

Antonio Carlos da Silva Junior

**RECONHECIMENTO AUTOMÁTICO DE PADRÕES EM DISLEXIA:
UMA ABORDAGEM BASEADA EM FUNÇÕES VISUAIS DE LEITURA
E APRENDIZADO DE MÁQUINA**

Dissertação apresentada à Universidade
Federal de São Paulo – Escola Paulista
de Medicina para obtenção do título de
Mestre em Ciências.

São Paulo

2019

Antonio Carlos da Silva Junior

**RECONHECIMENTO AUTOMÁTICO DE PADRÕES EM DISLEXIA:
UMA ABORDAGEM BASEADA EM FUNÇÕES VISUAIS DE LEITURA
E APRENDIZADO DE MÁQUINA**

Dissertação apresentada à Universidade Federal de São Paulo – Escola Paulista de Medicina para obtenção do título de Mestre em Ciências, área de Gestão e Informática em Saúde.

Orientador:

Prof. Dr. Felipe Mancini

Coorientadores:

Prof. Dr. Paulo Schor

Dr^a. Emanuela C. R. Gonçalves

São Paulo

2019

Ficha catalográfica elaborada pela Biblioteca Prof. Antonio Rubino de Azevedo,
Campus São Paulo da Universidade Federal de São Paulo, com os dados fornecidos pelo(a)
autor(a)

Junior, Antonio Carlos da Silva

RECONHECIMENTO AUTOMÁTICO DE PADRÕES EM DISLEXIA: UMA ABORDAGEM BASEADA EM FUNÇÕES VISUAIS DE LEITURA E APRENDIZADO DE MÁQUINA / Antonio Carlos da Silva Junior. - São Paulo, 2019. IX, 44f.

Dissertação (Mestrado) - Universidade Federal de São Paulo, Escola Paulista de Medicina. Programa de Pós-Graduação em Gestão e Informática em Saúde.

Título em inglês: Automatic recognition of dyslexic patterns: an approach based on visual reading functions and machine learning

1. Aprendizado de Máquina. Dislexia. Extração de Características. Geração sintética de Dados. Classificação. 2. Aprendizado de Máquina. 3. Dislexia. 4. Aprendizado de Máquina Não Supervisionado. 5. Aprendizado de Máquina Supervisionado. 6. Modelos Logísticos.

UNIVERSIDADE FEDERAL DE SÃO PAULO
ESCOLA PAULISTA DE MEDICINA
PROGRAMA DE PÓS-GRADUAÇÃO EM
GESTÃO E INFORMÁTICA EM SAÚDE

Coordenadora da Câmara de Pós-graduação e Pesquisa da Escola Paulista de Medicina: Profa. Dra. Monica Levy Andersen, livre docente

Coordenador do programa: Prof. Dr. Ivan Torres Pisa, livre docente

Antonio Carlos da Silva Junior

**RECONHECIMENTO AUTOMÁTICO DE PADRÕES EM
DISLEXIA: UMA ABORDAGEM BASEADA EM FUNÇÕES
VISUAIS DE LEITURA E APRENDIZADO DE MÁQUINA**

Presidente da banca:

Prof. Dr. Felipe Mancini

Banca examinadora:

Prof. Dr. Olival Cardoso do Lago

Prof. Dr. Rodrigo Filev Maia

Prof. Dr. Paulo Bandiera Paiva

DEDICATÓRIA

Dedico este trabalho ao meus pais, Valdecy e Antonio e a minha esposa Renata.

AGRADECIMENTOS

Gostaria de agradecer ao meus pais, pela paciência e pelo apoio no decorrer desta jornada.

Gostaria de agradecer a minha esposa Renata por sua paciência durante este período e por estar sempre do meu lado e por me apoiar nesse meu sonho de me tornar pesquisador.

Gostaria de agradecer ao meu orientador prof. Dr. Felipe Mancini por estar sempre presente desde o início, por sua amizade e por me trazer as oportunidades que eu precisava para viabilizar este trabalho.

Gostaria de agradecer a Dra. Emanuela, que não só me cedeu os dados de sua pesquisa como fomos juntos nesta jornada de decifrar os enigmas do olhar de disléxicos.

Gostaria de agradecer ao meu sogro e sogra que sempre me apoiaram nos momentos que eu estava mais precisando de alguém para conversar e por sempre saber o momento que eu devia me afastar para poder voltar a pesquisa com fôlego renovado.

Agradeço também a Wilma Honório que me ajudou demais na finalização deste trabalho por sua amizade e pelas inúmeras vezes que me ajudou a acreditar que eu conseguiria chegar até o fim.

Também agradeço a compreensão dos meus amigos e familiares que me afastei neste período.

Também agradeço a todos meus colegas de mestrado que estavam nessa empreitada e compartilharam comigo as alegrias e desesperos da vida acadêmica.

“A menos que modifiquemos a nossa maneira de pensar, não seremos

*capazes de resolver os problemas causados pela forma
como nos acostumamos a ver o mundo”*

(Albert Einstein)

SUMÁRIO

Dedicatória				
			6	Agradecim
entos8	Lista			de
Figuras14	Lista			de
Quadros17	Lista	de	abreviaturas	e
siglas				
			18	RES
UMO				
			22	ABST
RAC			241	
INTRODUÇÃO			262	
OBJETIVOS302.1				Objetivo
Geral312.2				Objetivos
Específicos313				REFERECIAL
TEÓRICO323.1		Aprendizado		de
Máquina333.2	Dislexia	e	Aprendizagem	de
Máquina333.3		<i>Overfitting</i>		e
<i>Underfitting</i> 373.4		Remoção		de
ruído393.5	Geração	de	Dados	Sintéticos
(SMOTE)393.6		Seleção		de
Características403.7		Extração		de
Características				413.8
Algoritmos				
				4
13.8.1	Mapas			Auto-Organizáveis
(SOM)				
				4
13.8.2	Árvore			de
Decisão				
				4
23.8.3	Árvore	de	decisão:	Árvore
				de

Aleatória

4

33.8.4Árvore

de

decisão:

C4.5

4

33.8.5Regressão:

Regressão

Logística

4

43.8.6Kernel:

Máquina

de

Vetores

de

Suporte

(SVM)

4

43.8.7Probabilístico:

Naive

Bayes

4

53.8.8Probabilístico:

Rede

Bayesiana453.9

Comparação

dos

algoritmos

4

63.9.1

Curva

ROC

4

63.9.2Curva

de

Calibração474

MATERIAIS

E

MÉTODOS494.1

Base

de

dados

514.2

Etapa 1

544.3

Etapa 2

554.4

Softwares

555

RESULTADOS

575.1

Etapa

1

5

85.1.1Clusters

e

Árvore

com

1L

		5
85.1.2 <i>Clusters</i>	e	árvore
com 3L		595.2
Etapa		2
		6
15.2.1 Variáveis	e	<i>Outliers</i>
Removidos		
		6
15.2.2 Geração	Sintética	de
Dados		
		6
35.2.3		Características
Selecionadas		
		6
35.2.4		
		Classific
ações		
		6
45.2.5 Curva		de
Calibração		646
DISCUSSÃO		666.1
Etapa 1		676.2
Etapa 2		687
		CONSIDERAÇÕES
FINAL	Erro!	Indicador
	não	definido.
REFERÊNCIAS		40
Anexo 1: Aprovação do Comitê de Ética e Pesquisa		79

LISTA DE FIGURAS

Figura 1: A curva de velocidades de leitura em ppm por tamanho de letra em logMAR baseado em O'Brien et al	7
Figura 2: No gráfico (a) temos um exemplo <i>underfit</i> , no (b) um exemplo de classificação adequada e no (c) um exemplo de <i>overfit</i>	9
Figura 3: Comparação entre variância e viés com gráfico de alvo	9
Figura 4: a) dados com 5 pontos com a linha de regressão, (b) linha de regressão com um <i>outlier</i>	10
Figura 5: Exemplo de geração sintética gerada pelo SMOTE	11
Figura 6: Mapa auto-organizável convergindo sobre área do espectro de informações (em azul)	12
Figura 7: Gráfico de dispersão de exemplo para árvore de decisão, os grupos estão coloridos da seguinte forma: W1 em cinza, W2 em vermelho, W3 em laranja e W4 em azul	13
Figura 8: Exemplo de árvore de decisão classificando os dados da Figura 7	14
Figura 9: Exemplo de SVM	16
Figura 10: Exemplo de grafo acíclico dirigido	17
Figura 11: Exemplo de quatro curvas ROC	18
Figura 12: Um exemplo de comparação das curvas de calibração de dois algoritmos (A1 e A2)	19
Figura 13: Diagrama do metodo em duas etapas do estudo.	21
Figura 14: Representação de uma sentença em uma única linha do MNREAD-P1L	22
Figura 15: Representação de uma sentença de três linha do MNREAD-P3L	22
Figura 16: Cluster baseados nas leituras do MNREAD-P1L.	27
Figura 17: Árvore de decisão baseada nos <i>clusters</i> das leituras do MNREAD-P1L	28
Figura 18: <i>Clusters</i> baseados nas leituras do MNREAD-P3L	29
Figura 19: Árvore de decisão baseada nos <i>clusters</i> das leituras do MNREAD-P3L	30
Figura 20: <i>Boxplots</i> com <i>outliers</i>	31
Figura 21: Gráfico de dispersão de dados de disléxicos (em azul) e não-disléxicos	

(em vermelho), com dados de MVL e AL, antes (esquerda) e depois (direita) do uso do SMOTE

32

Figura 22: Curvas de calibração dos algoritmos

34

LISTA DE QUADROS

Quadro 1: : Variáveis de funções visuais de leitura coletadas no banco de dados

54Quadro 2: Atributos selecionados para cada algoritmo levando em conta a melhor AUC. 63Quadro 3: Tabela comparativa dos algoritmos sobre, sensibilidade, especificidade e AUC 64Quadro 4: Comparação de desempenho de uma regressão logística com e sem a geração sintética de dados.64

LISTA DE ABREVIATURAS E SIGLAS

1L	MNREAD-P em uma linha
3L	MNREAD-P em uma linha
AL	Acuidade de leitura
AM	Aprendizado de máquina
ATCL	Acurácia no tamanho crítico de letra
AUC	Área sob a curva ROC
AV	Acuidade visual
Delta_LMVL_TCL	Diferença do tamanho de letra no MVL e no TCL
Delta_MVL_TCL	Diferença de velocidade de leitura entre a máxima e no TCL
DP	Desvio padrão
ET	<i>Eye tracker</i>
FFNN	<i>Feed-forward neural network</i>
FVL	Funções visuais de leitura
ITCL	Interferência no sentido dos erros cometidos no TCL
LMVL	Tamanho da letra no MVL
M	Média
MNREAD	<i>Minnesota low vision reading test</i>
MNREAD-P	MNREAD adaptado para o português
MVL	Máxima velocidade de leitura
ppm	Palavras por minuto
RNN	<i>Recurrent neural network</i>
ROC	<i>Receiver operating curve</i>
SMOTE	<i>Synthetic minority over-sampling technique</i>
SOM	<i>Self-organized maps</i>
SVM	<i>Support vector machine</i>
TCL	Tamanho crítico de letra
TCLE	Termo de consentimento livre e esclarecido

TF	Tamanho da fonte
VL	Velocidade de leitura
VTCL	Velocidade no tamanho crítico de letra

RESUMO

INTRODUÇÃO: Dislexia do desenvolvimento é uma disfunção neurológica que afeta a habilidade de leitura, que se não tratado pode levar a problemas de aprendizado e impactando negativamente o aumento de vocabulário. O diagnóstico da dislexia é complexo e feito por exclusão. Alguns estudos avaliaram dados de movimento ocular em conjunto com técnicas de aprendizado de máquina (AM) para classificar a dislexia. Outro estudo levanta a hipótese de padrões de funções visuais de leitura (FVL) para compreensão da dislexia. Entretanto, o estudo de FVL em conjunto de técnicas de AM ainda não foi explorado. **OBJETIVO GERAL:** Aplicar técnicas de aprendizado de máquina (AM) para explorar e auxiliar o diagnóstico de disléxicos a partir das funções visuais de leitura (FVL). **OBJETIVOS ESPECÍFICOS:** Explorar os dados de FVL de disléxicos e não-disléxicos, a partir de extração de características e classificá-los utilizando AM. **MATERIAL E MÉTODOS:** Esta dissertação foi executada em duas etapas: uma quantitativa e exploratória e uma quantitativa e correlacional. A primeira etapa explorou os dados de FVL de disléxicos de duas bases, uma de leituras de textos em 1 linha (1L) e outra de 3 linhas (3L). Foi aplicado o algoritmo de mapas auto-organizáveis em cada base para separá-los em *clusters* que foram então enviados para uma Árvore de Decisão para extrair as regras que regem cada um dos grupos. A segunda etapa utilizou dados de leituras de 3L e foi realizada uma seleção de *outliers*. Com os dados restantes foi aplicada geração sintética de dados com o algoritmo SMOTE. Então foi aplicada uma técnica de seleção de características tendo a melhor área sob a curva ROC (AUC) como alvo para cada um dos cinco algoritmos selecionados. Eles foram comparados pela AUC e acurácia. Todos também foram comparados pela sua curva de calibração. **RESULTADOS:** Na primeira etapa, a avaliação da base de 1L, resultou em uma aglomeração de 1 *cluster* de controles e 3 de disléxicos. Somente disléxicos obtiverem $MVL < 140,72$ ppm, já na avaliação de 3L foram obtidos 3 *clusters* de disléxicos e 1 de controle. Neste somente disléxicos tiveram uma Velocidade de Leitura no Tamanho Crítico de Letra (VTCL) inferior a 112,71 ppm. Na segunda etapa foram gerados dados sintéticos para cada grupo ter 100 registros. Na seleção de característica a Acuidade de Leitura (AL) foi selecionada em 4 dos 5 algoritmos. A Regressão Logística obteve a melhor AUC (0,999) e acurácia (99%) além de ter obtido a melhor curva de calibração. **CONCLUSÃO:** Na primeira etapa o fato de a MVL ter sido tão determinante na separação dos *clusters* com 1L e o VTCL no de 3L pode indicar que o efeito de *crowding* teve algum impacto no teste de 3L. O fato de AL ter sido selecionado em 4 das 5 seleções de características, a torna uma variável importante para o diagnóstico e estudo da dislexia. O algoritmo de Regressão Logística obteve os melhores resultados sendo indicado para classificação de disléxicos com base em FVL.

Palavras-chave: Aprendizado de Máquina. Dislexia. Extração de Características. Geração sintética de Dados. Classificação.

ABSTRACT

INTRODUCTION: Developmental dyslexia is a neurological disorder that affects reading ability, that when left untreated can lead to learning problems and negatively impacting vocabulary increase. The diagnosis of dyslexia is complex and made by exclusion. Some studies evaluated eye movement data in conjunction with machine learning (ML) techniques to classify dyslexia. Another study raises the hypothesis of visual reading function patterns (VRF) for dyslexic differentiation. The study of VRF in combination of ML techniques has not been explored. **GENERAL OBJECTIVE:** To apply ML techniques to explore and assist the diagnosis of dyslexics from VRF. **SPECIFIC OBJECTIVES:** To explore dyslexic and non-dyslexic VRF data with feature extraction and to classify dyslexic and non-dyslexic using ML. **MATERIAL AND METHODS:** This dissertation has two steps: a quantitative and exploratory and a quantitative and correlational. The first step explored two dyslexic VRF datasets, one of 1-line (1L) text readings and the other of 3-line (3L) text readings. The self-organizing map algorithm was applied to each base to separate them into clusters that were then sent to a decision tree to extract the rules characterize each of the groups. The second step used data from 3L readings. The outliers was selected by a specialist. With the remaining data, the SMOTE algorithm was applied. Then a feature selection technique was applied having the best area under the ROC curve (AUC) as target for each of the five selected algorithms. They were compared by AUC and accuracy. All were also compared by their calibration curve. **RESULTS:** In the first step, the 1L base evaluation resulted in a clustering of 1 cluster of controls and 3 of dyslexics. Only dyslexics obtained Maximum reading speed MRS <140.72 ppm, while in the 3L evaluation, 3 dyslexic clusters and 1 control were obtained. In this only dyslexics had reading speed at critical read size (RSCPS) of less than 112.71 ppm. In the second step, synthetic data were generated for each group to have 100 records. In feature selection the reading acuity (RA) was selected in 4 of the 5 algorithms. Logistic regression obtained the best AUC (0.999) and accuracy (99%) and obtained the best calibration curve. **CONCLUSION:** In the first step, the fact that MRS was so determinant in the separation of the 1L clusters and the RSCPS in the first one. It may indicate that the crowding effect had some impact on the 3L test. The fact that RA has been selected in 4 of the 5 feature selections may be an important variable for the diagnosis and study of dyslexia. The logistic regression algorithm obtained the best results and was indicated for VRF-based dyslexic classification.

Keywords: Machine Learning. Dyslexia. Feature Extraction. Synthetic Data Generation. Classification.

1 INTRODUÇÃO

Dislexia do desenvolvimento é uma disfunção de origem neurológica que afeta a habilidade de leitura, caracterizada por dificuldade significativa e persistente de aprendizagem escolar - especialmente ao ler - afetando a acurácia, fluência e compreensão leitora. Mesmo o portador de dislexia tendo recebido educação adequada, não apresentar nenhuma desordem de desenvolvimento intelectual, não possuir problemas de visão e nem carência de proficiência na linguagem ou adversidade psicossocial, o disléxico apresenta *performance* leitora marcadamente inferior ao esperado.^[1-3]

Quando não tratada, a dislexia pode levar a problemas de compreensão e menor disposição para a leitura, dificultando o aumento do vocabulário,^[2] além do baixo aproveitamento escolar - o que pode levar a diversos impactos na sua vida adulta.^[4]

O diagnóstico da dislexia é complexo, pois o mesmo é realizado por exclusão e tem uma característica multidisciplinar incluindo oftalmologistas, psicopedagogos, fonoaudiólogos e psicólogos,^[1,5] que torna o diagnóstico demorado e custoso. Recentemente em novos estudos promissores ao trazerem novos indicadores de dislexia estudaram o comportamento do olhar durante a leitura. Estes estudos utilizaram um equipamento de captura do olhar (*Eye Tracker*), que é capaz de coletar os dados do movimento do olhar e os traduz em dados de fixações e sacadas.^[6-8]

Uma outra métrica menos explorada, mas não menos importante no estudo diagnóstico de dislexia são as Funções Visuais de Leitura (FVL). Na FVL são avaliadas velocidades e acuidades de leitura.^[2,9] Esta abordagem possui uma vantagem em relação ao estudo do movimento ocular. Para coleta de FVL não é necessário um *Eye Tracker* para aquisição dos dados. Desta forma, a FVL é um processo menos oneroso para, por exemplo, auxiliar ao diagnóstico em dislexia.

O Aprendizado de Máquina (AM) é um ramo da inteligência artificial que aplica sistematicamente algoritmos para sintetizar os relacionamentos subjacentes entre dados e informações de forma que utilize recursos extraídos dos dados como entradas e retorna o conhecimento adquirido como saída.^[10-12]

Alguns trabalhos recentes foram feitos utilizando dados de movimento do olhar em conjunto com o AM. Rello & Ballesteros^[6] avaliaram a variação do

comportamento do olhar em disléxicos e não-disléxicos alterando a fonte da letra e aplicaram uma técnica de AM para esta avaliação. Lustig^[7] coletou dados de movimento do olhar e os avaliou com três técnicas de AM distintas. Benfatto^[8] realizou uma pesquisa utilizando uma base de um estudo de *coorte* e utilizou uma técnica de AM para avaliar os dados do movimento do olhar e predizer a chance de desenvolvimento da dislexia.

Foi executado uma busca no *PubMed* utilizando os seguintes termos *MeSH* “*dyslexia and machine learning*” e não foi encontrado nenhum trabalho que aplicasse o uso de AM com teste de linguagem baseados em leitura que avaliasse FVL. Especificamente, o uso de AM aplicado a FVL ainda não foi explorado de forma que esta dissertação tem como objetivo aplicar técnicas de AM em dados FVL de leitura de disléxicos com os objetivos de extrair características dos dados e classificar disléxicos com a mesma base para que possam auxiliar o estudo e diagnóstico de dislexia.

2 OBJETIVOS

2.1 Objetivo Geral

Aplicar técnicas de Aprendizado de Máquina (AM) para explorar e auxiliar o diagnóstico da dislexia a partir das Funções Visuais de Leitura (FVL).

2.2 Objetivos Específicos

- Explorar os dados de FVL de disléxicos e não-disléxicos a partir de uma técnica de extração de características.
- Classificar disléxicos e não-disléxicos utilizando AM.

3 REFERECIAL TEÓRICO

3.1 Aprendizado de Máquina

Aprendizado de Máquina (AM) é uma subárea da inteligência artificial, que possui como objetivo transferir o processo de aprendizado característico do pensamento humano para o computador.^[13,14] O AM foi proposto pela Ciência da Computação, e é visto como uma faceta do campo de reconhecimento de padrões, provenientes de diferentes trabalhos desenvolvidos nas Engenharias.^[11]

O AM é caracterizado pela sua capacidade de utilizar a “experiência” para melhorar sua *performance* ou fazer previsões precisas. Neste sentido, a “experiência” refere-se a dados coletados de forma eletrônica, que são disponibilizados para futura análise. O aprendizado é a tarefa realizada por algoritmos que possuem propriedades intrínsecas de aprendizagem em base de dados.^[14]

O AM é utilizado de diversas maneiras, e cada dia novas formas de aplicações surgem.^[14]

Algumas destas aplicações de AM são:

- Classificação: Onde cada registro possui uma categoria e o algoritmo deve associar os registros a essa categoria.
- Regressão: Cada registro possui um valor real a ser predito pelo algoritmo.
- Ranqueamento: O algoritmo tem que ordenar os registros de acordo com algum critério.
- Clusterização: O algoritmo separa os itens em regiões homogêneas.

De acordo com a disponibilização e características dos dados e os objetivos a serem atingidos pelo algoritmo de AM, pode-se caracterizar alguns cenários de aprendizado para se avaliar o tipo de algoritmo que deve ser utilizado.

- Aprendizado supervisionado: Os dados possuem rótulos indicando um alvo correspondente (uma classe ou um valor a ser predito) e o algoritmo supervisionado utiliza estes rótulos para aprender os padrões dos dados e ser capaz de atingir o alvo especificado.^[11,14,15]

Este é o cenário mais comum em aplicações de classificação, regressão e ranqueamento.

- **Aprendizado não supervisionado:** O algoritmo recebe somente dados sem rótulos e os registros são agrupados de acordo com similaridades entre si^[11,14,16] e o algoritmo faz previsões para os dados não vistos. Clusterização e redução de dimensionalidade são problemas comumente encontrados nestes cenários.
- **Aprendizado semi-supervisionado:** O algoritmo recebe dados de treinamento contendo dados com e sem rótulos e faz previsões a partir destes.^[14] Este tipo de aprendizado costuma ser utilizado em problemas de classificação, regressão e ranqueamento.
- **Inferência transdutiva:** o algoritmo recebe dados de treinamento contendo dados com e sem rótulos como o semi-supervisionado, porém este deve classificar somente estes dados de treinamento.^[14]

3.2 Dislexia e Aprendizagem de Máquina

Dislexia do desenvolvimento é uma disfunção de origem neurológica que afeta a habilidade de leitura, caracterizada por uma dificuldade significativa e persistente de aprendizagem escolar relacionada com a leitura, como acurácia e fluência de leitura e, compreensão leitora. O que pode ser indicado por uma *performance* marcadamente 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]

Em 2004 Lima e Silva^[2] realizaram um estudo de prevalência da dislexia em alunos do 3º ano do ensino fundamental de quatro escolas particulares, submetidos a avaliações fonoaudiológicas, pedagógicas, psicológicas, neurológicas e audiométricas. Foi constatada uma prevalência de 12,1%, com um intervalo de confiança de 95% entre 7,4 e 19%.^[17]

Alguns estudos foram realizados reunindo os dados de movimento do olhar

de disléxicos e avaliados com técnicas de AM para estimar sua *performance* em classificar disléxicos e desta forma verificar se o movimento do olhar pode ser utilizado como indicador de dislexia.

Rello & Ballesteros^[6] utilizaram uma base com 1.135 leituras de 12 textos em espanhol com fontes diferentes (de 97 leitores sendo 48 disléxicos), capturadas por meio de um rastreador de olhar (*Eye Tracker- ET*). Formando uma base com 12 variáveis: idade, tipo de fonte utilizada, itálico, *serif*, fonte específica para dislexia, preferência de fonte do participante, número de visitas na área de interesse, média da duração na área de interesse, somatória de visitas no texto todo, média de fixações, número de fixações, soma de todas as fixações. Com os dados coletados foi utilizado um reconhecedor automático de padrões chamado de “Máquina de Vetores de Suporte” (*Support Vector Machine - SVM*) para classificar os sujeitos de pesquisa como disléxicos ou não-disléxicos obtendo uma acurácia de 80,18% no diagnóstico automático de dislexia.

Lustig^[7] realizou uma pesquisa utilizando ET, onde extraiu dados de 126 amostras de leituras de texto advindos da versão sueca de testes para leitura PISA, de 18 sujeitos de pesquisa (9 disléxicos). Os dados coletados para avaliação foram 7: Média do tempo de fixação, frequência de fixação, média do tamanho das sacadas para frente, frequência das sacadas para frente, estabilidade da fixação, média de amplitudes de sacadas, frequência de sacadas de regressão. A partir destes dados foram efetuados testes com três classificadores de padrões - SVM, *Feed-Foward Neural Network* (FFNN) e *Recurrent Neural Network* (RNN), os métodos empregados obtiveram acurácia na classificação de 83%, 83%, 78% respectivamente.

Benfatto^[8] realizou uma pesquisa utilizando uma base de um estudo de *coorte* que acompanhou 2.165 estudantes de 8-9 anos de 1989 a 2010, o teste consistiu na leitura de uma página branca de papel com um texto em alto contraste com um texto distribuído em 8 linhas e 10 sentenças com média de 4,6 palavras. A partir destes dados foram separados 5% dos participantes com *performance* inferior, resultando em uma base de 185 participantes separados em 88 com baixo risco de dislexia e 97 com alto risco de dislexia, resultando numa base com 168 variáveis. Que por sua vez foi utilizado uma SVM com seleção aleatória de características para identificar os participantes com alto e baixo risco obtendo uma acurácia de

95,6%, com uma sensibilidade de 95,5% e especificidade de 95,7%.

Uma outra técnica que vem sendo utilizada para estudar os impactos da dislexia são as FVL que se apresentam segundo as velocidades de leitura e tamanhos de letra.

O protocolo MNREAD (*Minnesota Low Vision Reading Test*)^[18] foi desenvolvido especificamente para avaliar FVL, onde são apresentados uma sequência de textos, montados com um vocabulário controlado, em tamanhos decrescentes (em escala logMAR¹) e em cada leitura é medido a velocidade em palavras por minuto (ppm). Na Figura 1 é apresentado um exemplo de medidas de velocidade de leitura deste teste, onde pode ser extraído a máxima velocidade de leitura (MVL), caracterizado pela maior velocidade obtida em todos os textos, a linha máxima de velocidade de leitura (LMVL), caracterizado pelo tamanho da letra durante a maior velocidade de leitura, o tamanho crítico de letra (TCL) caracterizado pelo tamanho de letra em que a velocidade de leitura começa a decair, a velocidade de leitura durante o tamanho crítico de letra (VTCL) e a acuidade de leitura (AL) o menor tamanho de letra lido.

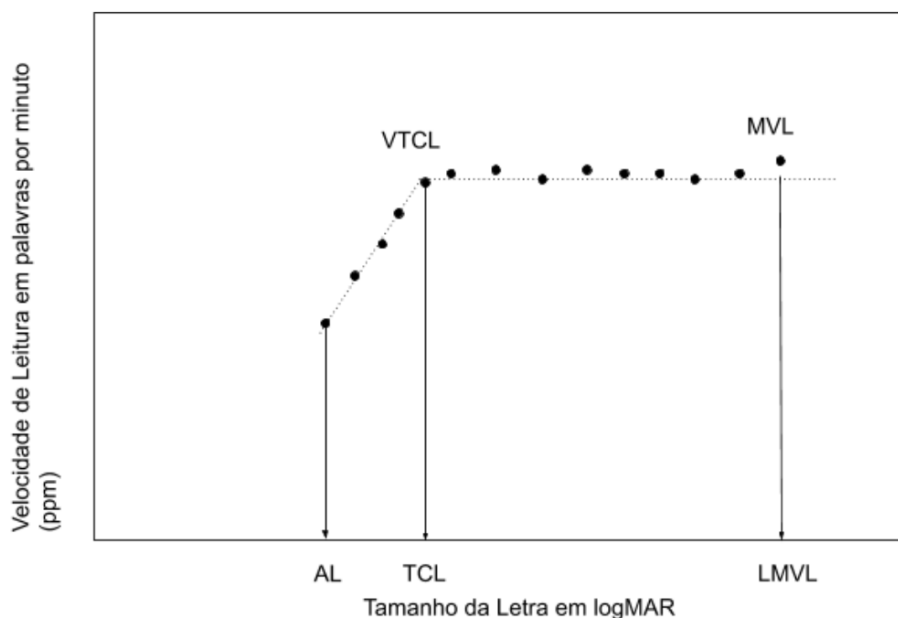


Figura 1: A curva de velocidades de leitura em ppm por tamanho de letra em logMAR

Fonte: o autor – modificado^[9]

¹ logMAR é uma unidade de medida para tamanho de letras que representam o logaritmo do ângulo de visão, a maior letra de uma tabela de acuidade visual possui logMAR=1,0 ou 50 minutos de arco [19,20]

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]

Beth A. O'Brien, Mansfield, & Legge, em 2005^[9] realizaram um estudo onde disléxicos e não-disléxicos apresentaram perfis similares de curvas em VL e TF o que implica, assim sendo, que qualitativamente, VL em disléxicos possui a mesma dependência qualitativa em relação com a diminuição ou aumento do TF. Contudo, apesar das similaridades entre as curvas, duas características são diferentes: como esperado, a velocidade de leitura foi menor para o grupo disléxicos; adicionalmente, o TCL foi maior para disléxicos que para o grupo controle, indicando assim que eles precisam de fontes maiores para atingir a mesma VL que o grupo de controle. Portanto parece pertinente escolher a investigação das FVL e desempenho de leitura em disléxicos e não-disléxicos com o propósito de desenvolver uma ferramenta de suporte para a identificação precoce de dislexia em crianças.

3.3 *Overfitting e Underfitting*

Overfit e *underfit* são dois problemas a serem enfrentados ao fazer classificação com AM, e que possuem uma maior tendência de ocorrer com base de dados pequenas. O *underfit* ocorre quando o algoritmo não alcança uma acurácia satisfatória, os dados apresentados no treino geram uma *performance* inferior ao esperado. A Figura 2 (a) exibe a ocorrência de *underfit* no treinamento. 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, como pode ser visto na Figura 2 (c).

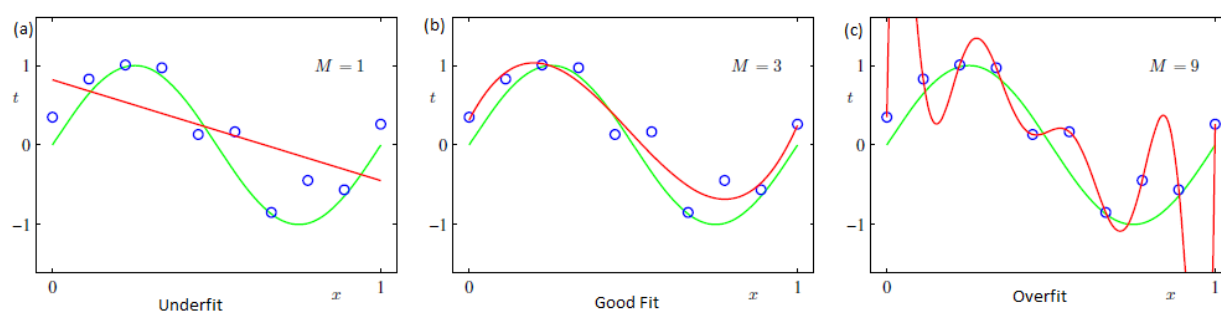


Figura 2: No gráfico (a) temos um exemplo *underfit*, no (b) um exemplo de classificação adequada e no (c) um exemplo de *overfit*

Fonte: [11]

De um modo geral, algoritmos que estão em *underfit* possuem uma baixa variância e alto viés e algoritmos que estão em *overfit* possuem uma alta variância e um baixo viés. Esse comportamento é devido ao algoritmo mais simples possuir um viés mais alto e um algoritmo complexo uma variância mais alta.^[21,22] A Figura 3 apresenta a comparação entre variância e viés.

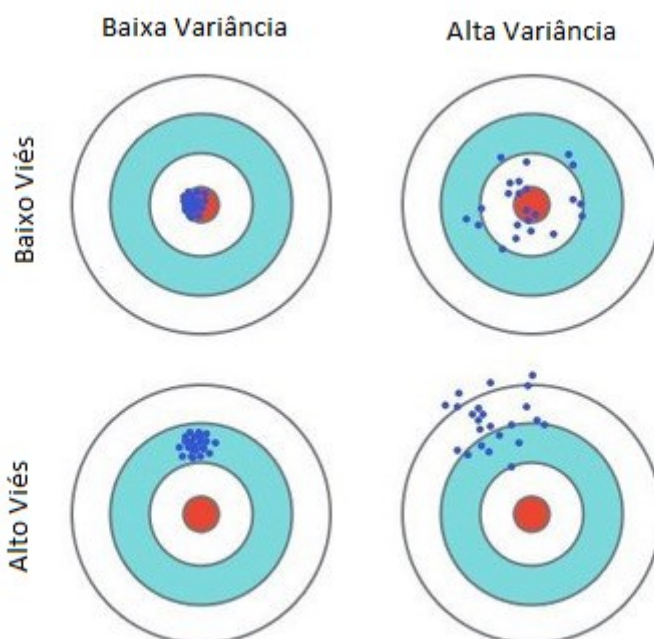


Figura 3: Comparação entre variância e viés com gráfico de alvo

Fonte: o autor - modificado^[23]

3.4 Remoção de ruído

Uma avaliação de cada variável para encontrar características que são irrelevantes ou sem sentido na classificação, variáveis que tenham pouco (tenha entre 1 a 3 registros com informações diferentes em uma classe), nenhuma variação (todas as variáveis iguais em uma classe), que possuam dados sem sentido ou que sejam dependentes entre si também devem ser removidas.^[23,24]

Outliers podem impactar negativamente o resultado dos algoritmos, especialmente dos mais simples, como os de regressão. A Figura 4 exibe uma regressão em uma base de 5 pontos com e sem *outlier*.

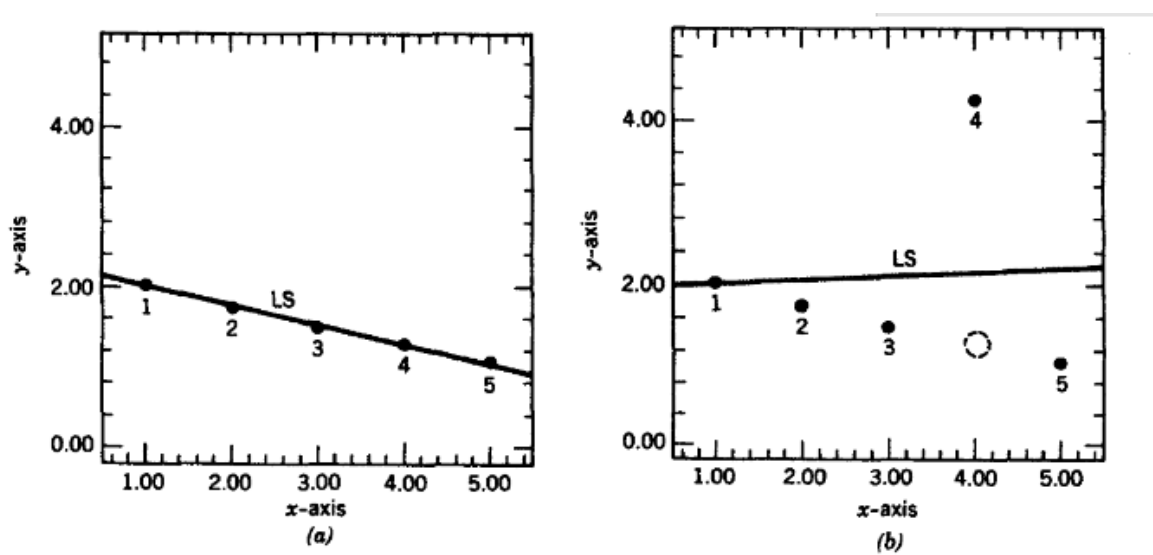


Figura 4: (a) dados com 5 pontos com a linha de regressão, (b) linha de regressão com um *outlier*

Fonte: ^[25]

Portanto para avaliação dos *outliers*, foram utilizados *boxplots* com o objetivo de avaliá-los junto com profissional especialista da área e remover os que não forem uma forte representação de disléxicos ou não-disléxicos, pois os mesmos podem impactar negativamente na classificação^[23,25,26] e geração de dados sintéticos.

3.5 Geração de Dados Sintéticos (SMOTE)

O *Synthetic Minority Over-sampling Technique* (SMOTE)^[27] 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. A Figura 5 exibe a técnica de geração sintética - SMOTE.

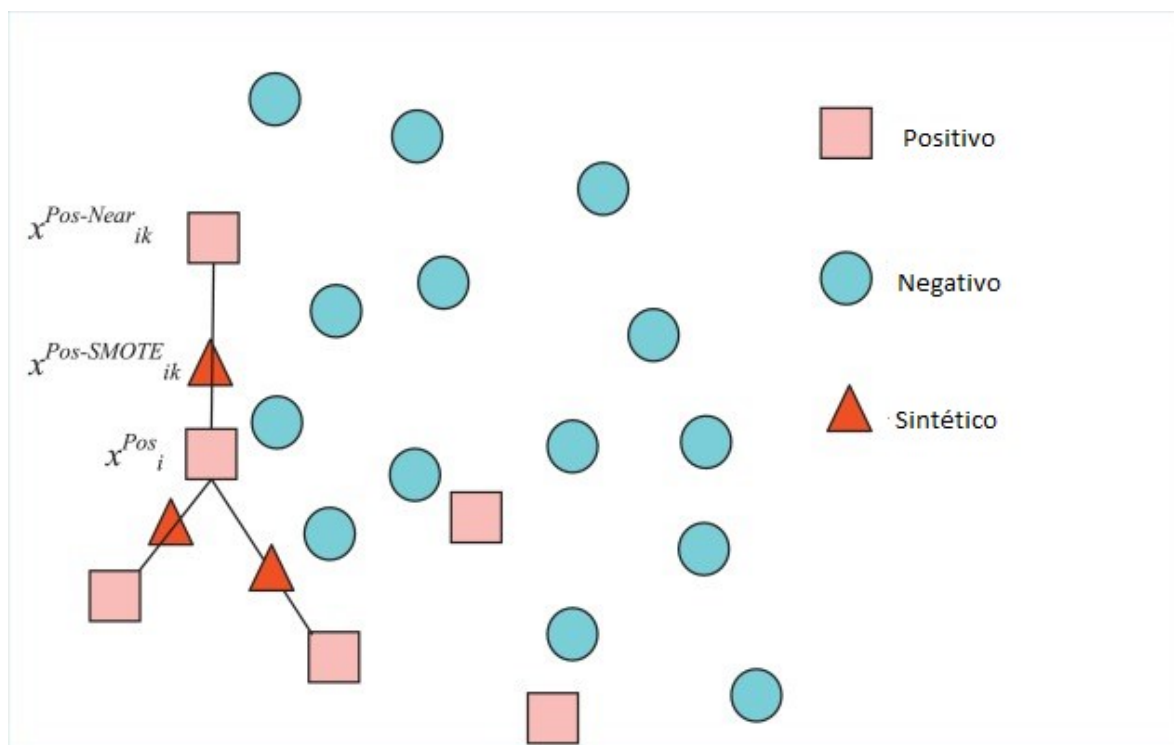


Figura 5: Exemplo de geração sintética gerada pelo SMOTE

Fonte: o autor - modificado [28]

Para execução do algoritmo SMOTE,^[27] foi executado para aumentar a base para 100 registros em cada grupo. Depois de gerado foi aplicado o algoritmo *Randomize*, que mistura os dados gerados para garantir a aleatoriedade dos dados durante a execução dos algoritmos de AM.

3.6 Seleção de Características

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.^[23,29]

Depois de gerado os dados, será utilizado o algoritmo *WrapperSubsetEval* que avalia grupos de variáveis por um método selecionado, então avalia de acordo com um indicador de performance e gera a estimativa usando uma validação cruzada.

Portanto 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*.^[30]

3.7 Extração de Características

Extração de Características é a representação do dado original em um subconjunto de dados que contém as informações discriminatórias para facilitar a tarefa de classificação por outros algoritmos computacionais ou por humanos.^[11,21,31–33]

3.8 Algoritmos

Esta seção discorre sobre os algoritmos de AM utilizados nesta dissertação.

3.8.1 Mapas Auto-organizáveis (SOM)

O algoritmo de mapas auto-organizáveis (SOM) é um classificador que aplica um tipo especial de redes neurais baseada em aprendizado competitivo. Nesta rede é possível transformar uma informação multidimensional em grupos usualmente de uma ou duas dimensões. Para aprender, os neurônios competem entre si para serem ativados, de forma que somente um neurônio de saída seja ativado para um determinado grupo.^[11,15] Depois de várias iterações os neurônios se distribuem por todo o espectro da informação como apresentado na Figura 6, podendo assim extrair características de um padrão de entrada específico.^[22]

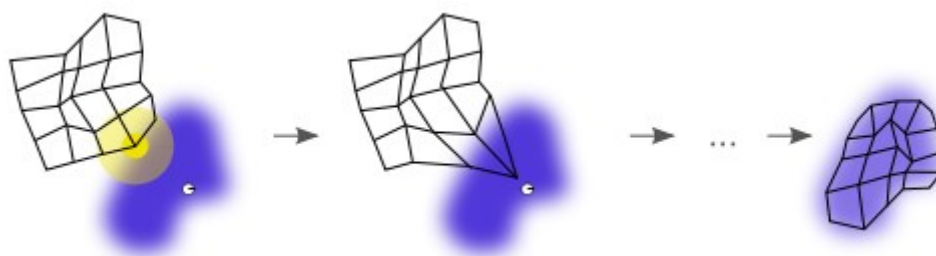


Figura 6: Mapa auto-organizável convergindo sobre área do espectro de informações (em azul)

Fonte:^[15]

3.8.2 Árvore de Decisão

Os algoritmos de Árvore de Decisão geram um grafo onde os nós (ramos ou galhos) partem de uma única origem e seguem até o final onde se encontram os resultado (folhas), que são usados para classificar os dados.^[34] Um exemplo de espalhamento de dados é apresentado na Figura 7 e um exemplo de Árvore de Decisão é mostrado na Figura 8 onde foram classificados os dados da Figura 7.

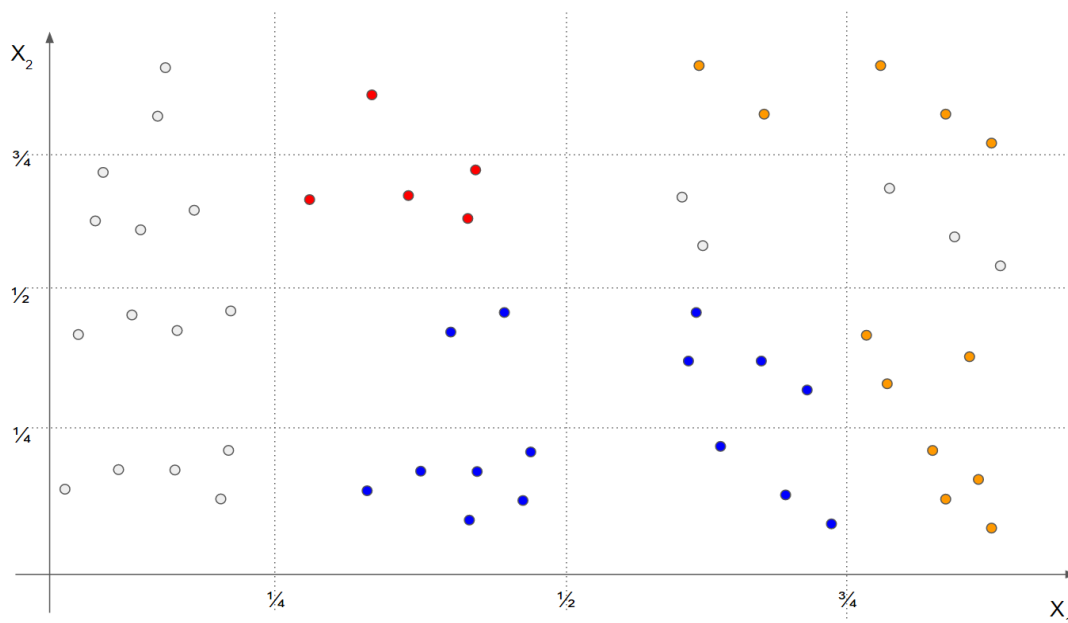


Figura 7: Gráfico de dispersão de exemplo para árvore de decisão. Os grupos estão coloridos da seguinte forma: W1 em cinza, W2 em vermelho, W3 em laranja e W4 em azul

Fonte: o autor

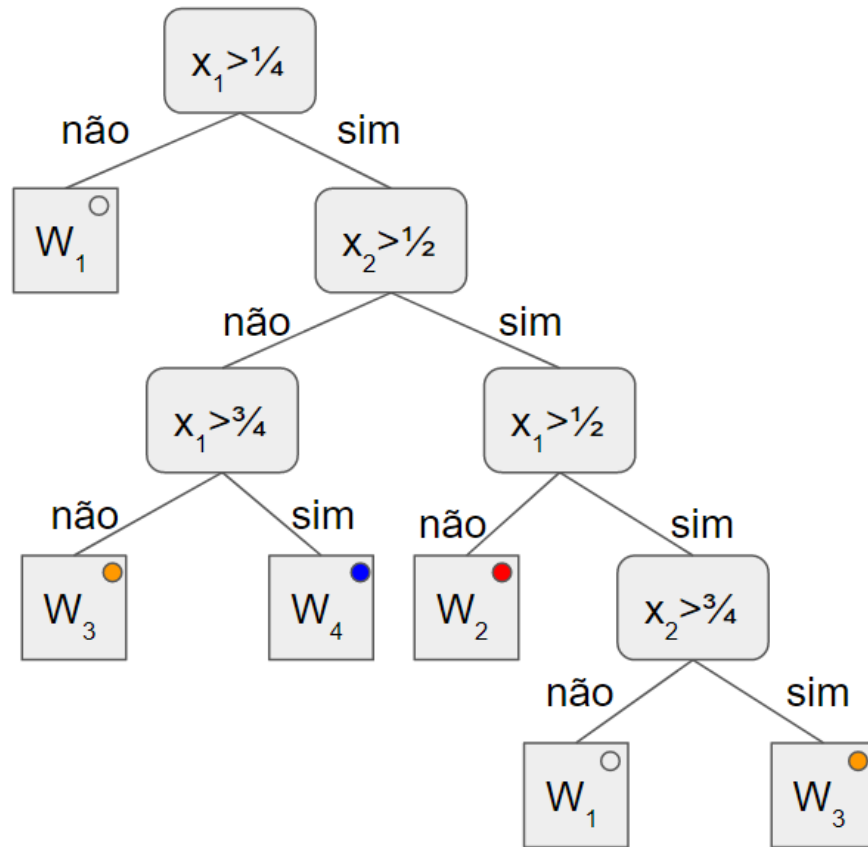


Figura 8: Exemplo de árvore de decisão classificando os dados da Figura 7

Fonte: o autor - modificado^[16]

3.8.3 Árvore de decisão: Árvore Aleatória

O algoritmo de Árvore Aleatória seleciona para criar em cada nó um subconjunto aleatório de todas as variáveis selecionando a de menor entropia e maior ganho de informação para então definir o ponto de decisão, até chegar em uma folha.^[35–37]

3.8.4 Árvore de decisão: C4.5

O algoritmo C4.5 descende do algoritmo ID3 (*Induction of Decision Trees*), onde ambos usam a técnica de entropia máxima para encontrar os nós que serão criados ^[11,34,38]. No entanto, o C4.5 se diferencia do ID3 por conseguir lidar tanto com dados contínuos como discretos. Assim, para lidar com os atributos contínuos o algoritmo C4.5 utiliza do ganho de informação para dividir as listas em maior igual

ou menor ao corte. O C4.5 também consegue classificar registros incompletos. Faz-se isto atribuindo uma interrogação “?” para os dados faltantes. Assim, este atributo não é usado para os cálculos de entropia. Após isto, este algoritmo repassa a árvore removendo galhos (nós de decisão) por folhas (decisão). O C4.5 também executa um processo de poda que calcula a taxa de erro de cada ramo e o substitui por uma folha se for menor que a taxa de erro estimada.

3.8.5 Regressão: Regressão Logística

A regressão logística é um algoritmo de classificação binária que retorna uma distribuição condicional de probabilidades que pode ser utilizado para definir a probabilidade de uma variável independente influenciar uma variável dependente. Este algoritmo faz seu aprendizado utilizando uma função de verossimilhança.^[39,40]

3.8.6 *Kernel*: Máquina de Vetores de Suporte (SVM)

A Máquina de Vetores de Suporte (SVM) é tida como a técnica de aprendizado de *Kernel*.^[15] Os métodos de *Kernel* trabalham em dois passos. Primeiramente é realizada a transformação de um conjunto de padrões não separados linearmente em um novo conjunto que possui uma maior chance de ser separado linearmente. Em seguida é executada uma função para classificação da distância entre o *Kernel* e os dados.^[11,15] Assim, é possível classificar os conjuntos como “positivo” ou “negativo” de acordo com a linha de *Kernel*. Para problemas não lineares, muitos algoritmos usam um conceito conhecido como truque de *Kernel*.^[15] Essa estratégia baseia-se na ideia de trocar a função de *Kernel* bidimensional por uma função multidimensional, fazendo com que o algoritmo classifique dados não-lineares. A SVM se utiliza desta premissa das funções de *Kernel*. E além disso, adiciona vetores que possuem uma amostra dos dados como pode ser visto na Figura 9. A margem é definida perpendicularmente a linha de decisão e a classificação do dado é definida pela sua proximidade com as bordas de decisão a posição das linhas é definida baseado num subconjunto de informações conhecidos como vetores de suporte. Esta abordagem melhora a classificação dos dados.^[11,15,16]

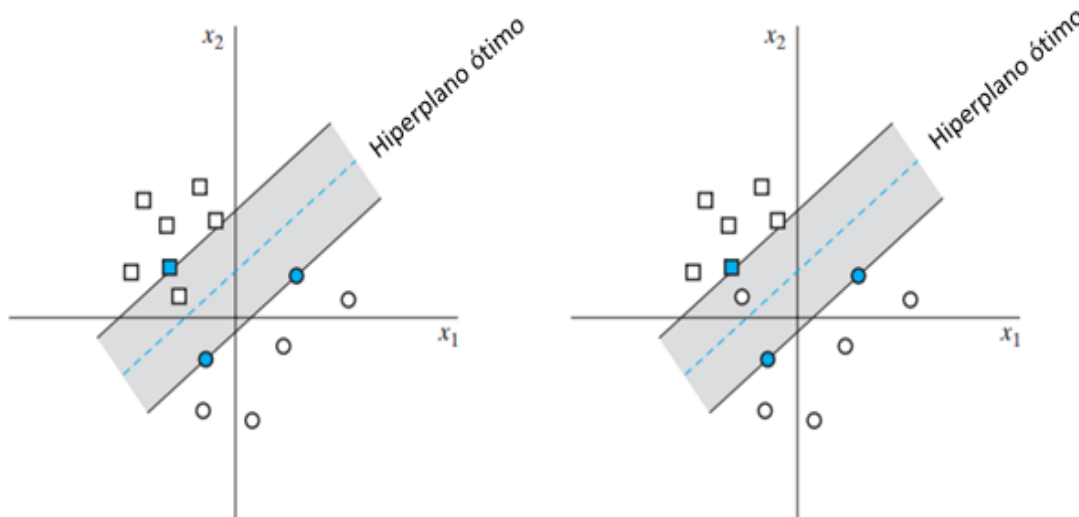


Figura 9: Exemplo de SVM

Fonte:^[11]

3.8.7 Probabilístico: *Naive Bayes*

O *Naive Bayes* é um algoritmo que separa cada rótulo da variável dependente, para em seguida calcular a probabilidade de cada variável independente, então cada probabilidade é multiplicada para cada rótulo da variável dependente. Então os mesmos são divididos pelo total de registros. Este processo resulta na probabilidade e de um evento acontecer (variável dependente) a partir de N condições (variáveis independentes).^[40]

3.8.8 Probabilístico: *Rede Bayesiana*

São classificadores que utilizam técnicas de probabilidade propostas por Thomas Bayes no século XVIII. É um modelo gráfico-probabilístico composto por uma rede estruturada na forma de um grafo acíclico dirigido, o qual representa a fatoração das probabilidades das junções de todas as variáveis,^[41] como pode ser visto na Figura 10.

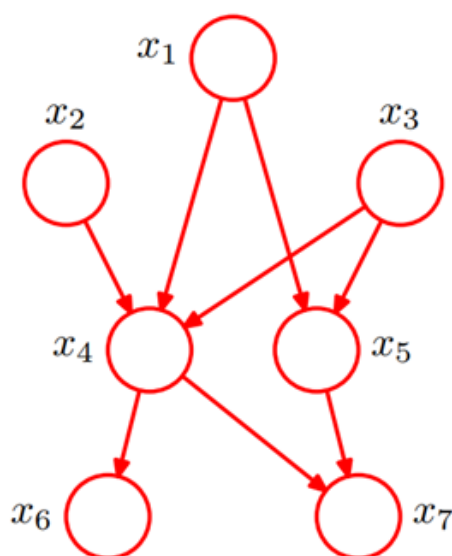


Figura 10: Exemplo de grafo acíclico dirigido

Fonte:[11]

3.9 Comparação dos algoritmos

3.9.1 Curva ROC

A curva ROC (*Receiver Operating Characteristic Curve*) consiste na curva formada da probabilidade de acerto de verdadeiros e a probabilidade de acertos falsos, também denominados sensibilidade e especificidade respectivamente.^[42] Desta maneira pode se dizer que uma curva que se assemelha a uma linha reta não possui poder discriminatório, portanto, pode se calcular a área desta curva não discriminatória e a curva representando a performance do teste em questão.^[42,43] Esta medida é conhecida como área sob a curva ROC (AUC), ela pode variar de 0 a 1, sendo 0 nenhum poder discriminatório e 1 uma curva perfeita em exemplo disto pode ser visto na Figura 11. Nesta figura a curva D possui uma AUC de 0, as curvas B e C possuem algum poder discriminatório com uma AUC entre 0 e 1, por fim a curva A possui uma discriminação perfeita tendo uma AUC igual a 1.

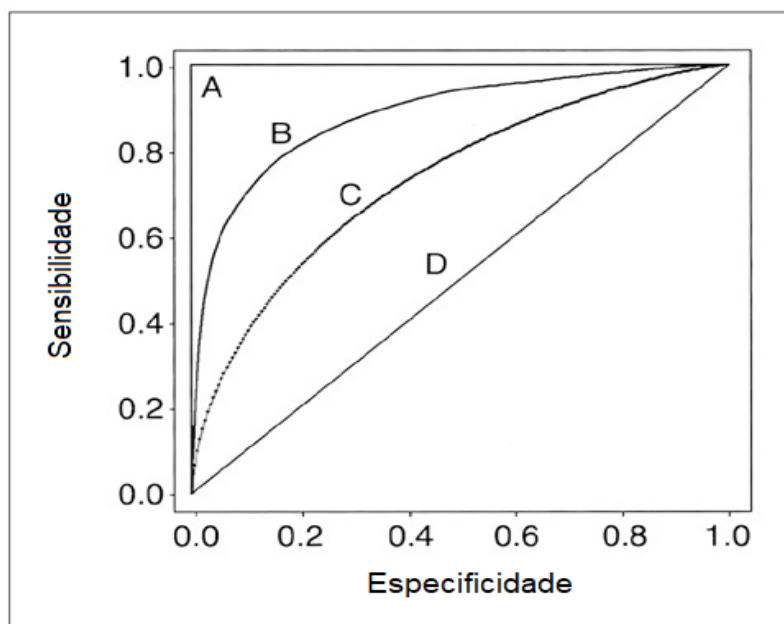


Figura 11: Exemplo de quatro curvas ROC

Fonte: O autor - modificado^[44]

A AUC têm se mostrado uma métrica mais confiável para comparação de técnicas de AM,^[28,30,42] especialmente em dados de saúde que possuam poucos dados ou dados desbalanceados.

3.9.2 Curva de Calibração

Para verificar se o algoritmo tem a capacidade de atribuir probabilidades acuradas, deve-se avaliar a habilidade do mesmo representar a verdadeira probabilidade de o evento de interesse ocorrer. Uma forma de se realizar esta avaliação é com a curva de calibração.^[30,45]

Para se construir uma Curva de Calibração, a probabilidade estimada do evento de interesse ocorrer é ordenada e dividida em grupos, utilizando os percentis da probabilidade estimada. Estas probabilidades são então comparadas com os percentis das probabilidades observadas ao se testar o algoritmo, ambas as probabilidades variam de 0 a 100%.^[30,45–47]

Desta forma, um algoritmo perfeitamente calibrado irá apresentar uma reta de 45° e, a variação acima desta curva representa que a probabilidade observada está acima da probabilidade estimada. Da mesma maneira, sob a linha diagonal a probabilidade observada encontra-se abaixo da probabilidade estimada^[30,45–47]. Isto

significa que, por exemplo, se um algoritmo possui 50% de chance de acerto, e quando é medido ele acertar 50% das vezes, significa que ele está perfeitamente calibrado. Um algoritmo perfeitamente calibrado não significa que ele é um algoritmo acurado, mas que ele apresenta um resultado condizente com sua probabilidade estimada.^[45] Como pode ser visto no exemplo da Figura 12, que apresenta a comparação das curvas de calibração de dois algoritmos. A linha tracejada que representa a classificação perfeita e pode-se, a partir de uma avaliação visual, identificar que o algoritmo A2 se aproxima mais da linha tracejada. Portanto é o algoritmo A2 que está melhor calibrado.

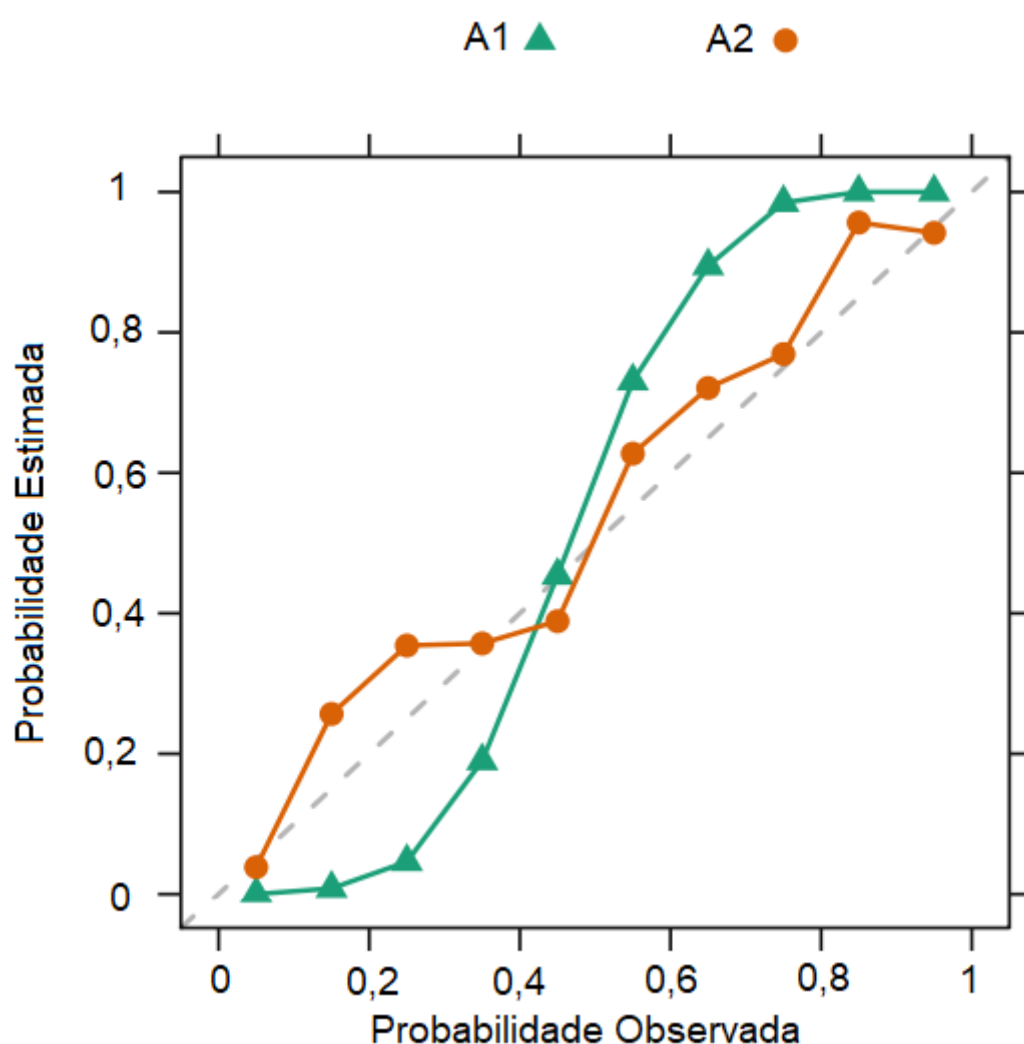


Figura 12: Um exemplo de comparação das curvas de calibração de dois algoritmos (A1 e A2)

Fonte: o autor – Modificado ^[45]

4 MATERIAIS E MÉTODOS

Este estudo utilizou uma base de dados contendo dados de FVL de disléxicos e não-disléxicos provenientes de tabelas de leitura (MNREAD-P) com frases independentes entre si apresentadas em uma linha (1L) e três linhas (3L). Este estudo possui duas etapas: uma de caráter quantitativo e exploratório e uma segunda etapa de caráter quantitativo e correlacional. Na Figura 13 é apresentado um diagrama com o método do estudo: Na primeira etapa, foi realizada uma extração de características com os dados de 1L e 3L separadamente. Inicialmente foi utilizada a técnica de Mapas Auto-organizáveis (SOM) para agrupar os registros em *clusters*, em seguida para interpretar as regras que regem cada *cluster* do SOM foi aplicado um algoritmo de Árvore de Decisão (*Random Tree*). Então os *clusters* e as árvores resultantes deste processo foram interpretados por um especialista da área para identificar as características que possuíam a possibilidade de auxiliar no estudo e diagnóstico de dislexia.

Na segunda etapa foram utilizados somente dados de 3L. Inicialmente foram removidos os ruídos (variáveis altamente correlacionadas e sem mudança de valores dentro de uma classe) da base de dados. Os *outliers* foram selecionados por um especialista da área de acordo com avaliação da necessidade imperativa de o mesmo continuar na base, caso contrário foram removidos, pois os mesmos impactam muito na classificação e especialmente na geração sintética de dados. Com os dados provenientes desta seleção foram gerados dados sintéticos para que ambas as classes possuissem 100 registros.

Foram selecionados 5 (cinco) algoritmos supervisionados e, para cada um deles foi realizada uma seleção de características para selecionar o conjunto de variáveis que possuíam a maior área sob a curva ROC (AUC). Então cada um dos algoritmos foi executado com validação cruzada *leave-one-out* e os resultados desta execução foram colocados em um quadro de desempenhos contendo a sensibilidade, especificidade, acurácia e AUC (Quadro 3). Após as execuções foi gerado uma curva de calibração para comparar a probabilidade esperada do algoritmo e a obtida com a validação cruzada.

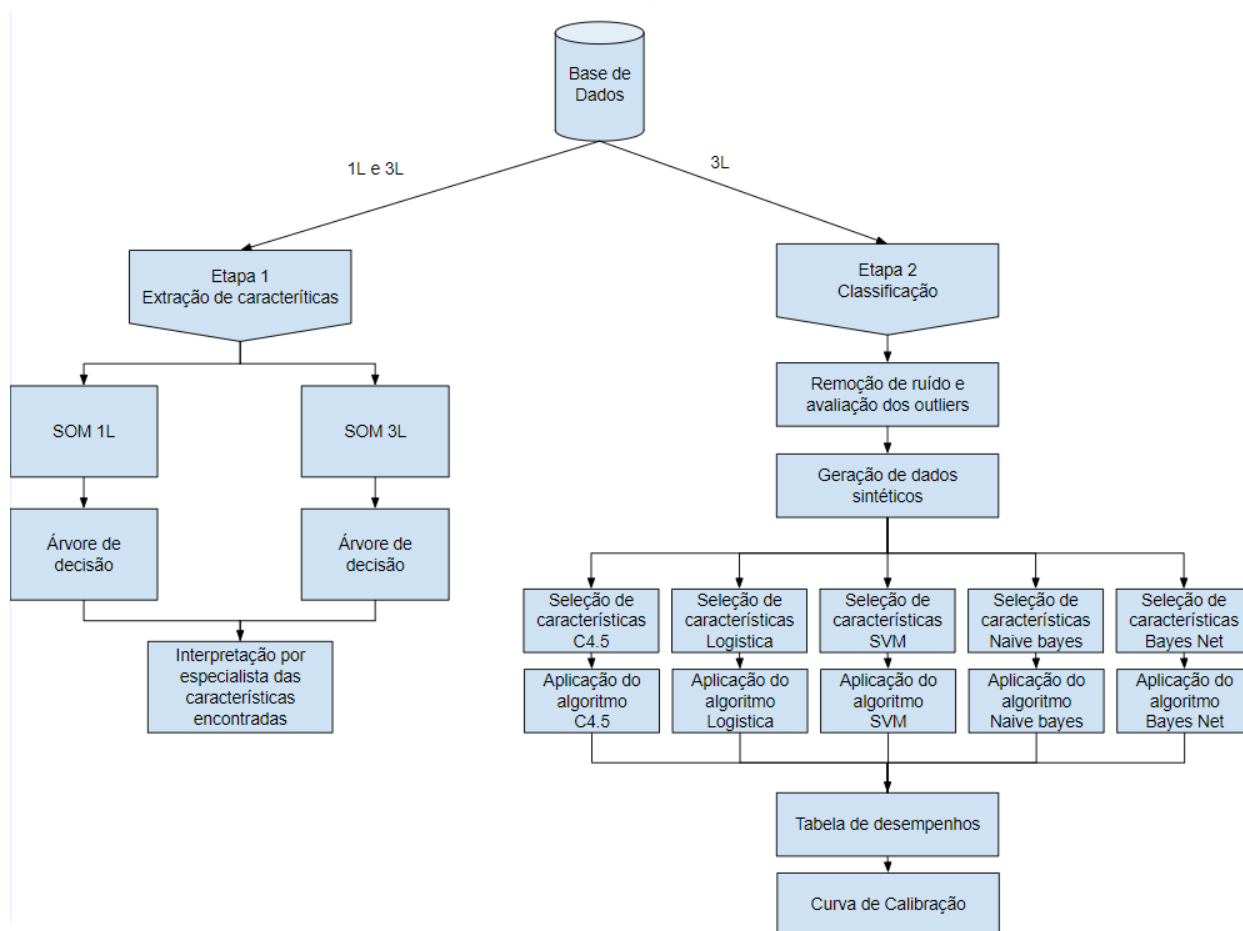


Figura 13: Diagrama do método em duas etapas do estudo.

Fonte: o autor

4.1 Base de dados

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

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 e 24 disléxicos 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 e sem antecedentes oftalmológicos (e.g., estrabismo ou cirurgia prévia). Todos os 18 participantes do grupo controle ($M = 23.67$ anos, $DP = 8.84$) foram considerados

aptos para participação no experimento. No grupo disléxico, 3 dos 24 participantes foram incapazes de realizar a tarefa de leitura do experimento, sendo, portanto, analisados os dados de 20 participantes ($M = 15.05$ anos, $DP = 6.55$). Todos os participantes, ou seus responsáveis legais, assinaram Termo de Consentimento Livre e Esclarecido (TCLE) previamente à coleta de dados. Essa coleta foi aprovada pelo Comitê de Ética em Pesquisa da Universidade Federal de São Paulo (CEP-597982/2015) e realizada no Laboratório de Bioengenharia Ocular no Departamento de Oftalmologia e Ciências Visuais da Unifesp, e o uso dos dados para esta pesquisa foi aprovado pelo mesmo comitê de ética (CEP-0899/2017) conforme Anexo 1.

Neste estudo, a tabela de leitura MNREAD (adaptada para português brasileiro MNREAD-P^[48] - tabela comumente usada na avaliação para adaptação de auxílios ópticos em pacientes com baixa visão) foi transformada e apresentada em formato digital. Dois formatos de tabela foram apresentados aos participantes de forma randômica e contrabalanceada:

1 - MNREAD – P1L - sentença apresentada em somente uma linha (Figura 14)

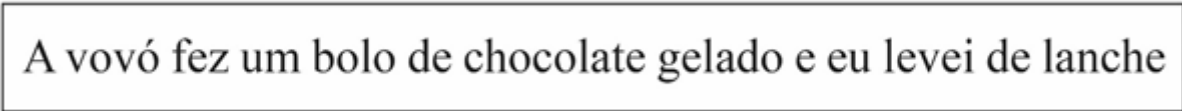


Figura 14: Representação de uma sentença em uma única linha do MNREAD-P1L

Fonte:^[49]

2 - MNREAD – P3L - sentença distribuída e apresentada em 3 linhas diferentes (Figura 15).

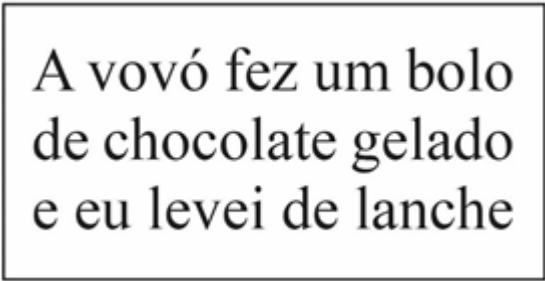


Figura 15: Representação de uma sentença de três linha do MNREAD-P3L

Fonte:^[49]

Cada formato de tabela consiste de 13 tentativas com sentenças diferentes, porém equivalentes em dificuldade.^[9] 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 entre 0.0 a 1.0 logMAR, apresentadas de forma decrescente (i.e., do maior, 1.0 logMAR, para o menor, 0.0 logMAR). A variação de tamanho foi de 0.1 logMAR por tentativa, e adotada a fonte Times New Roman.

Antes dos testes de leitura, cada participante teve sua AV monocular e binocular analisada utilizando a Tabela “*Lea Numbers Near Vision Card*” (Referência 251000) em formato físico (i.e., impresso em um quadro de poliéster 20.3cm x 25.4cm). O teste da AV foi realizado na mesma distância do teste de leitura, a 40cm entre o olho do participante e o estímulo.

Neste estudo, a tarefa consistiu na leitura da tabela MNREAD-P nos dois formatos, MNREAD – P1L e MNREAD – P3L.

A leitura dos dois formatos de tabela, MNREAD-P1L e MNREAD-P3L, foi feita de forma randomizada e contrabalanceada entre grupos. Antes do início da tarefa de leitura em ambos experimentos, todos os participantes receberam as mesmas instruções sobre cada tabela e tarefa a ser realizada:

2. MNREAD-P3L: “Várias frases serão apresentadas, uma de cada vez. Cada frase é apresentada em três linhas. Você deverá ler as frases em voz alta, como lê normalmente. Após você ter acabado de ler uma sentença, é apresentada uma estrela de fixação, você deve olhá-la para que a próxima sentença apareça na tela.”

1. MNREAD-P1L: “Várias frases serão apresentadas, uma de cada vez. Cada frase é apresentada em uma única linha. Você deverá ler as frases em voz alta, como lê normalmente. Após você ter acabado de ler uma sentença, é apresentada uma estrela de fixação, você deve olhá-la para que a próxima sentença apareça na tela.”

A coleta resultou em um banco de dados com as seguintes variáveis (Quadro 1).

Quadro 1: Variáveis de funções visuais de leitura coletadas no banco de dados

Variável	Descrição
GRUPO	Controle ou Dislético
AL	Acuidade de leitura que consiste no menor tamanho de letra lido
VL	Velocidade média de leitura
LMVL	Linha máxima de leitura consiste no tamanho da letra no momento da máxima velocidade de leitura
MVL	Máxima velocidade de leitura
TCL	Tamanho crítico de letra
VTCL	Velocidade de leitura no tamanho crítico de letra
Delta_MVL_TCL	Diferença de velocidade de leitura entre a máxima e no TCL
Delta_VTCL_TCL	Diferença do tamanho de letra na maior velocidade de leitura e no tamanho crítico de letra
ATCL	Acurácia no tamanho crítico de letra, consiste no número de palavras lidas erroneamente sobre o total de palavras
ITCL	Se houve Interferência no sentido dos erros cometidos no tamanho crítico de letra

4.2 Etapa 1

Esta etapa tem como objetivo principal explorar os dados de FVL coletados com estímulos de uma (1L) e três linhas (3L) separadamente e encontrar informações relevantes para o estudo e o diagnóstico da dislexia utilizando extração de características.

Para a tarefa de extração de características foram selecionados o algoritmo não supervisionado SOM e o algoritmo supervisionado árvore aleatória para encontrar as regras que regem os clusters agrupados pelo SOM.

4.3 Etapa 2

Esta etapa tem como objetivo classificar automaticamente disléxicos com uma base contendo dados de FVL coletada com estímulos de três linhas (3L) lidas por 21 disléxicos e 18 não-disléxicos.

Para avaliação dos outliers foram utilizados gráficos de boxplot.

Para a tarefa de classificação foram selecionados 5 algoritmos supervisionados, 4 indicados para pequenas bases por serem mais simples e possuírem uma variância menor.^[11,28,29] (C4.5, regressão logística, Máquina de vetores de suporte - SVM, *Naive Bayes*) e 1 algoritmo que em testes preliminares obteve melhores resultados com os dados originais (*Bayes Net*)

Para seleção de características foi utilizado o `wrapperSubsetEval` com busca exaustiva em cada algoritmo.

Os algoritmos foram comparados pela sua área sob a curva ROC (AUC) e pela sua curva de calibração.

4.4 Softwares

Para executar o algoritmo SOM foi utilizado MATLAB² com *Som Toolbox*. Os dados foram normalizados. A estrutura do mapa configurado para 2x2; *lattice* em hexa (formato hexagonal); e os rótulos em freq. (frequência).

Para executar o SVM, Rede Bayesiana, C4.5, Árvore Aleatória, *Naive Bayes* e Regressão Logística foi utilizado o *Weka*³ nas seguintes configurações respectivamente todos utilizando como opção de teste: *leave-one-out*.

Na execução do SVM foi utilizado a implementação do classificador no Weka chamado SMO com parâmetro de complexidade de 1,0; parâmetro de tolerância de 0,001; e *epsilon* de arredondamento de erro de 1,0E-12; tipo de filtro configurado para "*Normalize training data*" (Normalizar os dados de treinamento); número de *folds* para gerar a calibração do modelo configurado para -1 (usar os dados de treinamento); semente de aleatoriedade de 1; a função de *Kernel* utilizada foi "*Poly*

² Disponível em <https://www.mathworks.com/products/matlab.html>

³ Disponível em <https://www.cs.waikato.ac.nz/ml/weka/>

Kernel” com expoente de 1,0 e tamanho de *cache* de 250007; e a função de calibração usada foi “*logistic*” com *ridge* de 1,0E-8; número máximo de iterações configurada em -1; e número de casas decimais configurado para 4.

Na execução da Rede Bayesiana foi utilizado a implementação do classificador no *Weka* chamado *BayesNet* com “*useADTree*” em falso; algoritmo de busca setado para “K2” com número de pais para 1; o tipo de pontuação configurado para “*BAYES*”; estimador configurado para “*Simple Estimator*” com *alpha* configurado para 0,5.

Na execução do C4.5 foi utilizado a implementação do classificador no *Weka* chamado J48 com fator de confiança de 0,25; e número mínimo de instâncias por folha configurado em 2.

Na execução da Árvore Aleatória foi utilizado a implementação do classificador no *Weka* chamado *RandomTree* com o número de atributos escolhidos aleatoriamente setado em 0; o número mínimo total de pesos das instâncias em uma folha configurado em 1,0; a proporção da variância configurado para 0,001; e semente setado para 1.

Na execução do *Naive bayes* foi utilizado a implementação do classificador no *Weka* chamado *Naive Bayes* configurado com número de casas decimais para 2.

Na execução da Regressão Logística foi utilizado a implementação do classificador no *Weka* chamado *Logistic* configurado o número de casas decimais para 4 e o *ridge* para 1E-8.

Para gerar as curvas de calibração foi utilizado a biblioteca do *Weka* de curva de calibração em cada algoritmo da etapa 2, e foram exportados e importados para o *MS Excel* 2010⁴ para fazer um gráfico unificado.

Para gerar os *boxplots* na avaliação dos *outliers* foi utilizado o *software* JASP⁵.

⁴ Disponível em <https://www.microsoft.com/pt-br/download/details.aspx?id=28518>

⁵ Disponível em <https://jasp-stats.org/>

5 RESULTADOS

Neste capítulo são apresentados os resultados desta dissertação.

5.1 Etapa 1

Nessa seção são apresentados os resultados da etapa 1.

5.1.1 *Clusters e Árvore com 1L*

Para avaliar os *clusters* agrupados pelo algoritmo SOM, cada *cluster* foi denominado disléxico ou controle de acordo com a maior incidência de casos agrupados dentro do mesmo. O teste feito com o protocolo MNREAD-P1L resultou no C1 como controle e C2, C3 e C4 como disléxico como apresentado na Figura 16, representando os clusters com rótulo disléxico com o fundo cinza claro e os com rótulo de controle com o fundo cinza escuro.

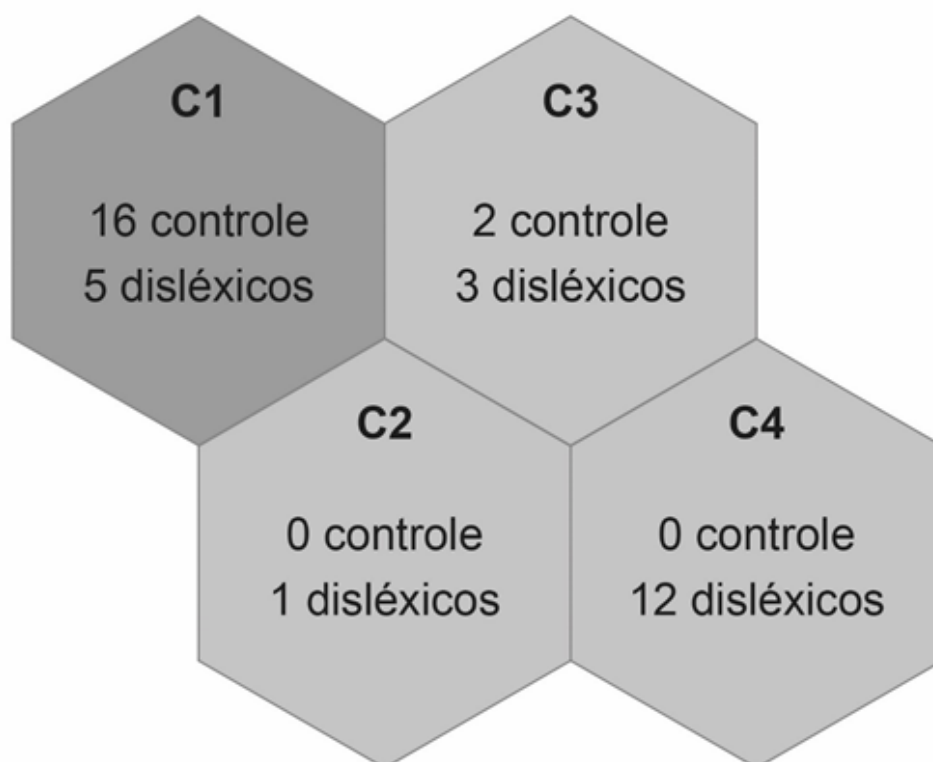


Figura 16: Cluster baseados nas leituras do MNREAD-P1L.

Fonte: o autor

O algoritmo *Random Tree* foi utilizado para avaliar quais variáveis são mais importantes para determinar cada *cluster* como apresentado na Figura 17.

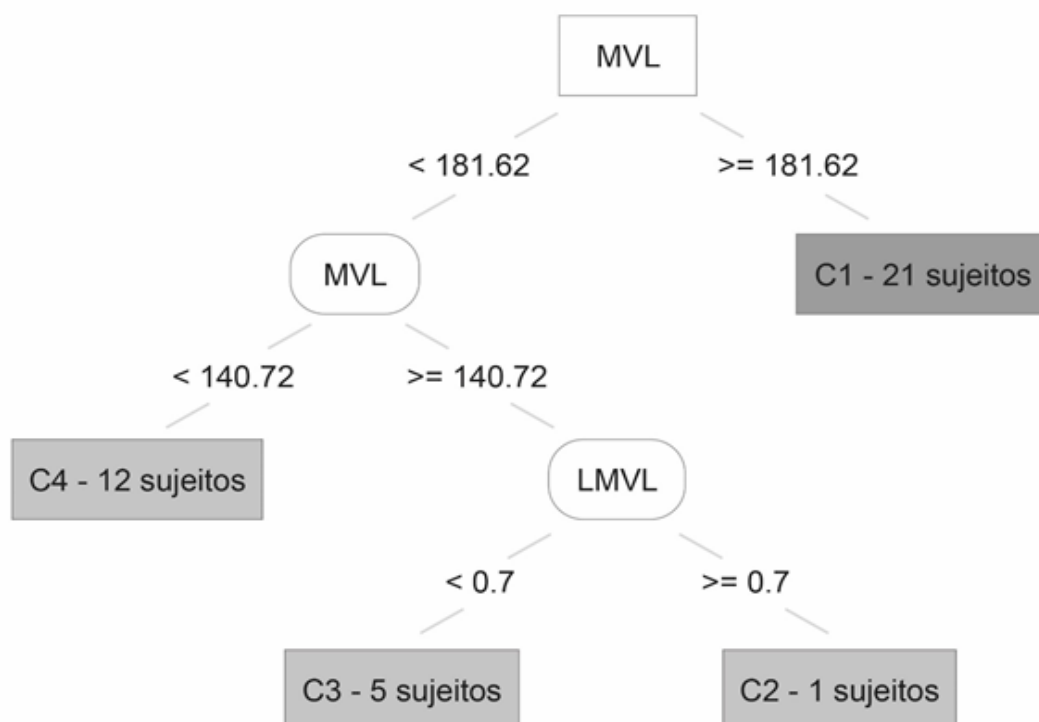


Figura 17: Árvore de decisão baseada nos *clusters* das leituras do MNREAD-P1L

Fonte: o autor

É possível observar, na Figura 17, que o $MVL > 181,62$ ppm separa a maioria dos não-disléxicos (*cluster* 1: 16 não-disléxicos e 5 disléxicos). Em outras palavras, não-disléxicos tendem ter velocidade de leitura maior do que disléxicos. Da mesma forma, sujeitos que foram agrupados no C4 (Figura 16, 12 disléxicos) tiveram $MVL < 140,72$ ppm. Os sujeitos agrupados no C2 e C3 (Figura 16) estão em uma região nebulosa do SOM e não se pode obter conclusões sobre estes sujeitos.

5.1.2 *Clusters* e árvore com 3L

O mesmo método que foi aplicado para as leituras de 1L foram feitas para o de 3L resultando em C1, C2 e C3 como controle e C4 como disléxico como pode ser visto na Figura 18, representando os *clusters* com rótulo disléxico com o fundo cinza claro e os com rótulo de controle com o fundo cinza escuro.

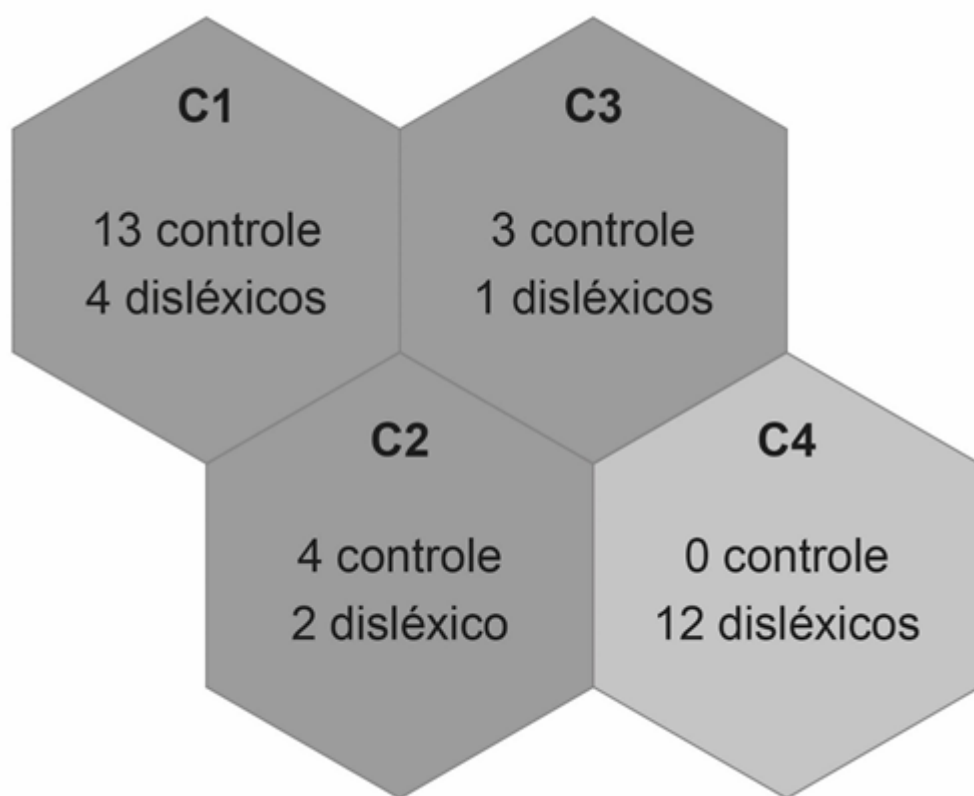


Figura 18: Clusters baseados nas leituras do MNREAD-P3L

Fonte: o autor

O algoritmo *Random Tree* foi executado resultando no diagrama apresentado na Figura 19.

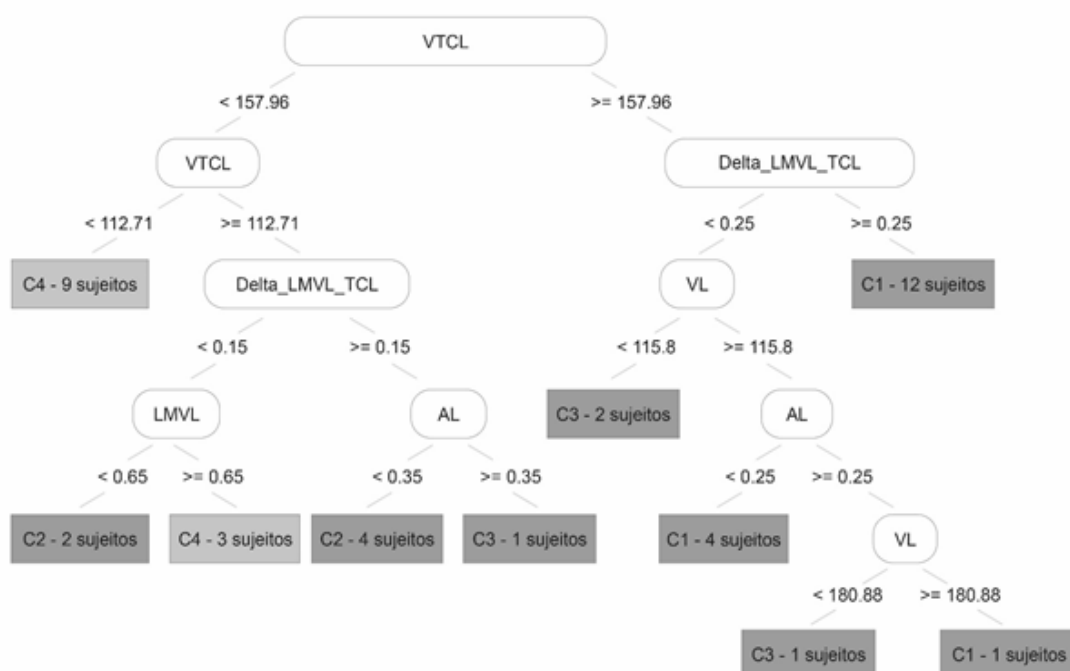


Figura 19: Árvore de decisão baseada nos *clusters* das leituras do MNREAD-P3L

Fonte: o autor

Nesta árvore apresentada na Figura 19 o valor de $VTCL < 157,96$ ppm determina que a velocidade de leitura do tamanho crítico de letra é uma variável importante para prever o comportamento de disléxicos, considerando que todos os sujeitos do C4 (Figura 18) estão do lado esquerdo da árvore.

5.2 Etapa 2

Nesta seção são apresentados os resultados da etapa 2.

5.2.1 Variáveis e *Outliers* Removidos

Neste processo foram removidas as variáveis de ATCL e ITCL pois nenhum participante do grupo controle cometeu erros na leitura, portanto também não havia alterações no significado.

As variáveis Delta_MVL_TCL e Delta_VTCL_TCL foram removidas neste critério pois são medidas obtidas pela subtração de duas outras variáveis já presentes na base.

Os outliers foram detectados utilizando boxplots de cada variável que permaneceu na base, como apresentado na Figura 20 com os dados de controle em vermelho e os de disléxicos em azul.

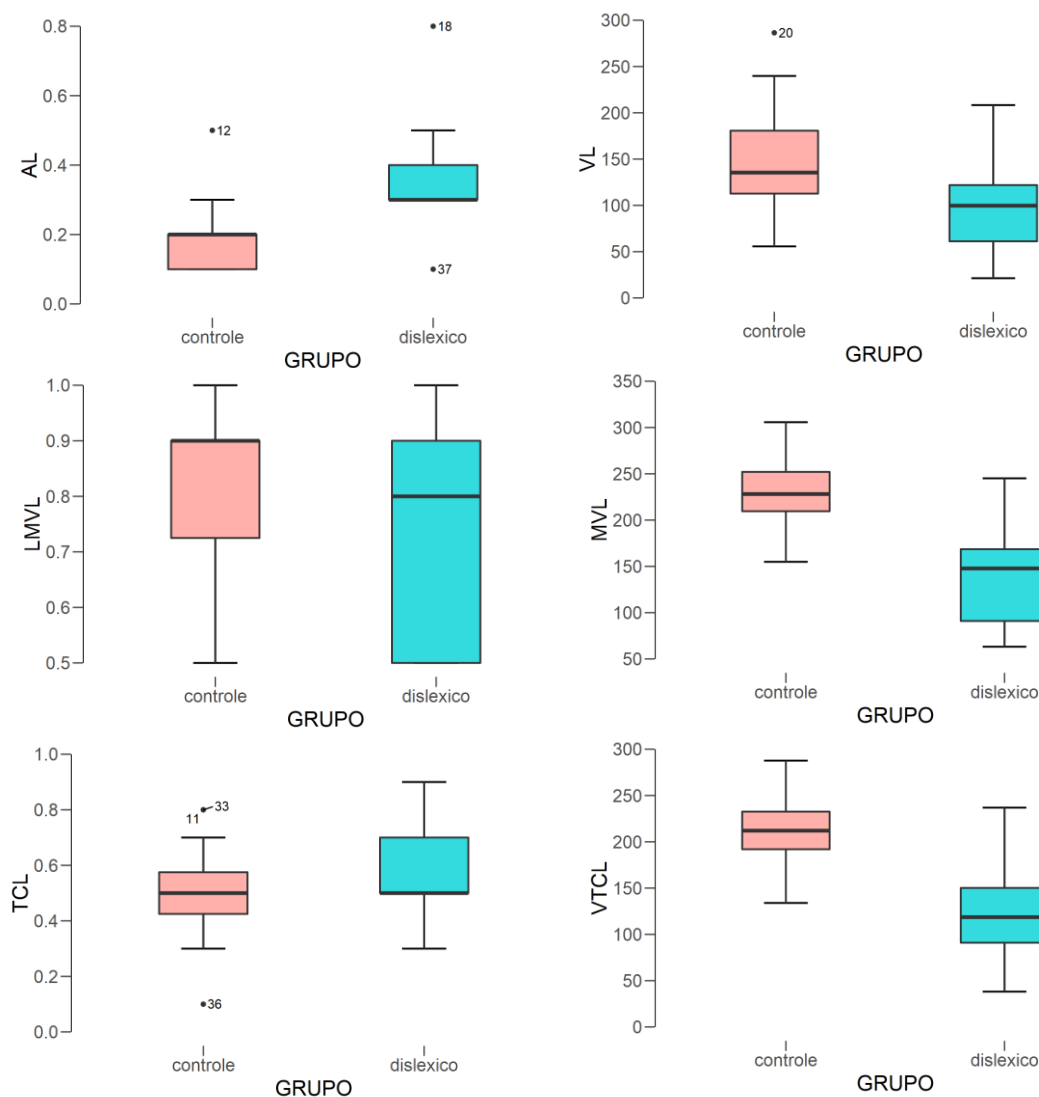


Figura 20: Boxplots com outliers

Fonte: o autor

Cada *outlier* 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. Nesta avaliação foram removidos um disléxico e quatro controles resultando em uma base com 20 disléxicos e 14 não-disléxicos.

5.2.2 Geração Sintética de Dados

O algoritmo SMOTE foi utilizado para aumentar a quantidade de registros - até 100 registros para cada grupo. Assim, o SMOTE foi parametrizado para selecionar 2 vizinhos próximos no cálculo de cada novo registro. Desta maneira, foi aumentada a quantidade de registros do grupo disléxico em 400%, e em 615% para o grupo não-disléxico. Como exemplo, pode-se ver a distribuição desta geração sintética de dados na base na Figura 21 – os dados em azul são de disléxicos e em vermelho os dados de não disléxicos, antes e depois da aplicação do SMOTE.

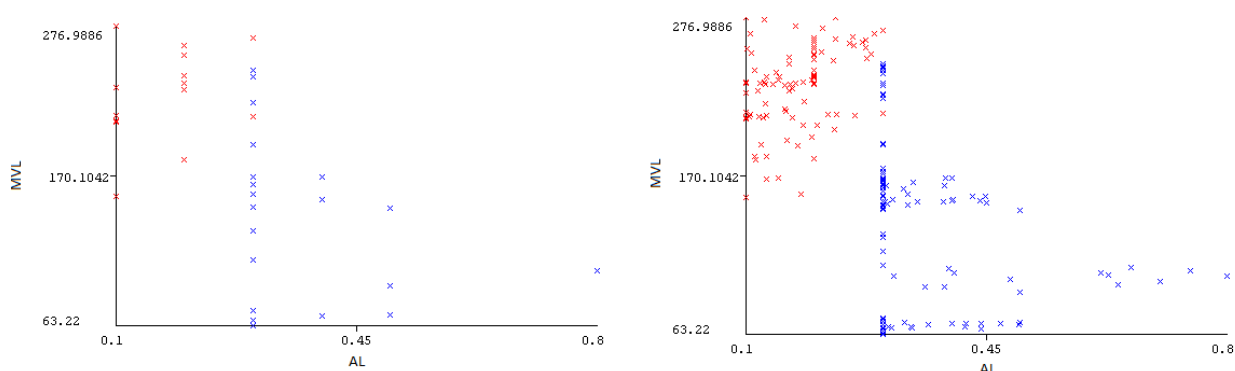


Figura 21: Gráfico de dispersão de dados de disléxicos (em azul) e não-disléxicos (em vermelho), com dados de MVL e AL, antes (esquerda) e depois (direita) do uso do SMOTE

Fonte: o autor

5.2.3 Características Seleccionadas

No Quadro 2 estão apresentadas as Características Seleccionadas

Quadro 2: Atributos seleccionados para cada algoritmo levando em conta a melhor AUC.

Algoritmo	Atributos seleccionados
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

5.2.4 Classificações

Abaixo é apresentado o Quadro 3, a Tabela comparativa dos algoritmos sobre sensibilidade, especificidade, acurácia e AUC

Quadro 3: Tabela comparativa dos algoritmos sobre sensibilidade, especificidade, acurácia e AUC

Algoritmo	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

Todos os algoritmos selecionados tiveram bons desempenhos. A Regressão Logística alcançou o melhor desempenho, apresentando AUC com uma diferença de 0,002 acima, se comparado com a Rede Bayesiana. Então, uma execução com e sem o SMOTE foi feita, para verificar o efeito total da aplicação da geração sintética de dados no resultado final. Estes podem ser visto no Quadro 4.

Quadro 4: 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

5.2.5 Curva de Calibração

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

na Figura 22. Neste a linha pontilhada representa a calibração perfeita quanto mais próximo desta linha melhor calibrado está o algoritmo.

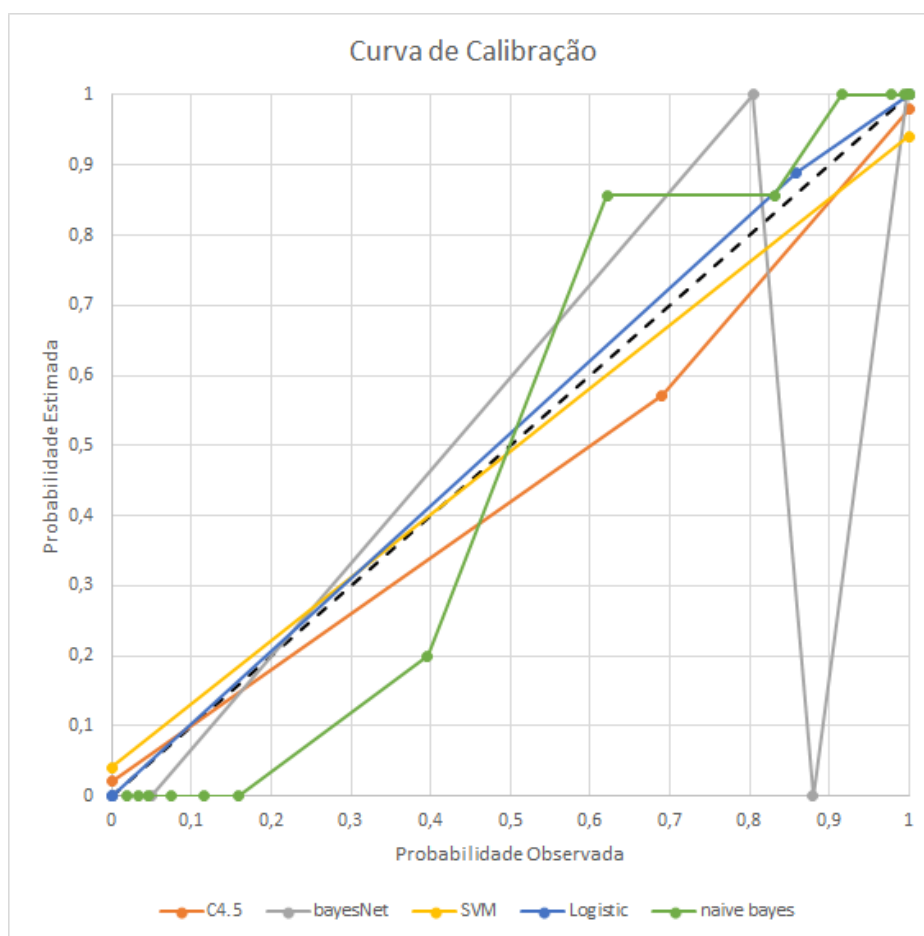


Figura 22: Curvas de calibração dos algoritmos

Fonte: o autor

6 DISCUSSÃO

Nesta seção é realizada a discussão desta dissertação.

6.1 Etapa 1

Normalmente, velocidade de leitura (VL) é estável por vários tamanhos de estímulos para um mesmo indivíduo. Conforme o tamanho da letra diminui, existe um ponto onde a velocidade de leitura começa a diminuir - este ponto é conhecido como tamanho crítico de letra (TCL). A acuidade de leitura (AL) é alcançada quando um sujeito lê o estímulo no menor tamanho angular possível sem cometer erros significantes. De acordo com a literatura, o tamanho de letra correspondente ao TCL é normalmente duas vezes maior que a AL.^[18] Portanto, VL é utilizado como métrica que permite a derivação de outras variáveis que reflete tanto aspectos visuais como não visuais.^[50]

No estudo do “Efeito do tamanho da letra na velocidade de leitura em disléxicos”^[9] as curvas de velocidade de leitura por tamanho da letra em disléxicos seguiram os mesmos padrões que não-disléxicos. Não obstante, as curvas de leituras de disléxicos apresentaram tamanhos maiores de letras e velocidade de leitura mais lenta por tamanho de letra abaixo do TCL. O presente estudo corrobora com estes achados, uma vez que existiram diferenças significantes entre os grupos em termos de máxima velocidade de leitura, acuidade de leitura e, especialmente, tamanho crítico de letra. Disléxicos precisam de tamanho de letra maior para atingir o mesmo desempenho que leitores não-disléxicos. É por isso que nós recomendamos fortemente a investigação de tamanho crítico de letra para cada indivíduo, especialmente nos dias atuais, onde o uso de mídias digitais (i.e., *tablets* etc.) tem sido cada vez mais comum no processo educacional. Portanto, a árvore de decisão apresentada na Figura 17, mostra a importância da máxima velocidade de leitura em diferenciar ambos os grupos. O uso conjunto dos algoritmos SOM e árvore de decisão no processo de extração de características foi uma ferramenta poderosa que recomendamos o seu uso em paralelo com a análise da estatística clássica. Isto permitiu identificar variáveis importantes que devem ser consideradas para identificar pessoas com distúrbios relacionados a aprendizagem. Neste estudo,

a extração de características corroborou ao fato da velocidade de leitura e tamanho crítico de letra são chaves em predize-los. Como mostrado na Figura 17, a importância do TCL é traduzida pela distinção de VTCL. Como reportado Castro et al,^[48] acuidade de leitura também apareceu como uma variável importante. Disléxicos apresentaram uma menor velocidade de leitura, um maior tamanho crítico de letra e maior acuidade de leitura. Interessantemente não houve diferença significativa observada entre os 2 protocolos apresentados na análise estatística convencional. Estes achados nos intrigaram, uma vez que o efeito de *crowding* pode ter sido mais pronunciado no MNREAD-P3L. Contudo, o fato de as sentenças não ter nenhum contexto entre elas pode ter influenciado o comportamento dos disléxicos.

6.2 Etapa 2

A avaliação dos *outliers* foi importante para remover registros que pudessem impactar negativamente na classificação e ainda poderiam influenciar o SMOTE em gerar mais dados não representativos. Por outro lado, os registros que eram representantes específicos dos seus respectivos grupos também foram importantes na geração de dados sintéticos, melhorando o espectro dos dados.

A execução do SMOTE gerou um espectro mais completo da amostra, que pode ser visto na Figura 21. Assim, como esperado, uma “nuvem” de pontos foi gerada em direção do *outlier*, criando um espaço de instâncias mais definido nesta área específica. A execução do SMOTE antes da seleção de atributos ocorreu com o objetivo de evitar que esses atributos fossem selecionados por *overfitting*, e desta forma prejudicando a classificação.

Observando o Quadro 2, é possível averiguar que os 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 mesma foi selecionada 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. Especificamente, o classificador que obteve a melhor AUC (0,999) foi a regressão logística.

O ganho de performance do melhor classificador (regressão logística) com a aplicação da geração sintética de dados, apresentado no Quadro 4, foi de 10,77 pontos percentuais de acurácia e 0,025 na AUC, demonstrando o impacto do SMOTE na melhora de performance da classificação.

Observando a curva de calibração dos algoritmos é possível averiguar 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 - indica que o mesmo não garante a mesma performance em dados desconhecidos, não sendo indicado para classificação de disléxicos com base em FVL.

7 CONCLUSÃO

Esta dissertação extraiu características em funções visuais de leitura de disléxicos a partir da aplicação das técnicas SOM e árvore aleatória. A partir deste estudo, foi verificado que, por meio do uso do protocolo MNREAD-P1L, disléxicos possuem máxima velocidade de leitura menor que 140.72 palavras por minuto, se comparado com não-disléxicos. Também foi detectado que, por meio do uso do protocolo MNREAD-P3L, a VTCL também é menor nos disléxicos obtendo valores inferiores a 112.71 palavras por minuto.

Além disto, foi detectado que disléxicos também possuem problemas relacionados ao tamanho crítico de letra, quando apresentado especificamente textos com 3 linhas (MNREAD-P3L). Tal análise sugere interferência de *crowding* para funções visuais de leitura em disléxicos. Estudos futuros podem explorar o *crowding* em funções visuais de leitura em disléxicos.

Esta dissertação também abordou a classificação de disléxicos e não-disléxicos com geração de dados sintéticos e seleção de características. A partir desta abordagem, foi possível averiguar que o fato do atributo AL estar presente na seleção de atributos de 4 de 5 algoritmos indica que essa é uma característica importante na diferenciação entre disléxicos e não-disléxicos. Desta forma se tornando uma característica candidata para ser utilizada precocemente em triagens de dislexia.

Neste estudo o atributo AL ficou bem evidente pois os participantes foram selecionados pelo critério de inclusão de possuir uma acuidade visual de no máximo 0,3 logMAR. Os participantes não-disléxicos obtiveram uma AL menor ou igual a 0,3 logMAR, já os disléxicos obtiveram uma AL maior ou igual 0,3 logMAR.

Mais estudos são necessários para se avaliar este efeito de a AL ter uma tendência de ser maior que a acuidade de visual em disléxicos.

Todos os algoritmos avaliados para fazer a classificação de disléxicos obtiveram uma acurácia na classificação superior a 0,95%. Neste sentido, destaque-se 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 de o algoritmo ter a mesma performance em dados desconhecidos, a regressão logística demonstrou-se bem mais próxima da calibração perfeita, se

comparada com os outros algoritmos. Este cenário reforça o bom desempenho da regressão logística 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 FVL.

8 REFERÊNCIAS

1. ICD-11 - Mortality and Morbidity Statistics [Internet]. [citado 2019 nov 16];Available from: <https://icd.who.int/browse11/l-m/en#/http://id.who.int/icd/entity/1008636089>
2. Lima e Silva NML. A Prevalência da dislexia em alunos do Ensino Fundamental de escolas particulares." 1 (2004): 1. Universidade Federal de Santa Maria 2004;1:1.
3. Lyon G, Shaywitz SE, Shaywitz BA. A definition of dyslexia. *Ann of Dyslexia* 2003;53(1):1–14.
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;18(3):216–228.
5. Como é feito o Diagnóstico? – ABD | Associação Brasileira de Dislexia [Internet]. [citado 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 [Internet]. In: *Proceedings of the 12th Web for All Conference*. New York, NY, USA: ACM; 2015 [citado 2017 jun 4]. página 16:1–16:8.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 [citado 2016 set 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;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;28(3):332–49.
10. Awad M, Khanna R. *Efficient learning machines: theories, concepts, and applications for engineers and system designers*. Berkley? Apress Open; 2015.
11. Bishop C. *Pattern Recognition and Machine Learning*. New York: Springer; 2007.
12. Alpaydin E. *Introduction to Machine Learning*. MIT Press; 2009.
13. Rezende SO, organizador. *Sistemas inteligentes: fundamentos e aplicações*. 1. ed. Barueri, SP: Ed. Manole; 2003.

14. Mohri M, Rostamizadeh A, Talwalkar A. Foundations of machine learning. Second edition. Cambridge, Massachusetts: The MIT Press; 2018.
15. Haykin S. Neural Networks: A Comprehensive Foundation. 2 edition. Upper Saddle River, N.J.: Prentice Hall; 1998.
16. Theodoridis S, Koutroumbas K. Pattern recognition. 4. ed. Amsterdam: Elsevier Acad. Press; 2009.
17. Rotta NT, Ohlweiler L, Santos Riesgo R dos. Transtornos da aprendizagem: abordagem neurobiológica e multidisciplinar [Internet]. Porto Alegre: Bookman; 2007 [citado 2019 fev 3]. Available from: <http://site.ebrary.com/id/10707248>
18. Calabrèse A, Owsley C, McGwin G, Legge GE. Development of a Reading Accessibility Index Using the MNREAD Acuity Chart. JAMA Ophthalmol 2016;134(4):398–405.
19. Bailey IL, Lovie JE. New design principles for visual acuity letter charts. Am J Optom Physiol Opt 1976;53(11):740–5.
20. Bailey IL, Lovie-Kitchin JE. Visual acuity testing. From the laboratory to the clinic. Vision Research 2013;90:2–9.
21. Krishnan R, Sivakumar G, Bhattacharya P. Extracting decision trees from trained neural networks. Pattern Recognition 1999;32(12):1999–2009.
22. Ultsch A, Lötsch J. Machine-learned cluster identification in high-dimensional data. Journal of Biomedical Informatics 2017;66:95–104.
23. Corrales DC, Corrales JC, Ledezma A. How to Address the Data Quality Issues in Regression Models: A Guided Process for Data Cleaning. Symmetry 2018;10(4):99.
24. Xiong H, Gaurav Pandey, Steinbach M, Vipin Kumar. Enhancing data analysis with noise removal. IEEE Transactions on Knowledge and Data Engineering 2006;18(3):304–19.
25. Rousseeuw PJ, Leroy AM. Robust Regression and Outlier Detection. John Wiley & Sons; 2005.
26. Van den Broeck J, Argeseanu Cunningham S, Eeckels R, Herbst K. Data Cleaning: Detecting, Diagnosing, and Editing Data Abnormalities. PLoS Med [Internet] 2005 [citado 2019 set 23];2(10). Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1198040/>
27. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. jair 2002;16:321–57.
28. 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;425:76–91.

29. Raudys SJ, Jain AK. Small sample size effects in statistical pattern recognition: recommendations for practitioners. *IEEE Trans Pattern Anal Machine Intell* 1991;13(3):252–64.
30. Santos HG dos. Comparação da performance de algoritmos de machine learning para a análise preditiva em saúde pública e medicina [Internet]. 2018 [citado 2019 set 13];Available from: <http://www.teses.usp.br/teses/disponiveis/6/6141/tde-09102018-132826/>
31. He Y, Gao B, Sophian A, Yang R. Chapter 2 - Magnetic Sensor Based Pulsed Eddy Current for Defect Detection and Characterization [Internet]. In: He Y, Gao B, Sophian A, Yang R, organizadores. *Transient Electromagnetic-Thermal Nondestructive Testing*. Butterworth-Heinemann; 2017 [citado 2019 nov 2]. página 7–35.Available from: <http://www.sciencedirect.com/science/article/pii/B9780128127872000022>
32. Giannakopoulos T, Pikrakis A. Chapter 4 - Audio Features [Internet]. In: Giannakopoulos T, Pikrakis A, organizadores. *Introduction to Audio Analysis*. Oxford: Academic Press; 2014 [citado 2019 nov 2]. página 59–103.Available from: <http://www.sciencedirect.com/science/article/pii/B9780080993881000042>
33. Malone J, McGarry K, Wermter S, Bowerman C. Data mining using rule extraction from Kohonen self-organising maps. *Neural Comput & Applic* 2006;15(1):9–17.
34. Wu X, Kumar V, Quinlan JR, Ghosh J, Yang Q, Motoda H, et al. Top 10 algorithms in data mining. *Knowl Inf Syst* 2008;14(1):1–37.
35. Mishra AK, Ratha BK. Study of Random Tree and Random Forest Data Mining Algorithms for Microarray Data Analysis. 2016;3.
36. Kalmegh S. Comparative Analysis of WEKA Data Mining Algorithm RandomForest, RandomTree and LADTree for Classification of Indigenous News Data. *IJETAE* 2015;5(1):595.
37. RandomTree [Internet]. [citado 2019 nov 1];Available from: <http://weka.sourceforge.net/doc.dev/weka/classifiers/trees/RandomTree.html>
38. Quinlan JR. C4.5: Programs for machine learning. 2014.
39. Tolles J, Meurer WJ. Logistic Regression: Relating Patient Characteristics to Outcomes. *JAMA* 2016;316(5):533–4.
40. Azuaje F. Witten IH, Frank E: Data Mining: Practical Machine Learning Tools and Techniques 2nd edition. *BioMed Eng OnLine* 2006;5(1):51.
41. Lee S-M, Abbott PA. Bayesian networks for knowledge discovery in large datasets: basics for nurse researchers. *Journal of Biomedical Informatics* 2003;36(4):389–99.

42. Bradley AP. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition* 1997;30(7):1145–59.
43. Fan J, Upadhye S, Worster A. Understanding receiver operating characteristic (ROC) curves. *CJEM* 2006;8(01):19–20.
44. Park SH, Goo JM, Jo C-H. Receiver Operating Characteristic (ROC) Curve: Practical Review for Radiologists. *Korean J Radiol* 2004;5(1):11.
45. Kuhn M, Johnson K. *Applied Predictive Modeling* [Internet]. New York: Springer-Verlag; 2013 [citado 2019 set 24]. Available from: <https://www.springer.com/gp/book/9781461468486>
46. Vuk M, Curk T. ROC Curve, Lift Chart and Calibration Plot. :20.
47. Zupan B, Demšar J, Kattan MW, Beck JR, Bratko I. Machine learning for survival analysis: a case study on recurrence of prostate cancer. *Artificial Intelligence in Medicine* 2000;20(1):59–75.
48. Castro CTM de, Kallie CS, Salomão SR. Development and validation of the MNREAD reading acuity chart in Portuguese. *Arquivos Brasileiros de Oftalmologia* 2005;68(6):777–83.
49. Gonçalves E. Percepção visual durante a leitura um estudo das funções visuais e movimentação ocular em disléxicos e não-disléxicos. 2019;
50. Virgili G, Pierrottet C, Parmeggiani F, Pennino M, Giacomelli G, Steindler P, et al. Reading Performance in Patients with Retinitis Pigmentosa: A Study Using the MNREAD Charts. *Invest Ophthalmol Vis Sci* 2004;45(10):3418–24.

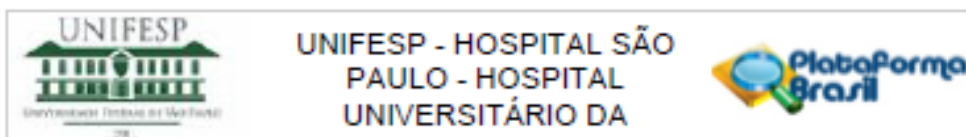
APÊNDICE I: FOLHA DE IDENTIFICAÇÃO DA DISSERTAÇÃO

UNIVERSIDADE FEDERAL DE SÃO PAULO ESCOLA PAULISTA DE MEDICINA PROGRAMA DE PÓS-GRADUAÇÃO EM GESTÃO E INFORMÁTICA EM SAÚDE

Pró-reitora:	Profa. Dra. Lia Rita Azeredo Bittencourt Pró-reitoria de Pós-graduação e Pesquisa - UNIFESP
Coordenadora da câmara:	Profa. Dra. Monica Levy Anderson Câmara de Pós-graduação e Pesquisa da - EPM - UNIFESP
Coordenador do curso:	Prof. Dr. Ivan Torres Pisa, livre docente
Curso:	Mestrado acadêmico
Título do projeto:	RECONHECIMENTO AUTOMÁTICO DE PADRÕES EM DISLEXIA: UMA ABORDAGEM BASEADA EM FUNÇÕES VISUAIS DE LEITURA E APRENDIZADO DE MÁQUINA
Aluno:	Antonio Carlos da Silva Junior
Orientação:	Prof. Dr. Felipe Mancini
Coorientação:	Prof. Dr. Paulo Schor e Dr ^a . Emanuela C. R. Gonçalves
Matrícula:	127773 – Período 01/10/2017 a 31/12/2019
CEP:	Comitê de Ética em Pesquisa da UNIFESP 0899/2017
Linha de pesquisa:	Registro, recuperação e relacionamento de dados em saúde
Grupo de pesquisa:	Bioengenharia Ocular http://dgp.cnpq.br/dgp/espelhogrupo/5894644663535064

Apoio financeiro:	
----------------------	--

ANEXO 1: APROVAÇÃO DO COMITÊ DE ÉTICA E PESQUISA



PARECER CONSUBSTANCIADO DO CEP

DADOS DO PROJETO DE PESQUISA

Título da Pesquisa: RECONHECIMENTO AUTOMÁTICO DA DISLEXIA: UMA ABORDAGEM BASEADA NO COMPORTAMENTO OCULAR E CLASSIFICADORES DE PADRÕES

Pesquisador: ANTONIO CARLOS DA SILVA JUNIOR

Área Temática:

Versão: 1

CAAE: 72290017.1.0000.5505

Instituição Proponente: Universidade Federal de São Paulo

Patrocinador Principal: UNIVERSIDADE FEDERAL DE SAO PAULO

DADOS DO PARECER

Número do Parecer: 2.267.112

Apresentação do Projeto:

Nº CEP: 0899/2017

Dislexia é um distúrbio de aprendizagem com origem neurológica. Este transtorno é caracterizado pela dificuldade específica e significativa no desenvolvimento das habilidades de leitura e escrita, e que não pode ser descrito por baixa idade mental, acuidade visual ou educação inadequada. Isto pode afetar a habilidade de compreensão, o reconhecimento da escrita e a desempenho em tarefas que necessitem de leitura. Este transtorno é complexo, demorado e oneroso de ser diagnosticado porque é multiprofissional e realizado por exclusão. Para auxiliar no diagnóstico de disléxicos, eye trackers (rastreadores do olhar, ET) podem ser utilizados para a coleta e análise dos dados do comportamento ocular durante a leitura. Este projeto tem como objetivo aplicar algoritmos computacionais para extrair características e classificar padrões oculares da leitura de disléxicos a partir dados provenientes de rastreadores de olhar. Trata-se, em um primeiro momento, de um estudo exploratório e quantitativo. Assim, será utilizada uma base de dados do comportamento ocular de disléxicos durante a leitura. A partir desta base, será aplicado um algoritmo conhecido como mapas auto-organizáveis (self-organizing maps, SOM) e árvore de decisão para a extração de características de dados do olhar. Em um segundo momento este projeto assume uma

Endereço: Rua Francisco de Castro, 55
Bairro: VILA CLEMENTINO **CEP:** 04.020-050
UF: SP **Município:** SAO PAULO
Telefone: (11)5571-1062 **Fax:** (11)5539-7162 **E-mail:** cep@unifesp.edu.br



UNIFESP - HOSPITAL SÃO
PAULO - HOSPITAL
UNIVERSITÁRIO DA



Continuação do Parecer: 2.207.112

característica correlacional e quantitativa. Neste caso, serão testados e comparados, a partir da análise dos valores de sensibilidade e especificidade, diversos classificadores de padrões para a tarefa de reconhecimento automático de disléxicos. Será resultante deste estudo uma avaliação dos valores de sensibilidade e especificidade alcançados a partir da aplicação de técnicas de classificação de padrões com dados do comportamento ocular de disléxicos. Espera-se que esta pesquisa possa contribuir com a detecção de variáveis de comportamento ocular relevantes para a classificação de disléxicos. Além disto, este projeto também almeja disponibilizar um classificador que aumente o ferramental para o diagnóstico de dislexia em falantes do português brasileiro.

Objetivo da Pesquisa:

.1 Objetivo geral Aplicar algoritmos computacionais para extrair características e classificar padrões oculares da leitura de disléxicos a partir dados provenientes de rastreadores do olhar (eye tracker, ET). .2 Objetivos específicos A partir de dados provenientes de ET, este trabalho tem como objetivos específicos: Aplicar o algoritmo de mapa auto organizável (self-organizing maps, SOM) para identificar padrões de comportamentos oculares em disléxicos; Utilizar algoritmos de classificação de padrões para reconhecer automaticamente os sujeitos disléxicos; Implementar uma ferramenta computacional que classifique automaticamente sujeitos disléxicos

Avaliação dos Riscos e Benefícios:

Segundo o pesquisador: Riscos: A presente pesquisa envolve riscos mínimos

Benefícios: Pretende-se disponibilizar um classificador que aumente o ferramental para o diagnóstico de dislexia em falantes do português brasileiro.

Comentários e Considerações sobre a Pesquisa:

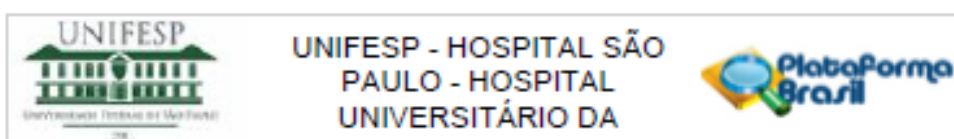
Projeto de pesquisa apresentado à Universidade Federal de São Paulo, para matrícula no curso de Mestrado no Programa de Pós-Graduação de Gestão e Informática em Saúde, pelo Departamento de Informática em Saúde, Campus São Paulo.

Orientador: Prof. Dr. Felipe Mandini

Coorientador: Profa. Dra. Marlina Navarro; Prof. LD. Paulo Schor

MÉTODO: Este projeto tem em um primeiro momento a característica de um estudo exploratório e quantitativo. Assim, será utilizada uma base de dados do comportamento ocular de disléxicos durante a leitura. A partir desta base, será aplicado um algoritmo conhecido como mapas auto-organizáveis (self-organizing maps, SOM) e árvore de decisão para a extração de características de

Endereço: Rua Francisco de Castro, 55
Bairro: VILA CLEMENTINO CEP: 04.020-050
UF: SP Município: SÃO PAULO
Telefone: (11)5571-1062 Fax: (11)5530-7162 E-mail: cep@unifesp.edu.br



Continuação do Parecer: 2.207.112

dados do olhar. Em um segundo momento este projeto assume uma característica correlacional e quantitativa. Neste caso, serão testados e comparados, a partir da análise dos valores de sensibilidade e especificidade, diversos classificadores de padrões para a tarefa de reconhecimento automático de disléxicos. **RESULTADOS ESPERADOS:** Será resultante deste estudo uma avaliação dos valores de sensibilidade e especificidade alcançados a partir da aplicação de técnicas de classificação de padrões com dados do comportamento ocular de disléxicos. Espera-se que esta pesquisa possa contribuir com a detecção de variáveis de comportamento ocular relevantes para a classificação de disléxicos. Além disto, este projeto também almeja disponibilizar um classificador que aumente o ferramental para o diagnóstico de dislexia em falantes do português brasileiro. Este trabalho analisará dados provenientes do grupo de pesquisa Bioengenharia Ocular da UNIFESP, coordenado pelo Prof. LD. Paulo Schor - também coordenador deste projeto. Cuja aquisição prévia dos dados, aprovado por meio do CEP número CAAE 20491913.8.0000.5505.

Considerações sobre os Termos de apresentação obrigatória:

Documentos obrigatórios apresentados: Folha de Rosto *folhaDeRosto.pdf*; Projeto Detalhado / Brochura Investigador Projeto_Antonio_Carlos_da_Silva_Junior20170619.pdf cadastro - CEP1.pdf

Recomendações:

Nada consta

Conclusões ou Pendências e Lista de Inadequações:

Sem inadequações

Considerações Finais a critério do CEP:

Parecer do relator acatado pelo colegiado

O CEP informa que a partir desta data de aprovação, é necessário o envio de relatórios parciais (anualmente), e o relatório final, quando do término do estudo.

Este parecer foi elaborado baseado nos documentos abaixo relacionados:

Tipo Documento	Arquivo	Postagem	Autor	Situação
Informações Básicas do Projeto	PB_INFORMACOES_BASICAS_DO_PROJETO_961154.pdf	31/07/2017 19:24:29		Aceito
Projeto Detalhado / Brochura	Projeto_Antonio_Carlos_da_Silva_Junior20170619.pdf	31/07/2017 19:23:27	ANTONIO CARLOS DA SILVA JUNIOR	Aceito

Endereço: Rua Francisco de Castro, 55
 Bairro: VILA CLEMENTINO CEP: 04.020-050
 UF: SP Município: SÃO PAULO
 Telefone: (11)5571-1062 Fax: (11)5530-7162 E-mail: cep@unifesp.edu.br



UNIFESP - HOSPITAL SÃO
PAULO - HOSPITAL
UNIVERSITÁRIO DA



Continuação do Parecer: 2.207.112

Investigador	Projeto_Antonio_Carlos_da_Silva_Junio r20170619.pdf	31/07/2017 19:23:27	ANTONIO CARLOS DA SILVA JUNIOR	Acelto
Outros	CEP1.pdf	31/07/2017 19:18:07	ANTONIO CARLOS DA SILVA JUNIOR	Acelto
Outros	folhaDeRosto2.pdf	31/07/2017 19:17:43	ANTONIO CARLOS DA SILVA JUNIOR	Acelto
Folha de Rosto	folhaDeRosto.pdf	31/07/2017 19:15:45	ANTONIO CARLOS DA SILVA JUNIOR	Acelto

Situação do Parecer:

Aprovado

Necessita Apreciação da CONEP:

Não

SAO PAULO, 20 de Setembro de 2017

Assinado por:
Miguel Roberto Jorge
(Coordenador)

Endereço: Rua Francisco de Castro, 55
Bairro: VILA CLEMENTINO CEP: 04.020-050
UF: SP Município: SAO PAULO
Telefone: (11)5571-1062 Fax: (11)5539-7162 E-mail: cep@unifesp.edu.br



COMITÊ DE ÉTICA EM PESQUISA



PREENCHIMENTO PELO CEP

Nota Técnica: _____ Data da relatoria: ____/____/____ Relator: _____

Nº CEP: _____/2017 CAAE: _____, 5505

ESTE DOCUMENTO DEVERÁ SER COLOCADO NA PLATAFORMA BRASIL. DESDE 01/01/2017 NÃO PRECISA MAIS ENTREGAR NO CEP.

IDENTIFICAÇÃO DO PESQUISADOR PRINCIPAL

Antonio Carlos da Silva Junior (Especialização) E-MAIL: tymonex@hotmail.com CEL: 11995141441 CPF: 31375628801

VÍNCULO INSTITUCIONAL DO PESQUISADOR PRINCIPAL

Aluno de Pós-Graduação CAMPUS: São Paulo DEPTO/LOCAL: Informática em Saúde - São Paulo CHEFE/RESP: Prof. Dr. Meide Silva Anção E-MAIL CHEFE/RESP: tymonex@hotmail.com

INFORMAÇÕES SOBRE O PROJETO DE PESQUISA

CARACTERÍSTICA: Retrospectivo/Retrospectivo ORIENTADOR: Prof. Dr. Felipe Mancini E-MAIL: fmancini@gmail.com
TÍTULO: RECONHECIMENTO AUTOMÁTICO DA DISLEXIA: UMA ABORDAGEM BASEADA NO COMPORTAMENTO OCULAR E CLASSIFICADORES DE PADRÕES
OBJETIVO ACADÊMICO: Mestrado LOCAL/CENTRO DE PESQUISA: Rua Botucatu nº 740 - 3º andar - Sala 307
04023-062 - São Paulo - SP
(11) 5578-4848 ramal 1267
Prof. Dr. Meide Silva Anção

INFORMAÇÕES ADICIONAIS (A PESQUISA TERÁ OU FARÁ USO)

HSP: Não OGM: Não RADIOISÓTOPO/RADIOATIVO: Não PATENTE: Não BIORREPOSITÓRIO: Não BIOBANCO: Não
BIOBANCO/INFO: FONTE DE RECURSOS: Institucional (UNIFESP) TOTAL DE GASTOS PREVISTOS (R\$): Até 1000,00

CIÊNCIA DE PROCEDIMENTO(S)

Os eventuais itens a seguir são providências que devem ser tomadas pelo PESQUISADOR RESPONSÁVEL.

ATENÇÃO: Este projeto de pesquisa só será recebido pelo Comitê de Ética, se TODOS OS ITENS a seguir, estiverem satisfeitos!

- ☐ FOLHA DE ROSTO (gerada na Plataforma Brasil) assinada pelo pesquisador responsável e pelo chefe do departamento ou pelo diretor do campus envolvido, digitalizada e anexada na Plataforma Brasil.
- ☐ CÓPIA DIGITALIZADA DESTA DOCUMENTO (com as devidas assinaturas) anexada na Plataforma Brasil.
- ☐ Projeto cadastrado na Plataforma Brasil, enviado ao CEP/UNIFESP e com status "Em Recepção e Validação Documental".
- ☐ Este documento deverá ser assinado pelo orientador, trata-se de projeto de MESTRADO.

ASSINATURAS

São Paulo 21.07.2017

Antonio Carlos da Silva Junior
CPF: 31375628801
Pesquisador Responsável

Prof. Dr. Felipe Mancini
Orientador

Prof. Dr. Meide Silva Anção
Chefe do Departamento de
Informática em Saúde EPM/Unifesp