Utilização de Modelos Computacionais Baseados em Classificadores Para Predição da Dislexia em Crianças

Ronieri Nogueira de Sousa[2], **Roney Nogueira de Sousa**[1], **Rhyan Ximenes de Brito**[1], **Janaide Nogueira de Sousa Ximenes**[2]

[1]Instituto Federal de Educação, Ciência e Tecnologia do Ceará (IFCE)
Av. Tabelião Luiz Nogueira de Lima S/N – Tianguá – CE – Brasil

[2]Faculdade IEducare (FIED) – Rua Conselheiro João Lourenço, 406 - CEP 62320-000 – Tianguá – CE – Brasil

{nsronieri,nogueiraroney453,rxbrito,nogueirajanaide}@gmail.com

Abstract. *Dyslexia is one of the most common learning difficulties in classrooms. Thus, the study aimed to classify children with or without dyslexia through the application of Computational Intelligence (CI) techniques. The methodology used a public database and the application of neural architectures, Multilayer Perceptron (MLP), Radial Basis Function (RBF) and Extreme Learning Machine (ELM) and statistical classifiers, Support Vector Machine (SVM), Random Forest (RF) and K-Nearest Neighbors (K-NN), as well as k-fold, SMOTE and z-score normalization techniques. The results showed that the SVM classifier had the best average hit rate with 98.03% accuracy.*

Resumo. *A dislexia é uma das dificuldades de aprendizagem mais comum nas salas de aula. Dessa forma o estudo teve como finalidade a classificação de crianças com ou sem dislexia através da aplicação de técnicas de Inteligência Computacional (IC). Para a metodologia utilizou-se de uma base de dados pública e da aplicação das arquiteturas neurais, Multilayer Perceptron (MLP), Radial Basis Function (RBF) e Extreme Learning Machine (ELM) e dos classificadores estatísticos, Support Vector Machine (SVM), Random Forest (RF) e K-Nearest Neighbors (K-NN), assim como das técnicas k-fold, SMOTE e normalização z-score. Os resultados demonstraram que o classificador SVM obteve a melhor taxa média de acerto com 98,03% de acurácia.*

1. Introdução

Segundo [Jerônimo and Duarte 2016] a dislexia é uma das dificuldades de aprendizagem mais comum nas salas de aula, tendo como ponto de partida a compreensão, das quatro habilidades fundamentais da linguagem verbal: leitura, escrita, fala e escuta. Possui uma incidência em torno de 2% a 5% da população, sendo mais comum em alunos do sexo masculino em 25% a 49% dos casos.

Com base nessa premissa foi realizado um estudo comparativo entre os modelos neurais *Multilayer Perceptron* (MLP) e *Radial Basis Function* (RBF), *Extreme Learning Machine* (ELM) e os classificadores *Support Vector Machine* (SVM), *Random Forest* (RF) e *K-Nearest Neighbors* (K-NN) com o objetivo de comparar os resultados baseados em procedimentos quantitativos, com ênfase no treinamento das redes, de forma a auxiliar

na classificação de crianças com ou sem o transtorno da dislexia, utilizando um *dataset* público composto por amostras de crianças com e sem o Transtorno da Dislexia, intitulado *"Predicting Risk of Dyslexia"link* [1].

A motivação do trabalho em questão deve-se a necessidade do aprimoramento de técnicas que garantam a identificação precoce do transtorno de forma que se possa tomar medidas que visem a redução dos impactos na aprendizagem do indivíduo, assim como na condução de um acompanhamento psicopedagógico adequado.

Este trabalho está organizado em 8 seções: a Seção 2 apresenta os trabalhos relacionados, Seção 3 nos mostra a fundamentação teórica sobre o transtorno da dislexia, a Seção 4 apresenta os modelos neurais MLP, RBF e ELM, a Seção 5 aborda os classificadores estatísticos SVM, RF e K-NN a Seção 6 aborda os materiais e métodos utilizados no trabalho, Seção 7 apresenta os resultados e discussões, por fim, a Seção 8 apresenta as considerações finais e trabalhos futuros.

2. Trabalhos Relacionados

Esta seção um revisão bibliográfica de diferentes abordagens do uso de técnicas de Inteligência Computacional Aplicada (ICA) como ferramenta de auxílio no diagnóstico de transtornos.

[de Brito et al. 2020a] realizou um trabalho a qual foi realizado um estudo através da implementação e análise das redes neurais ELM e MLP, comparando as acurácias resultantes de treinamentos com dados de adolescentes com ou sem o Transtorno do Espectro Autista. Com relação aos resultados a rede MLP sem normalização obteve a melhor média atingindo 89,70%.

[Araujo et al. 2019] apresentou um sistema baseado em regras fuzzy para um prédiagnóstico da doença esquizofrenia. Teve como base de sua metodologia pesquisas bibliográficas e simulações com pacientes fictícios, onde foram utilizadas como variáveis de entrada sintomas característicos da doença. Os resultados se demonstraram promissores na medida que foram bem próximos dos observados por profissionais da saúde.

[Ceravolo et al. 2019] realizou um trabalho em que propõe a análise de movimentos oculares no processo de leitura utilizando *wavelets* e algoritmos de aprendizado de máquina a fim de identificar leitores que são propensos a dislexia. Obteve como resultado uma acurácia de 97,30% com o algoritmo RF.

[Asvestopoulou et al. 2019] propuseram um instrumento de triagem para a dislexia. Utilizou uma base de dados composta por algumas informações de movimentos oculares de 69 crianças. Obteve como acurácia de 97,10% utilizando o SVM e redução das características a partir do método de LASSO (*Least Absolute Shrinkage and Selection Operator*).

[de Brito et al. 2020b] realizaram um estudo por meio da implementação das redes neurais artificiais (RNAs), MLP e RBF a partir de uma base de dados composta por 292 amostras de crianças com ou sem autismo de um banco de dados público[2], com validação cruzada (*k-fold*) e normalização *z-score*. Os resultados foram analisados consi-

[1] https://www.kaggle.com/luzrello/dyslexia

[2] https://archive.ics.uci.edu/ml/machine-learning-databases/00419/

derando as características e os comportamentos diferentes das RNAs utilizadas, atingindo-se taxas médias de acerto para a acurácia, com 86,26% para a MLP e 83,12% para a RBF.

Os autores desse estudo trazem como diferencial com relação aos demais trabalhos de [Asvestopoulou et al. 2019], [Ceravolo et al. 2019], [Araujo et al. 2019], [de Brito et al. 2020a], a utilização de técnicas de aprendizagem de máquina como as métricas de avaliação (acurácia, sensibilidade, especificidade e *f1-score*), normalização (*z-score*), validação cruzada (*k-fold*), SMOTE entre outras, que podem contribuir com o rastreio da dislexia, auxiliando no diagnóstico desse transtorno que atinge várias crianças, jovens e até adultos.

3. O Transtorno da Dislexia

A dislexia é um transtorno de aprendizagem que se caracteriza por dificuldades em ler, interpretar e escrever. Sua causa tem sido pesquisada e várias teorias tentam explicar o porquê da dislexia [Cândido 2013]. Já [Deuschle and Cechella 2009] afirma que a dislexia é um distúrbio específico de leitura, ocasionado pela interrupção ou malformação nas conexões cerebrais que ligam zonas anteriores com zonas mais posteriores do córtex cerebral.

A dislexia pode ser considerada como um transtorno genético e hereditário da linguagem, evidenciado pela dificuldade de decodificar o estímulo escrito ou o símbolo gráfico e geralmente é classificada em diferentes graus: leve, médio e severo [Pain 2012]. Destaca-se por apresentar grandes dificuldades na ligação entre símbolos gráficos e a sua interpretação, às vezes mal reconhecidos, e fonemas, muitas vezes, mal identificados [Mendes et al. 2021].

4. Redes Neurais Artificiais

Essa seção apresenta um resumo sobre as arquiteturas de redes neurais artificiais utilizadas no trabalho como mecanismos de apoio a aprendizagem de máquina.

4.1. *Multilayer Perceptron* (MLP)

Diferindo das redes *Perceptron* convencionais, a MLP apresenta uma camada intermediária escondida, o que a caracteriza como uma rede multicamadas, apresentando uma arquitetura *feedforward*, ou seja, o fluxo do conhecimento segue em uma única direção [Caldas et al. 2020].

4.2. *Radial Basis Function* (RBF)

De acordo com [Assis et al. 2019] as redes RBF utilizam funções de base radiais como funções de ativação ao invés de funções sigmoides como as RNAs tipo *Perceptron*. A camada de entrada conecta os dados de entrada até a camada escondida. A camada escondida realiza o cálculo em cada neurônio da função de base radial dos dados de entrada e propaga o sinal para a camada de saída.

4.3. *Extreme Learning Machine* (ELM)

Segundo [de Oliveira 2017] o objetivo da *extreme learning machine* é promover um modelo de aprendizado unificado para diferentes tipos de arquiteturas de redes. Em sua essência, a ELM é caracterizada por não exigir que os pesos sinápticos dos neurônios de sua camada ocultam sejam otimizados.

5. Classificadores Estatísticos

Esta seção apresenta uma revisão teórica sobre os classificadores estatísticos utilizados nos treinos e testes neste trabalho.

5.1. *Support Vector Machine* (SVM)

De acordo com [Brito and Ling 2019] o SVM é uma técnica de aprendizado de máquina supervisionado que o seu funcionamento básico consiste em encontrar o hiperplano que separa um conjunto de dados de treinamento em duas classes. Neste processo, o SVM busca o hiperplano que maximiza a distância entre as classes.

5.2. *Random Forest* (RF)

O RF é um algoritmo de aprendizagem de máquina baseado em árvores de decisão. Elas são treinadas isoladamente na tentativa de encontrar um modelo para resolver o mesmo problema diminuindo a variância. Mostra-se muito eficiente quando se busca analisar um grande volume de dados [de Alvarenga Júnior 2018].

5.3. *K-Nearest Neighbors* (K-NN)

Segundo [Enriquez et al. 2020] o K-NN é um algoritmo de mineração de dados baseado em aprendizagem supervisionada. Ele apresenta vantagem de fácil compreensão e implementação em plataformas computacionais, podendo ser utilizado em aplicações industriais, pois pode gerar bons resultados em tarefas de classificação com dados bem caracterizados.

6. Materiais e Métodos

Para o emprego dos algoritmos utilizou-se da linguagem *Python* em sua versão 3.7 e um banco de dados público intitulado *Predicting Risk of Dyslexia* obtido através do *link* [3], é composto por 3644 amostras, que foram avaliadas com relação a 98 atributos. Assim esta seção está dividida em duas subseção a saber: (i) base de dados; e (ii) pré-processamento, treinamento e teste.

6.1. Base de Dados

A base de dados foi projetada baseada em 32 exercícios linguísticos apropriados para inclusão em um teste gamificado, o teste foi conduzido com o auxílio de 3.644 participantes, dentre estes, 392 possuíam diagnóstico de dislexia realizado por um profissional.

Os exercícios gamificados foram desenvolvidos usando dois métodos. Em primeiro lugar, alguns exercícios foram baseados em uma análise empírica de erros de escrita por pessoas com dislexia, foram anotados os erros com características linguísticas gerais, bem como com informação fonética e visual.

Em segundo lugar, foram elaborados exercícios de teste para atingir processos cognitivos específicos, diferentes tipos de conhecimento e dificuldades inerentes à leitura. Cada exercício aborda três ou mais dos seguintes indicadores relacionados à dislexia, que são diferentes tipos de habilidades de linguagem, memória operacional, funções executivas e processos perceptivos.

[3] https://www.kaggle.com/luzrello/dyslexia

6.2. Pré-processamento, Treinamento e Teste

A base de dados foi pré-processada, removendo-se amostras com atributos ausentes assim como aquelas que não estavam relacionadas à classificação, houve também o descarte de atributos que apresentavam redundância. Seguiu-se com a normalização utilizando a técnica *z-score* e o balanceamento garantindo a proporcionalidade entre as amostras.

Para o balanceamento de dados foi utilizado do método *SMOTE*, que consiste em gerar dados sintéticos da classe minoritária a partir de vizinhos. Além disso foi utilizado o método de validação cruzada *k-fold*, com o k=10 *folds*. Para a classificação baseada nas arquiteturas utilizadas, salienta-se que foram testados vários hiper parâmetros de forma a ajustá-los através da estratégia *Randomized Search* a atingirem os melhores resultados.

7. Resultados e Discussões

Os resultados obtidos e analisados tiveram como base as taxas médias de acertos de acurácia, sensibilidade, especificidade e *f1-score* adquiridos nos treinamentos e testes realizados com os classificadores, assim como na discussão dos resultados encontrados por meio das amostras do banco de dados trabalhado. A Tabela 1 mostra os hiperparâmetros testados que obtiveram os melhores resultados com os respectivos classificadores.

Tabela 1. Hiperparâmetros Classificadores

Classificador	Hiperparâmetros
MLP	2 camadas ocultas com 15 neurônios cada, função de ativação: tangente hiperbólica, *learning rate* $= 1$, *batch size* $= 0,8$ e, número de *epochs* $= 1200$
RBF	2 camadas ocultas com 15 neurônios cada
ELM	1 camada oculta com 45 neurônios
SVM	*kernel* linear, constante de relaxamento $C = 0,91$ e, *KernelScale*: 2, *one-vs-one*
RF	número de *seeds* $= 2$, número de *trees*$= 600$
K-NN	$K = 75$

A Tabela 2 traz os resultados obtidos a partir da utilização de vários classificadores e algumas métricas de avaliação. Com base na mesma, pode-se obervar que o classificador SVM obteve a melhor taxa média de acerto para a acurácia em relação aos demais com 98,03% , já a MLP obteve o pior resultado com 90,80%.

Tabela 2. Taxa Média de Acertos Por Modelo

Métricas	MLP	RBF	ELM	SVM	RF	K-NN
Acurácia	$90,80\%$	$92,85\%$	$92,79\%$	$98,03\%$	$96,61\%$	$91,65\%$
Sensibilidade	$97,28\%$	$97,17\%$	$90,97\%$	$96,37\%$	$98,27\%$	$90,97\%$
Especificidade	$85,79\%$	$89,98\%$	$91,79\%$	$99,21\%$	$95,51\%$	$91,79\%$
F1-score	$91,12\%$	$92,84\%$	$92,84\%$	$97,77\%$	$95,87\%$	$91,26\%$

Para a sensibilidade o classificador RF obteve 98,27% apresentando-se como o melhor resultado frente aos demais, por outro lado pode-se perceber que o pior resultado foi adquirido com o ELM e o K-NN ambos com 90,97% taxa média de acerto.

Com a métrica especificidade, percebeu-se que o melhor resultado atingido foi com o classificador SVM com 99,21% de taxa média frente aos demais, com o MLP atingindo o pior resultado com 85,79%.

Por fim com a métrica *f1-score* observou-se que os classificadores testados atingiram resultados acima dos 90% em comparação as com demais, com ênfase para o melhor resultado com a SVM atingindo 97,77% e 91,12% para a MLP como o pior resultado. Salienta-se ainda que a melhor média geral entre todas as métricas testadas foi da acurácia com 93,78%.

8. Considerações Finais e Trabalhos Futuros

Este trabalho teve como finalidade realizar um estudo comparativo entre modelos neurais e os classificadores estatísticos. Para os testes foram utilizados dados reais gerando resultados que podem ser utilizados por profissionais da área de educação como os psicopedagogos.

Nos resultados gerados percebeu-se que o classificador SVM de forma geral obteve as melhores médias de acertos para as métricas de avaliação utilizadas com ênfase para a acurácias, atingindo 98,03%, especificidade com 99,21% e sensibilidade e *f1-score* com 97,77%. Para a sensibilidade percebeu-se que RF obteve o melhor desempenho com 98,27%, por outro lado a MLP atingiu os piores na acurácia com 90,80%, especificadade 85,79% e *f1-score* com 91,12%.

Como trabalho futuro é sugerido a realização de um estudo com outras técnicas de ICA, como as redes de aprendizagem profunda comparando com os resultados atingidos pelos classifcadores utilizados neste trabalho a fim de verificar qual abordagem melhor se aplica ao problema proposto.

Referências

Araujo, M. D. A., Moreira, L. Y. M. R., and de Brito, R. X. (2019). Modelo computacional com fuzzy como recurso auxiliador na predição da esquizofrenia em adultos. In *Anais da VII Escola Regional de Computação Aplicada à Saúde*, pages 199–204. SBC.

Assis, A. D., Coutinho, T. M., Torres, L. C., and Braga, A. P. (2019). Filtragem linear e abordagem geométrica baseada em grafos para o ajuste de parâmetros e reduçao de complexidade de redes neurais rbf.

Asvestopoulou, T., Manousaki, V., Psistakis, A., Smyrnakis, I., Andreadakis, V., Aslanides, I. M., and Papadopouli, M. (2019). Dyslexml: Screening tool for dyslexia using machine learning. *arXiv preprint arXiv:1903.06274*.

Brito, D. and Ling, L. L. (2019). Hand vein biometric recognition approaches based on wavelet, svm, articial neural network and image registration. *Inteligencia Artificial*, 22(63):101–120.

Caldas, G., Maia, E. J. Q., Lima, N. C. d. A., et al. (2020). Rede mlp para auxílio ao diagnóstico do transtorno do espectro autista em crianças e adolescentes. *Anais do Encontro de Computação do Oeste Potiguar ECOP/UFERSA (ISSN 2526-7574)*, (4).

Cândido, E. d. C. (2013). Psicopedagogia para a dislexia nas séries iniciais do ensino fundamental. *Especialização em Psicopedagogia. Universidade Cândido Mendes. Rio de Janeiro: RJ*.

Ceravolo, I., Brasil, A., and Komati, K. (2019). Classificação de dislexia a partir de movimentos oculares durante a leitura usando aprendizado de máquina e wavelets. In *Anais do XVI Encontro Nacional de Inteligência Artificial e Computacional*, pages 880–891. SBC.

de Alvarenga Júnior, W. J. (2018). Métodos de otimização hiperparamétrica: um estudo comparativo utilizando árvores de decisão e florestas aleatórias na classificação binária.

de Brito, R., Fernandes, C. A., and de Sousa Ximenes, J. (2020a). Avaliação de rnas durante treinamento supervisionado para classificação de adolescentes com autismo. In *Anais da VIII Escola Regional de Computação do Ceará, Maranhão e Piauí*, pages 53–60, Porto Alegre, RS, Brasil. SBC.

de Brito, R. X., Fernandes, C. A. R., and Amora, M. A. B. (2020b). Análise de desempenho com redes neurais artificiais, arquiteturas mlp e rbf para um problema de classificação de crianças com autismo. *iSys-Revista Brasileira de Sistemas de Informação*, 13(1):60–76.

de Oliveira, L. F. d. R. (2017). *Comparação de Desempenho Entre os Modelos Neurais Ageis ELM e WISARD*. PhD thesis, Universidade Federal do Rio de Janeiro.

Deuschle, V. P. and Cechella, C. (2009). O déficit em consciência fonológica e sua relação com a dislexia: diagnóstico e intervenção. *Revista Cefac*, 11:194–200.

Enriquez, A. R. S. et al. (2020). Diagnóstico de falhas em transformadores de potência através de análise de gases dissolvidos usando rede neural artificial.

Jerônimo, S. D. and Duarte, F. J. (2016). Dificuldade de aprendizagem: A importância do fazer do professor para inclusão do aluno disléxico. *Editora Realize*.

Mendes, H. M. D., Mendes, N. R. S., and Soares, J. C. (2021). Dislexia: Dificuldades de aprendizagem-um olhar sobre a dislexia. *RECIMA21-Revista Científica Multidisciplinar-ISSN 2675-6218*, 2(3):337–350.

Pain, S. (2012). Diagnóstico e tratamento dos problemas de aprendizagem. In *Diagnóstico e tratamento dos problemas de aprendizagem*, pages 85–85.