada um dos pares de pontos de clusters distintos.

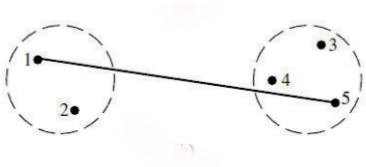

Figura 23. Encontrando as distâncias euclidianas nos clusters

VII. Conclusão e Trabalhos Futuros

Através deste trabalhos podemos perceber que os métodos de aprendizagem são muito eficientes para problemas de classificação, mas que é necessário entender as abordagens de modo que fique evidente em qual situação é utilizada cada método, tornando assim os resultados mais precisos e melhores. Além disso, os extratores tem um papel muito importante de selecionar as características que serão relevantes para a análise, entretanto, há vários tipos de extratores cada qual com um objetivo específico.

Em trabalhos futuros espera-se aplicar outras metodologias, como por exemplo: Aprendizagem Profunda, buscando-se outros tipos de abordagem talvez até mais precisas. Além disso, espera-se aprender modos de melhorar os dados para classificação a fim trazer melhoras para o resultado final, como por exemplo: métodos de segmentação no caso de imagens, e métodos de ajustes finos no caso de dados.

Referências

[1] P. Norvig, *Inteligência Artificial (Em Portuguese do Brasil)*. Elsevier, 2013.

[2] A. Géron, *Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly Media, 2017.

[3] B. Rao and M. Babu, "A critical study of big data techniques and predictive analytics algorithms," *International Journal of Computer Sciences and Engineering*, vol. 6, pp. 695–700, 12 2018.

[4] B. Settles, "Active learning literature survey," University of Wisconsin–Madison, Computer Sciences Technical Report 1648, 2009.

[5] R. Pires, H. F. Jelinek, J. Wainer, E. Valle, and A. Rocha, "Advancing Bag-of-Visual-Words Representations for Lesion Classification in Retinal Images," 10 2016.

[6] M. da Saúde. Malária: o que é, causas, sintomas, tratamento, diagnóstico e prevenção. [Online]. Available: http://saude.gov.br/saude-de-a-z/malaria

[7] R. O. Stehling, M. A. Nascimento, and A. X. Falcão, "A compact and efficient image retrieval approach based on border/interior pixel classification," in *Proceedings of the eleventh international conference on Information and knowledge management - CIKM '02*. ACM Press, 2002. [Online]. Available: https://doi.org/10.1145/584792.584812

[8] Z. Guo, L. Zhang, and D. Zhang, "Rotation invariant texture classification using LBP variance (LBPV) with global matching," *Pattern Recognition*, vol. 43, no. 3, pp. 706–719, Mar. 2010. [Online]. Available: https://doi.org/10.1016/j.patcog.2009.08.017

[9] S. A. Chatzichristofis and Y. S. Boutalis, "FCTH: Fuzzy color and texture histogram - a low level feature for accurate image retrieval," in *2008 Ninth International Workshop on Image Analysis for Multimedia Interactive Services*. IEEE, 2008. [Online]. Available: https://doi.org/10.1109/wiamis.2008.24

[10] M. A. Stricker and M. Orengo, "Similarity of color images," in *Storage and Retrieval for Image and Video Databases III*, W. Niblack and R. C. Jain, Eds. SPIE, Mar. 1995. [Online]. Available: https://doi.org/10.1117/12.205308