Universidade Estadual de Londrina

ARTHUR ALEXANDRE ARTONI

APLICAÇÃO DE APRENDIZADO DE MÁQUINA NO AUXÍLIO AO DIAGNÓSTICO DO TRANSTORNO DO ESPECTRO AUTISTA

Londrina
2020

ARTHUR ALEXANDRE ARTONI

APLICAÇÃO DE APRENDIZADO DE MÁQUINA NO AUXÍLIO AO DIAGNÓSTICO DO TRANSTORNO DO ESPECTRO AUTISTA

Dissertação apresentada ao Programa de Mestrado em Ciência da Computação da Universidade Estadual de Londrina para obtenção do título de Mestre em Ciência da Computação.

Orientador: Profa. Dra. Cinthyan Renata Sachs Camerlengo de Barbosa

Londrina
2020

Ficha de identificação da obra elaborada pelo autor, através do Programa de Geração Automática do Sistema de Bibliotecas da UEL

AR792 Artoni, Arthur Alexandre .
APLICAÇÃO DE APRENDIZADO DE MÁQUINA NO AUXÍLIO AO DIAGNÓSTICO DO TRANSTORNO DO ESPECTRO AUTISTA / Arthur Alexandre Artoni. - Londrina, 2020.
69 f. : il.

Orientador: Cinthyan Renata Sachs Camerlengo de Barbosa.
Dissertação (Mestrado em Ciência da Computação) - Universidade Estadual de Londrina, Centro de Ciências Exatas, Programa de Pós-Graduação em Ciência da Computação, 2020.
Inclui bibliografia.

1. Transtorno do Espectro Autista - Tese. 2. Aprendizado de Máquina - Tese. 3. Diagnóstico com Inteligência Artificial - Tese. I. Barbosa, Cinthyan Renata Sachs Camerlengo de . II. Universidade Estadual de Londrina. Centro de Ciências Exatas. Programa de Pós-Graduação em Ciência da Computação. III. Título.

CDU 519

ARTHUR ALEXANDRE ARTONI

APLICAÇÃO DE APRENDIZADO DE MÁQUINA NO AUXÍLIO AO DIAGNÓSTICO DO TRANSTORNO DO ESPECTRO AUTISTA

Dissertação apresentada ao Programa de Mestrado em Ciência da Computação da Universidade Estadual de Londrina para obtenção do título de Mestre em Ciência da Computação.

BANCA EXAMINADORA

_____

Orientadora: Profa. Dra. Cinthyan Renata Sachs Camerlengo de Barbosa
Universidade Estadual de Londrina – UEL

_____

Prof. Dr. Marcelo Morandini
Universidade de São Paulo – USP

_____

Prof. Dr. Evandro Bacarin
Universidade Estadual de Londrina – UEL

_____

Prof. Dr. Vitor Valério de S. Campos
Universidade Estadual de Londrina – UEL

Londrina, 27 de fevereiro de 2020.

Este trabalho é dedicado às crianças adultas
que, quando pequenas, sonharam em se
tornar cientistas.

AGRADECIMENTOS

Aos meus pais e a toda minha família pelo incentivo e apoio.

A todos os meus amigos de fora e dentro da UEL.

A minha orientadora Cinthyan Renata Sachs Camerlengo de Barbosa, por todas as oportunidades, pelo conhecimento, pela experiência e conselhos que vão além dos estudos.

Ao professor Dr. Sylvio Barbon, que sempre esteve disponível para me ajudar com boas ideias e sugestões.

Aos professores membros da banca examinadora pelo tempo, correções, sugestões e contribuições para este trabalho.

A CAPES por me ajudar a financiar meus estudos durante o programa.

A UEL, local onde estudo desde o início da graduação.

Ao Dr. Fadi Thabtah da Universidade Huddersfield que disponibilizou as Bases de Dados utilizadas neste trabalho.

Meu muito obrigado!

"O dia de hoje jamais acontecerá novamente. Mas uma boa ação pode fazê-lo durar para sempre.
(Talmud)

ARTONI, A. A.. **Aplicação de Aprendizado de Máquina no Auxílio ao Diagnóstico do Transtorno do Espectro Autista**. 2020. 69f. Dissertação (Mestrado em Ciência da Computação) – Universidade Estadual de Londrina, Londrina, 2020.

RESUMO

Mesmo com os constantes avanços da medicina, o diagnóstico de transtornos mentais é um desafio para profissionais da área. Dentre esses está o Transtorno do Espectro Autista (TEA), que é uma patologia muito comum que afeta a parte comportamental, social e de comunicação do indivíduo. Porém identificá-lo é complexo, uma vez que não existem exames de imagens ou de sangue capazes de apontar o TEA e seu diagnóstico é feito apenas por análise comportamental. Diversas técnicas podem ser utilizadas, como o uso de escalas diagnósticas que contém questionários específicos formulados por especialistas, que servem como guia no processo de diagnóstico. Neste trabalho o Aprendizado de Máquina (AM) foi empregado em três bases de dados contendo resultados dos testes AQ-10 para adultos, adolescentes e crianças; além de outras características que poderiam influenciar no diagnóstico do TEA. Experimentos foram realizados nas bases de dados a fim de elencar quais atributos seriam realmente relevantes para o diagnóstico do TEA utilizando AM. Para a seleção dos atributos, a *Random Forest* foi utilizada como ferramenta para fazer um ranqueamento de 23 atributos presentes na base. Em duas das três bases de dados foi possível reduzir o número de atributos para apenas 5, mantendo uma Acurácia acima de 0.9. Na outra Base de Dados para manter o mesmo nível de Acurácia, o menor número de atributos utilizado foi 7. A *Suport Vector Machine* se destacou dos demais algoritmos usados neste trabalho, obtendo resultados superiores em todos os cenários.

Palavras-chave: Transtorno do Espectro Autista. Aprendizado de Máquina. Diagnóstico.

ARTONI, A. A.. **Machine Learning Application to Support Autism Spectrum Disorder Diagnosis**. 2020. 69p. Master's Thesis (Master in Science in Computer Science) – State University of Londrina, Londrina, 2020.

ABSTRACT

Despite the medicine constant advancement, mental disorder diagnosis is a challenge for medical professionals. Autism Spectrum Disorder (ASD) is one of the most common pathologies affecting the patient's behavioral, social and communication skills and its identification is complex because there are no images or blood tests available for ASD diagnosis, only through behavioral analysis. Several techniques, such as the use of diagnostic scales with specific questionnaires are made by specialists as a guide in the diagnostic process. In this work, Machine Learning (ML) was used in three datasets containing AQ-10 test results for adults, adolescents and children, and other characteristics that could influence the diagnosis of ASD. Experiments were performed using databases in order to determine what attributes are really relevant for an ASD diagnosis using ML. In order to select the attributes, the Random Forest was used to build a ranking of 23 attributes present in the database. In two of the three databases, it was possible to reduce the number of attributes to only 5, keeping an accuracy above 0.9. In the other Database to maintain the same level of Accuracy, the lowest number of attributes used was 7. The Support Vector Machine stood out from the other algorithms used in this work, obtaining superior results in all scenarios.

Keywords: Autism Spectrum Disorder. Machine Learning. Diagnosis.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ABC	*Autism Behavior Checklist*
ADTree	*Alternating-Decision Tree*
AM	Aprendizado de Máquina
APA	*American Psychiatric Association*
AQ	*Autism Quocient*
ASIEP-3	*Autism Screening Instrument for Educational Planning - Third Edition*
ASM	*Acoustic Segment Model*
ASSQ	*Autism Spectrum Screening Questionnaire*
CARS	*Childhood Autism Rating Scale*
CHART	*Checklist for Autism in Toddler*
CV	*Cross Validation*
DBD-ES	*Developmental Checklist-Early Screen*
DNN	*Deep Neural Network*
DSM	*Diagnostic and Statistical Manual of Mental Disorders*
ESAT	*Early Screening for Autistic Traits*
FN	Falso Negativo
FP	Falso Positivo
GI	Ganho de Informação
IA	Inteligência Artificial
IDE	*Integrated Development Environment*
IG	*Information Gain*
K-nn	*K-Nearst Neighbor*
LDA	*Linear Discriminant Analysis*
LMT	*Logistic Model Trees*

LR	*Linear Regression*
LVQ	*Learning Vector Quantization*
M-CHART	*Modified Checklist for Autism in Toddlers*
M-CHART-R	*Modified Checklist for Autism in Toddlers, Revised*
McRBFN	*Meta-cognitive Radial Basis Function Network*
ML	*Machine Learning*
MLP	*Multilayer Perceptron*
NB	*Naive Bayes*
NSCH	*National Survey of Children's Health*
PCA	*Princial Components Analysis*
Q-CHART	*Quantitative Checklist for Autism in Toddler*
RBF	*Radial Basis Function*
RNA	Rede Neural Artificial
RF	*Random Forest*
SBCL	*Child 28 Behavior Checklist*
SCQ	*Social Communication Questionnaire*
SRS	*Social Responsiveness Scale*
SVC	*Support Vector Classification*
SVM	*Support Vector Machine*
TEA	Transtorno do Espectro Autista
VP	Verdadeiro Positivo
VN	Verdadeiro Negativo

SUMÁRIO

1 INTRODUÇÃO

Ainda não existem definições totalmente precisas sobre o Transtorno do Espectro Autista (TEA). Dentre as várias definições existentes sobre o autismo, atualmente, uma das mais aceitas é que o TEA pode ser considerado como uma síndrome neuropsiquiátrica que causa déficits comportamentais, emocionais, comunicativos e em especial na capacidade do indivíduo se relacionar com outras pessoas [1][2][3][4][5].O TEA também pode ser definido como um transtorno global do desenvolvimento, o qual afeta a parte neurológica do paciente, manifestando-se na infância, geralmente até os três anos de idade, acometendo principalmente o comportamento social, o desenvolvimento da função comunicativa e a percepção do indivíduo [6][7].

Ainda nos dias de hoje, os profissionais da área da saúde enfrentam uma grande dificuldade para realizar o diagnóstico de doenças como o TEA. Isso acontece em decorrência da inexistência de exames laboratoriais que possam confirmar o diagnóstico e sua grande variação sintomática [8]. Assim, é um profissional especialista no assunto, como neurologistas e psiquiatras quem faz a análise comportamental do indivíduo fechando o diagnóstico. Essa busca avaliar o grau de desenvolvimento do paciente baseado em uma série de fatores como: excessos comportamentais, cuidados pessoais, comunicação, habilidades sociais, dentre outros [9].

Uma maneira de guiar esse processo é utilizar questionários específicos para diagnóstico do TEA desenvolvidos por especialistas, os quais são conhecidos como escalas diagnósticas [8]. Essas servem como um guia para auxiliar o diagnóstico possuindo características próprias, mas em geral são formadas por questões objetivas respondidas por um especialista da área, baseando-se na observação do indivíduo e em entrevistas com o paciente e/ou seus responsáveis.

O Manual Estatístico e Diagnóstico da Associação Americana de Psiquiatria, conhecido pela sigla DSM (*Diagnostic and Statistical Manual of Mental Disorders*) atualmente é a mais conhecida referência para o diagnóstico de transtornos mentais. Esse manual foi desenvolvido para diagnosticar diversos tipos de transtornos mentais. A versão mais recente foi lançada em 2013 e recebe o nome de DSM-V [10]. Outros exemplos de escalas conhecidas são: CARS (*Childhood Autism Rating Scale*), criada por Schopler e Reichler em 1971 [11][12] e a escala M-CHAT (*Modified Checklist for Autism in Toddlers*) que foi elaborada pela *American Healthcare System* modificando a escala CHAT a fim de deixar os resultados mais exatos [13]. Diferente do DSM, essas escalas foram desenvolvidas com foco exclusivo no diagnóstico do TEA.

Existem alguns tipos diferentes de escalas diagnósticas e alguns deles são volta-

dos para especialistas da área, enquanto outros são chamados de escalas autoaplicáveis que podem ser utilizadas por pessoas sem conhecimentos específicos na área. Uma dessas escalas utilizáveis por qualquer pessoa é o *Autism Quocient* (AQ), que é um teste autoaplicável criado por Baron et al. [14], cujo objetivo era ser a primeira escala de diagnóstico de autismo que poderia ser aplicada sem a necessidade de um especialista. Tal teste será abordado no próximo capítulo.

Contudo, vários fatores podem retardar ou mesmo impedir o diagnóstico correto e, por consequência, o início do tratamento do TEA. Dentre esses fatores podem-se destacar: a demora na detecção dos primeiros sintomas, grande variação desses sintomas, falta de treinamento dos profissionais da área de saúde e também a dificuldades de acesso aos serviços de saúde [10][15].

Uma das possíveis alternativas para auxiliar os profissionais da área médica é empregar técnicas de *Machine Learning* (ML), ou em português *Aprendizado de Máquina* (AM), como um mecanismo de apoio à decisão, a fim de auxiliar a realização do diagnóstico ou pré-diagnóstico do TEA [16][17]. A aprendizagem de máquina consiste em uma técnica de Inteligência Artificial (IA) aplicada em vários ramos do conhecimento como: Sistemas Especialistas, Ferramentas de Apoio ao Diagnóstico, Processamento de Linguagem Natural, Reconhecimento de Padrões etc. Nessa técnica, o computador simula ou realiza um estudo do comportamento dos dados. Para isso, ele busca um novo conhecimento ou habilidade e organiza os dados, possibilitando a realização de uma melhora progressiva do seu próprio rendimento [18][19][20].

Tipicamente o diagnóstico do TEA utilizando AM é encarado como um problema de classificação, no qual um modelo é construído se baseando em dados previamente classificados de uma Base de Dados [21]. O uso do AM pode oferecer soluções bem eficientes para o diagnóstico do TEA, uma vez que possui modelos matemáticos e métodos computacionais capazes de lidar com grandes volumes de dados e regras. Esse é um dos pontos da grande contribuição deste trabalho.

1.1 Objetivos

Esta seção formaliza os objetivos deste trabalho.

1.1.1 Objetivo principal

Uma vez que o Aprendizado de Máquina oferece métodos computacionais que podem ser utilizados para criar modelos que solucionem problemas da vida real, o objetivo principal desta dissertação é desenvolver um modelo de AM para diagnóstico do Transtorno do Espectro Autista.

1.1.2 Objetivos específicos

Os objetivos específicos deste trabalho são:

1. Verificar a viabilidade de utilizar dados de escalas diagnósticas para criação de um modelo de aprendizado de máquina para a identificação do TEA;
2. Realizar levantamentos teóricos e experimentos que indiquem quais atributos causam mais impacto na construção de um modelo de Aprendizado de Máquina;
3. Selecionar o mínimo possível de atributos necessários para que o modelo funcione com pelo menos 90% de acurácia.

1.2 Organização do trabalho

Este trabalho está organizado da seguinte forma: o Capítulo 2 apresenta os conceitos e métodos que serviram como base para o desenvolvimento deste trabalho e dentre eles estão: o conceito do TEA, escalas diagnósticas, Aprendizado de Máquina e os algoritmos usados nos experimentos; o Capítulo 3 traz um levantamento de trabalhos publicados de 2012 até o começo de 2018 que utilizaram Aprendizado de Máquina para a identificação do TEA, os quais utilizaram vários tipos diferentes de bases de dados e algoritmos; o Capítulo 4 descreve a parte dos experimentos realizados durante o desenvolvimento deste trabalho, indo desde a proposta de um modelo de aprendizado até a aplicação dos algoritmos; o Capítulo 5 aponta os resultados após a realização dos experimentos para cada Base de Dados e modelo gerado, além de uma discussão sobre a acurácia desses modelos; o Capítulo 6 expõe as conclusões obtidas após o desenvolvimento do trabalho e possíveis trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

Este capítulo apresenta a fundamentação teórica do trabalho, os conceitos sobre o Transtorno do Espectro Autista e Aprendizado de Máquina.

2.1 Transtorno do Espectro Autista

O Transtorno do Espectro Autista atualmente é considerado como uma perturbação do desenvolvimento neurológico, afetando em especial, a maneira como seu portador compreende o mundo ao seu redor [22]. Existem vários tipos de classificações para o TEA. Para Lampreia [4], o autismo pode ser divido em três níveis, sendo o nível 1 o mais leve, o nível 2 considerado moderado e o nível 3 o mais severo. De acordo com Didehbani et al. [23], o primeiro nível tende a ter dificuldades em processar sinais sociais, tornando-se oprimido e ansioso em interações sociais, especialmente quando se trata de pessoas desconhecidas e dificuldades de expor pensamentos e emoções. O segundo nível apresenta graves problemas na sua comunicação, seja verbal ou não, além de uma extrema dificuldade de aceitar mudanças. O terceiro nível mostra um comprometimento mais grave em interações sociais, precisando de um apoio para sua comunicação ser funcional.

O termo Autismo foi utilizado pela primeira vez por Bleuler em 1911, ao se referir a um grupo de indivíduos que teriam perdido o contato com a realidade. Em 1943, o termo Autismo foi novamente utilizado dessa vez por Kanner [24] para descrever o comportamento de 11 crianças que segundo ele apresentavam “incapacidade de se relacionar de maneira normal com pessoas e situações, desde o princípio de suas vidas”. Em 1944, Asperger descreveu várias características semelhantes em indivíduos com dificuldade de comunicação e interação social, mas que possuíam inteligência normal. Graças aos estudos de Asperger, o TEA inicialmente começou a ser chamado de Síndrome de Asperger. Apenas os casos considerados leves de autismo e que não apresentavam atraso de linguagem receberam esse nome em homenagem ao pesquisador que descobriu essas características em comum nessas crianças. Cabe ainda ressaltar que hoje todos os indivíduos com TEA, dos mais leves aos mais severos são chamados de autistas [25] [7].

O TEA é uma patologia relativamente comum, embora as fontes de dados variem quanto a sua incidência (para Gomes et al. [1], estima-se que 1 a cada 88 nascidos vivos apresenta autismo, enquanto para Skafidas et al. [26] afeta 1 a cada 150) pode-se dizer que ele afeta aproximadamente 1% dos nascimentos no planeta, sendo o sexo masculino quatro vezes mais propenso que o feminino [10]. De acordo com Gomes et al. [1], no ano de 2010 no Brasil existiam cerca de 500 mil pessoas com autismo.

Quanto mais cedo o TEA é diagnosticado e o tratamento iniciado, melhores pode-

rão ser os resultados a essas pessoas. Devido às características do desenvolvimento físico da criança, em especial do cérebro, quanto antes o tratamento for iniciado, menores podem ser as expressões sintomática do TEA no indivíduo [10].

2.2 Escalas diagnósticas

Escalas diagnósticas constituem métodos específicos, desenvolvidos por especialistas para o diagnóstico de diversos tipos de transtornos mentais que não são diagnosticados com exames laboratoriais, seja por não existirem esses exames ou por se tratarem de exames muito complexos e demorados [8].

Um exemplo desses manuais é o DSM [27], apontado anteriormente, que engloba vários tipos de transtornos incluindo o TEA. Esse foi publicado pela primeira vez em 1952 pela *Associação Americana de Psiquiatria*[1] com o objetivo de criar uma melhor maneira de diagnosticar diversos transtornos mentais. A versão inicial era composta por 130 páginas que mostravam 106 transtornos mentais [28]. No DSM-IV trouxe a *tríade* de sintomas para o TEA que modela déficits em: - interação social; - comunicação; - interesses restritos e comportamentos repetitivos [29].

Desde então esse manual esteve sobre constante evolução, gerado pelo avanço na medicina psiquiátrica. Atualmente (lançado em 2013) o DSM está na versão 5 [27], conhecido como DSM-V e possui 947 páginas abordando mais de 300 tipos de transtornos e distúrbios psiquiátricos englobando critérios diagnósticos e indicações de tratamentos. A partir desse momento todos os casos são diagnosticados com diferentes níveis de gravidade e um único Espectro Autista [30]. O "autismo" é classificado como Transtorno do Neurodesenvolvimento, que recebe o nome de **Transtorno do Espectro do Autismo (TEA)** [7]. Em Guedes [31] é chamada a atenção que no DSM-V troca se a tríade diagnóstica por uma **díade** sendo composta por: - comportamentos repetitivos e interesses restritos; - e déficits sociais e comunicativos [29].

O site da Associação Americana de Psiquiatria[2] disponibiliza informações sobre o manual, além de incluir uma lista de sugestões propostas para serem utilizadas na criação de uma nova versão. Além desses manuais, que englobam vários tipos de patologias, existem diversas escalas diagnosticas que foram desenvolvidas especificamente para o TEA.

Escalas diagnósticas geralmente são formadas por questionários de análise comportamental do indivíduo, que podem ser aplicados por ele mesmo, caso esse esteja apto a responder às perguntas, ou pelos pais e/ou responsáveis que possam fornecer as informações de maneira correta ao profissional da área médica. O tamanho e o público alvo dessas escalas são fatos importantes a serem levados em consideração no processo de escolha [32].

[1] <https://www.apa.org/>
[2] <https://www.psychiatry.org/psychiatrists/practice/dsm>

Também existem escalas diagnósticas autoaplicáveis, ou seja, que não necessitam da presença de um especialista para a aplicação. O *Autism Quocient* (AQ) foi a primeira escala diagnostica autoaplicável devolvida por Baron et al. [14] em 2001. Ela contém 50 questões objetivas subdividida em 5 grupos de 10 questões cada. Cada grupo aborda aspectos nos quais indivíduos com TEA possuem algum tipo de dificuldade sendo eles: habilidades sociais, falta de concentração, atenção aos detalhes, comunicação e imaginação.

Apesar do AQ ser um teste de fácil realização, o número de questões utilizadas é alto, o que acaba por dificultar a correta realização do teste por indivíduos comuns. Para melhorar esse cenário, uma versão alternativa baseada no AQ foi apresentada em 2012 por Allison et al. [33]. Essa nova versão diminuiu o número de questões do teste para apenas 10, tendo, ainda, sido desenvolvido quatro novas versões do teste: para crianças com menos de quatro anos, crianças acima de quatro anos, adolescentes e adultos. O AQ-10 tornou-se uma solução mais atrativa que o AQ, pois o menor número de itens torna a realização do teste mais fácil e rápida.

Para medir o desempenho em problemas de classificação, as predições são marcadas como Verdadeiro Positivo (VP), Verdadeiro Negativo (VN), Falso Positivo (FP) e Falso Negativo (FN). O Verdadeiro Positivo se refere ao número de indivíduos com TEA corretamente classificados pelo teste. O Falso Positivo indica o número de indivíduos sem TEA que foram classificados como se tivessem autismo. O Falso Negativo indica os indivíduos com TEA classificados como não tivessem TEA. Por fim, o Verdadeiro Negativo é o número de indivíduos sem TEA classificados corretamente como sem [34].

É chamada de *sensibilidade* a capacidade que um modelo tem de predizer os Verdadeiros Positivos [35]. Nesse caso a capacidade de que os testes têm de diagnosticar corretamente um indivíduo com TEA é calculada pela equação 2.1.

$$Sensibilidade = \frac{VP}{VP + FN} \tag{2.1}$$

A *especificidade*, por sua vez, refere-se à capacidade do teste de predizer corretamente os Verdadeiros Negativos [35]. Ou seja, diagnosticar corretamente quais indivíduos não possuem TEA. Seu valor é obtido pela equação 2.2.

$$Especificidade = \frac{VN}{VN + FP} \tag{2.2}$$

A *acurácia* é a medida estatística que avalia a eficácia de uma classificação binária, ou seja, o quão bem ela consegue classificar corretamente as duas classes [35]. Vide equação 2.3.

$$Acurácia = \frac{VP + VN}{VP + VN + FP + FN} \tag{2.3}$$

A Figura 1 mostra as possíveis classificações dos indivíduos pelo teste. Essa representação é conhecida como *Matriz de Confusão*, na qual se compara o valor real de cada instância com o valor predito [35]. É uma medição de desempenho nesse caso para o problema de classificação de aprendizado de máquina, tendo uma tabela com quatro combinações diferentes de valores previstos e reais.

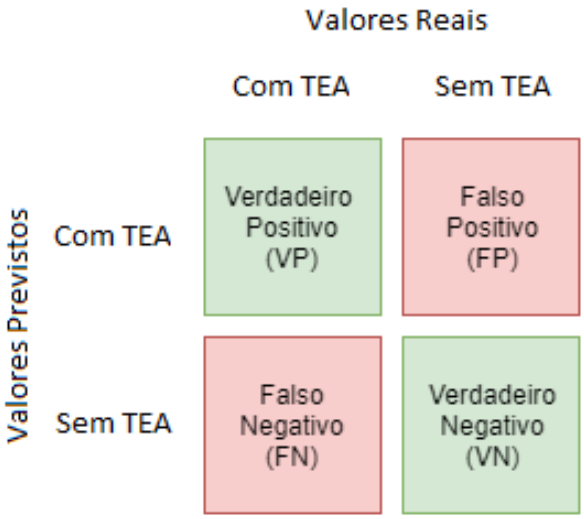


Figura 1 – Possíveis classificações dos testes

As Tabelas 1 e 2 mostram algumas escalas diagnósticas do TEA, dando informações sobre número de questões, público alvo, faixa etária indicada, tempo médio de realização do teste, sensibilidade e especificidade dos resultados [34].

Tabela 1 – Testes convencionais

Teste	Itens	Público Alvo	Faixa Etária	Tempo min	Sensibilidade	Especificidade
CHAT (Checklist for Autism in Toddlers)	14	crianças	18-24 meses	8 - 15	40%	98%
M-CHAT (Modified Checklist for Autism in Toddlers)	23	crianças	16-30 meses	10 - 20	95-97%	99%
M-CHAT-R (Modified Checklist for Autism in Toddlers, Revised)	20	crianças	16-30 meses	10 - 20	ND	ND
Q-CHAT (Quantitative Checklist for Autism in Toddler)	25	crianças	18-24 meses	15 - 20	88%	91%
Q-CHAT-10	10	crianças	19-24 meses	5 - 10	91%	89%
ABC (Autism Behavior Checklist)	57	crianças	3-14 anos	20 -30	77%	91%
(CARS)-2 Childhood Autism Rating Scale	15 * 2	crianças	V1: até 6 anos V2: 6-13 anos	10 - 20	81%	87%
DBD-ES (Developmental Checklist-Early Screen)	17	crianças	18-48 meses	10 - 15	83%	48%
ESAT (Early Screening for Autistic Traits)	14	crianças	16-30 meses	10 - 15	88%	14%
ASIEP-3 (Autism Screening Instrument for Educational Planning - Third Edition)	47	crianças	2-13 anos	ND	100%	81%

Tabela 2 – Testes híbridos

Teste	Itens	Público Alvo	Faixa Etária	Tempo min	Sensibilidade	Especificidade
ASSQ (Autism Spectrum Screening Questionnaire)	27	crianças e adolescentes	7-16 anos	10 - 15	91%	86%
AQ(Autism Spectrum Quotient)	50	adultos	>18 anos	20-30	93%	52%
AQ-10-Adult	10	adultos	>18 anos	5-10	77%	74%
AQ-Adolescent	50	adolescentes	12-15 anos	20-30	ND	ND
AQ-Child	50	crianças	4-9 anos	20-30	95%	95%
AQ-10-Adolescent	10	adolescentes	12-15 anos	5 - 10	ND	ND
AQ-10-Child	10	crianças	4-11 anos	5-10	ND	ND
SCQ (Social Communication Questionnaire)	40	crianças e adolescentes	<4 anos	10 - 20	58-62%	93-100%
SRS (Social Responsiveness Scale)	65	crianças e adolescentes	4-18 anos	20-30	67%	78%
SRS-2 (Social Responsiveness Scale)	65	crianças e adolescentes	4-18 anos	20-30	78%	94%
CBCL (Child 28 Behavior Checklist)	118	crianças e adolescentes	6-18 anos	25-40	75%	82%

A maioria dos testes das Tabelas 1 e 2 apresenta um nível de acerto acima de 70%, o que pode ser considerado um bom grau de acerto. Alguns testes não possuem valores definido para a Sensibilidade e/ou Especificidade.

2.3 Aprendizado de Máquina

Em um mundo imerso em diversos tipos de tecnologias, é natural procurarmos maneiras para utilizá-las em benefício da qualidade de vida das pessoas. Existem algumas maneiras de aplicar a computação na resolução dos problemas do mundo real, como, por exemplo, por meio da utilização da Inteligência Artificial (IA) que consiste na criação de programas que possam apresentar características de inteligência, ou seja, capacidade para aprender a executar uma tarefa simples ou resolver problemas complexos [18].

O Aprendizado de Máquina, denotado neste trabalho por AM, é uma área da

IA que se desenvolveu com estudos de Reconhecimento de Padrões. Com a aplicação de métodos matemáticos e estatísticos, os algoritmos de AM aprendem automaticamente por meio de um processo de treino e de teste a realizar uma determinada tarefa, gerando assim um modelo de aprendizado [36]. Em geral, o AM é aplicado aos problemas de Classificação e de Regressão.

Classificação refere-se a um problema cujo objetivo final do modelo seja prever variáveis de valores discretos, ou seja, apresenta um número finito de respostas possíveis [36]. Dessa forma, pode-se classificar algo em duas ou mais categorias previamente definidas. Para isso é necessário uma base de dados já classificados, para ser usada no processo de treinamento do algoritmo. Esse valor é conhecido como *Classe* ou *Rótulo*. Com isso, os algoritmos geram modelos de conhecimento capazes de classificar novas entradas [37]. Neste trabalho o objetivo é aplicar AM utilizando a classificação para identificar se um indivíduo possui TEA ou não.

A *regressão*, por sua vez, é utilizada em problemas cujo resultado final deva apresentar variáveis de natureza contínuas, ou seja, não existe um número finito de valores possíveis [36]. A regressão emprega métodos matemáticos e estatísticos que relacionam variáveis explanatórias e as dependentes. Pode ser utilizada para predizer, por exemplo, qual será o reajuste em um plano de saúde, desde que seja fornecida uma Base de Dados anteriores, contendo as variáveis para explicar como o reajuste é calculado.

O AM é dividido em duas categorias: o aprendizado supervisionado e o não supervisionado. O *aprendizado não supervisionado* consiste em um algoritmo no qual nenhum tipo de resposta esperada é dado para a resolução do problema, cujo objetivo é descobrir por si mesmo a melhor maneira de reconhecer os padrões de dados [18]. Por outro lado, o *aprendizado supervisionado* funciona de maneira oposta ao não supervisionado. Antes da aplicação do algoritmo, um rótulo ou classe é aplicado aos dados, a fim de mostrar uma resposta esperada para o problema. Esse aprendizado funciona basicamente com duas fases, na qual a primeira consiste no treinamento, em que uma parte dos dados da base é usada para criar um modelo. Após o treino, outra parte dos dados que não foi previamente utilizada, é usada para avaliar o desempenho. Após essas etapas, é possível aos algoritmos modelarem uma solução para identificar se o indivíduo possui ou não TEA com base nos exemplos fornecidos [18].

2.4 Algoritmos de AM utilizados

Esta seção apresenta uma introdução aos algoritmos de Aprendizado de Máquina utilizados para o desenvolvimento deste trabalho. Esses algoritmos foram escolhidos com base em uma revisão sistemática realizada sobre o uso de AM no diagnóstico do TEA e pode ser visto mais detalhes na Seção 3.1 e também no tipo de base de dados.

2.4.1 *Principal Compenents Analysis* - PCA

Principal Compenents Analysis (PCA) ou em português, Análise dos Componentes Principais é um método de análise de dados baseado na correlação entre as variáveis. O algoritmo emprega um modelo matemático utilizando a álgebra linear para a transformação dos dados. A PCA converte o conjunto de variáveis original do problema em um novo conjunto de variáveis, obtido pela combinação linear das variáveis originais, chamado de *Principal Componentes* (PC) ou em português Componentes Principais, que são nomeados como PC1, PC2, ..., PCn [38].

Os PCs tendem a concentrar a maior parte da informação do problema em um número menor de variáveis. Dessa forma, eles provêm uma maneira de reduzir a dimensionalidade do problema sem perda significativa das informações. A PCA é um dos métodos mais comuns para realizar a análise de informações de um problema, uma vez que redução de dimensionalidade desse pode facilitar a compreensão dos dados [39].

Além disso, sua representação gráfica possibilita a observação dos espaços de dados podendo indicar algumas características do problema, como por exemplo, o seu grau ou quais tipos de algoritmos de AM podem ter uma maior eficiência [38].

2.4.2 *k-Nearst Neighbor* - k-NN

O *k-Nearst Neighbor* (k-NN) é um algoritmo que pode ser utilizado para problemas de regressão e classificação. O k-NN funciona considerando que cada amostra, presente na Base de Dados, é um ponto em um espaço composto por n-dimensões, em que n é igual ao número de atributos [40].

Após a criação do classificador, toda vez que uma nova amostra é apresentada ao modelo, ela tem sua classificação feita por meio da observação de seus k-vizinhos. Para realizar essa observação é calculada a distância (normalmente usa-se a *Distância Euclidiana*, mas outros tipos de cálculos de distância podem ser utilizados) entre a nova amostra e seus k mais próximos vizinhos. No caso de classificação, a classe com maior número de vizinhos é atribuída a ela e, em caso de regressão, é atribuída uma média dos valores dos k vizinhos [40].

Um exemplo de k-NN com k = 5 pode ser visto na Figura 2, no qual o elemento representado com um losango dourado será classificado com seus 5 vizinhos mais próximos. Nesse caso 3 dos 5 vizinhos pertencem à Classe 2. Portanto a nova entrada será classificada como Classe 2 .

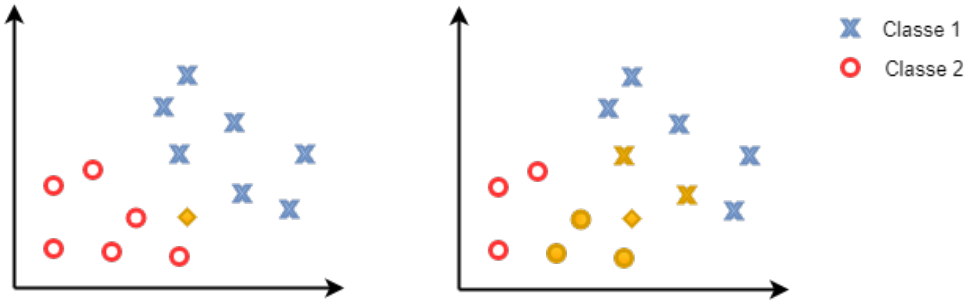


Figura 2 – Exemplo kNN com k = 5.

2.4.3 Árvore de Decisão J48

A J48 é uma implementação em Linguagem de Programação Java baseada do modelo C4.5, proposto por Ross Quinlan em 1986 [41], com o objetivo de fornecer um modelo de classificação que pudesse ser utilizado em diversos tipos de problemas sem a necessidade de modificações, ou seja, um modelo genérico.

As árvores de decisão são muito populares em trabalhos que utilizam AM devido a sua fácil compreensão. O modelo gerado em formato de árvore, no qual apenas é necessário seguir os galhos para encontrar a resposta correta, é muito simples se comparado a outros algoritmos de AM [42].

Diferente de uma árvore utilizada em estruturas de dados, na qual cada nó armazena uma informação, *na árvore de decisão* cada nó contém um teste ou uma condição e os ganhos mostram as possíveis saídas desse teste e apenas as raízes da árvore exibem o resultado final da classificação [41].

A construção dessas árvores é feita empregando métricas, como Ganho de Informação que cada atributo dá ao problema. Dessa forma, ao se aplicar o ganho de informação, um problema mais complexo é decomposto em problemas menores e mais simples [42].

Na Figura 3 há exemplo de uma árvore de decisão, criado por [43], em que a árvore mostra um processo de decisão para indicar se um paciente está dou não doente. Para isso, são feitos questionamentos sobre sintomas que o paciente possui e a cada resposta dada uma nova pergunta é feita baseada nas respostas anteriores.

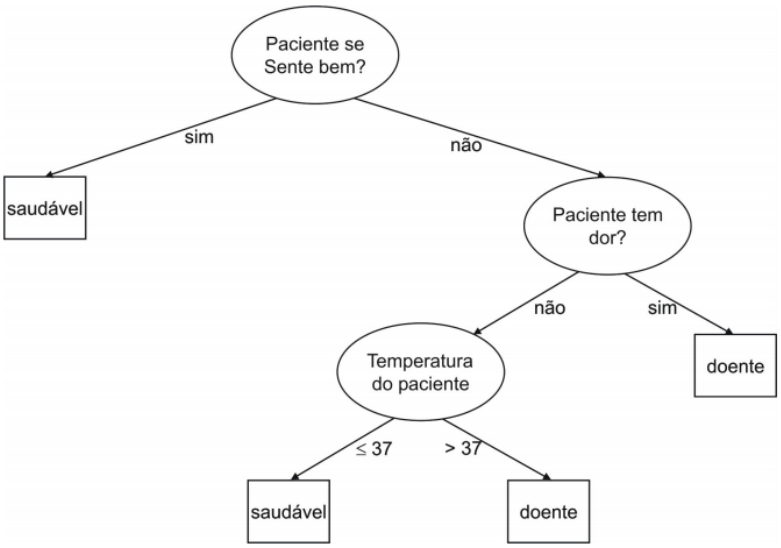


Figura 3 – Exemplo da estrutura de uma árvore de decisão para indicar se um paciente está ou não doente

2.4.4 *Random Forest* - RF

Random Forest é um algoritmo proposto por Breiman [44], baseado no modelo de agregação de ideias, criado para oferecer uma boa resolução genérica de problemas complexos. Trata-se de um algoritmo do tipo *Ensemble* que utiliza aleatoriedade e uma combinação de múltiplas árvores de decisão do tipo *Classification And Regression Tree* (CART). Essas árvores são criadas utilizando amostragens geradas da separação da Base de Dados inicial em diversos subconjuntos por meio de uma técnica chamada de *bootstrap aggregating (bagging)* [44].

A primeira parte é o processo de *bootstrap*. Nele os dados de um conjunto original são divididos em n subconjuntos (*subsets*) independentes montados de maneira aleatória por reamostragem com reposição e sem repetição, ou seja, é possível que uma amostra da Base de Dados original possa aparecer em todos os subconjuntos, porém apenas uma vez em cada [44].Um exemplo desse processo pode ser visto na Figura 4, no qual um conjunto de dados originais foi dividido em três subconjuntos de dados.

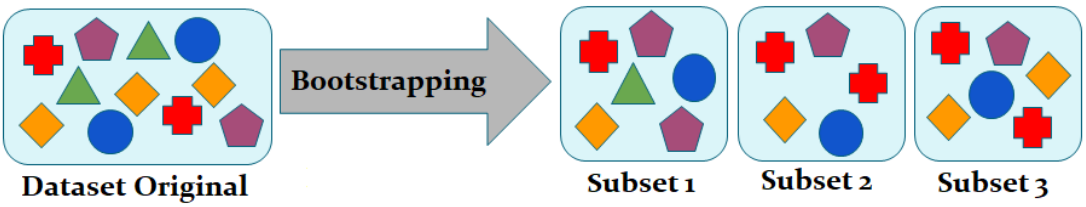


Figura 4 – Um modelo de separação de amostras usado pela RF.

Após a construção desses subconjuntos de dados, os mesmos são submetidos individualmente a um algoritmo de árvore de decisão CART gerando então o mesmo número de

modelos que de subconjuntos. Ná próxima etapa do processo o *aggregating* ocorre quando uma amostra é submetida ao modelo, seja de classificação ou regressão.

Caso o modelo seja de classificação, cada árvore gerada irá apresentar uma classificação para a entrada e será efetuado um processo de "votação", no qual a classificação final, dada a entrada, será igual à majoritária [45].

Já em um problema de regressão, em vez de votação, são efetuados cálculos dos resultados, baseados em valores, tais como final da média, moda ou mediana das árvores [45].

O processo de cálculo de importância é realizado durante a construção das árvores do modelo de aprendizado, no qual os atributos mais relevantes tendem a se alocar nos nós mais próximos à raiz da árvore. Dessa forma, é possível saber quais são os atributos que mais se destacam na Base de Dados para a resolução do problema [46].

A RF utiliza o Índice Gini que mede o grau de impureza dos atributos para realizar o processo de separação das Classes. Assim sendo, quanto menor for esse valor de Gini, maior a importância desse atributo. *Mean Decrease Gini* é uma métrica que faz parte do algoritmo e que serve exatamente para mostrar o valor da média de decaimento do Gini, destarte, nesse caso, quanto maior for essa métrica, maior será a importância do atributo [47].

2.4.5 *Support Vector Machine* - SVM

A *Support Vector Machine* (SVM) funciona traçando hiperplanos que visam maximizar a distância mínima entre as Classes, criando uma região de fronteira entre elas. Pode ser aplicada aos problemas de classificação e regressão. Além disso, possui outras funções como a detecção de *outliers* e vantagens em ralação a outras técnicas de AM, conseguindo gerar bons resultados com problemas de classificação não lineares, com poucas amostras para treinamento ou problemas com alta dimensionalidade [48].

Inicialmente, o foco da SVM era a solução de problemas binários e, por meio dos hiperplanos, buscava a separação máxima entre as classes [48]. Para obter essa separação ideal entre as classes são utilizados os vetores de suporte que são traçados nas regiões de fronteira e após a criação desses vetores de suporte um hiperplano ideal para divisão entre as classes é gerado. Um exemplo de uma SVM aplicada a um problema linearmente separável entre duas classes pode ser visto na Figura 5, em que os hiperplanos em azul e vermelho representam as regiões de fronteiras entre as classes 1 e 2, respectivamente, e o hiperplano em verde seria o ideal para a separação das classes.

Contudo, nem sempre os problemas são de natureza linear. Dessa forma a SVM perdia muito de seu potencial quando utilizada em problemas mais complexos. Para solucionar esse problema, um sistema de alteração de *kernels* foi adicionado. Na SVM podem

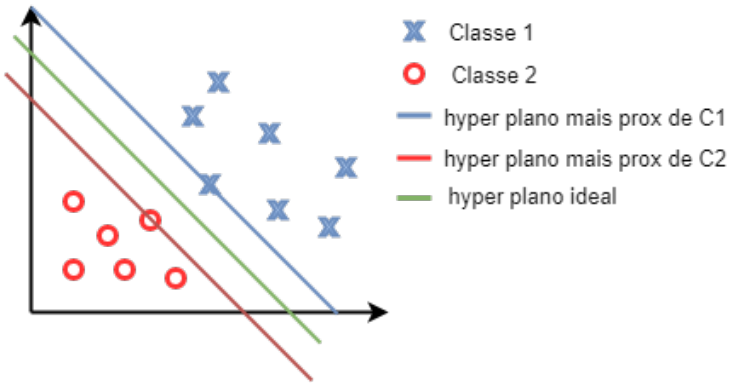


Figura 5 – Uma SVM sendo utilizada para separar 2 classes.

ser utilizados diferentes tipos de *kernels*, como por exemplo, linear, polinomial, radial e sigmoidal [49]. Além disso, é possível criar novos tipos ou fazer alterações nos modelos mais clássicos caso necessário.

Os *kernels* representam o núcleo da SVM, ou seja, o tipo de função matemática que será utilizada para calcular a região de fronteira entre as amostras. É chamada região de fronteira a área na qual ocorreria a máxima separação entre as classes. Em uma SVM Linear, esse núcleo utiliza equações de retas para gerar as fronteiras entre as classes. É importante salientar que os problemas de classificação possuem características próprias que podem fazer com que um problema se adapte melhor com um dos *kernels* em questão, melhorando assim significativamente a eficácia do modelo gerado [49].

Um problema pode ser dito não linear quando um hiperplano não é capaz de dividir de uma maneira satisfatória os dados de treinamento de um problema [49]. Nesse caso, em vez de utilizar um hiperplano, é possível usar outros tipos de fronteira, como podem ser vistos na Figura 6, na qual uma fronteira curva é a mais eficaz para separar os dados do que qualquer hiperplano.

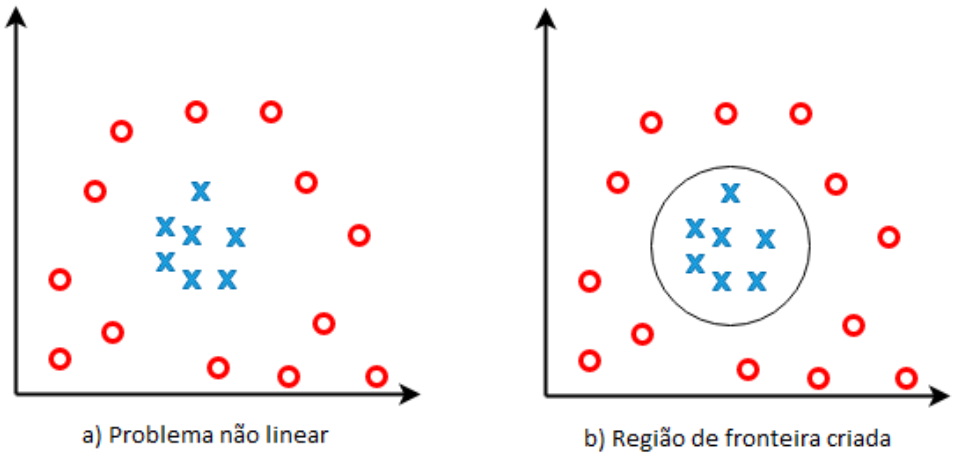


Figura 6 – Uma SVM para um problema de dados não lineares.

A alteração de *kernels* possibilita essa flexibilização para criar a região de fronteira que melhor se adapte ao problema [49].

3 TRABALHOS RELACIONADOS

A fim de obter um maior entendimento sobre como o AM vem sendo utilizado para o diagnóstico do TEA, foi realizada uma revisão sistemática sobre o assunto.

3.1 Revisão Sistemática

Uma revisão sistemática é uma metodologia de pesquisa científica, que tem por objetivo identificar, avaliar e interpretar as pesquisas mais relevantes disponíveis em uma dada área de conhecimento ou tema [50]. Para realizar essa revisão sistemática foram realizadas três etapas: Planejamento, Desenvolvimento e Resultados.

3.1.1 Planejamento

A etapa de planejamento de uma revisão sistemática refere-se à elaboração do roteiro pelo qual o trabalho irá se guiar. Essa etapa ocorre antes do inicio do trabalho e devem incluir: objetivos da revisão, as perguntas que devem ser respondidas, as palavras-chave, a estratégia para obter os dados e critérios de inclusão e exclusão [50].

3.1.2 Objetivos e Perguntas a Serem Respondidas

As perguntas selecionadas que devem ser respondidas pelos artigos selecionados são:

1. Quais os benefícios de utilizar AM para diagnóstico do TEA?
2. Qual o tipo mais utilizado de AM para diagnosticar TEA?
3. Quais os Algoritmos de AM mais utilizados no diagnóstico de TEA?
4. Quais tipos de base de dados vêm sendo usados para o diagnóstico do TEA?

3.1.3 Bases de Pesquisa

Os dados presentes nesta revisão sistemática foram retirados das seguintes bases de pesquisa:

1. ScienceDirect[1];
2. IEEE Xplore Digital Library[2];

[1] <sciencedirect.com>
[2] <ieeexplore.ieee.org>

3. ACM Digital Library[3];
4. Nature Publishing Group (Nature research)[4];
5. SchoolarGoogle[5];
6. Springer[6].

3.1.4 Palavras-Chave e Sinônimos

Uma das partes mais importantes para uma boa obtenção de dados é a escolha de boas palavras-chave. Para tal, algumas pesquisas preliminares foram feitas, a fim de estabelecer as melhores palavras-chave. Após isso, as seguintes palavras-chave foram definidas:

- *machine learning to improve autism spectrum disorder* (ASD) *diagnosis*;
- *machine learning and ASD*;
- *machine learning and ASD diagnosis* ;
- *use of machine learning to diagnose ASD.*

3.1.5 Critérios de Inclusão e Exclusão

Os critérios para inclusão e exclusão dos trabalhos na revisão sistemática são descritos a seguir.

Critérios para inclusão:

- escritos em Línguas Inglesa ou Portuguesa;
- ser o foco do artigo, ou seja o uso de Aprendizado de Máquina para diagnosticar Transtorno do Espectro Autista;
- artigos com no máximo seis anos de publicação.

Critérios para exclusão:

- artigos que exploram a utilização de AM para problemas diferentes do TEA;
- artigos que não expõe claramente os tipos de base de dados e/ou algoritmos utilizados;

[3] <dl.acm.org>
[4] <nature.com>
[5] <scholar.google.com>
[6] <rd.springer.com>

- artigos com mais de seis anos de publicação.

Essas buscas foram descartadas, uma vez que o objetivo dessa revisão sistemática é mostrar como o tema proposto vem sendo trabalhado atualmente.

3.1.6 Desenvolvimento

Na etapa de desenvolvimento também foi seguido o modelo proposto por Kitchenham et al. [50] e foram executadas as seguintes ações:

1. Busca dos artigos nas bases de conhecimento anteriormente mencionadas;
2. Remoção dos artigos duplicados;
3. Leitura dos *abstracts* dos artigos para aplicação dos critérios de inclusão e exclusão;
4. Leitura completa de todos os artigos selecionados.

3.1.7 Resultados

No total foram levantados 37 trabalhos publicados entre 2012 e 2018 que atenderam aos critérios definidos. A Figura 7 mostra uma distribuição temporal dos artigos selecionados, apontando quantos desses foram publicados em cada ano. Importante salientar que o número de artigos de 2018 refere-se apenas aos publicados nos primeiros seis meses do ano, uma vez que esse levantamento foi feito no meio do ano de 2018.

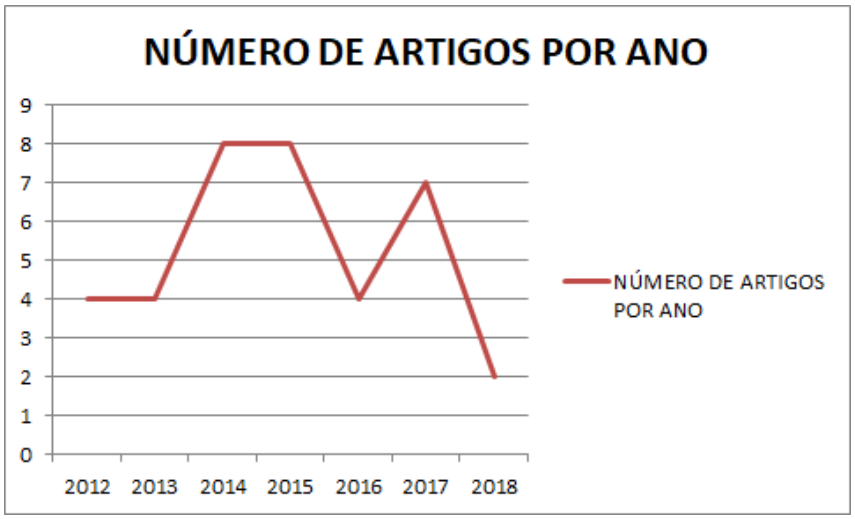


Figura 7 – Número de artigos sobre a temática pulicados por ano.

Esses 37 trabalhos utilizaram estratégias variadas buscando propor modelos de AM para diferentes tipos de bases de dados e algoritmos.

Na Tabela 3 mostra a referência, ano de publicação, o tipo de Base de Dados e os Algoritmos de AM utilizado(s) para cada um dos trabalhos.

Tabela 3 – Trabalhos selecionados que utilizam AM para diagnóstico do TEA.

Trabalho	Base de Dados	Classificador(es)
Alwakeel et al. (2015) [51]	Obtidos por sensores wireless para aquisição de dados comportamentais	Criado pelos autores
Martinez et al. (2012) [52]	Imagens Faciais	SVM
Yahata et al. (2016) [53]	Imagens de Ressonância Magnética	Combinação de Sparse CanonicalCorrelation Analysis (L1-SCCA) e Simple Linear Regression (SLR)
Mythili et al. (2014) [54]	Comportamentais	Rede Neural, SVM e Lógica Fuzzy
Vidhusha et al. (2015) [55]	Imagens de Ressonância Magnética	Learning Vector Quantization (LVQ) e SVM
Bone et al. (2015) [16]	Comportamentais	Alternating-Decision Tree (ADTree)
Wang et al. (2015) [56]	Imagens Olhos	SVM
Vigneshwaran et al. (2013) [57]	Imagens Ressonância Magnética	Meta-cognitive Radial Basis Function Network (McRBFN)
Thabtah (2017) [21]	Comportamentais	ADTree, Neural Networks
Haker et al. (2016) [58]	Imagens Ressonância Magnética	Algoritmos Bayesianos
Jamal et al. (2014) [59]	Imagens Ressonância Magnética	SVM
Duda et al. (2017) [60]	Comportamentais	Support Vector Classification (SVC), Logistic Regression with l1 Regularization (Lasso), Logistic Regression with l2 Regularization (Ridge) e Linear Discriminant Analysis (LDA).
Jiang et al. (2013) [61]	Genéticos	Random Forest
Cicchetti et al. (2014) [62]	Comportamentais	SVM

Tabela 3 – Trabalhos selecionados que utilizam AM para diagnóstico do TEA.

Trabalho	Base de Dados	Classificador(es)
Hazlett et al. (2017) [63]	Imagens de Ressonância Magnética e dados físicos	Deep Learning
Lee et al. (2013) [64]	Sonoros	K-nn, SVM, Deep Neural Network (DNN), Acoustic Segment Model (ASM)
Obafemi et al. (2015) [65]	Imagens Faciais	SVM, RF e MLP
Emerson et al. (2017) [66]	Imagens de Ressonância Magnética	SVM
Heinsfeldet al. (2018) [67]	Imagens de Ressonância Magnética	Deep learning
Deshpande et al. (2013) [68]	Imagens de Ressonância Magnética	SVM
Liu et al. (2016) [69]	Imagens Faciais	K-means e SVM
Segovia et al. (2014) [70]	Imagens de Ressonância Magnética	SVM
Ghiassian et al. (2013) [71]	Imagens de Ressonância Magnética	Naive Bayes (NB), K-nn e SVM
Garberson et al. (2017) [72]	Comportamentais e Vídeos	Random Forest
Thabtah (2018) [73]	Revisão Sistemática	Revisão Sistemática
Zhou et al. (2014) [74]	Imagens de Ressonância Magnética	SVM, Bayes network (BayesNet)
Ecker et al. (2015) [75]	Imagens de Ressonância Magnética	SVM
Retico et al. (2014) [76]	Imagens de Ressonância Magnética	SVM
Skafidas et al. (2014) [26]	Genéticos	
Kosmicki et al. (2015) [77]	Comportamentais	ADTree, Functional Tree, SVM, Logistic Model Trees (LMT), Logistical Regression (LR), Naive Bayes (NB), NBTree e Random Forest
Duda et al. (2014) [78]	Comportamentais	Criou seu próprio

Tabela 3 – Trabalhos selecionados que utilizam AM para diagnóstico do TEA.

Trabalho	Base de Dados	Classificador(es)
Wall et al. (2012) [79]	Comportamentais	15 – se destacando ADTree e LADTree.
Bone et al. (2016) [80]	Comportamentais	SVM
Crippa et al. (2015) [81]	Comportamentais	SVM
Wall et al. (2012) [17]	Comportamentais	ADTree
Bekerom (2017) [82]	Comportamentais	Naive Bayes, SVM, J48 e RF
Michaelson et al. (2012) [83]	Genéticos	Random Forest

3.2 Análise geral dos trabalhos

Definir como um problema será modelado por meio de AM começa com a decisão dos tipos de dados que serão utilizados, sendo que a eficácia de um bom modelo preditivo está diretamente ligada à escolha adequada de dados para treinamento do modelo.

Dentre os 37 trabalhos citados anteriormente foram encontrados vários tipos de dados que foram usados em conjunto com AM, a fim de buscar novas soluções para o diagnóstico do TEA. Esses tipos de dados foram:

- **Dados de Escalas Diagnósticas**: dados produzidos pela utilização de escalas diagnósticas que avaliam o comportamento do indivíduo em questão. Cada escala diagnóstica possui suas particularidades na maneira de como se chega ao resultado final. Algoritmos de AM irão fornecer modelos baseando-se nos algoritmos aplicados. Um exemplo é o uso de algoritmos de Árvore de Decisão para criar um modelo de aprendizado, o qual será gerado em forma de uma árvore, ou seja, criará uma nova abordagem para avaliar o teste;
- **Dados de Imagem de Ressonância Magnética**: trata-se do emprego de imagens do cérebro de pessoas com e sem TEA. Utilizando algoritmos de AM aplicado a essas imagens busca-se encontrar diferenças entre um cérebro com e sem o transtorno do autismo;
- **Dados de Imagens Faciais**: assim como nos dados com imagens de ressonância magnética, as imagens faciais são utilizadas com o objetivo de identificar caracte-

rísticas faciais que sejam únicas aos indivíduos com TEA e que possam ser usadas para auxiliar o diagnóstico;

- **Dados Genéticos**: possuem caraterísticas dos genes de indivíduos que têm ou não TEA, a fim de encontrar quais características genéticas podem auxiliar no diagnóstico;
- **Dados de Áudio**: contém vários registros feitos em áudio e convertidos para dados, a fim de serem usados para o diagnóstico do TEA.

Conforme mostra a Tabela 3, a abordagem mais utilizada para trabalhar com o TEA foi a classificação com os algoritmos SVM, RF, *Naive Bayes* [82], RNA's e Árvores de Decisão.

Após a escolha dos tipos de dados, é necessário definir como esses serão trabalhados, ou seja, quais algoritmos serão aplicados para a geração de um modelo preditivo. Nos trabalhos foram variados tipos de algoritmos os que mais se destacaram, como: Árvores de Decisão, *Naive Bayes*, RF, Redes Neurais e SVM's. Dentre os algoritmos aplicados na geração dos modelos de aprendizado de máquina, o mais usado foi a SVM que esteve presente em 23 dos 37 trabalhos. Isso se deve ao fato de que a SVM é capaz de gerar modelos preditivos muito poderosos com diversos tipos de dados, o que a torna um algoritmo extremamente versátil, podendo trabalhar todos os tipos de bases de dados mencionados anteriormente [48].

A maior parte dos artigos usou bases de dados de escalas ou imagens para diagnosticar o TEA. Em Bekerom [82], a base de dados da *National Survey of Children's Health* (NSCH) contendo vários tipos de dados relativos às várias doenças, não apenas o TEA. Essa Base de Dados foi composta por 95577 amostras e 367 atributos. Contudo, como apenas uma pequena porcentagem desses indivíduos (cerca de 2000) possuía TEA empregaram-se técnicas de *under-sampling* as quais reduzem o número de amostras, para criar uma nova base dados que continha cerca de 50% de indivíduos com TEA e 50% sem.

Depois da criação dessa nova base, dos 367 atributos anteriores, Bekerom [82] utilizou-se de apenas atributos que já estavam presentes na literatura, sendo eles: dificuldade de aprendizagem, atraso no desenvolvimento, problemas com fala ou linguagem, peso ao nascer, ser prematuro, realizar atividades físicas, envolvimento com atividades religiosas e índice de massa corporal, ou seja, utilizou uma mistura de atributos físicos e comportamentais que estariam ligados ao TEA. Após esse processo de preparação foram aplicados os seguintes algoritmos de AM: Naive Bayes, SVM, J48 e RF, obtendo um acerto da média de 85%.

Em Obafemi et al. [65] aplicou-se AM com dados de imagens de estruturas faciais de pessoas, em especial crianças, com o objetivo de procurar características nas faces que poderiam ser exclusivas dos indivíduos com TEA. Para isso eles mesmos selecionaram

crianças, cujo diagnóstico do TEA já tivesse sido realizado e aplicaram algoritmos para detalhar as características faciais de cada indivíduo criando, assim, uma Base de Dados. Após esse processo, os dados foram submetidos a três algoritmos de AM, sendo eles: SVM, RF e *Neural Networks Multilayer Perceptron* (MLP), descrevendo dois cenários distintos.

No primeiro cenário foram empregadas 171 métricas faciais com todos esses atributos e o modelo com mais acerto foi o gerado pela MLP, com 93.55%. A SVM obteve 91,.94% e a RF 88.71%. No segundo cenário foi reduzido o número de atributos para 31 por meio de uma união matemática resultante da saída dos atributos mais relevantes para os três algoritmos de seleção.

Após a redução dos atributos, nesse segundo cenário, houve um ganho de desempenho nos modelos gerados pela SVM e pela RF, aumentando para 95.16% e 91.94%, respectivamente, enquanto o desempenho da MLP não sofreu alteração com a redução no número de atributos [65].

4 METODOLOGIA E EXPERIMENTOS

Este capítulo busca descrever a metodologia utilizada e os experimentos realizados ao longo deste trabalho.

4.1 Modelo proposto

Este trabalho propõe a utilização de AM para criar três modelos de aprendizado: para adultos, adolescentes e crianças. A Figura 8 mostra o fluxo realizado durante o desenvolvimento deste trabalho.

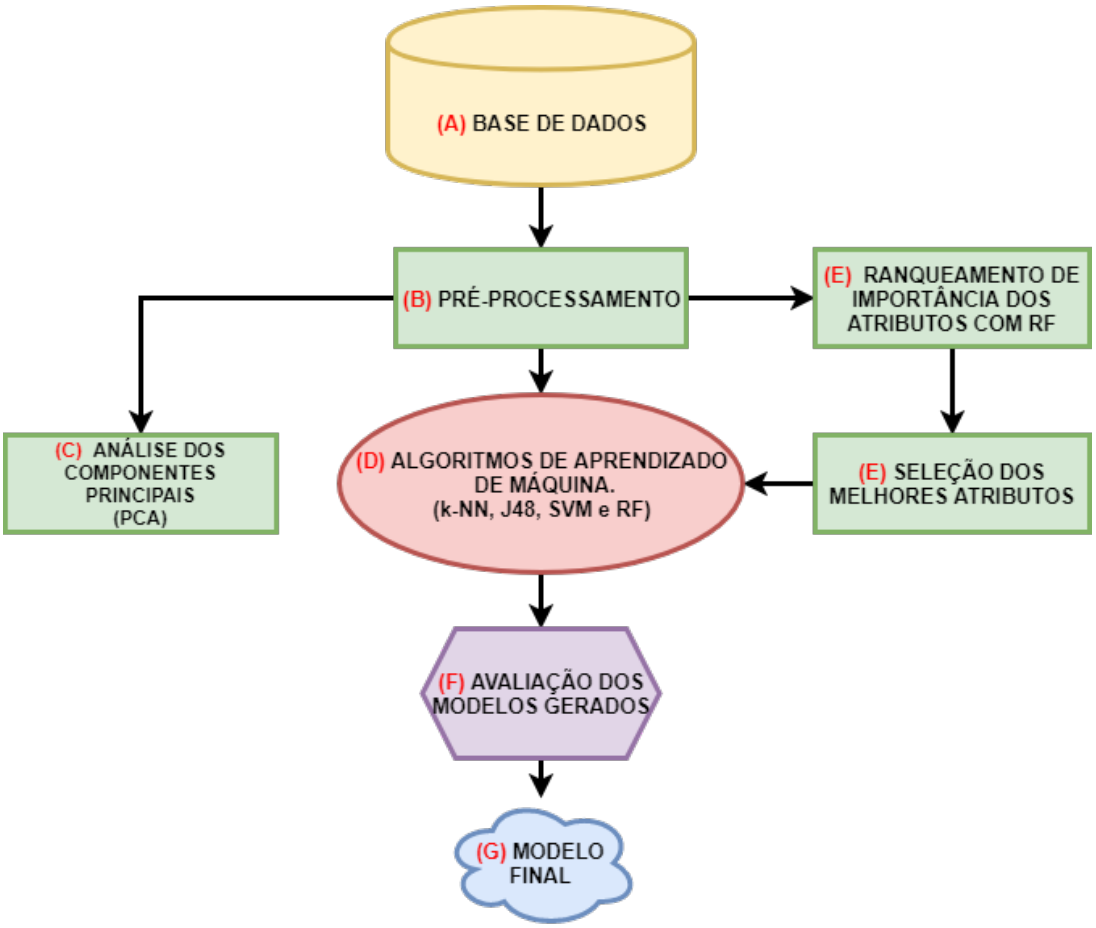


Figura 8 – Fluxo de desenvolvimento seguido durante o trabalho.

O diagrama começa com a Base de Dados, definida como **(A)**. Neste trabalho as três bases de dados foram submetidas ao mesmo processo separadamente, portanto esse fluxo do diagrama se repete três vezes. Os dados das bases de dados foram para o pré-processamento **(B)**, etapa responsável por verificar a existência de dados faltantes e ruídos. Após o pré-processamento foi executada a PCA **(C)** para fornecer uma melhor visão dos dados. Os dados do pré-processamento passaram por um processo de ranqueamento de

importância e também de seleção dos melhores atributos baseados nesse ranqueamento **(E)**.

Os dados resultantes apenas de **(B)** e os dados de **(E)** foram submetidos aos algoritmos de AM **(D)** que geraram modelos para ambas entradas. Após a construção dos modelos, sua eficácia foi comparada **(F)** e o modelo com melhor desempenho foi escolhido **(G)**.

4.2 Ferramentas utilizadas

Para o desenvolvimento deste projeto foi utilizada a Linguagem de Programação R[1] com a interface RStudio. Essa foi escolhida por se tratar de um ambiente gratuito desenvolvido para computação estatística e gráfica, e oferecer uma grande variedade de ferramentas que trabalha com AM, além de ser compatível com vários Sistemas Operacionais.

O RStudio[2] é uma IDE *(Integrated Development Environment)* desenvolvida para a Linguagem de Programação R que apresenta maior facilidade para trabalhar com recursos como: *console* integrado, editor que destaca a sintaxe e facilita a execução do código, fácil instalação de bibliotecas, ferramenta para visualização de gráficos, etc.

4.3 Bases de Dados

As bases de dados utilizadas durante o desenvolvimento deste trabalho foram disponibilizadas à pesquisa pelo Dr. Fadi Thabtah da Universidade Huddersfield no Reino Unido e estão disponíveis para *download*[3]. Essas contêm dados do teste AQ-10 para adultos, adolescentes e crianças, além de outras características, que estariam relacionadas com o TEA e, por conseguinte, poderiam ter algum tipo de impacto positivo se usadas para o seu diagnóstico [21].

Essas bases são recentes e ainda pouco exploradas em trabalho científicos. Sua primeira versão foi disponibilizada no final de 2017 e a versão utilizada neste trabalho foi do final de 2018. A diferença entre as duas versões é a quantidade de amostras em cada Base de Dados. A maior quantidade de dados possibilita um estudo mais aprofundado e genérico dos dados.

A Figura 9 mostra o número de amostras com e sem TEA e o total em cada Base de Dados.

[1] <https://www.r-project.org/about.html>
[2] <https://www.rstudio.com/products/rstudio/>
[3] <http://fadifayez.com/autism-datasets/>

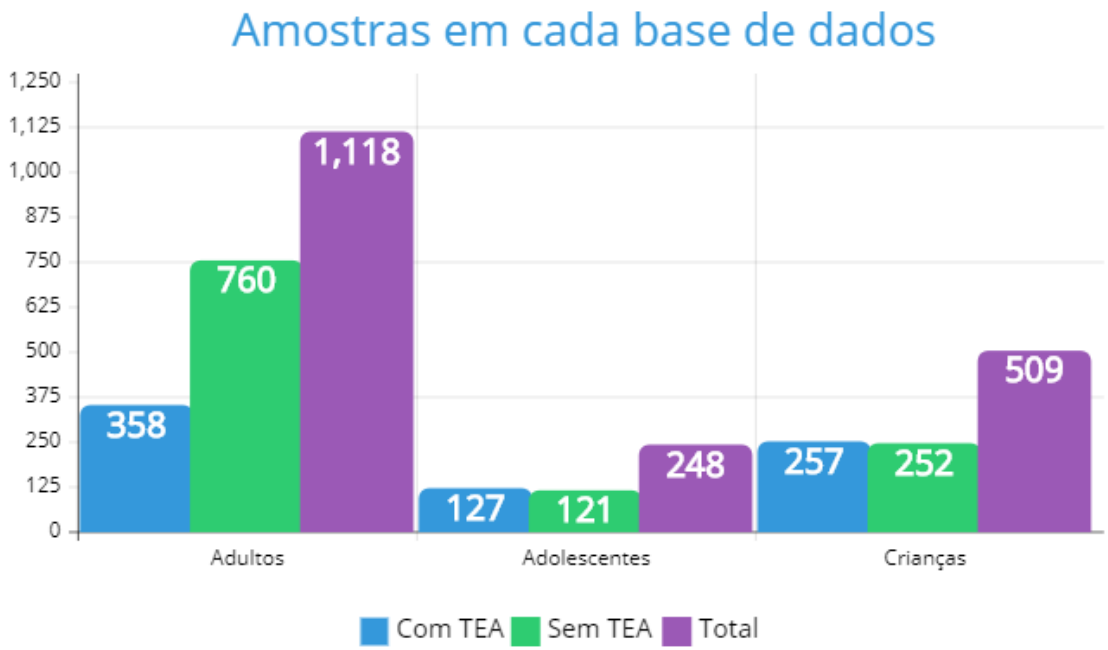


Figura 9 – Número de amostras em cada Base de Dados.

Todos os atributos presentes nas três Bases de Dados estão detalhados na Tabela 4 contendo o nome do atributo e o seu tipo.

Tabela 4 – Atributos presentes nas bases de dados

Atributo	**Tipo**
Caso nº	Inteiro (único)
Idade	Inteiro
Gênero	Texto
Etnia	Texto
Nascido com icterícia?	Booleano (Sim ou Não)
Casos de TEA na família?	Booleano (Sim ou Não)
Quem está completando o teste?	Texto
País de residência	Texto
Idioma	Texto
Usou aplicativo de triagem antes?	Booleano (Sim ou Não)
Tipo de método de triagem	Inteiro (0,1,2,3)
Questões [A1-A10]	Binário (0, 1)
Resultado do teste	Inteiro
Classe	Booleano (Sim ou Não)

A Tabela 5 mostra a equivalência dos atributos A1 - A10, que representam as questões de 1 a 10 nos testes AQ-10 para Adultos, Adolescentes e Crianças. Importante salientar que caso o valor desses atributos seja 1 não significa necessariamente que o

usuário considerou a afirmação verdadeira. O valor 1 ou 0 é atribuído de acordo com as regras do teste AQ-10 para cada questionário.

Um exemplo dos dados presentes nas bases de dados pode ser visto na Tabela 6, que contêm 20 amostras de dados pertencentes à Base de Dados adolescentes antes de iniciar o pré-processamento dos dados

Tabela 5 – Equivalência dos atributos A1-A10 para as questões dos testes.

	AQ-10 Adulto	AQ-10 Adolescente	AQ-10 Criança
A1	Eu noto muitas vezes pequenos ruídos que passam despercebidos às outras pessoas	Ele/a repara sempre em padrões/categorias nas coisas.	Ele/a nota muitas vezes pequenos ruídos que passam despercebidos às outras pessoas
A2	Habitualmente, concentro-me mais na imagem ou situação no seu todo, do que nos seus pequenos detalhes.	Habitualmente, ele/a concentra-se mais na imagem ou situação no seu todo, do que nos seus pequenos detalhes.	Habitualmente, ele/a concentra-se mais na imagem ou situação no seu todo, do que nos seus pequenos detalhes.
A3	Acho fácil realizar mais de uma tarefa ao mesmo tempo.	Quando está num grupo social, ele/a consegue facilmente seguir conversas de várias pessoas.	Quando está num grupo social, ele/a consegue facilmente seguir conversas de várias pessoas.
A4	Em caso de interrupção, eu consigo muito rapidamente voltar ao que estava a fazer.	Em caso de interrupção, ele/a consegue muito rapidamente voltar ao que estava a fazer.	Ele/a consegue facilmente fazer mais do que uma coisa ao mesmo tempo
A5	Eu acho fácil 'ler nas entrelinhas' quando alguém está falando comigo.	Frequentemente, ele/a nota que não sabe como manter uma conversa.	Ele/a não sabe como manter uma conversa com os seus pares.
A6	Eu consigo identificar se alguém, que está me ouvindo, está ficando entediado.	Socialmente, ele/a é bom/boa conversador/a.	Socialmente, ele/a é bom/boa conversador/a.
A7	Durante a leitura de uma história, tenho dificuldades em perceber as intenções e as emoções das personagens.	Quando era mais novo/a, ele/a gostava de brincar a jogos de faz-de-conta com as outras crianças.	Durante a leitura de uma história, ele/a tem dificuldades em perceber as intenções e as emoções das personagens.
A8	Gosto de coletar informações sobre categorias/coisas (tipos de carros, tipos de pássaros, tipos de trains, tipos de plasntas etc).	Ele/a tem dificuldades em imaginar como seria ser outra pessoa.	Na pré-escola, ele/a gostava de brincar a jogos de faz de conta com as outras crianças.
A9	Percebo facilmente o que alguém está a pensar ou a sentir, apenas olhando para a sua cara.	Ele/a acha as situações sociais fáceis.	Ele/a percebe facilmente o que alguém está a pensar ou a sentir, apenas olhando para a sua cara.
A10	Acho difícil perceber as intenções das pessoas.	Ele/a tem dificuldades em fazer novos amigos.	Ele/a tem dificuldades em fazer novos amigos.

Tabela 6 – Exemplo de 20 amostras da Base de Dados Adolescentes.

Case.No	A1	A2	A3	A4	A5	A6	A7	A8	A9	A19	Age	Sex	Ethnicity	Jaundice	Family With ASD	Residence	Used App Before	Score	Screening_Type	Language	User	Class
19	1	1	0	1	0	0	1	1	0	0	15	m	asian	no	no	Indonesia	no	5	12-16 years	english	self	NO
22	0	1	1	1	0	1	1	0	1	0	15	m	middle eastern	no	no	Argentina	no	6	12-16 years	spanish	parent	NO
90	1	1	1	0	1	1	1	1	1	1	15	f	middle eastern	no	yes	Egypt	no	9	12-16 years	english	parent	YES
106	1	1	1	1	1	1	1	1	0	0	16	f	latino	no	no	Austria	no	8	12-16 years	english	health care professional	YES
109	1	1	1	1	1	1	1	1	1	1	15	f	aboriginal	no	no	Barbados	no	10	12-16 years	english	health care professional	YES
117	0	1	1	1	0	1	1	0	1	0	14	m	middle eastern	no	no	Bangladesh	no	6	12-16 years	farsi	parent	NO
119	1	0	0	0	1	0	0	1	0	1	14	f	middle eastern	no	no	Barbados	no	4	12-16 years	arabic	parent	NO
120	1	0	0	0	1	0	0	1	0	1	12	m	middle eastern	no	no	Armenia	no	4	12-16 years	urdu	parent	NO
121	1	0	0	0	1	0	0	1	0	1	15	f	middle eastern	no	no	Argentina	no	4	12-16 years	urdu	parent	NO
122	0	1	1	1	0	1	1	0	1	0	12	f	middle eastern	no	no	Austria	no	6	12-16 years	turkish	parent	NO
123	0	1	1	1	0	1	1	0	1	0	16	f	middle eastern	no	no	Anguilla	yes	6	12-16 years	turkish	parent	NO
127	1	0	1	0	0	0	0	1	0	1	16	m	asian	no	no	India	no	4	12-16 years	english	parent	NO
130	0	0	0	1	1	1	1	1	1	0	12	m	latino	yes	yes	Azerbaijan	no	6	12-16 years	english	relative	NO
131	0	0	0	1	1	1	1	1	1	0	15	m	black	yes	yes	Austria	no	6	12-16 years	swahili	parent	NO
132	1	0	0	0	1	0	0	1	0	1	15	m	aboriginal	yes	yes	Belize	no	4	12-16 years	english	relative	NO
133	0	0	0	0	0	0	0	0	1	1	15	m	hispanic	no	no	Austria	no	2	12-16 years	english	relative	NO
134	0	0	0	0	0	0	0	0	1	1	12	f	middle eastern	no	no	Afghanistan	no	2	12-16 years	arabic	parent	NO
135	1	0	0	0	1	0	0	1	0	1	15	f	middle eastern	no	no	Azerbaijan	no	4	12-16 years	arabic	health care professional	NO
136	0	0	0	0	0	0	0	0	1	1	12	f	middle eastern	no	no	Algeria	no	2	12-16 years	arabic	relative	NO
137	0	0	0	0	0	0	0	0	1	1	14	f	black	no	no	Angola	no	2	12-16 years	arabic	health care professional	NO

4.4 Pré-processamento dos Dados

Após obter as bases de dados, o primeiro passo é realizar a etapa de Pré-Processamento.

De acordo com Han e Kamber [84], bases de dados podem ser elementos bastante complexos por vários fatores como:

- **Tamanho**. O tamanho de uma Base de Dados é um dos fatores a ser levado em conta para a utilização da mesma. Existem bases de dados com tamanhos variados, desde poucos *bytes* até vários *Gigabytes*. Uma Base de Dados com poucos elementos pode não fornecer uma boa visão ou generalização para a solução de um problema, por outro lado, uma Base de Dados muito grande pode fazer com que o tempo demandado para o processo de aprendizagem de máquina seja maior do que o necessário para uma solução ideal;
- **Registos inconsistentes, faltantes e *Outliers***. É muito comum que existam em uma Base de Dados registros inconsistentes, sejam eles causados por erro humano ao montar a Base de Dados ou por problemas em sensores responsáveis por obtê-los. Esses registros podem ter vários formatos como ausência de valor ou valores que não poderiam existir, como uma idade de uma pessoa negativa ou muito elevada;
- **Registros duplicados**. Em alguns tipos de problemas a presença de registros duplicados pode acabar fazendo com que o processo de aprendizagem seja manipulado;
- **Transformação de dados**. Alguns algoritmos de aprendizado de máquina necessitam que os dados estejam em um formato específico como, por exemplo, valores numéricos. Outros algoritmos também podem exigir que os dados estejam escalonados, ou seja, na mesma escala.

Dessa forma o pré-processamento pode ser visto como a etapa responsável por preparar os dados para que possam ser submetidos ao processo de aprendizagem [84].

Nas bases de dados pertencentes a este trabalho, alguns dos atributos são utilizados apenas para um melhor controle dos dados e não apresentam nenhum tipo de relevância no diagnóstico do TEA. São eles:

- **Caso nº**. Trata-se de um identificador de indivíduos que foi utilizado durante o processo de aquisição dos dados. É um valor único e crescente;
- **Quem está completando o teste?** Também não possui nenhum valor para determinar se a pessoa possui ou não TEA. Trata-se apenas de um dado para armazenar quem estaria realizando o teste;

- **País de residência**. Também é outro atributo meramente estatístico usado para saber quantas pessoas de um determinado país realizaram o teste;
- **Idioma**. O AQ-10 está disponível em vários idiomas. Essa variável foi usada para mostrar em qual idioma a pessoa realizou o teste, não tendo também nenhum tipo de relevância para o diagnóstico;
- **Tipo de método de triagem**. Esse campo apenas armazena qual tipo de teste foi realizado, sendo 0 para crianças pequenas (*toddlers*), 1 para crianças (*children*), 2 adolescentes (*adolescents*) e 3 para adultos (*adults*);
- **Se foi usado o aplicativo de triagem antes**: Outra variável meramente estatística, já que o fato de ter usado antes ou não um aplicativo para triagem não é um fator para diagnóstico do TEA;
- **Idade**. A idade é meramente utilizada para selecionar o tipo correto de teste a ser aplicado para a pessoa.

Assim, além das questões relativas ao Teste AQ-10, outras características que poderiam estar relacionadas ao diagnóstico de TEA e, portanto, melhorar o desempenho de algoritmos de AM, são: sexo, etnia, nascido com icterícia (*jaundice*) e membro da família com TEA (*Family With ASD*).

O atributo Resultado (*Score*) refere-se à pontuação tradicional gerada após a correção do teste AQ-10. Com esse valor é possível determinar se a pessoa possui ou não TEA. Uma vez que é empregado o AM, esse valor torna-se desnecessário, pois o objetivo de usar o AM é exatamente substituir essa pontuação, usada anteriormente pelo teste, por um modelo de aprendizagem que será gerado pelos demais atributos e treinado com a classe que indica se o indivíduo possui ou não TEA.

O primeiro passo do pré-processamento foi fazer a remoção dos atributos de controle e do *Score*. Após isso, foi realizado um processo de verificação da integridade dos dados, o qual buscou garantir que os valores presentes na Base de Dados estejam de acordo com os especificados. Todos os registros foram verificados e, no caso dessa Base de Dados, não existiam valores faltantes ou inconsistentes.

4.5 Ranqueamento de Importância

O *Ranqueamento de Importância* pode ser descrito como um dos métodos de seleção de atributos. Nesse AM pode ser utilizada para auxiliar na escolha dos melhores atributos de uma Base de Dados [85]. Em outras palavras, esse processo busca remover o maior número possível de atributos que tenham uma baixa influência e que podem, em alguns casos, diminuir consideravelmente o desempenho do modelo.

Existem diversas maneiras e algoritmos capazes de estimar a importância da variável em uma Base de Dados. Uma dessas maneiras é por meio do cálculo do Ganho de Informação, ou em inglês *Information Gain* (IG), mecanismo esse utilizado por árvores de decisão [44]. O processo de construção de uma árvore vai da raiz até as folhas, com a utilização do IG na construção da árvore.

Quanto mais próximo o nó está da raiz, maior foi o seu IG calculado. Exatamente por essa razão, existe um sistema de poda em alguns modelos de árvore para retirar nós que vão muito fundo em sua construção e não trazem ganhos reais ao modelo.

Neste trabalho a RF foi utilizada para efetuar o processo de ranqueamento de importância, e a mesma utiliza a métrica de *Mean Decrease Gini*. Os valores resultantes do ranqueamento podem ser vistos na Tabela 7. Para uma melhor visualização da variação da *Mean Decrease Gini*, uma representação gráfica é mostrada na Figura 10.

Tabela 7 – Ranqueamento de importância dos atributos via RF

Atributo	Adulto	Adolescente	Criança
A1	22.166235	4.709013	17.237419
A2	18.346285	6.039242	7.719202
A3	36.201569	17.137619	15.500225
A4	38.465004	13.989011	47.037182
A5	68.412177	13.455042	19.639696
A6	78.111460	17.967928	21.681128
A7	27.178058	6.958421	14.983677
A8	16.347136	4.621953	19.818508
A9	79.709742	6.367678	22.634410
A10	22.972870	8.094193	21.679284
Sex	5.530077	2.262165	3.695835
Ethnicity	24.598594	10.074884	14.848777
Jaundice	2.915085	1.527666	3.571882
Family_ASD	4.671645	2.038788	2.506743

Para a Base de Dados Adultos, o ranqueamento mostrou uma maior variação do IG entre os atributos em relação as outras bases de dados. Isso se dá devido ao maior número de amostras na Base de Dados. Na Base Adolescentes, que possui o menor número de amostras, a variação do IG foi mais sutil, embora ainda releve que existam atributos com um grau de importância superior aos demais.

Após o ranqueamento das três bases de dados, fica claro que, com exceção do atributo *Ethnicity*, as questões relativas ao teste AQ-10 apresentaram um IG bem superior aos demais atributos.

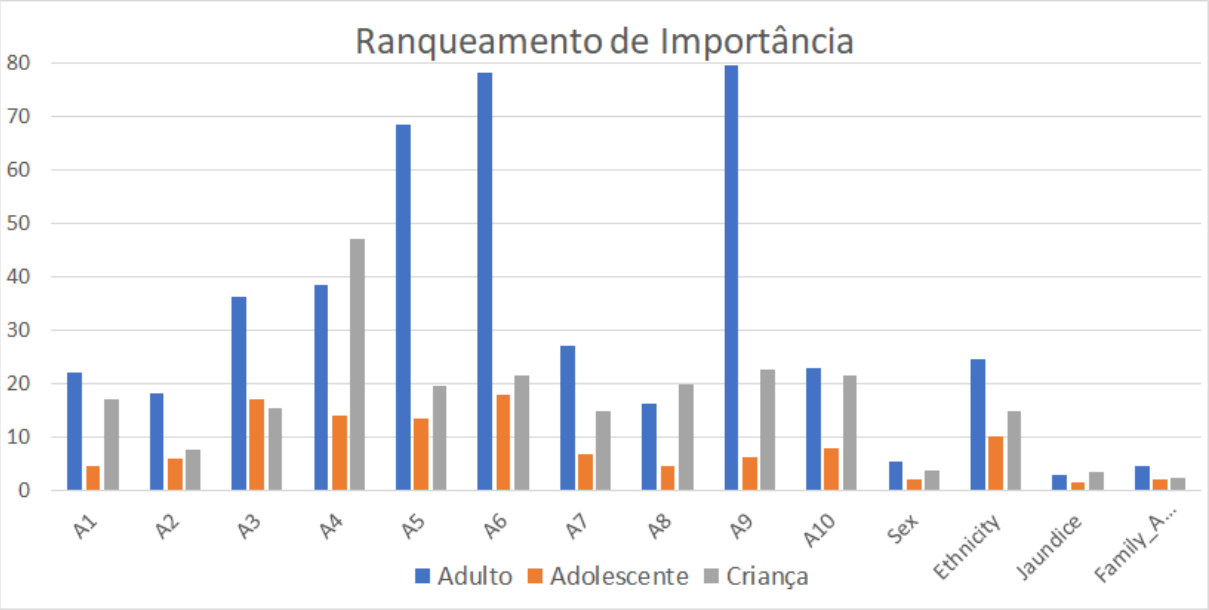


Figura 10 – Variação da *Mean Decrease Gini* gerada pela RF

4.6 Aplicação de AM

Os algoritmos de AM k-NN, J48, RF e SVM (linear, polinomial, radial e sigmoidal), descritos em 2.4.3, foram utilizados para a geração dos modelos de aprendizagem. Eles foram implementados com o pacote *CARET*[4] *(short for Classification And REgression Training)* e suas dependências, o qual foi escolhido por possibilitar a aplicação de vários tipos de algoritmos de AM, além de ser bastante utilizado para trabalhos nessa área.

Contudo, antes que os dados fossem submetidos aos algoritmos, eles passaram pelo processo de *Cross Validation* (CV), ou em português, *Validação Cruzada*.

A *Validação Cruzada* é uma técnica matemática que tem como objetivo evitar que a disposição dos dados na base possa estar arranjada de maneira a beneficiar alguma técnica específica de AM. Ela funciona dividindo a Base de Dados em K subconjuntos exclusivos e de tamanhos iguais, chamados de *folds* ou folhas, então um desses é utilizado para teste, enquanto os k-1 restantes são usados para treinar o modelo. Dessa forma, não é possível colocar os dados em uma disposição específica na Base de Dados [86]. Um exemplo de Validação Cruzada com 10 folhas pode ser visto na Figura 11.

Foi aplicada a técnica de *Repeated Cross Validation*, que possibilitou repetir o processo de CV quantas vezes forrem solicitadas. Dessa forma, cada vez que a função for repetida, irá gerar novos conjuntos de teste e de treino, garantindo assim que não exista nenhum tipo de arranjo entre as amostras das bases de dados [87]. Quando o processo de CV é repetido, oferece uma taxa de erro menor em relação à CV simples, devendo ela deve ser utilizada com cuidado, uma vez que o processo de se repetir a divisão de uma

[4] <https://cran.r-project.org/web/packages/caret/>

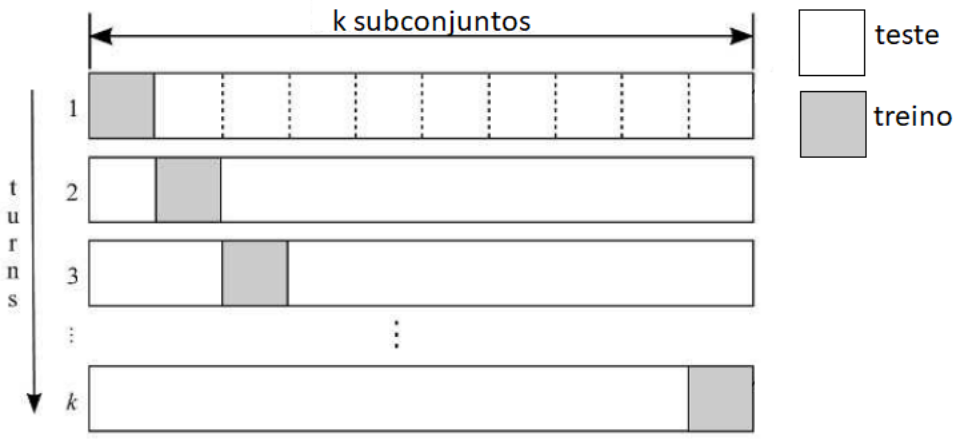


Figura 11 – Exemplo de Validação Cruzada com 10 folhas

Base de Dados pode ser muito custosa caso a mesma seja muito grande.

Os algoritmos foram aplicados utilizando a *Repeated Cross Validation* de 10 vezes com 10 *folds*. Como as bases de dados não são muito grandes, o processo de CV ser repetido por 10 vezes não torna a execução muito custosa.

5 RESULTADOS E DISCUSSÕES

Este capítulo expõe os resultados dos experimentos realizados e uma discussão sobre os mesmos. O trabalho utilizou três bases de dados que são a de adultos, adolescentes e de crianças, as quais continham respostas do Teste AQ-10, que foram submetidos a diferentes algoritmos de AM.

Com base no ranqueamento de importância dos atributos, foram elaborados três cenários para a aplicação de AM, que foram usados em cada uma das bases de dados, sendo eles:

- utilização de todos os atributos;
- uso apenas das questões do teste AQ-10;
- utilização das questões TOP 9 até TOP 5, de acordo com o resultado do ranqueamento.

As variáveis que representam do TOP 9 ao TOP 5 de cada base de dados estão detalhadas na Tabela 8.

Tabela 8 – Atributos pertencentes aos TOPs atributos.

	ADULTOS	ADOLESCENTES	CRIANÇAS
TOP 9	A9 - A6 - A5 - A4 -A3 A7 - A10 - A1 - A2	A6 - A3 - A4 - A5 - A10 A7 - A9 - A2 - A8	A4 - A9 - A6 - A10 - A8 A5 - A1 - A3 - A7
TOP 8	A9 - A6 - A5 - A4 -A3 A7 - A10 - A1	A6 - A3 - A4 - A5 - A10 A7 - A9 - A2	A4 - A9 - A6 - A10 - A8 A5 - A1 - A3
TOP 7	A9 - A6 - A5 - A4 -A3 A7 - A10	A6 - A3 - A4 - A5 - A10 A7 - A9	A4 - A9 - A6 - A10 - A8 A5 - A1
TOP 6	A9 - A6 - A5 - A4 -A3 A7	A6 - A3 - A4 - A5 - A10 A7	A4 - A9 - A6 - A10 - A8 A5
TOP 5	A9 - A6 - A5 - A4 -A3	A6 - A3 - A4 - A5 - A10	A4 - A9 - A6 - A10 - A8

Cada uma das bases de dados foi submetida a esses cenários e a métrica escolhida para a avaliação da performance foi a Acurácia, uma vez que as amostras nas bases de dados não são desbalanceadas.

Para a exibição dos resultados foi escolhido o gráfico em forma de radar, sendo as linhas coloridas os cenários elaborados e os vértices do gráfico os algoritmos de AM. Dessa forma, quanto mais próxima do vértice estiver uma das linhas, significa que o modelo gerado por aquele algoritmo obteve um melhor resultado para o mesmo. Os gráficos estão em uma escala que vai de 0.7 até 1. Essa escala foi adotada para uma melhor verificação da variação da performance entre os algoritmos.

5.1 Análise dos Componentes Principais

A PCA foi um dos primeiros experimentos realizados nas Bases de Dados e por meio dela é possível observar algumas características dos problemas, entre elas o grau.

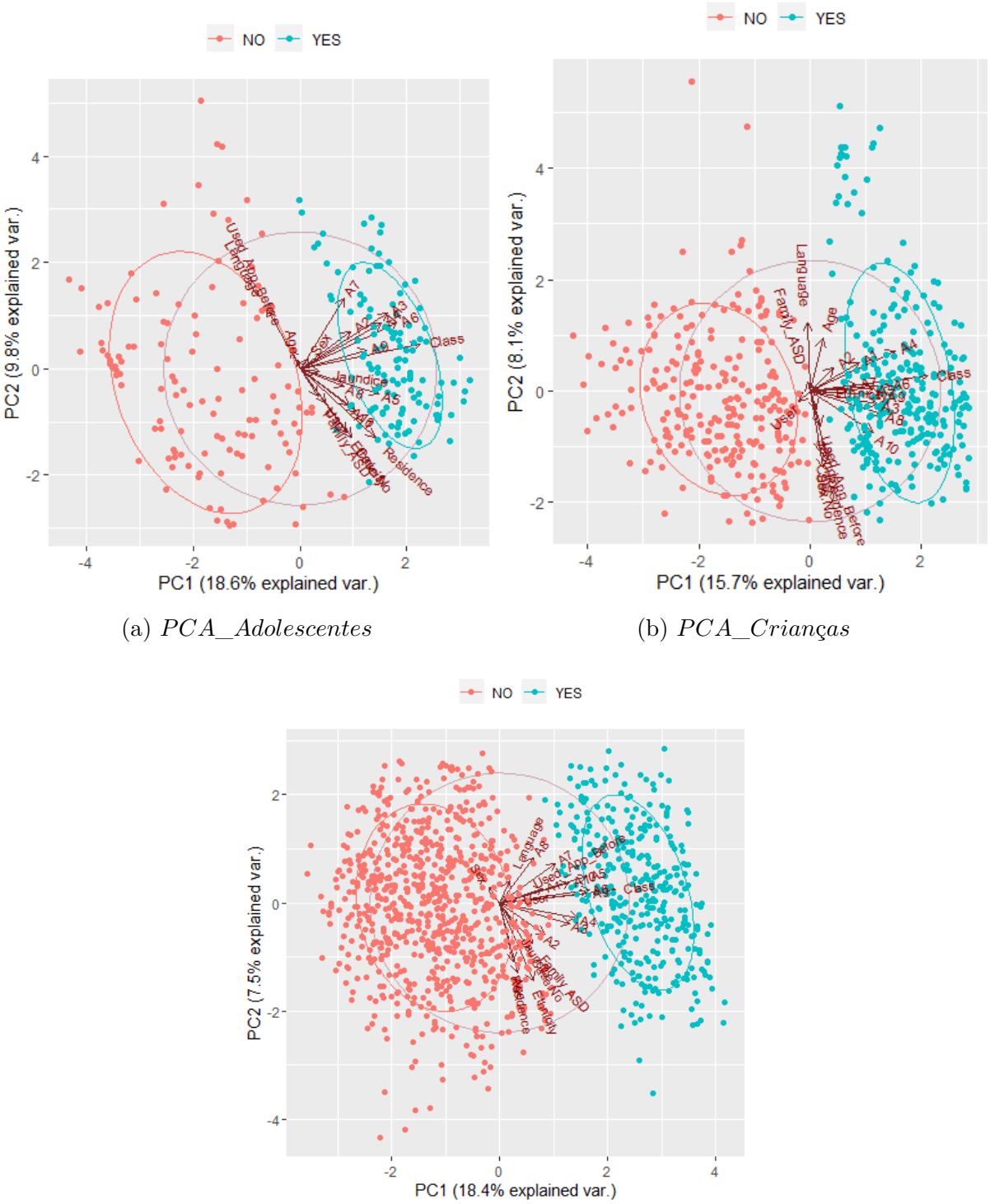
Existem diversos pacotes disponíveis para representação gráfica da PCA em *R* e neste trabalho o escolhido foi o *ggbiplot*[1], que foi desenvolvido especialmente para representação gráfica apresentando vários recursos para representação da PCA.

Na representação adotada, os eixos X e Y representam respectivamente os PC1 e PC2 de cada base de dados. Cada um desses possui um valor que representa o quanto esse PC contribui para a explicação do problema. Como esse só pode ter dois tipos de classificação, com ou sem TEA, amostras representadas em azul e vermelho representam respectivamente os indivíduos que foram diagnosticados com TEA e os que não foram. Caso existisse algum outro tipo possível de classificação, o mesmo seria representado no gráfico com outra cor.

Nessa biblioteca cada vetor é relativo a um dos atributos da base de dados que foram utilizados na PCA. O tamanho e o sentido do vetor representam o valor de sua correlação com as demais variáveis. Vetores que apontam para lados diferentes do gráfico apontam atributos que têm uma correlação negativa entre si, enquanto os que apontam para o mesmo lado apresentam uma correlação positiva.

Nas bases é possível observar que existe uma boa separação entre as classes. A partir dessa constatação, foi demonstrada a natureza linear do problema de classificação. Esse cenário era esperado uma vez que os testes AQ-10 seguem uma lógica baseada na pontuação para classificar os indivíduos. Contudo, foi possível analisar por meio dos vetores que os atributos não pertencentes ao AQ-10 estavam com uma correlação negativa em comparação aos demais atributos. A Figura 12 mostra a aplicação do algoritmo nas bases de dados.

[1] <https://github.com/vqv/ggbiplot>



(a) *PCA_Adolescentes*

(b) *PCA_Crianças*

(c) *PCA_Adultos*

Figura 12 – PCA das Bases de Dados

Esses atributos extras poderiam melhorar o desempenho dos algoritmos de AM para classificação [34]. Contudo, o resultado da PCA mostra que esses atributos poderiam na verdade atrapalhar o desempenho e não colaborar.

5.2 Base de Dados Adolescentes

A primeira é a Base de Dados Adolescentes. Ela contém no total 248 amostras, sendo 127 delas com TEA e 121 sem. É a menor dentre as três bases de dados selecionadas.

A Figura 13 mostra a Acurácia dos modelos gerados pelos algoritmos de AM quando aplicado na Base de Dados Adolescentes.

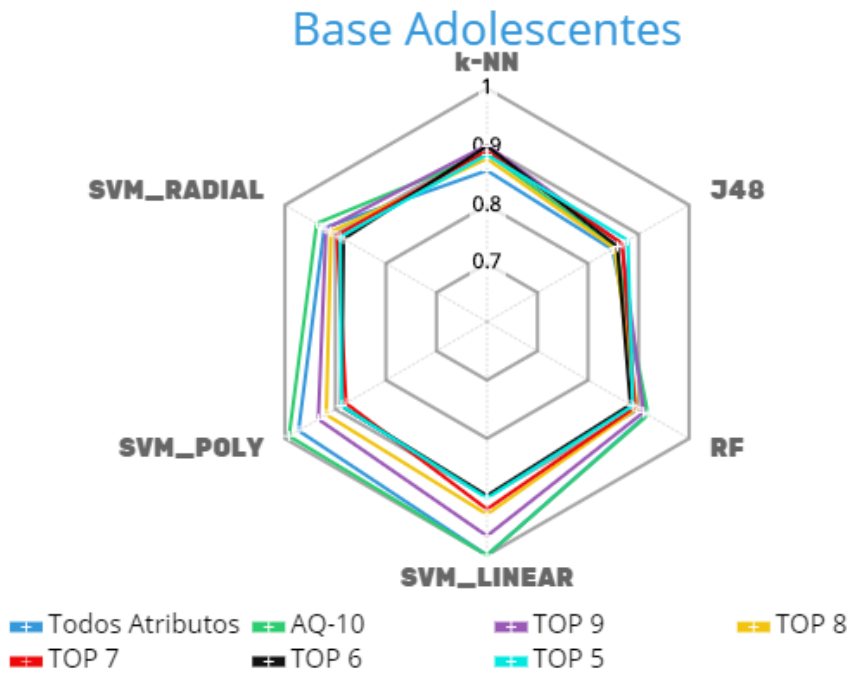


Figura 13 – Resultados dos Experimentos na Base de Dados Adolescentes.

Para essa Base de Dados o melhor desempenho do modelo se deu quando foram utilizadas apenas as questões referentes ao teste e foram apresentados pelas SVMs, sendo que a Linear obteve os melhores resultados chegando a uma Acurácia de 100%. Quando o número de questões foi reduzido para apenas 5, a SVM linear continuou a oferecer o melhor modelo preditivo com 0.9 de Acurácia.

5.3 Base de Dados Adultos

A Base de Dados Adultos é a maior em número de amostras, contendo 1118, e também a mais desbalanceada em questão à quantidade de indivíduos com e sem TEA (358 e 760, respectivamente). Contudo, essa diferença entre as classes não é o bastante para que a Base de Dados seja considerada desbalanceada, logo a medida de Acurácia ainda pode ser empregada.

A Figura 14 mostra a Acurácia dos modelos preditivos gerados pelos algoritmos de AM que foram aplicados à Base de Dados Adultos.

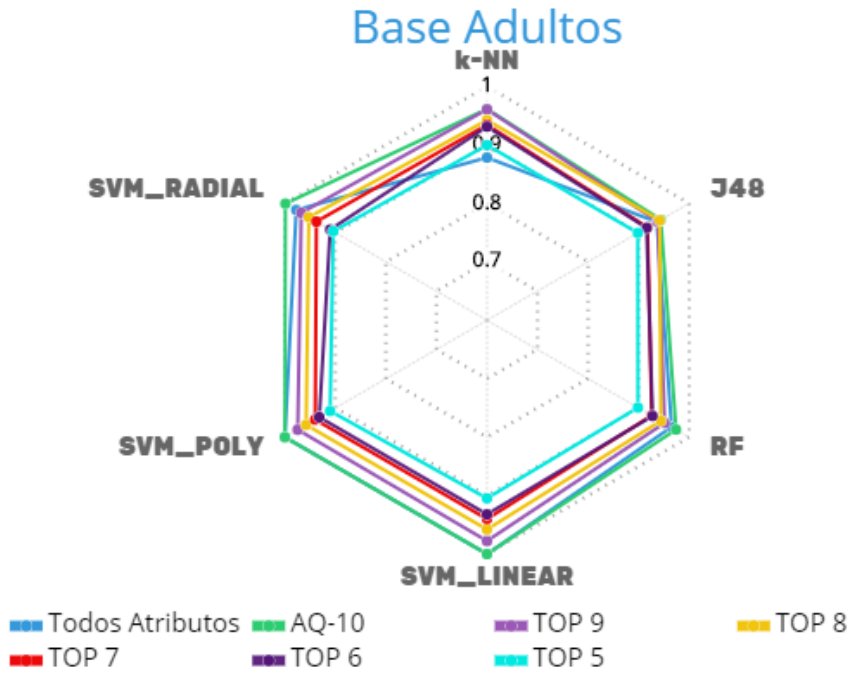


Figura 14 – Resultados dos Experimentos na Base de Dados Adultos.

Devido ao maior número de amostras na Base de Dados, os modelos construídos pelos algoritmos em todos os cenários geraram resultados mais próximos entre si. Todos os modelos mantiveram a Acurácia acima de 0.9 em quase todos os cenários, com exceção do cenário TOP 5, no qual k-NN, J48 e RF que ficaram com 0.89.

Assim como na Base de Dados anterior, ao utilizar apenas as questões do teste, os modelos apresentaram resultados superiores comparando com o uso da Base de Dados completa. As SVMs Linear e Polykernel mantiveram a acurácia em 100% utilizando todos os atributos e quando apenas as questões foram selecionadas. Contudo, os outros modelos tiveram um ganho de performance usando apenas as questões, o que demonstrou novamente a não relevância desses atributos extras.

Quando o número de atributos foi reduzido a apenas 5 questões, o melhor modelo foi o gerado pela SVM Polykernel chegando a 0.91 de Acurácia.

5.4 Base de Dados Crianças

A Base de Dados Crianças contém 509 amostras ficando em um patamar intermediário em relação às outras duas bases. É composta por 257 crianças com TEA e 252 sem.

A Figura 15 mostra a Acurácia dos modelos preditivos gerados pelos algoritmos de AM aplicados à Base de Dados Crianças.

O melhor cenário em termos de Acurácia, assim como nas bases anteriores, foi

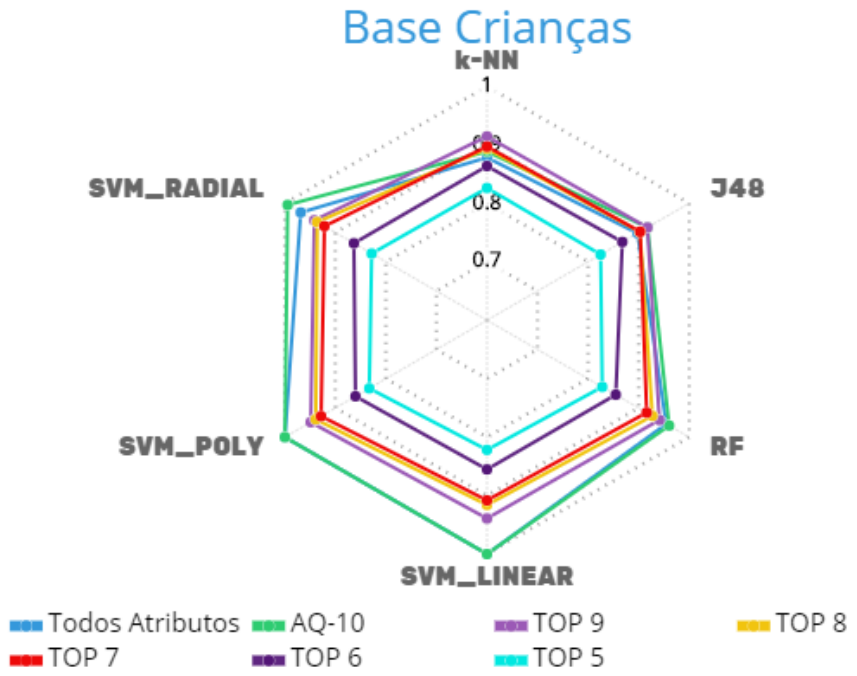


Figura 15 – Resultados dos experimentos na Base de Dados Crianças.

a utilização apenas das 10 questões relativas ao AQ-10 nos modelos de aprendizagem gerados pelas SVMs. Essa diferença fica mais evidente quando se compara o desempenho da SVM Radial para todos os atributos (em azul) e apenas para as 10 questões (em verde). O aumento no número de atributos novamente atrapalhou o processo de aprendizagem dos modelos.

Diferente das bases anteriores, a Base de Dados Crianças mostrou uma maior queda da Acurácia para cada questão removida, chegando a ficar próxima a 0.8 nos cenários de TOP 6 e TOP 5 em todos os modelos gerados. Nesse caso, a fim de manter a Taxa de Acurácia em torno de 0.9, o menor número possível de questões utilizadas seria 7, ou seja, o TOP 7.

Para o TOP 7 as SVMs Polykernel e Radial geraram o melhor modelo de aprendizagem, com 0.9283 e 0.9218, respectivamente.

5.4.1 Comparação de Performance

Os valores detalhados dos experimentos realizados neste trabalho podem ser vistos na Tabela 9. Ela mostra o desempenho de cada modelo de aprendizagem para cada cenário proposto e para cada Base de Dados.

Com os resultados pode-se observar que os demais atributos mencionados por Thabtah [21] não mostraram nenhum tipo de ganho quando se utilizou AM nessa Base de Dados.

Tabela 9 – Resultados detalhados dos modelos e cenários

	MODELOS	TODOS OS ATRIBUTOS	TODAS AQ-10	TOP 9	TOP 8	TOP 7	TOP 6	TOP 5
Adolescentes	K-NN	0.8591	0.8887	0.9025	0.8795	0.8928	0.9008	0.8863
	J48	0.8480	0.8545	0.8670	0.8521	0.8674	0.8582	0.8788
	RF	0.9174	0.9188	0.9100	0.8963	0.8930	0.8841	0.8885
	SVM_LINEAR	1	1	0.9669	0.9299	0.9210	0.8979	0.9020
	SVM_POLY	0.9730	0.9918	0.9339	0.9181	0.8791	0.8884	0.8878
	SVM_RADIAL	0.9220	0.9371	0.9186	0.9088	0.8964	0.8840	0.8906
Adultos	K-NN	0.9496	0.9614	0.9607	0.9423	0.9343	0.9312	0.8997
	J48	0.9362	0.9440	0.9381	0.9419	0.9177	0.9160	0.8987
	RF	0.9654	0.9742	0.9515	0.9442	0.9261	0.9276	0.8989
	SVM_LINEAR	1	1	0.9776	0.9579	0.9400	0.9320	0.9047
	SVM_POLY	1	1	0.9746	0.9580	0.9400	0.9319	0.9105
	SVM_RADIAL	0.9775	0.9990	0.9676	0.9526	0.9376	0.9103	0.9043
Crianças	K-NN	0.8783	0.8886	0.9144	0.8935	0.8972	0.8638	0.8263
	J48	0.8978	0.9176	0.9175	0.9027	0.9031	0.8679	0.8249
	RF	0.9549	0.9605	0.9412	0.9274	0.9161	0.8549	0.8281
	SVM_LINEAR	1	1	0.9391	0.9159	0.9081	0.8552	0.8218
	SVM_POLY	1	1	0.9489	0.9389	0.9283	0.8603	0.8330
	SVM_RADIAL	0.9687	0.9946	0.9418	0.9361	0.9218	0.8634	0.8282

Foi possível reduzir consideravelmente o número de atributos utilizados de 10 para 5 no caso de adultos e adolescentes e para 7 no caso de crianças, mantendo a acurácia acima de 0.9.

O k-NN, apesar de ser o mais simples dos algoritmos baseados em instâncias, entregou uma boa performance nos cenários propostos, mantendo uma acurácia sempre maior que 0.8 e, em alguns casos, chegando a obter uma performance superior a J48 e a RF.

Foi possível também notar que, conforme o número de atributos foi reduzido, mais os resultados da acurácia entre J48 e a RF se aproximaram. Isso era esperado, uma vez que com o mecanismo de *bagging* da RF, a redução na quantidade de atributos também afeta a geração dos novos subconjuntos pelo *bootstraping*. A menor quantidade de atributos acabou tornando os resultados mais parecidos. Dessa forma, a RF foi perdendo toda a efetividade de seu mecanismo principal.

Inicialmente, devido à natureza linear do problema, a SVM Linear obteve uma performance perfeita, mesmo com os atributos extras. Contudo, quando o número de atributos foi reduzindo, a natureza do problema começou a mudar, o que afetou a performance e tornou as SVMs Polykernel e Radial mais precisas.

5.5 Questões Selecionadas

A Tabela 10 mostra as questões necessárias para o modelo de AM proposto neste trabalho.

Tabela 10 – Questões necessárias para o modelo de aprendizagem

Base de dados	Questões Selecionadas
Adultos	Percebo facilmente o que alguém está a pensar ou a sentir, apenas olhando para a sua cara.
	Eu consigo identificar se alguém, que está me ouvindo, está ficando entediado.
	Eu acho fácil 'ler nas entrelinhas' quando alguém está falando comigo.
	Em caso de interrupção, eu consigo muito rapidamente voltar ao que estava a fazer.
	Acho fácil realizar mais de uma tarefa ao mesmo tempo
Adolescentes	Socialmente, ele(a) é bom/boa conversador(a).
	Quando está num grupo social, ele(a) consegue facilmente seguir conversas de várias pessoas.
	Em caso de interrupção, ele(a) consegue muito rapidamente voltar ao que estava a fazer.
	Frequentemente, ele(a) nota que não sabe como manter uma conversa.
	Ele(a) tem dificuldades em fazer novos amigos.
Crianças	Ele(a) consegue facilmente fazer mais do que uma coisa ao mesmo tempo.
	Ele(a) percebe facilmente o que alguém está a pensar ou a sentir, apenas olhando para a sua cara.
	Socialmente, ele(a) é bom(boa) conversador(a).
	Ele(a) tem dificuldades em fazer novos amigos.
	Na pré-escola, ele(a) gostava de brincar a jogos de faz de conta com as outras crianças.
	Ele(a) não sabe como manter uma conversa com os seus pares
	Ele(a) nota muitas vezes pequenos ruídos que passam despercebidos às outras pessoas.

Essas questões mostraram-se mais relevantes para os algoritmos de AM, uma vez que as mesmas foram selecionadas pelo ranqueamento de importância é natural deduzir que elas têm um papel de destaque se comparadas às demais para o processo diagnóstico. Portanto, esses padrões comportamentais podem servir como uma contribuição para futuros estudantes da área médica ou outros profissionais, como psicólogos, por exemplo.

5.6 Modelo Escolhido

Com base nos resultados descritos ao longo deste capítulo, foi possível elencar a SVM Linear como o algoritmo de AM que gerou o melhor modelo de aprendizagem nas três Bases de Dados.

Para o modelo obtido pela aplicação da SVM Linear na Base de Dados de adultos, composta por 1118 amostras e utilizando 5 preditores (variáveis), foi obtida uma acurácia de 0.9047. Na base de adolescentes, composta por 248 amostras e também utilizando 5 preditores, foi alcançada uma acurácia de 0.9020. Já na de crianças composta por 509 amostras foi necessário utilizar 7 preditores para se manter a acurácia acima de 0.9, a mesma ficou em 0.9081.

A fim de avaliar melhor a acurácia, é possível observar a matriz de confusão gerada pela média de desempenho do modelo. A Matriz de Confusão [35] mencionada na Seção 2.2 e ilustrada pela Figura 1, mostra o número de VP, VN, FP e FN do modelo de aprendizado. Ela funciona comparando o valor esperado para a classificação de cada instância com o valor predito para saber se ele foi ou não corretamente classificado [35].

A Tabela 11 mostra a Matriz de Confusão obtida pela SVM Linear nos cenários definidos como ideal. Nela a porcentagem de amostras previstas corretamente aparece destacada em verde e incorretas em vermelho. Importante salientar que como os modelos foram construídos com Validação Cruzada, os valores da matriz de confusão aparecem em porcentagem, pois a mesma é gerada pela média dos valores de todas as matrizes de confusão que foram obtidas durante o processo de treinamento e teste. Como a Validação Cruzada foi utilizada com $K = 10$ foram construídas dez matrizes de confusão, uma para cada iteração do algoritmo.

Tabela 11 – Comparação das Matrizes de Confusão dos modelos escolhidos.

		Sem TEA	Com TEA
Adultos	Classificado como sem TEA	63.6	5.2
	Classificado como com TEA	4.3	26.8
Adolescentes	Classificado como sem TEA	43.4	4.4
	Classificado como com TEA	5.4	46.8
Crianças	Classificado como sem TEA	45.4	5.1
	Classificado como com TEA	4.1	45.4

Com base nos resultados da Tabela 11 é possível observar que o erro dos modelos, tanto para FP como para FN fica em torno de 5%, confirmando a acurácia esperada do modelo que era de 90%. Contudo, os valores de FP e FN são próximos, o que indica que o modelo não tem uma dificuldade específica em classificar indivíduos com ou sem TEA. Caso esse valor fosse muito discrepante, como por exemplo, 1% de FP e 9% de FN, poderia

indicar que o modelo estaria mais propenso a classificar erroneamente um indivíduo sem TEA como um com diagnóstico, o que demonstraria um possível problema no modelo.

6 CONCLUSÕES

O diagnóstico do Transtorno do Espectro Autista pode ser muito complexo, mesmo com o crescente avanço da medicina e da tecnologia. Ainda hoje os profissionais da saúde enfrentam dificuldades para diagnosticar doenças no espectro autista e outras que são realizadas por análises comportamentais.

Durante o desenvolvimento deste trabalho ficou evidente a necessidade de estudar e desenvolver novos modelos computacionais que possam auxiliar e ajudar a diagnosticar um distúrbio tão comum, porém de difícil identificação como é o caso do TEA, ainda mais em um cenário que nem todos os profissionais da saúde têm o conhecimento e/ou recursos necessários para realizar esse diagnóstico, não contando com qualquer tipo de apoio.

Assim sendo, o problema do diagnóstico do TEA foi abordado utilizando técnicas de Aprendizado de Máquina para classificação de indivíduos com ou sem TEA. Para tal, foi proposto um modelo de AM utilizando três Bases de Dados compostas por 23 atributos formados por questões de escalas diagnósticas (AQ-10) em conjunto com outras características.

A respeito do Teste AQ-10 foi possível identificar que as questões possuem um grau de importância diferente entre elas, utilizando a *Random Forest* para o ranqueamento da importância das características. Também foi possível comparar o desempenho do teste com as perguntas mais importantes para manter uma acurácia acima de 0.9 com cinco das 10 questões nas bases de Dados adultos e adolescentes. Na Base de Dados crianças para continuar com performance acima dos 0.9 foi necessário manter 7 questões.

Além disso, o grau de acerto do modelo pode também sugerir que o teste apresenta questões desnecessárias para o diagnóstico do TEA, uma vez que utilizando apenas metade dos atributos em duas das três bases foi possível manter uma acurácia acima de 90%. Na Base de Dados de Crianças, os 90% só foram obtidos utilizando até o TOP 7 e no TOP 5 o valor ficava em cerca de 80%, o que não é uma performance tão ruim, embora seja bem inferior ao rendimento das demais bases de dados.

Com a realização dos experimentos no decorrer deste trabalho, ficou evidente que as características físicas e sociais adicionadas por Thabtah [21] não ofereceram nenhum ganho aos modelos gerados; muito pelo contrário, algumas delas foram influenciadas negativamente por esses dados. Uma vez que o objetivo deste trabalho foi utilizar apenas os atributos mais relevantes para a classificação dos dados, os irrelevantes foram descartados.

6.1 Trabalhos futuros

Como trabalhos futuros, além de testar outros algoritmos de AM, seria de suma importância estudar a viabilidade de utilizar técnicas de criação de amostras para as Bases de Dados, em especial para a de Adolescentes que não apresentaram performance satisfatória em relação às demais analisadas.

Além disso, seria interessante comparar o desempenho dos cenários propostos com outros algoritmos de ranqueamento de importância de atributos para verificar se uma mudança na seleção dos atributos causaria uma grande variação no desempenho dos modelos gerados.

Também poderia ser realizada uma análise nas amostras classificadas de maneira errada (*miss classification analysis*) para verificar qual o padrão dessas amostras classificadas incorretamente.

REFERÊNCIAS

[1] GOMES, P. et al. Autism in Brazil: a systematic review of family challenges and coping strategies. *Jornal de pediatria*, SciELO Brasil, v. 91, n. 2, p. 111–121, 2015.

[2] TOWLE, P. O.; PATRICK, P. A. Autism spectrum disorder screening instruments for very young children: A systematic review. *Autism research and treatment*, Hindawi Publishing Corporation, v. 2016, 2016.

[3] ALENCAR, C. N. et al. Rel085-cartilha educativa como importante ferramenta para a detecção precoce de sinais clínicos de autismo. *Anais do IV Congresso de Educação em Saúde da Amazônia*, 2015.

[4] LAMPREIA, A. R. S. *Perceções parentais sobre a perturbação do espetro do autismo: processo de diagnóstico, interferência e recursos*. Dissertação (Mestrado) — Faculdade de Psicologia, Universidade de Lisboa, 2015.

[5] CRANE, L. et al. Experiences of autism diagnosis: A survey of over 1000 parents in the united kingdom. *Autism*, SAGE Publications Sage UK: London, England, v. 20, n. 2, p. 153–162, 2016.

[6] SILVA, A.; GAIATO, M. B.; REVELES, L. T. *Mundo Singular: entenda o autismo*. Rio de Janeiro: Fontana, 2012.

[7] BRITO, M. M. V. A. *A contribuição do PECS no desenvolvimento da comunicação de uma aluna com perturbações do espectro do autismo*. Dissertação (Mestrado) — Universidade de Trás-os-Montes e Alto Douro, Mestrado em Ciências da Educação, 2015.

[8] FERREIRA, R. d. S. Autism testing: Uma ferramenta móvel no auxílio ao pré-diagnóstico do autismo. In: *Anais do XXII Conferência Internacional sobre Informática na Educação*. Fortaleza, Ceará - Brasil: Nuevas Ideas en Informática Educativa, 2010. v. 13, p. 178–187.

[9] WIGGINS, L. D. et al. Using standardized diagnostic instruments to classify children with autism in the study to explore early development. *Journal of autism and developmental disorders*, Springer, v. 45, n. 5, p. 1271–1280, 2015.

[10] ZANON, R. B.; BACKES, B.; BOSA, C. A. Identificação dos primeiros sintomas do autismo pelos pais. *Psicologia: Teoria e Pesquisa*, SciELO Brasil, v. 30, n. 1, p. 25–33, 2014.

[11] RELLINI, E. et al. Childhood Autism Rating Scale (CARS) and Autism Behavior Checklist (ABC) correspondence and conflicts with DSM-IV criteria in diagnosis of autism. *Journal of autism and developmental disorders*, Springer, v. 34, n. 6, p. 703–708, 2004.

[12] SCHOPLER, E. et al. Toward objective classification of childhood autism: Childhood Autism Rating Scale (CARS). *Journal of autism and developmental disorders*, Springer, v. 10, n. 1, p. 91–103, 1980.

[13] KLEINMAN, J. M. et al. The modified checklist for autism in toddlers: a follow-up study investigating the early detection of autism spectrum disorders. *Journal of autism and developmental disorders*, Springer, v. 38, n. 5, p. 827–839, 2008.

[14] BARON-COHEN, S. et al. The Autism-spectrum Quotient (AQ): Evidence from Asperger Syndrome/High-Functioning Autism, Malesand Females, Scientists and Mathematicians. *Journal of autism and developmental disorders*, Springer, v. 31, n. 1, p. 5–17, 2001.

[15] PENTEADO, F. A. O. et al. Software para auxílio ao diagnóstico de autismo. In: *Anais do VIII Congresso de extensão universitária da UNESP*. São Paulo: Universidade Estadual Paulista Júlio de Mesquita Filho, 2015. p. 1–4.

[16] BONE, D. et al. Applying machine learning to facilitate autism diagnostics: pitfalls and promises. *Journal of autism and developmental disorders*, Springer, v. 45, n. 5, p. 1121–1136, 2015.

[17] WALL, D. et al. Use of machine learning to shorten observation-based screening and diagnosis of autism. *Translational psychiatry*, Nature Publishing Group, v. 2, n. 4, p. e100, 2012.

[18] XUE, M.; ZHU, C. A study and application on machine learning of artificial intellligence. In: *International Joint Conference on Artificial Intelligence*. Pasadena, California (USA): IEEE, 2009. p. 272–274.

[19] WILSON, R.; OBIMBO, C. Self-organizing feature maps for user-to-root and remote-to-local network intrusion detection on the kdd cup 1999 dataset. In: IEEE. *2011 World Congress on Internet Security*. Londron, 2011. p. 42–47.

[20] ISHAK, W. H. W.; SIRAJ, F. Artificial intelligence in medical application: An exploration. *Health Informatics Europe Journal*, v. 16, 2002.

[21] THABTAH, F. Autism spectrum disorder screening: Machine learning adaptation and dsm-5 fulfillment. In: *Proceedings of the 1st International Conference on Medical and Health Informatics 2017*. New York, NY, USA: ACM, 2017. p. 1–6. ISBN 978-1-4503-5224-6. Disponível em: <http://doi.acm.org/10.1145/3107514.3107515>.

[22] GUEDES, N. P. da S.; TADA, I. N. C. A produção científica brasileira sobre autismo na psicologia e na educação. *Psicologia: teoria e pesquisa*, v. 31, n. 3, p. 303–309, 2015.

[23] DIDEHBANI, N. et al. Virtual reality social cognition training for children with high functioning autism. *Computers in Human Behavior*, Elsevier, v. 62, p. 703–711, 2016.

[24] KANNER, L. et al. Autistic disturbances of affective contact. *Nervous child*, New York, v. 2, n. 3, p. 217–250, 1943.

[25] KLIN, A. Autismo e síndrome de asperger: uma visão geral autism and asperger syndrome: an overview. *Revista Brasileira de Psiquiatria*, SciELO Brasil, v. 28, n. Supl I, p. S3–11, 2006.

[26] SKAFIDAS, E. et al. Predicting the diagnosis of autism spectrum disorder using gene pathway analysis. *Molecular psychiatry*, Nature Publishing Group, v. 19, n. 4, p. 504–510, 2014.

[27] ASSOCIATION, A. P. et al. *DSM-5: Manual diagnóstico e estatístico de transtornos mentais*. Porto Alegre: Artmed, 2014.

[28] GROB, G. N. Origins of DSM: A study in appearance and reality. *American Journal of Psychiatry*, v. 148, n. 4, p. 421–431, 1991.

[29] PORCIUNCULA, R. Investigação precoce do transtorno do espectro autista: Sinais que alertam para a intervenção. *ROTTA, NT; FILHO, CAB; BRIDI, FRS. Neurologia e aprendizagem: abordagem muldisciplinar.*, Porto Alegre: Artmed, p. 29–54, 2016.

[30] ROTTA, N. T.; OHLWEILER, L.; RIESGO, R. S. *Transtornos da aprendizagem: abordagem neurobiológica e multidisciplinar*. Porto Alegre: Artmed, 190f., 2016.

[31] GUEDES, D. F. *O uso das Tecnologias Digitais para a Alfabetização de Alunos com Transtorno do Espectro Autista: proposta de um Curso de Capacitação*. Dissertação (Mestrado) — Programa de Pós-Graduação em Ensino, Cornélio Procópio, 2019.

[32] THABTAH, F.; KAMALOV, F.; RAJAB, K. A new computational intelligence approach to detect autistic features for autism screening. *International Journal of Medical Informatics*, Elsevier, v. 117, p. 112–124, 2018.

[33] ALLISON, C.; AUYEUNG, B.; BARON-COHEN, S. Toward brief "red flags" for autism screening: the short autism spectrum quotient and the short quantitative checklist in 1,000 cases and 3,000 controls. *Journal of the American Academy of Child & Adolescent Psychiatry*, Elsevier, v. 51, n. 2, p. 202–212, 2012.

[34] ABDELJABER, F. *Detecting Autistic Traits using Computational Intelligence & Machine Learning Techniques*. Dissertação (Mestrado) — Department of Psychology, University of Huddersfield, January 2019. Disponível em: <http://eprints.hud.ac.uk/id/eprint/34844/>.

[35] ALPAYDIN, E. *Introduction to machine learning*. Cambridge, MA: MIT Press, 2009. 613 p.

[36] BISHOP, C. M. *Pattern recognition and machine learning*. New York, NY: Springer, 2006. (Information science and statistics). Softcover published in 2016. Disponível em: <http://cds.cern.ch/record/998831>.

[37] MICHIE, D. et al. Machine Learning. *Neural and Statistical Classification*, Technometrics, v. 13, 1994.

[38] RINGNÉR, M. What is principal component analysis? *Nature biotechnology*, Nature Publishing Group, v. 26, n. 3, p. 303, 2008.

[39] SABIN, J. G.; FERRÃO, M. F.; FURTADO, J. C. Análise multivariada aplicada na identificação de fármacos antidepressivos. Parte II: Análise por Componentes Principais (PCA) e o método de classificação simca. *Revista Brasileira de Ciências Farmacêuticas*, v. 40, n. 3, p. 387–396, 2004.

[40] CUNNINGHAM, P.; DELANY, S. J. K-nearest neighbour classifiers. *Multiple Classifier Systems*, Springer-Verlag, v. 34, n. 8, p. 1–17, 2007.

[41] SALZBERG, S. L. C4.5: Programs for machine learning. *Machine Learning*, Springer, v. 16, n. 3, p. 235–240, 1994.

[42] BARBON, A. P. et al. Storage time prediction of pork by computational intelligence. *Computers and Electronics in Agriculture*, v. 127, p. 368–375, 2016.

[43] MONARD, M. C.; BARANAUSKAS, J. A. Indução de regras e árvores de decisão. *Sistemas Inteligentes-Fundamentos e Aplicações*, v. 1, p. 115–139, 2003.

[44] BREIMAN, L. Random Forests. *Machine learning*, Springer, v. 45, n. 1, p. 5–32, 2001.

[45] BREIMAN, L. *Classification and Regression Trees*. Routledge, 2017. ISBN 9781351460484. Disponível em: <https://books.google.com.br/books?id=gLs6DwAAQBAJ>.

[46] GENUER, R.; POGGI, J.-M.; TULEAU-MALOT, C. Variable selection using Random Forests. *Pattern Recognition Letters*, Elsevier, v. 31, n. 14, p. 2225–2236, 2010.

[47] HAN, H.; GUO, X.; YU, H. Variable selection using mean decrease accuracy and mean decrease gini based on random forest. In: *7th IEEE International Conference on Software Engineering and Service Science*. Beijing, China: IEEE, 2016. p. 219–224.

[48] CORTES, C.; VAPNIK, V. Support-vector networks. *Machine Learning*, v. 20, n. 3, p. 273–297, Sep 1995. ISSN 1573-0565. Disponível em: <https://doi.org/10.1007/BF00994018>.

[49] LORENA, A. C.; CARVALHO, A. C. de. Uma introdução às support vector machines. *Revista de Informática Teórica e Aplicada*, v. 14, n. 2, p. 43–67, 2007.

[50] KITCHENHAM, B. Procedures for performing systematic reviews. *Keele University Technical Repor*, United Kingdom, v. 33, 08 2004.

[51] ALWAKEEL, S. S. et al. A machine learning based wsn system for autism activity recognition. In: IEEE. *14th International Conference on Machine Learning and Applications*. Miami, Florida, USA, 2015. p. 771–776.

[52] MARTINEZ, A.; DU, S. A model of the perception of facial expressions of emotion by humans: Research overview and perspectives. *Journal of Machine Learning Research*, v. 13, n. May, p. 1589–1608, 2012.

[53] YAHATA, N. et al. A small number of abnormal brain connections predicts adult autism spectrum disorder. *Nature communications*, Nature Publishing Group, v. 7, p. 11254–11261, 2016.

[54] MYTHILI, M.; SHANAVAS, A. M. A study on autism spectrum disorders using classification techniques. *International Journal of Computer Science and Information Technologies*, Citeseer, v. 5, n. 6, p. 7288–7291, 2014.

[55] VIDHUSHA, S.; ANANDHAN, K. Analysis and evaluation of autistic brain mr images using learning vector quantization and support vector machines. In: *International Conference on Industrial Instrumentation and Control*. Pune, India: IEEE, 2015. p. 911–916.

[56] WANG, S. et al. Atypical visual saliency in autism spectrum disorder quantified through model-based eye tracking. *Neuron*, Elsevier, v. 88, n. 3, p. 604–616, 2015.

[57] VIGNESHWARAN, S. et al. Autism spectrum disorder detection using projection based learning meta-cognitive rbf network. In: IEEE. *International Joint Conference on Neural Networks*. Dallas, TX, 2013. p. 1–8.

[58] HAKER, H.; SCHNEEBELI, M.; STEPHAN, K. E. Can bayesian theories of autism spectrum disorder help improve clinical practice? *Frontiers in psychiatry*, Frontiers, v. 7, p. 107–121, 2016.

[59] JAMAL, W. et al. Classification of autism spectrum disorder using supervised learning of brain connectivity measures extracted from synchrostates. *Journal of neural engineering*, IOP Publishing, v. 11, n. 4, p. 1–19, 2014.

[60] DUDA, M. et al. Crowdsourced validation of a machine-learning classification system for autism and ADHD. *Translational psychiatry*, Nature Publishing Group, v. 7, n. 5, p. e1133, 2017.

[61] JIANG, Y.-h. et al. Detection of clinically relevant genetic variants in autism spectrum disorder by whole-genome sequencing. *The American Journal of Human Genetics*, Elsevier, v. 93, n. 2, p. 249–263, 2013.

[62] WEE, C.-Y. et al. Diagnosis of autism spectrum disorders using regional and interregional morphological features. *Human brain mapping*, Wiley Online Library, v. 35, n. 7, p. 3414–3430, 2014.

[63] HAZLETT, H. C. et al. Early brain development in infants at high risk for autism spectrum disorder. *Nature*, Nature Publishing Group, v. 542, n. 7641, p. 348–363, 2017.

[64] LEE, H.-y. et al. Ensemble of machine learning and acoustic segment model techniques for speech emotion and autism spectrum disorders recognition. In: *Annual Conference of the International Speech Communication Association*. Lion, France: [s.n.], 2013. p. 215–219.

[65] OBAFEMI-AJAYI, T. et al. Facial structure analysis separates autism spectrum disorders into meaningful clinical subgroups. *Journal of autism and developmental disorders*, Springer, v. 45, n. 5, p. 1302–1317, 2015.

[66] EMERSON, R. W. et al. Functional neuroimaging of high-risk 6-month-old infants predicts a diagnosis of autism at 24 months of age. *Science translational medicine*, American Association for the Advancement of Science, v. 9, n. 393, p. 2882–2890, 2017.

[67] HEINSFELD, A. S. et al. Identification of autism spectrum disorder using deep learning and the abide dataset. *NeuroImage: Clinical*, Elsevier, v. 17, p. 16–23, 2018.

[68] DESHPANDE, G. et al. Identification of neural connectivity signatures of autism using machine learning. *Frontiers in human neuroscience*, Frontiers, v. 7, p. 670, 2013.

[69] LIU, W.; LI, M.; YI, L. Identifying children with autism spectrum disorder based on their face processing abnormality: A machine learning framework. *Autism Research*, Wiley Online Library, v. 9, n. 8, p. 888–898, 2016.

[70] SEGOVIA, F. et al. Identifying endophenotypes of autism: a multivariate approach. *Frontiers in computational neuroscience*, Frontiers, v. 8, p. 60, 2014.

[71] GHIASSIAN, S. et al. Learning to classify psychiatric disorders based on fmr images: Autism vs healthy and adhd vs healthy. In: *Proceedings of 3rd NIPS Workshop on Machine Learning and Interpretation in NeuroImaging*. Nevada, USA: Springer, 2013. p. 1–7.

[72] ABBAS, H. et al. Machine learning for early detection of autism (and other conditions) using a parental questionnaire and home video screening. In: *International Conference on Big Data (Big Data)*. Boston, MA, USA: IEEE, 2017. p. 3558–3561.

[73] THABTAH, F. Machine learning in autistic spectrum disorder behavioral research: A review and ways forward. *Informatics for Health and Social Care*, Taylor & Francis, p. 1–20, 2018.

[74] ZHOU, Y.; YU, F.; DUONG, T. Multiparametric mri characterization and prediction in autism spectrum disorder using graph theory and machine learning. *PloS one*, Public Library of Science, v. 9, n. 6, p. e90405, 2014.

[75] ECKER, C.; BOOKHEIMER, S. Y.; MURPHY, D. G. Neuroimaging in autism spectrum disorder: brain structure and function across the lifespan. *The Lancet Neurology*, Elsevier, v. 14, n. 11, p. 1121–1134, 2015.

[76] RETICO, A. et al. Neuroimaging-based methods for autism identification: a possible translational application? *Functional neurology*, CIC Edizioni internazionali, v. 29, n. 4, p. 231–239, 2014.

[77] KOSMICKI, J. et al. Searching for a minimal set of behaviors for autism detection through feature selection-based machine learning. *Translational psychiatry*, Nature Publishing Group, v. 5, n. 2, p. 1–7, 2015.

[78] DUDA, M.; KOSMICKI, J.; WALL, D. Testing the accuracy of an observation-based classifier for rapid detection of autism risk. *Translational psychiatry*, Nature Publishing Group, v. 4, n. 8, p. e424, 2014.

[79] WALL, D. P. et al. Use of artificial intelligence to shorten the behavioral diagnosis of autism. *PloS one*, Public Library of Science, v. 7, n. 8, p. e43855, 2012.

[80] BONE, D. et al. Use of machine learning to improve autism screening and diagnostic instruments: effectiveness, efficiency, and multi-instrument fusion. *Journal of Child Psychology and Psychiatry*, Wiley Online Library, v. 57, n. 8, p. 927–937, 2016.

[81] CRIPPA, A. et al. Use of machine learning to identify children with autism and their motor abnormalities. *Journal of autism and developmental disorders*, Springer, v. 45, n. 7, p. 2146–2156, 2015.

[82] BEKEROM, B. Using machine learning for detection of autism spectrum disorder. In: *26th Twente Student Conference on IT*. Netherlands: University of Twente, 2017. p. 1–7.

[83] MICHAELSON, J. et al. Whole genome sequencing in autism identifies hotspots for de novo germline mutation. *Cell*, v. 151, p. 1431–42, 12 2012.

[84] HAN, J.; PEI, J.; KAMBER, M. *Data mining: concepts and techniques*. [S.l.]: Elsevier, 2011.

[85] GUYON, I.; ELISSEEFF, A. An introduction to variable and feature selection. *Journal of machine learning research*, v. 3, p. 1157–1182, 2003.

[86] KOHAVI, R. et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: *Proceedings of 14th International Joint Conferences on Artificial Intelligence*. Montreal, Canada: IJCAI, 1995. v. 14, n. 2, p. 1137–1145.

[87] KIM, J.-H. Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. *Computational statistics & data analysis*, Elsevier, v. 53, n. 11, p. 3735–3745, 2009.

TRABALHOS PUBLICADOS PELO AUTOR

Trabalhos publicados pelo autor durante o programa.

Publicações principais do trabalho.

1. ARTONI, A. A.; PRECE, B.; SCARANTI, G.; BARBON JUNIOR, S, BARBOSA, C. R. S. C. Aplicação de Aprendizado de Máquina para Auxílio no Diagnóstico do Transtorno do Espectro Autista em Adultos, **XXIV Congresso Internacional de Informática Educativa (TISE)**, November, 2018, Volume 14, p. 167 - 173, Brasilia - DF, Brasil. (B5)

Publicações complementares.

1. ABONIZIO, H. Q.; ARTONI, A. A.; BARBOSA, C. R. S. C. Detecção Automática dos Heterônimos de Fernando Pessoa por Aprendizado de Máquina, **XII Symposium in Information and Human Language Technology and Collocates Events (STIL)** October, 2019, p. 144 - 153, Salvador – BA, Brasil. (B3)
2. PEREIRA, Y. H.; BARBOSA, C. R. S. C.; ARTONI, A. A.; MIONI, J. L. V. M ; BRANCHER, J. D. Interface em Linguagem Natural para uma Sublinguagem de Câncer de Pele, **XXI Computer on the Beach (COBT)**, September , 2020, Balneário Camboriú - SC, Brasil. (B4) (Aceito)