Programa de Pós-graduação em Gestão e Informática em Saúde

Escola Paulista de Medicina – Universidade Federal de São Paulo UNIFESP

Curso de Mestrado Acadêmico – Arguição de Dissertação de Mestrado

Reconhecimento automático de padrões em dislexia: Uma abordagem baseada em funções visuais de leitura e aprendizado de máquina

Aluno: Antonio Carlos da Silva Junior (acsjunior@unifesp.br)

Orientação: Prof. Dr. Felipe Mancini

Coorientação: Prof. Dr. Paulo Schor e Dra. Emanuela C. R. Gonçalves

São Paulo, 16 de dezembro de 2019

Programa de Pós-graduação em Gestão e Informática em Saúde

Escola Paulista de Medicina – Universidade Federal de São Paulo UNIFESP

Curso de Mestrado Acadêmico – Arguição de Dissertação de Mestrado

Curso:	Mestrado acadêmico do Programa de Pós-graduação em Gestão e Informática em Saúde EPM UNIFESP
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, Prof. Dr. Paulo Schor e Dra. Emanuela C. R. Gonçalves
Matrícula:	127773 – período 01/10/2017 a 30/12/2019
CEP:	Comitê de Ética em Pesquisa da UNIFESP 0899/2017 de 20/09/2017
Linha:	Registro, recuperação e relacionamento de dados em saúde
Grupo de pesquisa:	*Bioengenharia Ocular http://dgp.cnpq.br/dgp/espelhogrupo/5894644663535064*
Apoio:	

São Paulo, 16 de dezembro de 2019

Sumário

Introdução
- Dislexia
- Aprendizado de máquina
- Aprendizado de máquina e movimento ocular
- Função visuais de leitura

Objetivos

Métodos
- Base de dados
- Método etapa 1
- Método etapa 2

Resultados
- Resultados etapa 1
- Resultados etapa 2

Conclusões

INTRODUÇÃO

Dislexia

- A dislexia é uma dificuldade neurológica específica nos processamentos da linguagem[1,2]
  - reconhecer, reproduzir, associar e ordenar sons e formas das letras organizando-os corretamente
- Não pode ser descrito por:[1,2]
  - desordem do desenvolvimento intelectual, acuidade visual, desordem neurológica, educação inadequada, proficiência na linguagem ou adversidade psicossocial.
- Cerca de 10%[3] a 20%[4,5] das pessoas do mundo possui algum grau de dislexia
- No Brasil essa taxa é de aproximadamente 12%[6]

1. Lyon GR, Shaywitz SE, Shaywitz BA. A definition of dyslexia. Ann of Dyslexia 2003;53(1):1–14.
2. ICD-11 Version:2019
3. Rello L, Ballesteros M. Detecting Readers with Dyslexia Using Machine Learning with Eye Tracking Measures.
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. Maio de 2016;18(3):216–28.
5. Moody KC. Education Update - Dyslexia in the Prison Population
6. Rotta NT, Ohlweiler L, Santos Riesgo R dos. Transtornos da aprendizagem: abordagem neurobiológica e multidisciplinar

Diagnóstico da dislexia

- O diagnóstico de dislexia é complexo
  - Condição neurológica[1,2,3]
- De natureza multiprofissional[2,3]
- Por exclusão[2,3]
- Demorado e caro

1. Lyon GR, Shaywitz SE, Shaywitz BA. A definition of dyslexia. Ann of Dyslexia 2003;53(1):1–14.
2. ICD-11 Version:2019
3. Como é feito o Diagnóstico? – ABD | Associação Brasileira de Dislexia

O olhar como alternativa

- Alguns pesquisadores estão estudando o movimento do olhar e função visuais de leitura (FVL) para contribuir com os estudos de dislexia[1,2,3,4]
- O uso de técnicas de aprendizado de máquina têm apresentado resultados promissores e contribuído para o estudo da dislexia [1,2,3]

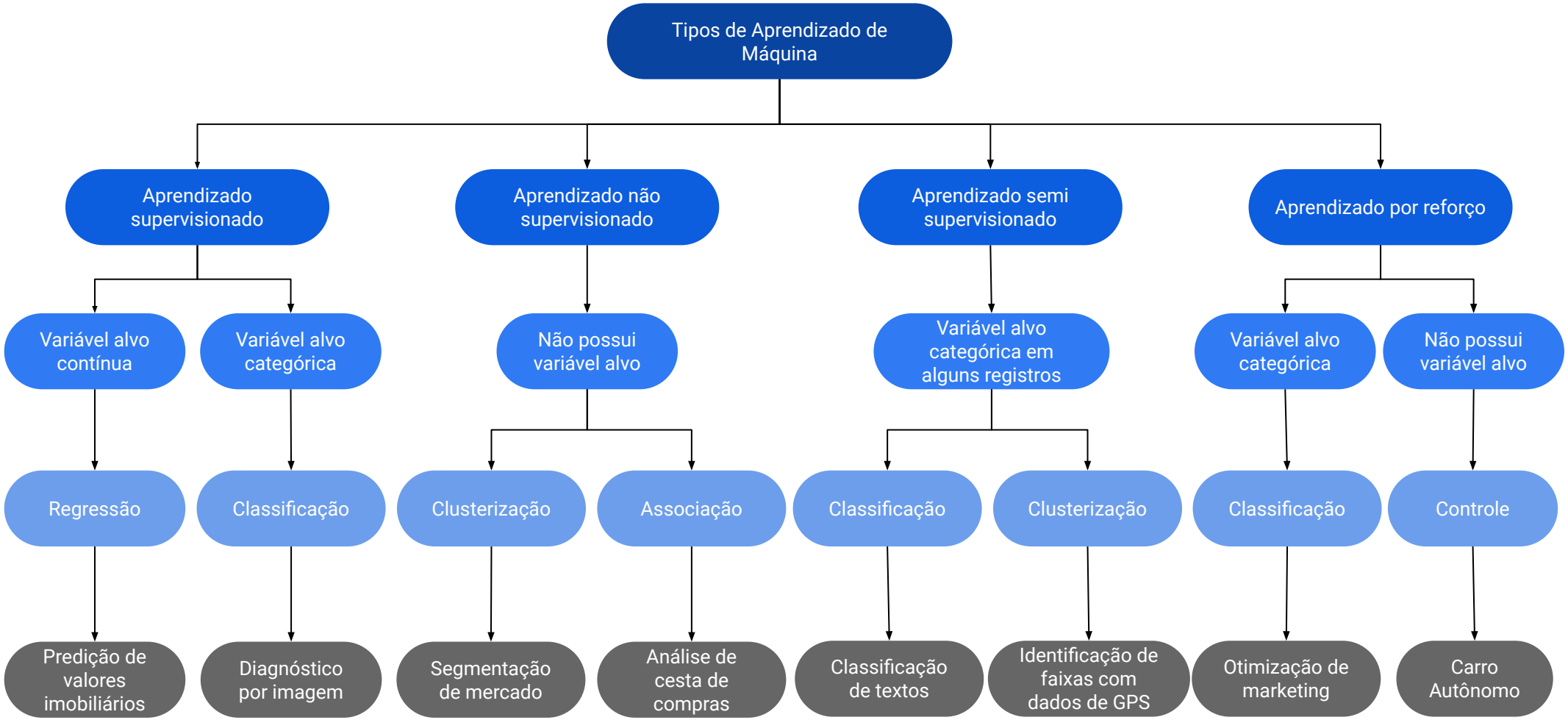
1. Rello L, Ballesteros M. Detecting Readers with Dyslexia Using Machine Learning with Eye Tracking Measures.
2. Lustig J. Identifying dyslectic gaze pattern : Comparison of methods for identifying dyslexic readers based on eye movement patterns
3. Benfatto MN, Seimyr GÖ, Ygge J, Pansell T, Rydberg A, Jacobson C. Screening for Dyslexia Using Eye Tracking during Reading
4. O'Brien BA, Mansfield JS, Legge GE. The effect of print size on reading speed in dyslexia.

Aprendizado de máquina

- É uma subárea da inteligência artificial[1,2,3]
- Têm capacidade de aprender com sua experiência e tomar decisões mais precisas[2]
  - Experiência: São os dados coletados
  - Aprendizado: Capacidade de algoritmos aprender com estes dados e fazer melhores previsões

1. Rezende SO, organizador. Sistemas inteligentes: fundamentos e aplicações. 1. ed. Barueri, SP: Ed. Manole; 2003.
2. Mohri M, Rostamizadeh A, Talwalkar A. Foundations of machine learning. Second edition. Cambridge, Massachusetts: The MIT Press; 2018.
3. Bishop C. Pattern Recognition and Machine Learning. New York: Springer; 2007.

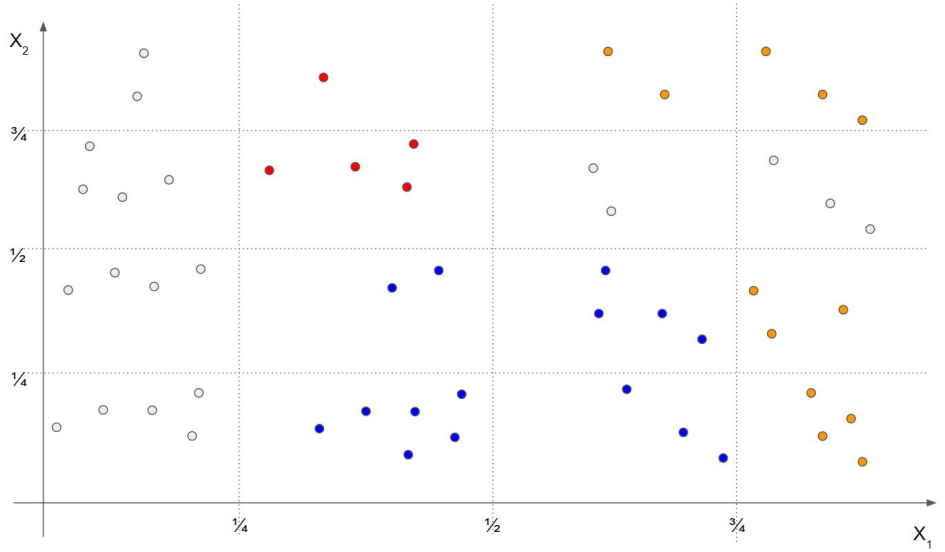
Aprendizado de máquina

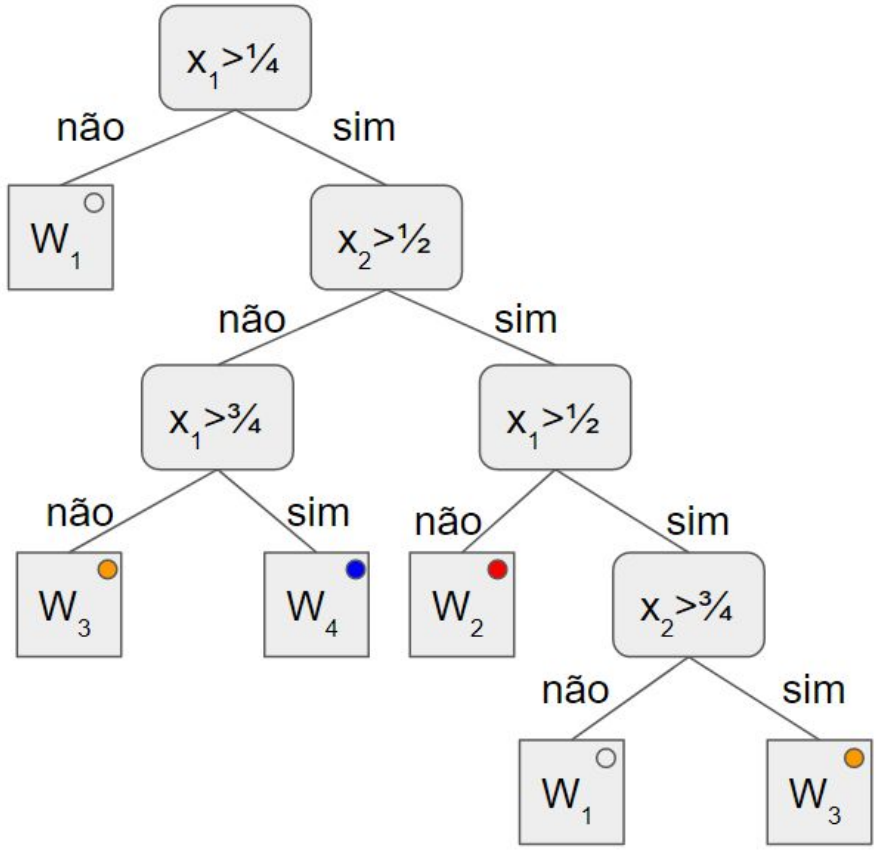


1. Ramasubramanian K, Singh AN. Machine learning using R. California: Apress; 2017

Aprendizado de máquina - Árvore de decisão

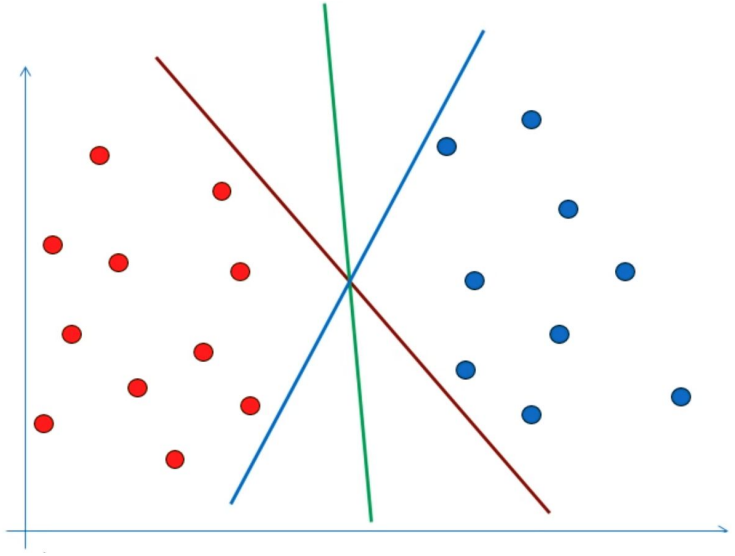
- Algoritmo supervisionado
- São grafos acíclicos
  - Nós de decisão (galhos)
  - Resultados (Folhas)
- Na construção da árvore
  - calcula-se entropia e ganho de informação

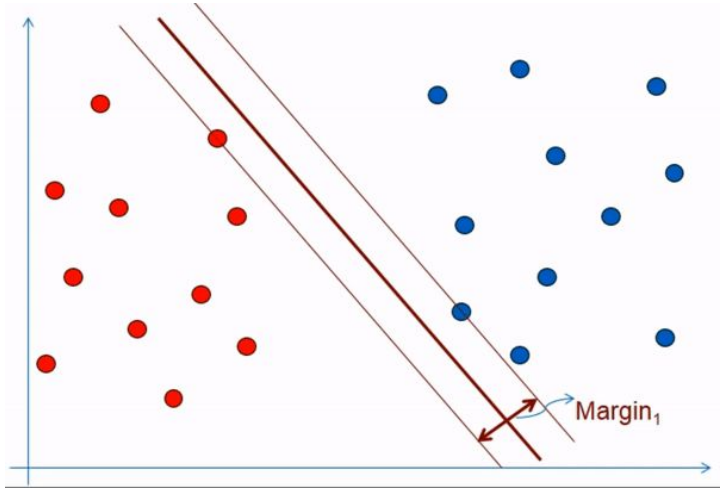




Aprendizado de máquina - *Support vector machine* (SVM)

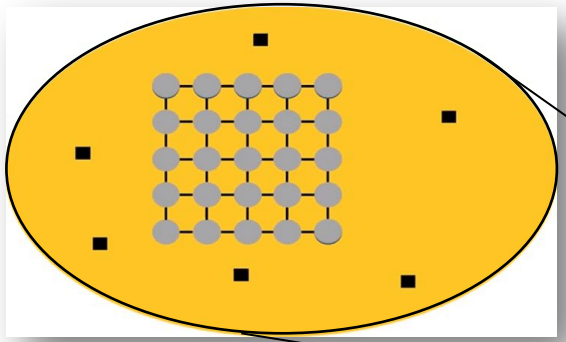
- Algoritmo supervisionado
- Dividir duas classes com uma linha
- Existem infinitas formas de dividir dois grupos com linhas
- Calcula-se as margens entre os resultados mais próximos e a linha de divisão
- Estas linhas paralelas são chamadas vetores de suporte

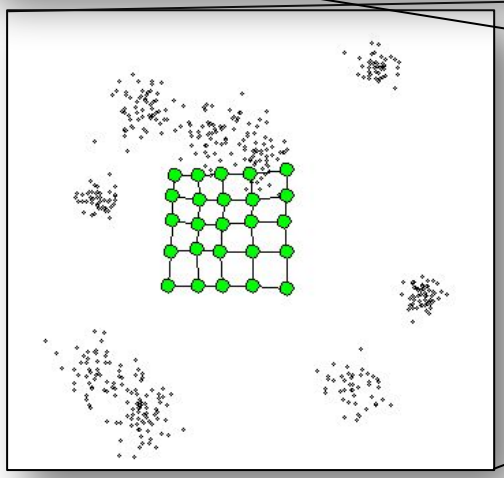


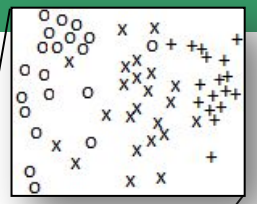


Aprendizado de máquina - Mapas auto-organizáveis (SOM)

- Algoritmo não supervisionado
- Aprendizado competitivo
  - Distância Euclidiana

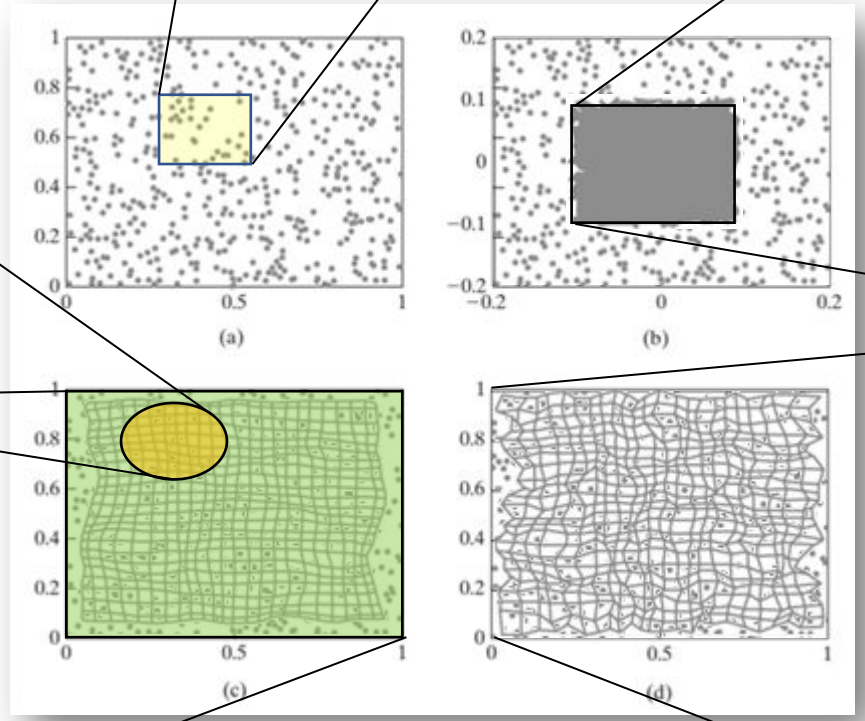






Variáveis das amostras

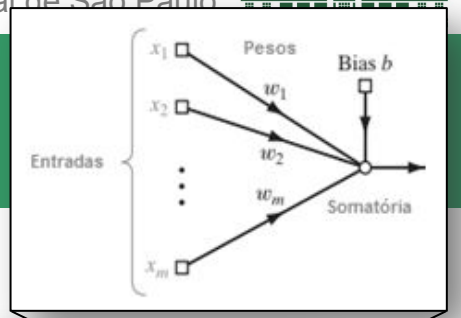
Pesos com valores iniciais

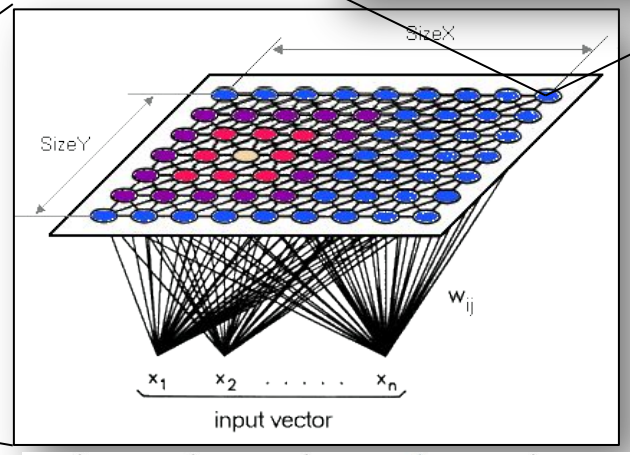


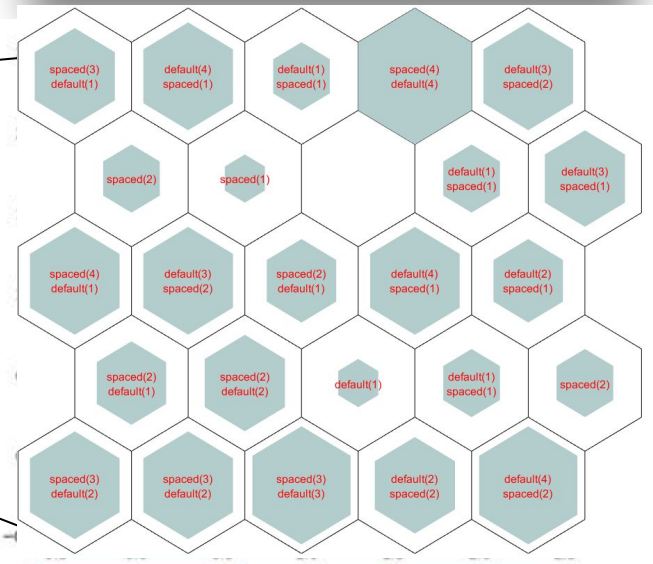
Fonte: adaptado de Haykin S. (3)

Processo de aprendizagem

Variáveis organizados em Clusters

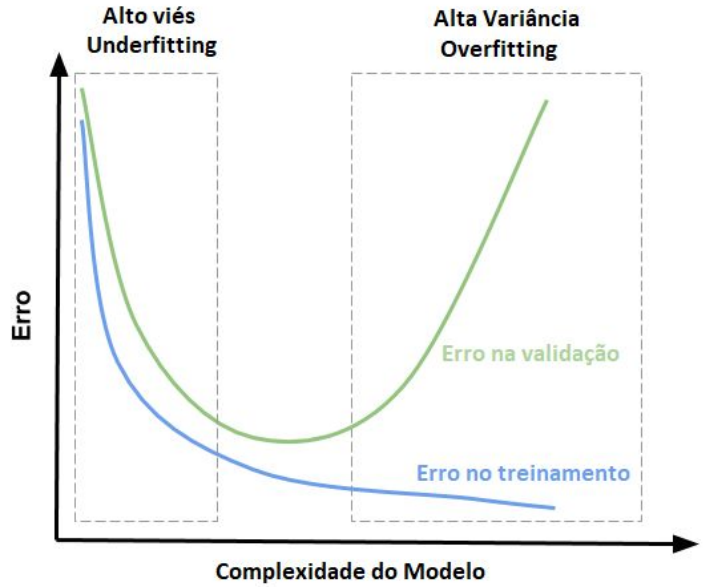


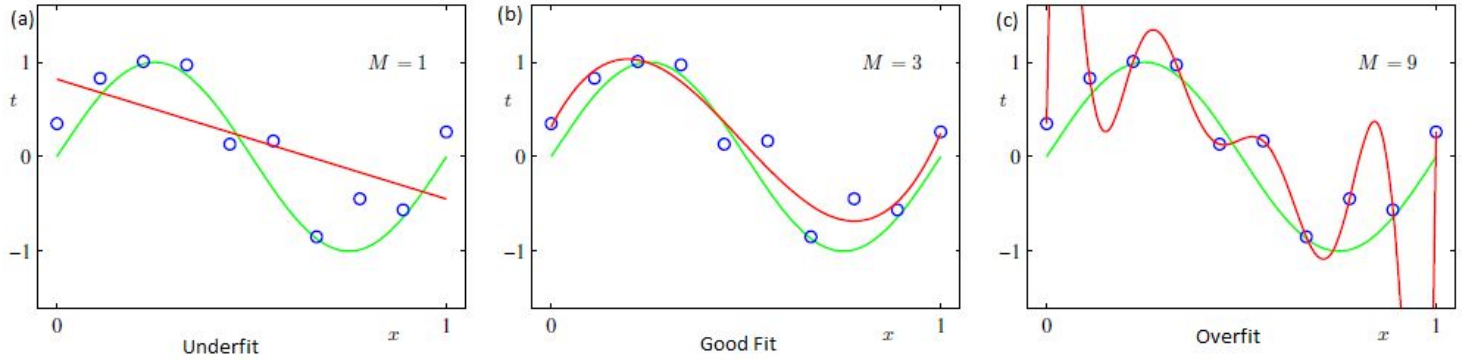




Aprendizado de máquina - *Underfitting e Overfitting*

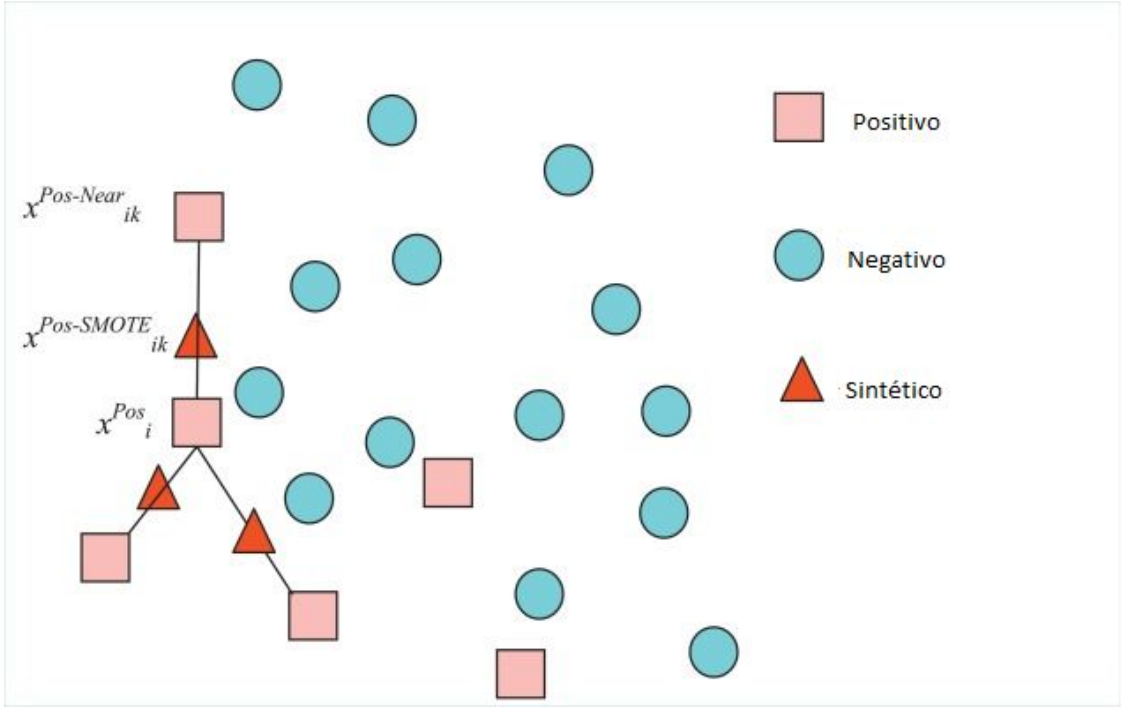
- Underfitting
  - Obtém erro alto no treinamento e erro alto na validação
  - O algoritmo não está aprendendo o padrão
  - Aumentar a quantidade de variáveis, utilizar um algoritmo mais complexo, treinar por mais tempo, conseguir mais dados
- Overfitting
  - Obtém baixo erro no treinamento e alto erro na validação
  - O algoritmo está se adaptando demais ao treino
  - Conseguir mais dados, diminuir a quantidade de variáveis, utilizar um modelo mais simples

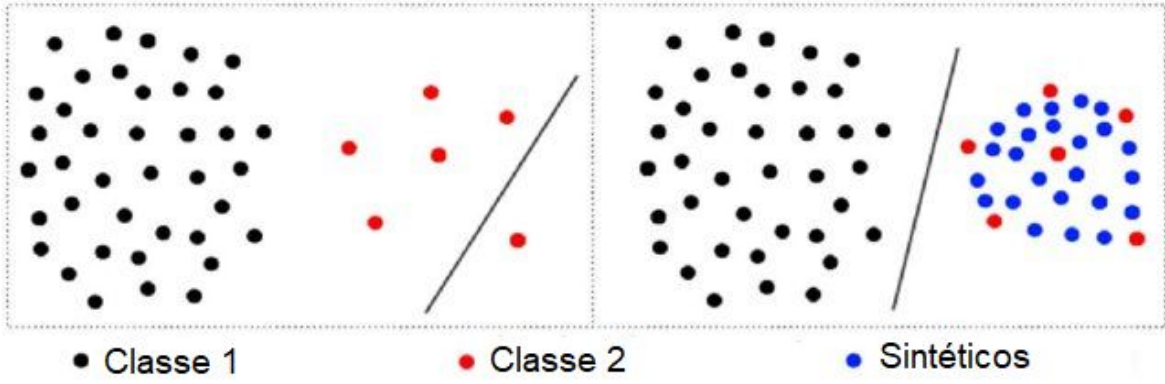




Geração sintética de dados (SMOTE)

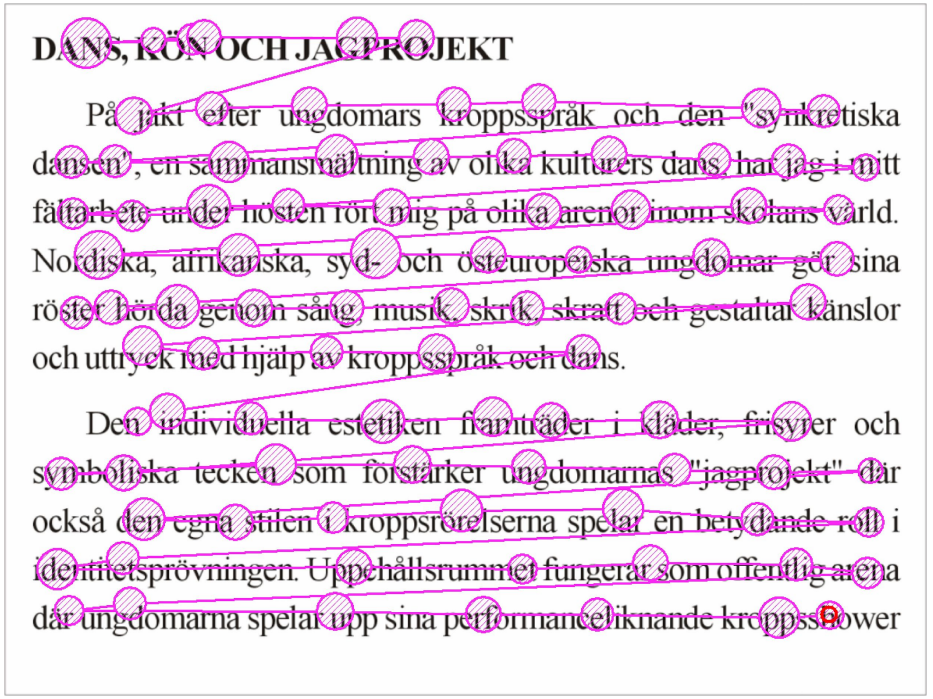
- Algumas situações é inviável conseguir mais dados
- SMOTE
  - Dados sintéticos são gerados aleatoriamente entre x vizinhos.
- Diminui a chance de underfitting e overfitting em problemas de classificação





Aprendizado de máquina e o movimento ocular

- Rello e Ballesteros, 2015 utilizaram dados do movimento do olhar e reconhecimento automático de padrões[1]
  - 49 não-disléxicos e 48 disléxicos
  - Linguagem: espanhol
    - Livro “Impostors” de lucas sanches
  - Tipos de fontes diferentes
  - Máquina de Vetores de Suporte (SVM)
    - 80,18% acurácia


DANS, KÖN OCH JAGPROJEKT

På jakt efter ungdomars kroppsspråk och den "synkretiska dansen", en sammansmältning av olika kulturers dans, har jag i mitt fältarbete under hösten rört mig på olika arenor inom skolans värld. Nordiska, afrikanska, syd- och östeuropeiska ungdomar gör sina röster hörda genom sång, musik, skrik, skratt och gestaltar känslor och uttryck med hjälp av kroppsspråk och dans.

Den individuella estetiken framträder i kläder, frisyrer och symboliska tecken som förstärker ungdomarnas "jagprojekt" där också den egna stilen i kroppsrörelserna spelar en betydande roll i identitetsprövningen. Uppehållsrummet fungerar som offentlig arena där ungdomarna spelar upp sina performanceliknande kroppsuttryck

1. Rello L, Ballesteros M. Detecting Readers with Dyslexia Using Machine Learning with Eye Tracking Measures.

Aprendizado de máquina e o movimento ocular

- Lustig, 2016 [1]
  - Linguagem: Sueco
  - 9 disléxicos e 9 não-disléxicos
    - Leituras baseadas no teste de leitura PISA em sueco
  - Três algoritmos aprendizado supervisionado
    - SVM
      - 83% de acurácia
    - *Feed Forward Neural Network* (FFNN)
      - 83% de acurácia
    - *Recurrent Neural Network* (RNN)
      - 78% de acurácia

1. Lustig J. Identifying dyslectic gaze pattern : Comparison of methods for identifying dyslexic readers based on eye movement patterns

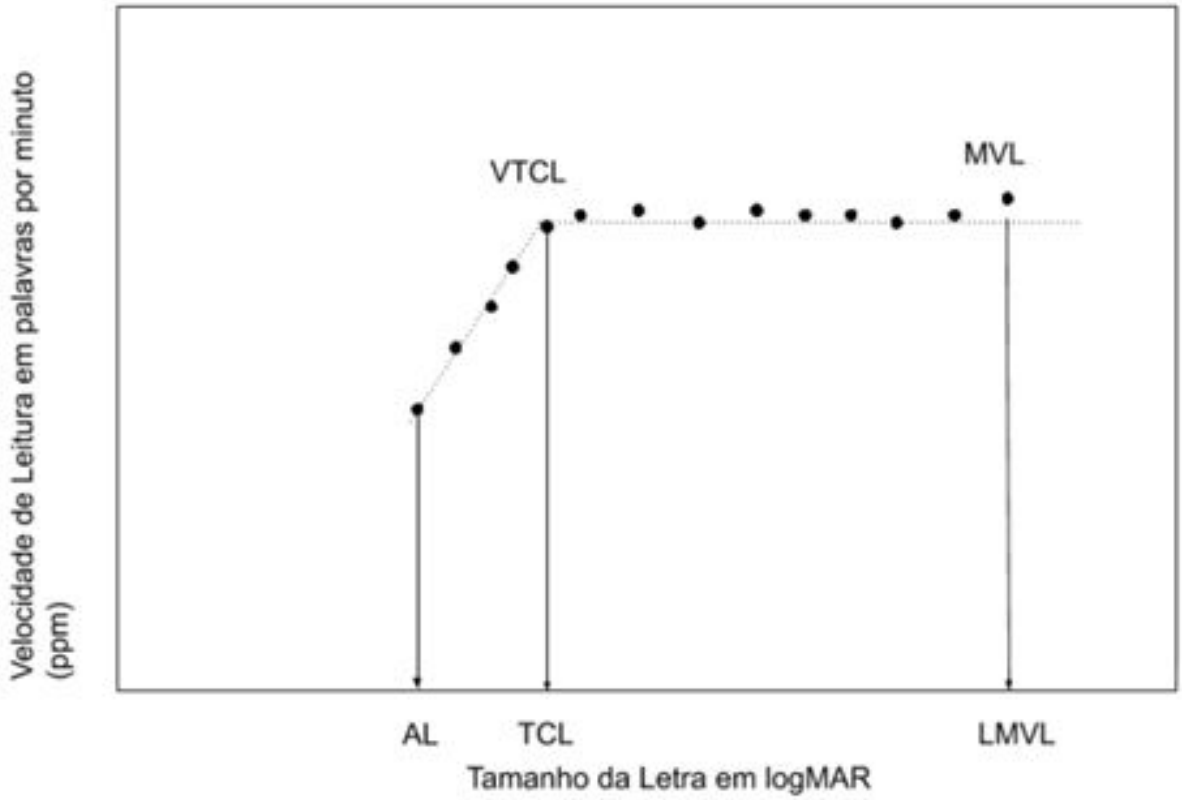
Aprendizado de máquina e o movimento ocular

- Benfatto et al, 2016 [1]
  - Uma base de um estudo de coorte
    - Foram selecionados 88 com baixo risco de dislexia e 97 com alto risco
    - Um texto em 8 linhas com 10 sentenças de 4,6 palavras em média
    - Linguagem: Sueco
  - SVM
    - 95,6% de acurácia
    - 95,5% sensibilidade
    - 95,7% especificidade

1. Benfatto MN, Seimyr GÖ, Ygge J, Pansell T, Rydberg A, Jacobson C. Screening for Dyslexia Using Eye Tracking during Reading

Funções visuais de leitura

- Um estudo de O'Brien BA, Mansfield JS, Legge avaliando funções visuais de leitura (FVL)[1]
  - Disléxicos e não disléxicos possuem perfis de curvas qualitativamente similares para velocidade de leitura (VL) e tamanho da fonte (TF)
  - VL foi menor para os disléxicos
  - TF foi maior para os disléxicos



1. O'Brien BA, Mansfield JS, Legge GE. The effect of print size on reading speed in dyslexia.

Não explorado

- O uso de técnicas de AM associado com FVL
- Uso de um algoritmo não supervisionado
- Linguagem: Português Brasileiro

UNIFESP

UNIVERSIDADE FEDERAL DE SÃO PAULO

1933

ESCOLA PAULISTA MEDICINA

1933

OBJETIVOS

Objetivos

- 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).
- 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.

UNIFESP
UNIVERSIDADE FEDERAL DE SÃO PAULO
1933

ESCOLA PAULISTA MEDICINA
1933

MÉTODO

Base de dados

- Dados utilizados provenientes do Grupo de Pesquisa Bioengenharia Ocular
- Voluntários que se apresentaram em 2 anos de coleta e passaram pelo critério de inclusão
  - Diferença interocular menor que 3 linhas
  - Acuidade visual igual ou melhor que 0,3 logMAR
  - Sem antecedentes oftalmológicos (ex: estrabismo ou cirurgia prévia)
- 24 sujeitos com dislexia (3 sujeitos não concluíram)
- 18 sujeitos sem dislexia
- Idade entre 8-45

Coleta

- Teste de leitura de Minnesota adaptado para o português (MNREAD-P)
  - Em dois formatos
    - Em um linha (1L)

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

    - Em três linhas (3L)

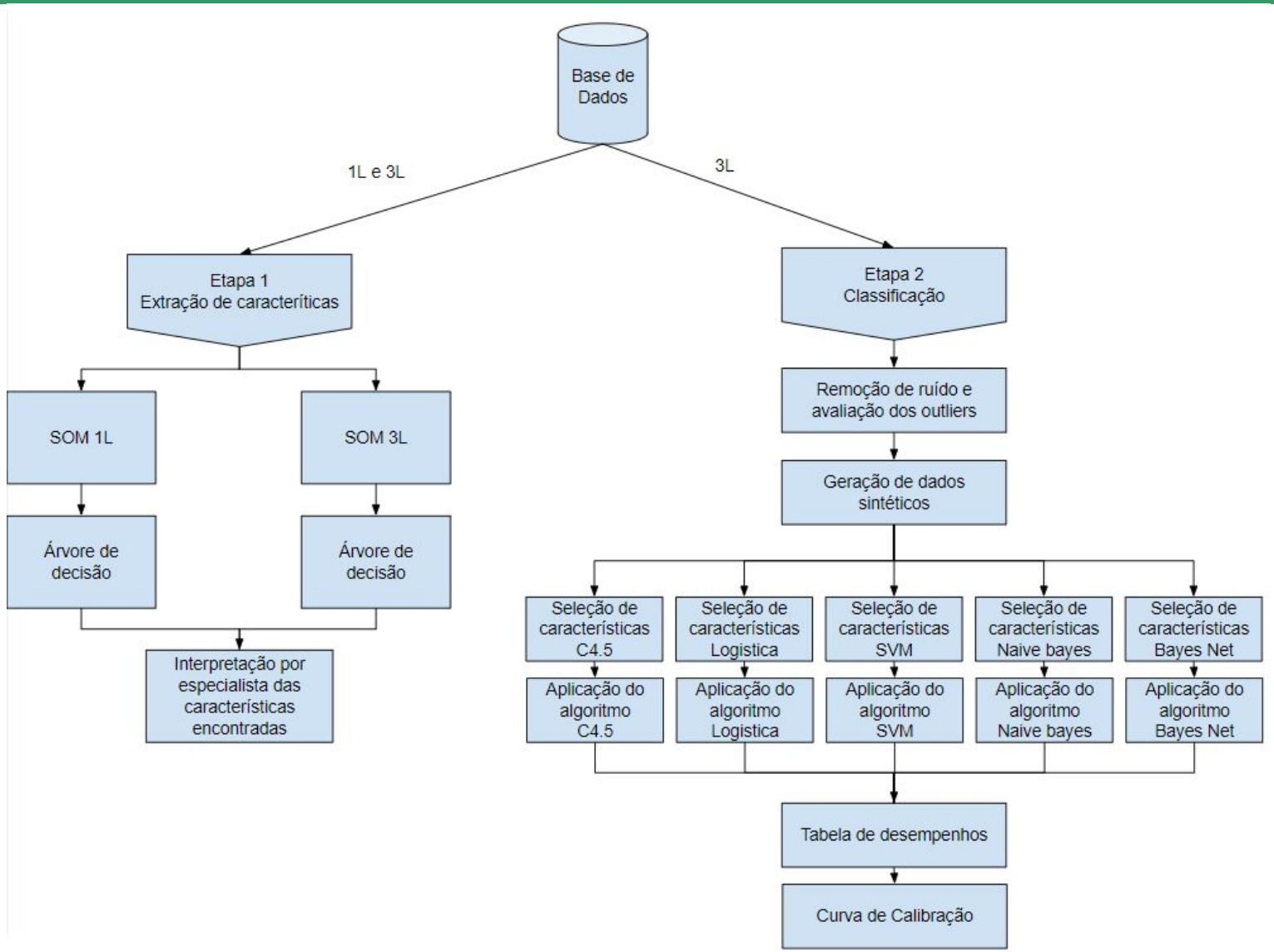
A vovó fez um bolo
de chocolate gelado
e eu levei de lanche

    - Feito em 13 tentativas com textos aleatórios
    - 2 de familiarização e 11 experimentais
    - A fonte decrescendo de 1,0 a 0,0 logMAR
      - Diminuindo 0,1 a cada texto
    - O teste foi feito com a fonte Times New Roman

Variáveis

Variável	Descrição
GRUPO	Controle ou Disléxico
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

Desenho metodológico



Método - Etapa 1

- Considerou os dados coletados de leituras com o estímulo de 1 linha (1L) e 3 linhas (3L)
- Algoritmo não supervisionado de mapas auto-organizáveis (SOM)
  - Para tarefa de clusterização
- Algoritmo supervisionado random tree
  - Para extrair as regras que regem os clusters
- Avaliação das árvores por especialista

Materiais - Etapa 1

- Para execução do SOM foi utilizado o software MATLAB com a biblioteca SOM toolbox
- Weka foi utilizado para execução do algoritmo random tree

Método - Etapa 2

- Utilizou somente os dados de 3L
- Remoção de ruído
  - Remoção de variáveis altamente correlacionadas e sem variação de valor
  - Seleção de outliers por avaliação de especialista
- Geração de dados sintéticos por Synthetic Minority Over-sampling Technique (SMOTE)
- Técnicas de seleção de atributos wrapperSubsetEval com busca exaustiva
- Aplicar Cinco algoritmos indicados para pequenas bases
  - Árvore de decisão - C4.5
  - Regressão – regressão logística
  - Kernel - SVM
  - Probabilístico - Naive Bayes
  - Probabilístico - BayesNet
- Comparar as técnicas pela área sob a curva ROC (AUC) e curva de calibração

Materiais Etapa 2

- Para encontrar os outliers foi utilizado o software JASP
- Para aplicação dos algoritmos foi utilizado o WEKA
- A curva de calibração foi gerado com o WEKA e aglomerados no Excel

UNIFESP

UNIVERSIDADE FEDERAL DE SÃO PAULO

1933

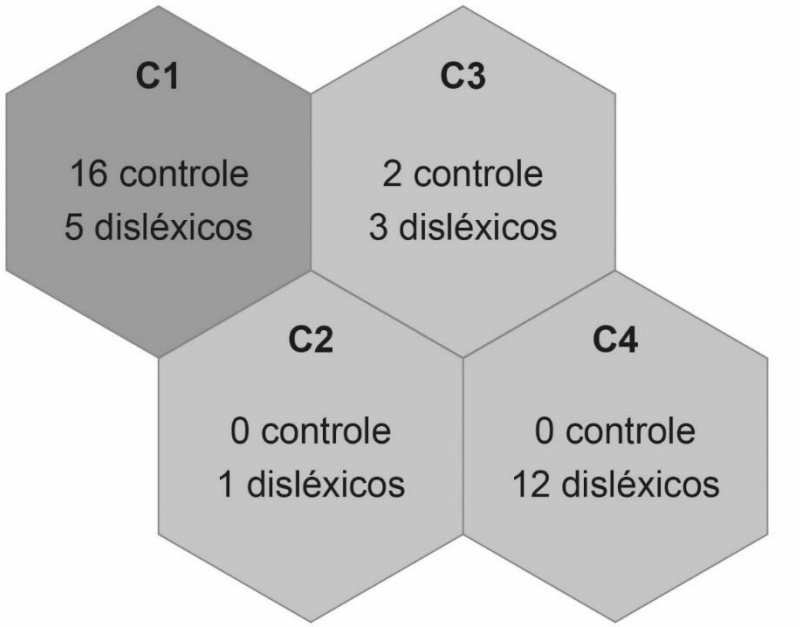
ESCOLA PAULISTA MEDICINA

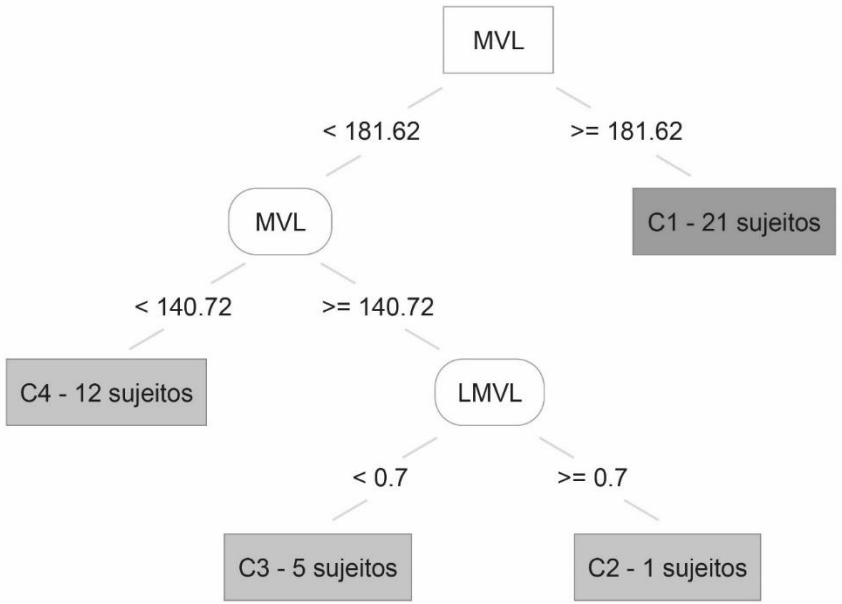
1933

RESULTADOS

Resultados etapa 1

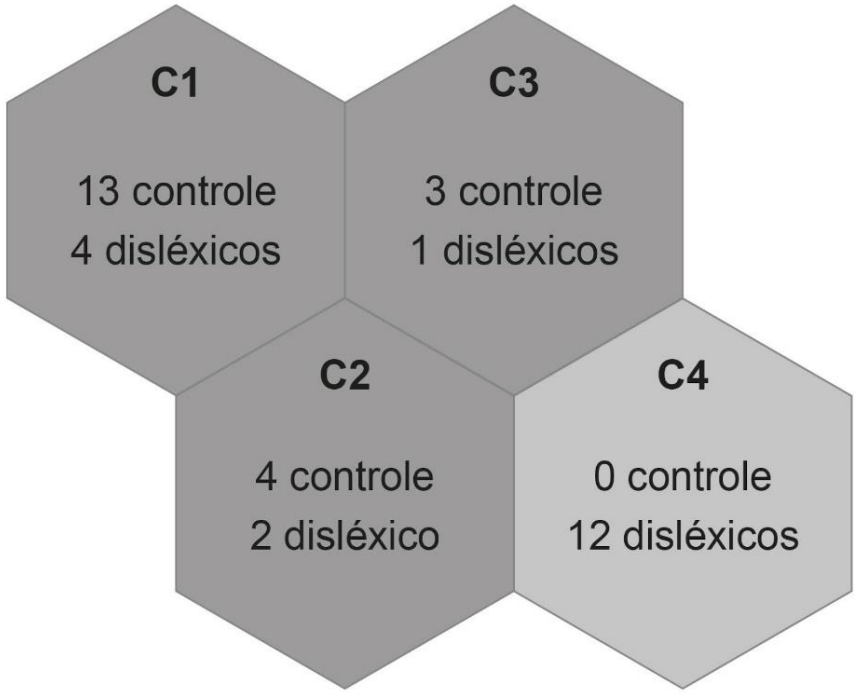
- Para leituras em 1L
- Os grupos disléxicos tiveram MVL inferior a 181,62 ppm e somente disléxicos obtiveram MVL inferior a 140,72 ppm

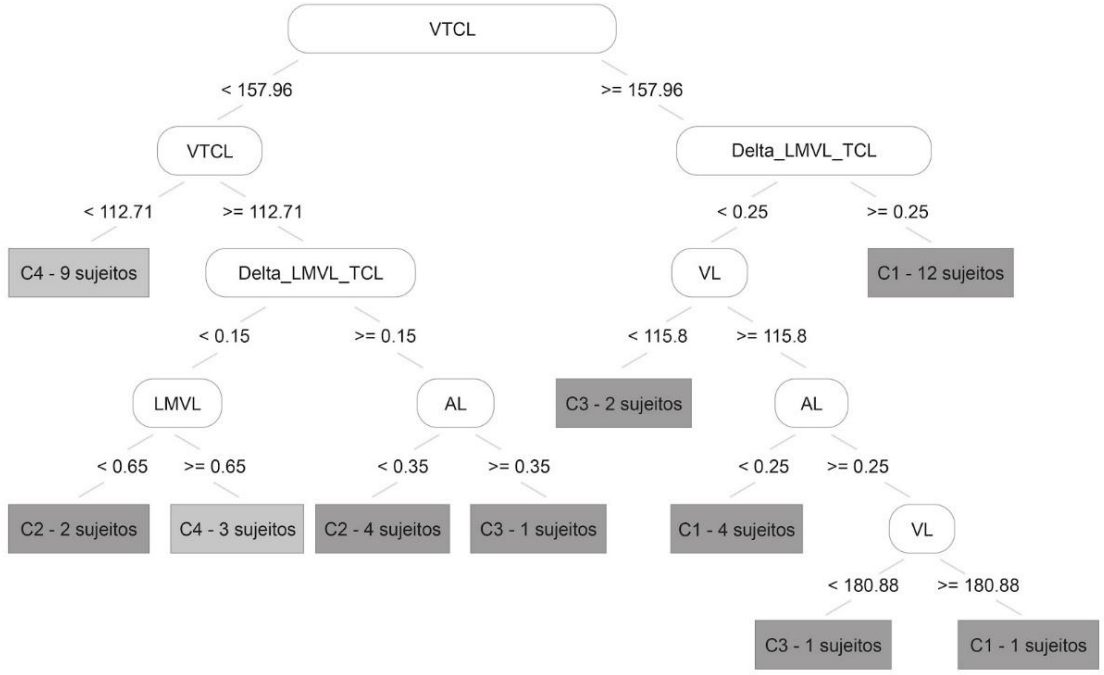




Resultados etapa 1

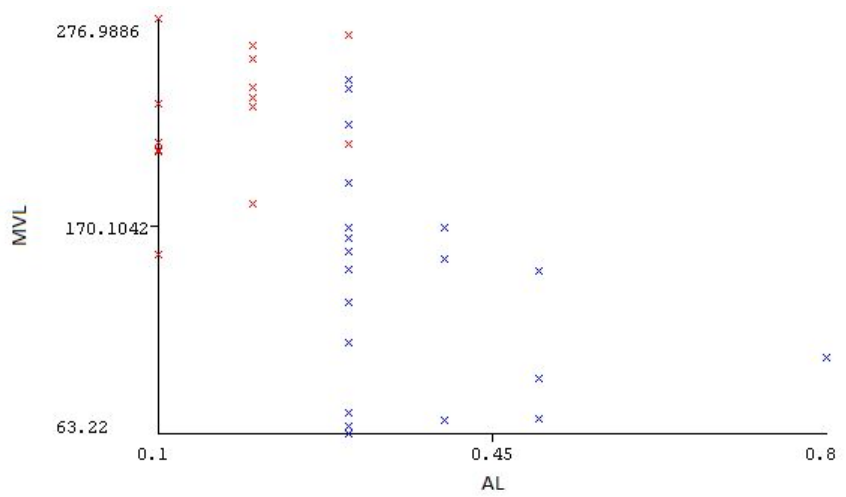
- Para leituras em 3L
- O grupo controle obteve VTCL igual ou superior a 157,96 ppm e somente disléxicos obtiveram VTCL inferior a 112,71 ppm

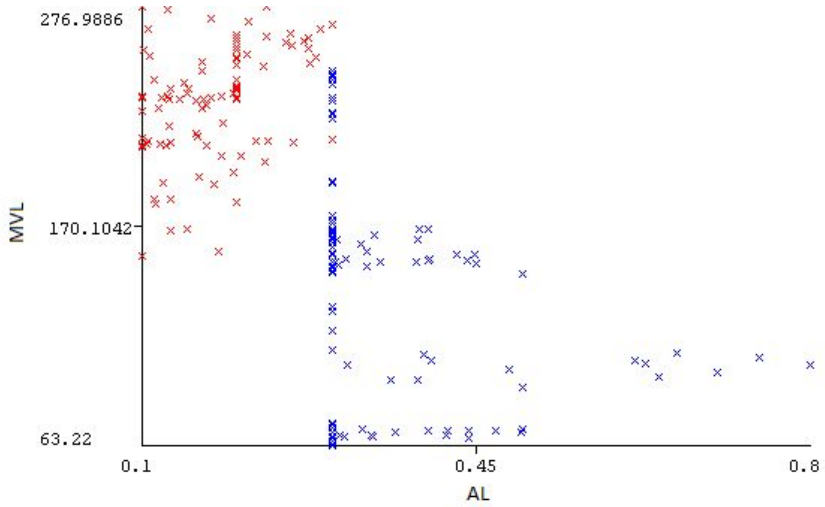




Resultados etapa 2

- Avaliação de outliers
  - Resultou em uma base com 20 disléxicos e 14 não-disléxicos
- Geração sintética de dados
  - Foram gerados dados sintéticos até ambos os grupos possuírem 100 registros





Resultados etapa 2

- Seleção de tributos

Algoritmo	**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

Resultados etapa 2

- Resultados dos algoritmos

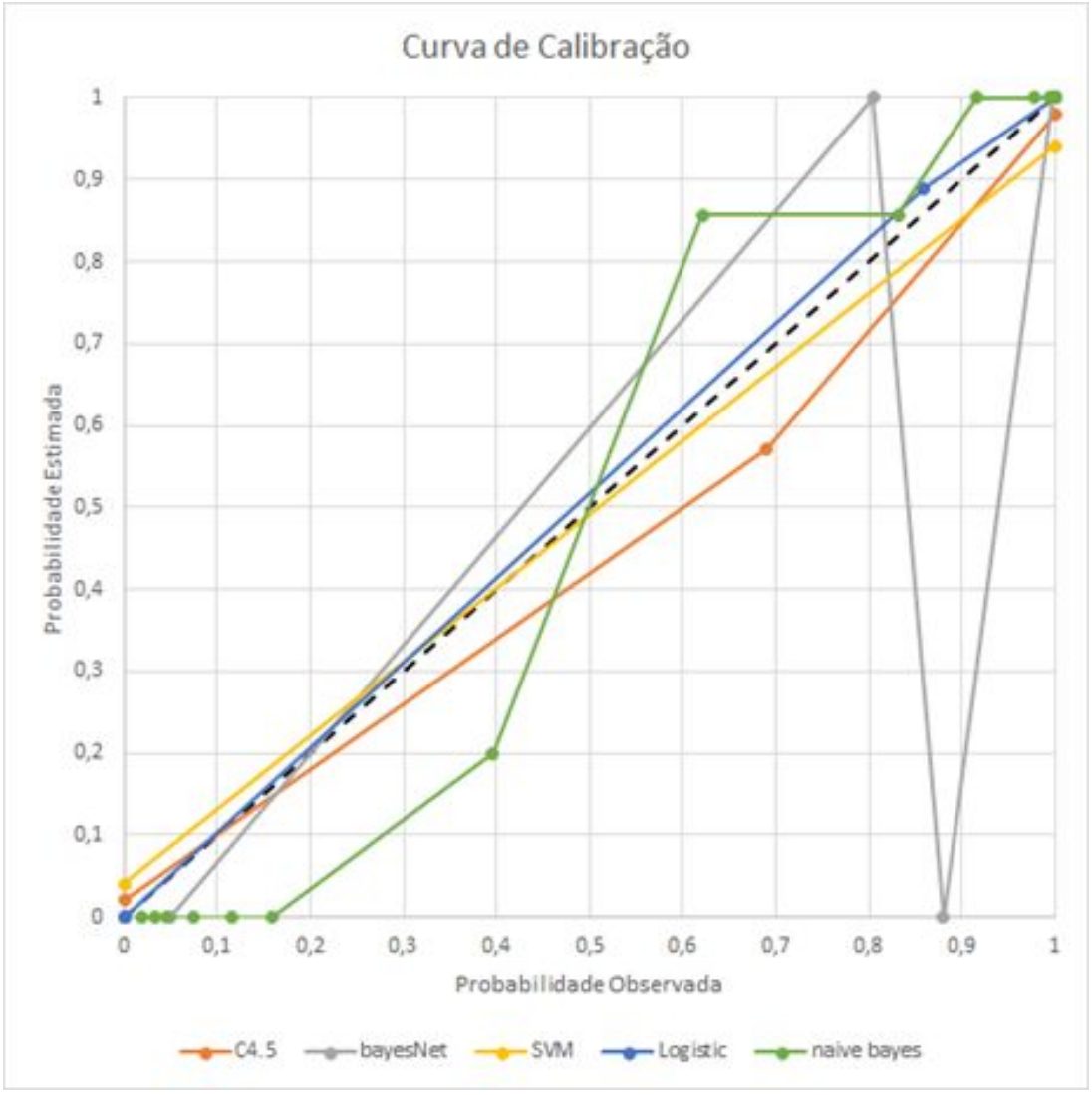
Algoritmo	Sensibilidade	Especificidade	Acurácia	AUC
C4.5	0,960	0,950	95,5%	0,978
Regressão Logística	0,990	0,990	99%	0,999
SVM	0,960	0,940	95%	0,950
Rede Bayesiana	1,000	0,980	99%	0,997
Naive Bayes	1,000	0,960	98%	0,996

Resultados etapa 2

- Impacto do SMOTE no algoritmo com melhor performance (regressão logística)

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

Resultados etapa 2



UNIFESP
UNIVERSIDADE FEDERAL DE SÃO PAULO
1933

ESCOLA PAULISTA MEDICINA
1933

CONCLUSÕES

Conclusões

- Para textos de 1L somente disléxicos obtiveram MVL<140,72 ppm
- Para textos de 3L somente disléxicos obtiveram VTCL<112,71 ppm
- A diferença entre as árvores de 1L e 3L pode ter sido causadas por efeito de *crownding*
- A acuidade de leitura foi selecionado para 4 dos 5 algoritmos como importante para melhorar a AUC
  - A regra de inclusão de acuidade visual de 0,3 logMAR tornou esta variável notável
  - Mais estudos sobre a diferença de AV e AL são necessários para entender melhor seus efeitos
- A geração sintética de dados em conjunto com a seleção de outliers foi capaz de fazer com que todos os algoritmos alcançarem uma acurácia superior a 95%
- A regressão logística obteve os melhores resultados (acurácia de 99% e AUC de 0,999)
- A regressão logística obteve a melhor calibração dentre todos os algoritmos avaliados

Publicações

- **Trabalho publicado em anais de evento (resumo)**
- Junior, A. C. S.; Gonçalves, E. C. R.; Meister, I. P.;Navarro, Martina; Mancini, Felipe; Eye Tracking and Artificial Neural Networks Application to Detection of Visual Patterns of Reading Performance in Dyslexic Persons In: XV Simpósio Brasileiro sobre Fatores Humanos em Sistemas Computacionais (IHC), 2016, São Paulo. **Qualis Conferência*: B2**
- **Pôster aprovado em evento**
- Antonio C. S. Junior, Emanuela C. R. Gonçalves, Izabel P. Meister, Clara R. B. Ávila, Martina Navarro, Felipe Mancini; Feature Extraction by Machine Learning from Dyslexic's Readings: AMIA Annual Symposium, 2018, San Francisco; **Qualis Conferência*: B2**
- **Artigos submetidos**
- Antonio C. S. Junior, Emanuela C. R. Gonçalves, Clara R. B. Ávila, Paulo Schor, Martina Navarro, Felipe Mancini, Exploring Dyslexia: An Approach to Exploring Visual Reading Functions by Feature Extraction; Health Information Science and Systems. **Fator de impacto: 4,5**
- Antonio C. S. Junior, Emanuela C. R. Gonçalves, Paulo Schor, Martina Navarro, Felipe Mancini, Geração de dados sintéticos para classificação de disléxicos por meio de aprendizado de máquina. Journal of Health Informatics; **Qualis: B3**

Programa de Pós-graduação em Gestão e Informática em Saúde

Escola Paulista de Medicina – Universidade Federal de São Paulo UNIFESP

Curso de Mestrado Acadêmico – Arguição de Dissertação de Mestrado

Reconhecimento automático de padrões em dislexia: Uma abordagem baseada em funções visuais de leitura e aprendizado de máquina

Aluno: Antonio Carlos da Silva Junior (acsjunior@unifesp.br)

Orientação: Prof. Dr. Felipe Mancini

Coorientação: Prof. Dr. Paulo Schor e Dra. Emanuela C. R. Gonçalves

Muito obrigado!

São Paulo, 16 de dezembro de 2019