JHI JOURNAL OF HEALTH INFORMATICS

Geração de dados sintéticos para classificação de disléxicos por meio de aprendizado de máquina

Synthetic data generation for classification of dyslexics by machine learning

Generación de datos sintéticos para clasificación de disléxicos mediante aprendizaje automático

Antonio Carlos da Silva Junior[1], Emanuela Cristina Ramos Gonçalves[2], Paulo Schor[3], Martina Navarro[4], Felipe Mancini[5]

Descritores: Dislexia; Aprendizado de Máquina; Leitura

RESUMO

Objetivo: Este estudo pretende aplicar a técnica de geração de dados sintéticos com auxílio de técnicas de limpeza de dados para a classificação de disléxicos e não - disléxicos. **Método:** Os outliers foram selecionados por especialista. Foi feito uma geração sintética de dados. para cada um de cinco algoritmos foram selecionados características com busca exaustiva. Cada algoritmo foi executado com as características selecionadas e então suas curvas de calibração foram comparadas. **Resultados:** A regressão logística se destacou como o melhor algoritmo, apresentando o resultado de 99% de acurácia e área sob a curva ROC de 0,999, além de ter obtido a melhor curva de calibração **Conclusão:** O uso da geração sintética de dados e seleção de características foram capazes de fazer todos os algoritmos avaliados obterem ótimos resultados na classificação de disléxicos e não disléxicos. A regressão logística foi selecionado como melhor algoritmo para classificação de disléxicos.

Keywords: Dyslexia; Machine Learning; Reading

ABSTRACT

Objective: This study aims to apply the synthetic data generation technique with the aid of data cleaning techniques for the classification of dyslexics and non - dyslexics. **Method:** Outliers were selected by specialist. Synthetic of data Generated. For each of five algorithms, characteristics were selected with exhaustive search. Each algorithm was executed with the selected characteristics and then their calibration curves were compared. **Results:** Logistic regression presented the best results with 99% accuracy and area under the ROC curve of 0.999, besides obtaining the best calibration curve. **Conclusion:** The use of synthetic data generation and feature selection were able to make all algorithms achieve excellent results in the classification of dyslexic and non - dyslexic. Logistic regression was selected as the best algorithm for dyslexic classification.

Descriptores: Dislexia; Aprendizaje Automático; Lectura

RESUMEN

Objetivo: Este estudio tiene como objetivo aplicar la técnica de generación de datos sintéticos con la ayuda de técnicas de limpieza de datos para la clasificación de disléxicos y no disléxicos. **Método:** los valores atípicos fueron seleccionados por especialistas. Se realizó una generación sintética de datos. Para cada uno de los cinco algoritmos, se seleccionaron características con búsqueda exhaustiva. Cada algoritmo se ejecutó con las características seleccionadas y luego se compararon sus curvas de calibración. **Resultados:** La regresión logística se destacó como el mejor algoritmo, presentando el resultado del 99% de precisión y área bajo la curva ROC de 0.999, además de obtener la mejor curva de calibración. **Conclusión:** El uso de la generación de datos sintéticos y la selección de Estas características lograron que todos los algoritmos evaluados obtuvieron excelentes resultados en la clasificación de disléxicos y no disléxicos. Se seleccionó la regresión logística como el mejor algoritmo para la clasificación disléxica.

[1] Mestre em Ciências, Programa de Pós-Graduação em Gestão e Informática em Saúde, Escola Paulista de Medicina, Universidade Federal de São Paulo, São Paulo (SP), Brasil.

[2] Doutora em Ciências, Programa de Pós-Graduação em Oftalmologia e Ciências Visuais, Escola Paulista de Medicina, Universidade Federal de São Paulo, São Paulo (SP), Brasil.

[3] Livre Docente, Departamento de Oftalmologia, Escola Paulista de Medicina, Universidade Federal de São Paulo, São Paulo (SP), Brasil.

[4] Doutora em Ciências, Department of Sport and Exercise Science, University of Portsmouth, Portsmouth, Hampshire, United Kingdom.

[5] Doutor em Ciências, Universidade Aberta do Brasil, Universidade Federal de São Paulo, São Paulo (SP), Brasil.

Autor Coorespondente: **Antonio Carlos da Silva Junior**
e-mail: **acsjunior@unifesp.br**

Artigo recebido: **15/04/2020**
Aprovado: **20/12/2020**

INTRODUÇÃO

Dislexia do desenvolvimento é uma disfunção de origem neurológica que afeta a habilidade de leitura caracterizado por uma dificuldade significativa e persistente de aprendizagem escolar relacionado com essa tarefa, afetando a acurácia de leitura, fluência de leitura e compreensão leitora. O que pode ser indicado por uma performance marcantemente inferior ao esperado pela idade cronológica mesmo o portador de dislexia tendo recebido educação adequada, não apresentar nenhuma desordem de desenvolvimento intelectual, não ter problemas de visão e nem carência de proficiência na linguagem ou adversidade psicossocial[1–3]. A dificuldade de leitura tem um grande impacto na vida dos disléxicos. Levando a uma dificuldade de compreensão da leitura, e por conseguinte, no aumento do vocabulário[3] e no aproveitamento escolar o que pode levar a muitos problemas na vida adulta[4].

O diagnóstico da dislexia é complexo, demorado e custoso pois o mesmo é feito por exclusão e tem uma característica multidisciplinar passando por oftalmologistas, psicopedagogos, fonoaudiólogos e psicólogos[1,5]. Em busca de novos indicadores de dislexia, uma nova abordagem começou a ser estudada, avaliar o comportamento ocular durante a leitura, utilizando um equipamento de captura do olhar (Eye Tracker), para capturar dados das fixações e sacadas e variáveis específicas de cada estudo então avaliadas conjuntamente com técnicas de aprendizado de máquina[6-8]. Uma outra métrica menos explorada são as Funções Visuais de Leitura (FVL[3,9] que possui uma vantagem de não necessitar de um Eye tracker para aquisição dos dados desta forma sendo uma alternativa menos onerosa.

A partir de avaliação de FVL foi levantado que a variação de velocidade de leitura medida durante a redução gradual do tamanho de letra entre disléxicos e não disléxicos têm similaridade qualitativa, porém os disléxicos apresentam em média um tamanho de letra maior nos resultados de leitura e uma velocidade de leitura menor[9].

Técnicas de Aprendizado de Máquina (AM) são, geralmente, utilizadas em grande quantidade de dados para realizar inferências no auxílio à tomada de decisões. Porém em algumas situações-problema a quantidade de dados disponíveis é escassa. Especificamente, na área da saúde essa é uma característica marcante, seja por restrição de tempo para coleta, ou por ser um procedimento invasivo, por ser muito caro conseguir novos registros, ou por se tratarem de dados confidenciais[10-11]. Dessa forma, aplicar técnicas de AM em uma pequena base é um desafio devido aos algoritmos desenvolvidos conseguirem melhores resultados com aumento na quantidade de dados no treinamento. Os problemas que podem ocorrer no treinamento de técnicas de AM especialmente quando se possui bases pequenas são o *underfit* e o *overfit*[10,12–15].

Para controlar estes problemas e selecionar uma técnica de AM que se ajuste melhor aos dados e que tenham um bom desempenho, a melhor solução é coletar mais dados. Porém, como apresentado anteriormente, nem sempre isso é possível. Portanto, para usar técnicas de AM em bases pequenas é preciso remover variáveis altamente correlacionadas ou que possuam valores únicos em um atributo e selecionar *Outliers* para garantir uma melhor performance do classificador[12,14,16-17].

Uma outra alternativa para se utilizar bases pequenas é a seleção de características para diminuir o número de dimensões. Para se obter uma boa classificação deve se ter um número de registros em torno de 10 vezes número de dimensões[12]. Além disso, para evitar o problema de overfit pode se utilizar uma técnica de geração sintética de dados para aumentar a quantidade de dados e diminuir a variância obtida pelo classificador.

Em uma busca no pubmed por "(Dyslexia) and (Machine Learning)" retornou 19 resultados, porém nenhum deles utilizou um dataset de FVL nem geração sintética de dados. Dessa maneira, esse estudo tem como objetivo principal aplicar a técnica de geração de dados sintéticos com auxílio de técnicas de limpeza de dados , utilizada em casos em que a coleta de amostragem é limitada para classificar disléxicos e não-disléxicos a partir de dados de FVL.

MÉTODO

O desenho metodológico deste estudo tem como objetivo melhorar a classificação de dados de uma base de dados com dados de FVL com poucos registros usando sensibilidade, acurácia e área sob a curva ROC (AUC) como métricas de avaliação.

Este estudo analisará dados provenientes do grupo de pesquisa Bioengenharia Ocular (http://dgp.cnpq.br/dgp/espelhogrupo/5894644663535064) que foram coletados da seguinte maneira.

Coleta de dados

Nesta pesquisa foram convidados voluntários tanto do ambulatório de transtornos de leitura e escrita do núcleo de ensino, assistência e pesquisa em escrita e leitura da UNIFESP e da Associação Brasileira de Dislexia, durante 2 anos. Neste período se apresentaram 18 não disléxicos (M = 23.67 anos, DP = 8.84) e 21 disléxicos M = 15.05 anos, DP = 6.55) que passaram no critério de inclusão que consistiu de uma avaliação da diferença de acuidade visual (AV) interocular menor do que 3 linhas, AV binocular melhor ou igual a 0.3 logMAR[18] e sem antecedentes oftalmológicos (e.g., estrabismo ou cirurgia prévia).

Neste estudo, foi utilizado o *Minnesota Low Vision Reading Test* adaptada para o português brasileiro (MNREAD-P)[19] protocolo validado para a avaliação de velocidade de leitura e FVL, geralmente utilizada para a adaptação de auxílios ópticos em pacientes com baixa visão. A Figura 1 exibe a curva de velocidade por tamanho de letras, onde o MVL consiste na maior velocidade de leitura atingida, o LMVL tamanho da letra durante MVL, o TCL é o ponto em que a velocidade de leitura começa a cair, VTCL é a velocidade de leitura obtida no TCL e AL é o menor tamanho de letra que o sujeito conseguiu ler.

Neste estudo, a tabela de leitura MNREAD-P foi transformada e apresentada em formato digital. As tabela foram apresentados aos participantes de forma randômica seguindo o formato MNREAD – P3L - sentença distribuída e apresentada em 3 linhas diferentes (Figura 2).

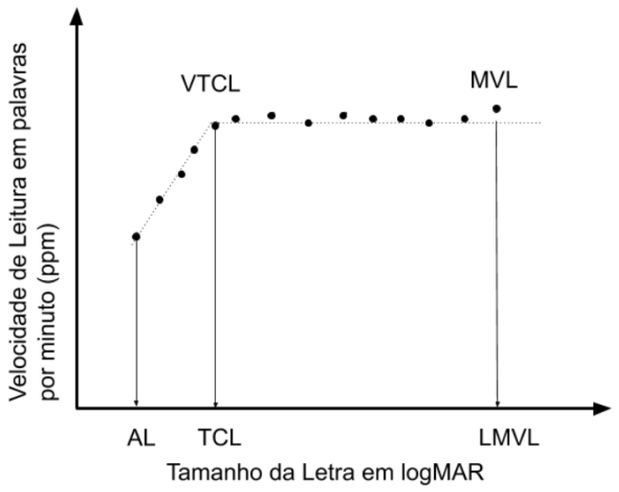

Figura 1 - A curva de velocidades de leitura em ppm por tamanho de letra em logMAR baseado em O'Brien et al[9]

A vovó fez um bolo
de chocolate gelado
e eu levei de lanche

Figura 2 - Representação de uma sentença de três linha do MNREAD-P3L

Para apresentar os estímulos foi utilizado o Psychopy[I] e através do mesmo aplicativo foram gravados os áudios de cada leitura feita pelos voluntários da pesquisa.

Os estímulos consistem de 13 tentativas com sentenças diferentes, porém equivalentes em dificuldade[19]. Das 13 tentativas, as 2 iniciais foram tentativas de familiarização de tamanho 1.0 logMAR, e as 11 tentativas experimentais subsequentes apresentavam tamanhos de fonte variando de 1.0 a 0.0 logMAR, apresentadas de forma decrescente. A variação de tamanho foi de 0.1 logMAR por tentativa, e adotada a fonte Times New Roman.

Agrupando estas variáveis com os tamanhos de letras resultaram nas variáveis AL: Acuidade de Leitura que consiste no menor tamanho de letra lido; VL: Velocidade de leitura em palavras por minuto (ppm); LMVL: Linha máxima de leitura em logMAR; MVL: Máxima velocidade de leitura em ppm; VTCL: Velocidade de leitura no tamanho crítico de letra em ppm. Esta coleta resultou em uma base com 10 variáveis.

Limpeza da base para geração sintética de dados e aplicação de técnicas de aprendizado de máquina

Para utilizar uma base pequena para aplicação de técnicas de ML ela deve estar livre de ruídos e os dados devem representar bem os grupos que fazem parte[20], especialmente que será utilizado uma técnica de geração sintética de dados que expandirá esta representação para outros registros[14-15]. Portanto foi feito uma seleção de outliers com o objetivo de evitar os principais problemas de classificação o *overfit* e o *underfit*.

Overfit e Underfit

Overfit e *underfit* são dois desafios a serem enfrentados que possuem uma maior tendência de ocorrer com base de dados pequenas. O *underfit* acontece quando o algoritmo não alcança uma acurácia satisfatória, os dados apresentados no treino gera uma performance inferior ao esperado. O *overfit* ocorre quando o algoritmo se ajusta aos dados do treino e possui uma performance insatisfatória ao generalizar para dados não utilizados no treino[15–17].

Remoção de outliers

Outliers podem impactar negativamente o resultado dos algoritmos, especialmente dos mais simples, como os de regressão.

Para avaliação dos *outliers*, foi utilizado o software JASP[II] para criar os boxplots com o objetivo de avaliá-los junto com profissional da área e remover os que não forem uma forte representação de disléxicos ou não disléxicos[17,21-22].

Geração sintética de dados

O *Synthetic Minority Over-sampling Technique* (SMOTE)[14] é uma técnica que gera uma quantidade pré-determinada de dados sintéticos em uma das classes do banco de dados, de acordo com um número informado de registros. É selecionado aleatoriamente um ponto intermediário entre dois objetos. Esse ponto é baseado no valor do atributo dos objetos envolvidos. Para execução do algoritmo SMOTE, foi utilizado o software WEKA[III].

Seleção de atributos

A seleção de atributos reduz o número de dimensões da base de dados de forma que encontre o número de atributos que dê o melhor resultado de classificação[12,17]. Depois de gerado os dados, será utilizado o algoritmo WrapperSubsetEval que avalia grupos de variáveis selecionados por um método selecionado, então avalia de acordo com um indicador de performance e gera a estimativa usando uma validação cruzada. Para cada algoritmo foi selecionado o conjunto de variáveis que obteve o melhor resultado da área sob a curva ROC avaliado com a avaliação cruzada leave-one-out[11,23].

Aplicação de técnicas de ML

Cada algoritmo foi executado com o seu subconjunto de variáveis selecionadas utilizando o método validação de performance leave-one-out para então comparar a acurácia, sensibilidade, especificidade e área sob a curva ROC de cada algoritmo[11].

$$Sensibilidade = \frac{VP}{P}$$

onde VP são os verdadeiros positivo, P o total de positivos.

[I] Ferramenta de código livre que permite a aplicação de experimentos de neurociencia,psicologia e psicofísica https://www.psychopy.org/
[II] JASP é uma ferramenta estatistica de código aberto https://jasp-stats.org/
[III] WEKA é uma ferramenta de Mineração de dados desenvolvida pela universidade de waikato https://www.cs.waikato.ac.nz/ml/weka/

$$Especificidade = \frac{VN}{N}$$

onde VN são os verdadeiros negativos, N o total de negativos.

$$Acurácia = \frac{VP + VN}{P + N}$$

onde VP são os verdadeiros positivo, P o total de positivos, VN são os verdadeiros negativos, N o total de negativos.

Validação das técnicas de ML

Sempre que modelos de classificação são utilizados para uma tarefa de predição é desejável que sejam estimados as verdadeiras probabilidades de predição da amostra, especialmente quando os modelos possuam desempenhos próximos, para verificar a aproximação da probabilidade prevista e a probabilidade observada se utiliza a curva de calibração[24]. Para esta tarefa foi utilizado o pacote de curva de calibração do WEKA[25]. Os dados das curvas de calibração foram exportados e importados no Excel para gerar um gráfico unificado com todas as curvas para facilitar a comparação visual entre as curvas.

RESULTADOS

Detecção, avaliação e remoção de outliers

Cada outlier (Figura 3) foi avaliado por profissional da área que coordenou e executou a coleta de dados com o objetivo de avaliar quais destes são representações essenciais de disléxicos e não-disléxicos que permaneçam na base. Assim, o especialista solicitou que fossem mantidos os seguintes sujeitos:

- Sujeito controle 36: Os parâmetros de leitura são considerados de normo-leitores.
- Sujeito disléxico 18: Representante importante dos disléxicos. Era esperado que houvesse outros disléxicos com grande dificuldade como ele, tanto que 3 disléxicos não foram incluídos na base pois não conseguiram terminar o teste, pois ficaram cansados de tentar ler durante a etapa de familiarização.

Geração sintética de dados com SMOTE

Após a remoção dos outliers que não representavam algo importante no grupo o algoritmo SMOTE foi

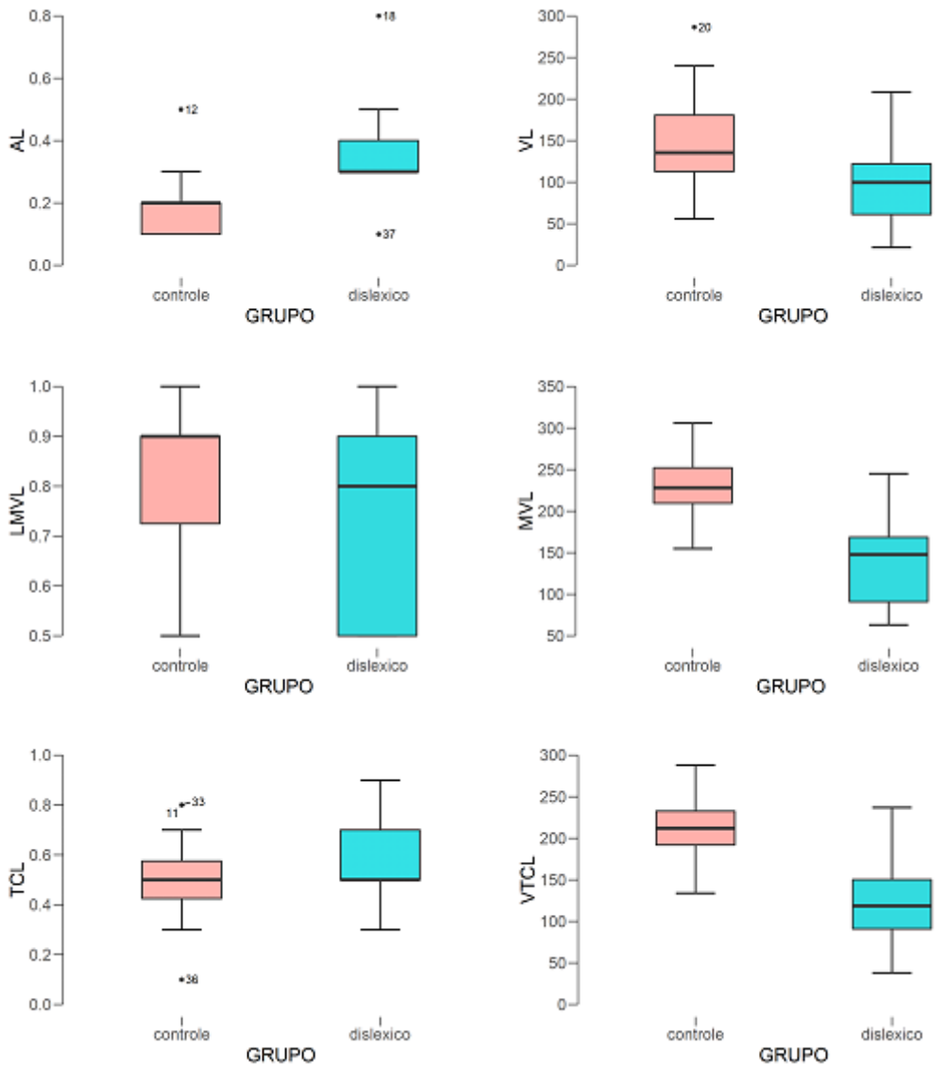

Figura 3 - Boxplot com outliers

utilizado para aumentar a quantidade de registros para 100 em cada grupo, para o mesmo foi parametrizado para selecionar 2 vizinhos próximos para calcular cada novo registro e para aumentar em 400% o grupo disléxico e em 615% o grupo não-disléxico.

Depois de aplicar o algoritmo SMOTE para balancear os 2 grupos em 100 registros, foi aplicado o algoritmo de seleção de caracteristicas WrapperSubsetEval com busca exaustiva resultando na Tabela 1

Após a seleção dos atributos cada algoritmo foi executado com seu próprio conjunto de atributos resultando na comparação de desempenho apresentado na Tabela 2.

Todos os algoritmos selecionados tiveram bons desempenhos com o regressão logística liderando com a maior AUC por uma diferença de 0,002 da Rede bayesiana. Então uma execução com e sem o SMOTE foi feita para verificar o efeito total da aplicação dessa geração sintética de dados no resultado final que pode ser visto na Tabela 3.

Em seguida da execução de cada algoritmo foi gerado a curva de calibração no weka, então os dados foram e compilados em 1 único gráfico que pode ser visto na Figura 4

DISCUSSÃO

A avaliação dos outliers foi importante para remover aqueles que não representavam bem os grupo que faziam parte que poderiam impactar negativamente na classificação e ainda poderiam influenciar o SMOTE em gerar mais dados não representativos, por outro lado, manter 2 deles que eram representantes importantes dos seus respectivos grupos foi também importante na geração de dados sintético.

Observando a Tabela 1 é possível averiguar que os

Tabela 1 - Atributos selecionados para cada algoritmo levando em conta a melhor AUC.

Algorítmo	Atributos selecionados
C4.5	VL, MVL, TCL
Regressão Logística	AL, VL, LMVL, MVL, VTCL
SVM	AL, LMVL, TCL
Rede Bayesiana	AL, LMVL, TCL
Naive Bayes	AL,VL

Tabela 2 - Tabela comparativa dos algorítimos sobre, sensibilidade, especificidade e AUC

Algorítmo	Sensibilidade	Especificidade	Acurácia	AUC
C4.5	0,960	0,950	95,5%	0,978
Regressão Logística	0,990	0,990	99%	0,999
SVM	0,960	0,940	95%	0,950
Rede Bayesiana	1,000	0,980	99%	0,997
Naive Bayes	1,000	0,960	98%	0,996

Tabela 3 - Comparação de desempenho de uma regressão logística com e sem a geração sintética de dados

	Sensibilidade	Especificidade	Acurácia	AUC
Com SMOTE	0,990	0,990	99%	0,999
Sem SMOTE	0,900	0,857	88,23%	0,975

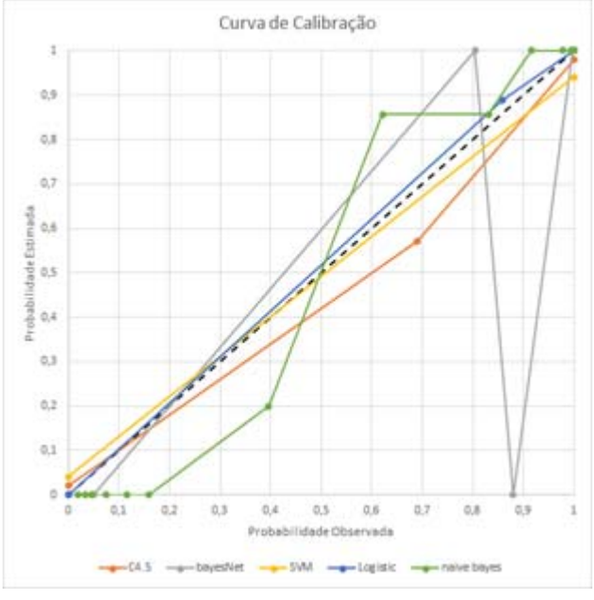

Figura 4 - Curvas de calibração dos algoritmos, onde a linha pontilhada define a calibração perfeita, quanto mais próximo desta linha melhor calibrado está o algoritmo

tamanhos de letras (AL, LMVL, TCL) foram importantes na classificação de todos algoritmos. Além disto, cabe ressaltar a importância da AL para classificação de disléxicos, tanto que a mesmo foi selecionado no melhor conjunto de 4 dos 5 algoritmos.

Todos algoritmos obtiveram uma classificação acima de 90% de acurácia e uma AUC superior à 0,90 sendo todos eles bons candidatos para criação de um classificador automático de disléxicos. Porém, o classificador que obteve a melhor AUC (0,999) foi a regressão logística, este algoritmo foi bastante favorecido pelo processo de seleção de outliers , pois o mesmo é mais impactado por eles, isto o fez tem um resultado excepcional.

Na comparação do ganho de performance do melhor classificador (regressão logística) com a aplicação da geração sintética de dados apresentado na Tabela 3 foi de 10,77 pontos percentuais de acurácia e 0,025 na AUC, demonstrando o impacto do SMOTE na melhora de performance da classificação.

Observado a curva de calibração dos algoritmos é possível observar que a regressão logística foi a que mais se aproximou de uma calibração perfeita o que lhe dá uma melhor confiança para aplicá-lo em dados desconhecidos. Já o algoritmo bayesNet que obteve uma classificação de 99% e uma AUC de 0,997 apresentou uma curva de calibração distante da calibração perfeita o que indica que o mesmo não garante a mesma performance em dados desconhecidos não sendo indicado para classificação de disléxicos com base em FVL.

Cabe ressaltar que a geração sintética de dados mitiga os problemas de classificação (ex. Overfitting) porem não os elimina pois esta técnica melhora os espectro de dados dentros dos registros presentes na base e não gera dados fora deste conjunto.

CONCLUSÃO

O uso de geração sintética de dados possibilita a extrapolação dos resultados para a população de forma mais robusta e demonstra ser uma poderosa ferramenta em casos onde a aquisição de novos registros seja inviável.

Todos os algoritmos avaliados para fazer a classificação de disléxicos obtiveram uma acurácia na classificação superior a 0,95% sendo todos aparentemente muitos bons, com destaque para a regressão logística que obteve os maiores valores (acurácia de 99% e AUC de 0,999). Quando confrontados com a curva de calibração para avaliar a chance do algoritmo ter uma mesma performance em dados desconhecidos a regressão logística se demonstrou bem mais próxima da calibração perfeita que os outros algoritmos reforçando seu bom desempenho na classificação de disléxicos. Observando o segundo melhor algoritmo no quesito AUC, o algoritmo de rede bayesiana obteve uma curva de calibração bem longe do ideal não sendo confiável para criação de um classificador automático de dislexia com base em FLV.

REFERÊNCIAS

1. ICD-11 - Mortality and Morbidity Statistics [Internet]. [cited 2019 Aug 8]. Available from: https://icd.who.int/browse11/l-m/en#/http://id.who.int/icd/entity/1008636089
2. Lyon GR, Shaywitz SE, Shaywitz BA. A definition of dyslexia. Ann of Dyslexia. 2003 Jan 1;53(1):1–14.
3. Silva NML de L e, Pedroso FS. A Prevalência da Dislexia em Alunos do Ensino Fundamental de Escolas Particulares. A Prevalência da Dislexia em Alunos do Ensino Fundamental de Escolas Particulares. 2004;1.
4. Snow PC. Elizabeth Usher Memorial Lecture: Language is literacy is language - Positioning speech-language pathology in education policy, practice, paradigms and polemics. International Journal of Speech-Language Pathology. 2016 May 3;18(3):216–28.
5. Como é feito o Diagnóstico? – ABD | Associação Brasileira de Dislexia [Internet]. [cited 2016 Nov 2]. Available from: http://www.dislexia.org.br/como-e-feito-o-diagnostico/
6. Rello L, Ballesteros M. Detecting Readers with Dyslexia Using Machine Learning with Eye Tracking Measures. In: Proceedings of the 12th Web for All Conference [Internet]. New York, NY, USA: ACM; 2015 [cited 2017 Jun 4]. p. 16:1–16:8. (W4A '15). Available from: http://doi.acm.org/10.1145/2745555.2746644
7. Lustig J. Identifying dyslectic gaze pattern/ : Comparison of methods for identifying dyslectic readers based on eye movement patterns [Internet]. 2016 [cited 2016 Sep 18]. Available from: http://kth.diva-portal.org/smash/record.jsf?pid=diva2%3A955646&dswid=-8238
8. Benfatto MN, Seimyr GÖ, Ygge J, Pansell T, Rydberg A, Jacobson C. Screening for Dyslexia Using Eye Tracking during Reading. PLOS ONE. 2016 Dec 9;11(12):e0165508.
9. O'Brien BA, Mansfield JS, Legge GE. The effect of print size on reading speed in dyslexia. J Res Read. 2005 Aug;28(3):332–49.
10. Pasini A. Artificial neural networks for small dataset analysis. J Thorac Dis. 2015 May;7(5):953–60.
11. Santos HG dos. Comparação da performance de algoritmos de machine learning para a análise preditiva em saúde pública e medicina [Internet] [text]. Universidade de São Paulo; 2018 [cited 2019 Sep 13]. Available from: http://www.teses.usp.br/teses/disponiveis/6/6141/tde-09102018-132826/
12. Raudys SJ, Jain AK. Small sample size effects in statistical pattern recognition: recommendations for practitioners. IEEE Trans Pattern Anal Machine Intell. 1991 Mar;13(3):252–64.
13. Wang T, Cao X, Xia T, Yang Z. Solving the small sample size problem in protein subcellular localization prediction. In: 2012 5th International Conference on BioMedical Engineering and Informatics [Internet]. Chongqing, China: IEEE; 2012 [cited 2019 Aug 16]. p. 915–8. Available from: http://ieeexplore.ieee.org/document/6513152/
14. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. jair. 2002 Jun 1;16:321–57.
15. Sun J, Lang J, Fujita H, Li H. Imbalanced enterprise credit evaluation with DTE-SBD: Decision tree ensemble based on SMOTE and bagging with differentiated sampling rates. Information Sciences. 2018 Jan 1;425:76–91.
16. Bishop C. Pattern Recognition and Machine Learning. New York: Springer; 2007. 738 p.
17. Corrales DC, Corrales JC, Ledezma A. How to Address the Data Quality Issues in Regression Models: A Guided Process for Data Cleaning. Symmetry. 2018 Apr;10(4):99.
18. Bailey IL, Lovie-Kitchin JE. Visual acuity testing. From the laboratory to the clinic. Vision Research. 2013 Sep 20;90:2–9.
19. Castro CTM de, Kallie CS, Salomão SR. Development and validation of the MNREAD reading acuity chart in Portuguese. Arquivos Brasileiros de Oftalmologia. 2005 Dec;68(6):777–83.
20. Aggarwal CC. Data Classification: Algorithms and Applications. CRC Press; 2014. 704 p.

21. Rousseeuw PJ, Leroy AM. Robust Regression and Outlier Detection. John Wiley & Sons; 2005. 354 p.
22. Van den Broeck J, Argeseanu Cunningham S, Eeckels R, Herbst K. Data Cleaning: Detecting, Diagnosing, and Editing Data Abnormalities. PLoS Med [Internet]. 2005 Oct [cited 2019 Sep 23];2(10). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1198040/
23. Falcão AEJ, Mancini F, Costa TM da, Hummel AD, Teixeira FO, Sigulem D, et al. InDeCS: Método automatizado de classificação de páginas Web de Saúde usando mineração de texto e Descritores em Ciências da Saúde (DeCS). Journal of Health Informatics [Internet]. 2009 Jul 27 [cited 2021 Feb 22];1(1). Available from: http://www.jhi-sbis.saude.ws/ojs-jhi/index.php/jhi-sbis/article/view/1
24. Kuhn M, Johnson K. Applied Predictive Modeling [Internet]. New York: Springer-Verlag; 2013 [cited 2019 Sep 24]. Available from: https://www.springer.com/gp/book/9781461468486
25. Frank E. calibrationCurve: VisualizePlugin for plotting class probability calibration curves [Internet]. [cited 2019 Sep 24]. Available from: http://weka.sourceforge.net/packageMetaData/calibrationCurve/index.html