

Autism Spectrum Disorder Diagnosis Assistance using Machine Learning

Auxílio ao Diagnóstico do Transtorno do Espectro Autista usando Aprendizado de Máquina

Arthur Alexandre Artoni^{1*}, Cinthyan Renata Sachs Camerlengo de Barbosa², Marcelo Morandini³

Abstract: Autism Spectrum Disorder (ASD) is a common but complex disorder to diagnose since there are no imaging or blood tests that can detect ASD. Several techniques can be used, such as diagnostic scales that contain specific questionnaires formulated by specialists that serve as a guide in the diagnostic process. In this paper, Machine Learning (ML) was applied on three public databases containing AQ-10 test results for adults, adolescents, and children; as well as other characteristics that could influence the diagnosis of ASD. Experiments were carried out on the databases to list which attributes would be truly relevant for the diagnosis of ASD using ML, which could be of great value for medical students or residents, and for physicians who are not specialists in ASD. The experiments have shown that it is possible to reduce the number of attributes to only 5 while maintaining an Accuracy above 0.9. In the other Database to maintain the same level of Accuracy, the fewer attribute numbers were 7. The Support Vector Machine stood out from the others algorithms used in this paper, obtaining superior results in all scenarios.

Keywords: Autism — Machine Learning — Diagnosis

Resumo: O Transtorno do Espectro Autista (TEA) é um transtorno comum, porém complexo de diagnosticar, uma vez que não existem exames de imagens ou de sangue capazes de apontar o TEA. Diversas técnicas podem ser usadas, como escalas diagnósticas que contêm 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 públicas 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, o que poderia ser de grande valia para estudantes ou residentes na área de medicina, além de médicos não especialistas em TEA. Os experimentos mostraram ser 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 foi 7. A *Support Vector Machine* se destacou dos demais algoritmos usados neste trabalho, obtendo resultados superiores em todos os cenários.

Palavras-Chave: Autismo — Aprendizado de Máquina — Diagnóstico

^{1,2}Programa de Pós-Graduação em Ciência da Computação da Universidade Estadual de Londrina (UEL), Londrina - Paraná, Brasil

³Programa de Pós-graduação em Sistemas de Informação Escola de Artes, Ciências e Humanidades da Universidade de São Paulo (EACH-USP), São Paulo - São Paulo, Brasil

*Corresponding author: arthurartoni@uel.br

DOI: <http://dx.doi.org/10.22456/2175-2745.126309> • Received: 31/07/2022 • Accepted: 12/11/2022

CC BY-NC-ND 4.0 - This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

1. Introdução

O Transtorno do Espectro Autista (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]. 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 [2].

Ainda nos dias de hoje, os profissionais da área da saúde enfrentam uma grande dificuldade para realizar o diagnóstico

de doenças como o TEA e não é diferente também com os residentes de medicina enquanto estão aprendendo essa ciência. Isso acontece em decorrência da inexistência de exames laboratoriais que possam confirmar com eficácia o diagnóstico, além da grande variação sintomática do TEA.

O processo de diagnóstico é feito com análise comportamental do indivíduo. Para guiar esse processo são utilizados questionários específicos para diagnóstico do TEA desenvolvidos por especialistas, os quais são conhecidos como escalas diagnósticas [3]. 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 Diagnóstico e Estatístico de Transtornos Mentais, conhecido pela sigla DSM (*Diagnostic and Statistical Manual of Mental Disorders*), da Associação Americana de Psiquiatria atualmente é a mais conhecida referência para o diagnóstico de transtornos mentais. A versão mais recente foi lançada em 2013 e recebe o nome de DSM-V [4]. Outros exemplos de escalas conhecidas são: CARS (*Childhood Autism Rating Scale*), criada por Schopler e Reichler em 1971 [5] [6] 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 [7]. Diferentemente do DSM, essas escalas foram desenvolvidas com foco exclusivo no diagnóstico do TEA.

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 e a grande variação desses, falta de treinamento dos profissionais da área de saúde e também as dificuldades de acesso aos serviços de saúde [4] [8].

Uma das possíveis alternativas para auxiliar os profissionais da área médica é empregar técnicas de Aprendizado de Máquina (AM), como um mecanismo de apoio à decisão no diagnóstico ou pré-diagnóstico do TEA [9] [10]. O AM consiste em uma área 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. Por meio desse, 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 [11].

Tipicamente o diagnóstico do TEA utilizando AM é encarado como um problema de classificação, no qual um modelo é construído baseando-se em dados previamente classificados de uma Base de Dados [12]. O uso do AM pode oferecer soluções bem eficientes para o diagnóstico do TEA, principalmente àqueles que estão ainda aprendendo tal diagnóstico tão complexo, uma vez que possui modelos matemáticos e métodos computacionais capazes de lidar com grandes volumes de dados e regras.

O trabalho empregou o conjunto de dados para avaliar o impacto da utilização de Dados Biológicos, estes estão presente na literatura médica como possivelmente relacionados com TEA, quando utilizados em conjunto com os testes diagnósticos já existentes podem realmente ajudar no processo diagnóstico [13].

O trabalho está organizado da seguinte maneira: a seção 2 aborda o referencial teórico que serviu como base para este trabalho; a seção 3 expõe os materiais e métodos que foram aplicados; a seção 4 mostra os resultados que foram obtidos seguindo os métodos propostos na seção 3; por fim, a seção 5 apresenta as conclusões obtidas e possíveis trabalhos futuros.

2. Referencial Teórico

Esta seção apresenta o referencial teórico 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 [14]. Existem vários tipos de classificações para o TEA. Para Lampreia [15], o autismo pode ser dividido em três níveis, sendo o nível 1 o mais leve, o 2 considerado moderado e o 3 mais severo. De acordo com Didehbani *et al.* [16], os indivíduos com autismo do primeiro nível tendem 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. Já os de segundo nível apresentam graves problemas na sua comunicação, seja verbal ou não, além de uma extrema dificuldade de aceitar mudanças. Finalmente, temos aqueles do terceiro nível que mostram 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 [17], 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 [17] 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 [17] descreveu várias características semelhantes em indivíduos com dificuldade de comunicação e interação social, mas que possuíam inteligência normal [17]. Graças aos estudos de Asperger, apenas os casos considerados leves de autismo e que não apresentavam atraso de linguagem receberam o nome de Síndrome de Asperger 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 [18] [19].

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.* [20] afeta 1 a cada 150 e pode-se dizer que ele afeta aproximadamente 1% dos nascimentos no planeta, sendo o sexo masculino quatro vezes mais propenso que o feminino [4]. De acordo com Gomes *et al.* [1], no ano de 2010 no Brasil existiam cerca de 500 mil pessoas com autismo. Seria acrescentada no Censo Demográfico do Instituto Brasileiro de Geografia e Estatística (IBGE) de 2020 tal informação para saber uma quantidade mais precisa de quantos indivíduos com TEA havia no Brasil, mas por conta da pandemia de Covid-19 isso não aconteceu. O Censo Demográfico 2022 está previsto quantificar as pessoas com TEA, porém em uma amostra de 10% da população. Isso pode não revelar o número exato de pessoas autistas, embora o censo seja o único instrumento capaz de estimar o número desse público.

Quanto mais cedo o TEA for diagnosticado, melhores poderão ser os resultados do tratamento. Isso se dá devido às características do desenvolvimento físico da criança, em especial do cérebro. Dessa forma quanto antes o tratamento for iniciado, menores podem ser as expressões sintomáticas do TEA no indivíduo [4].

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 por exames laboratoriais, seja por esses não existirem ou por se tratarem de exames muito complexos e demorados [3].

Um exemplo desses manuais é o DSM [21] 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 [22]. No DSM-IV trouxe a *triade* de sintomas para o TEA que modela déficits em: a) interação social; b) comunicação; c) interesses restritos e comportamentos repetitivos [23].

Desde então esse manual esteve sobre constante evolução, gerado pelo avanço na medicina psiquiátrica. Atualmente (lançado em 2013) o DSM está na versão 5 [21], 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 [24]. O “autismo” é classificado como Transtorno do Neurodesenvolvimento, que recebe o nome de Transtorno do Espectro Autista (TEA) [19]. Em Guedes [25] é chamada a atenção que no DSM-V troca-se a tríade diagnóstica por uma tríade sendo composta por: a) comportamentos repetitivos e interesses restritos; b) déficits

sociais e comunicativos [23].

O site da Associação Americana de Psiquiatria² 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 diagnósticas 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 [26].

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 diagnóstica autoaplicável desenvolvida por Baron *et al.* [27] 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, um novo teste baseado no AQ foi apresentada em 2012 por Allison *et al.* [28]. Esse diminuiu o número de questões do teste para apenas 10, além de apresentar quatro novas versões: 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.

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 Ciência da Computação na resolução dos problemas do mundo real, como, por exemplo, por meio da utilização da 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 [11].

O Aprendizado de Máquina, denotado neste trabalho por AM, é uma área da Inteligência Artificial (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 [29].

O AM é dividido em duas categorias: o aprendizado supervisionado e o não supervisionado. O *aprendizado não*

¹<<https://www.apa.org/>>

²<<https://www.psychiatry.org/psychiatrists/practice/dsm>>

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 [11]. 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 [11].

A *Classificação* foi aplicada neste artigo na criação dos modelos de aprendizado, uma vez que é adequada para tratar de problemas, cujo objetivo final do modelo seja prever variáveis de valores discretos, ou seja, apresenta um número finito de respostas possíveis [29]. Dessa forma, pode-se classificar algo em duas ou mais categorias previamente definidas. Para isso é necessária 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 [30]. Neste trabalho o objetivo é aplicar AM utilizando a classificação com Aprendizado Supervisionado para identificar se um indivíduo possui TEA ou não.

2.4 Trabalhos Relacionados

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.

A Tabela 1 mostra trabalhos que também fazem uso de Aprendizado de Máquina para diagnóstico do TEA. Nessa tabela podemos observar que os trabalhos utilizam diferentes tipos de bases de dados e de algoritmos.

Dentre os 16 trabalhos apresentados na Tabela 1, foram encontrados os seguintes tipos de bases de dados:

- **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. Já os modelos fornecidos pelos algoritmos de IA têm características próprias baseadas nos algoritmos. Um exemplo é o uso de algoritmos de Árvore de Decisão para criar um modelo de aprendizado, independentemente da escala que for utilizada como entrada, o algoritmo irá sempre gerar uma árvore de decisão na saída;
- **Dados de Imagem de Ressonância Magnética:** trata-se do emprego de imagens do cérebro de pessoas com

e sem TEA. Utilizando algoritmos de AM aplicados 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 características faciais que sejam únicas aos indivíduos com TEA e que possam ser usadas para auxiliar o diagnóstico;
- **Dados Genéticos:** possuem características 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.

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, utilizando variados tipos de algoritmos como: Árvores de Decisão, *Naive Bayes*, *Random Forest* (RF), Redes Neurais e *Support Vector Machine* (SVM).

Além dos trabalhos descritos, podemos destacar o trabalho de Bekerom [47], onde o mesmo utilizou uma Base de Dados (BD) da *National Survey of Children's Health* (NSCH) contendo vários tipos de dados relativos às várias doenças, não apenas ao TEA. Essa BD 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 *undersampling* que reduziram o número de amostras para criar uma nova base de dados que continha cerca de 50% de indivíduos com TEA e 50% sem.

Depois da criação dessa nova base de dados, dos 367 atributos anteriores, Bekerom [47] 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, religião e índice de massa corporal, ou seja, utilizou uma mistura de atributos físicos e comportamentais que poderiam estar 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%.

Por outro lado, Obafemi *et al.* [48] optou pela utilização de 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 BD. 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, onde o primeiro cenário foram empregadas

Tabela 1. Bases de dados e algoritmos utilizados

Artigo	Base de dados	Algoritmos
Zhao <i>et al.</i> (2019)[31]	Comportamentais biológicos	Support Vector Machine (SVM), Decision Tree (DT), Linear Discriminant Analysis (LDA), Random Forest (RF), <i>k</i> -nearest neighbors algorithm (KNN)
Alzubi <i>et al.</i> (2017) [32]	Genéticos	SVM, Naive Bayes (NB), LDA, KNN
Liu <i>et al.</i> (2020) [33]	Imagens ressonância magnética	SVM
Elnakieb <i>et al.</i> (2020) [34]	Imagens ressonância magnética	SVM, DT, KNN, Neural Network (NN)
Huang <i>et al.</i> (2019) [35]	Imagens ressonância magnética	SVM
ElNakieb <i>et al.</i> (2019) [36]	Imagens ressonância magnética	SVM, KNN, DT, NN, Deep Neural Network (DNN)
Mostafa, Tang and Wu (2019) [37]	Imagens ressonância magnética	LDA, Logistic Regression (LR), SVM, KNN, NN
Haputhanthri <i>et al.</i> (2019) [38]	Voz por EGG	SVM, LR, RF, NB
Wu <i>et al.</i> (2021) [39]	Imagens Vídeo	SVM
Sidhu (2019) [40]	Imagens ressonância magnética	Principal Component Analysis (PCA), Independent Component Analysis (ICA), SVM
Hasan, Jailani and Tahir (2018)[41]	Imagens 3D movimentos	KNN, SVM, NN
Vijayalakshmi <i>et al.</i> (2020) [42]	Comportamentais demográficos	NB, RF, LR
Akter <i>et al.</i> (2019) [43]	Comportamentais biológicos	SVM, DT, LR
Roopa and Prasad (2019) [44]	Imagens ressonância magnética	SVM, RF, DNN
Huang, Liu and Tan (2020) [45]	Imagens ressonância magnética	SVM
Aslam <i>et al.</i> (2021) [46]	Imagens EEG	Shallow Neural Network

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 [48].

Este trabalho buscou avaliar se as características previamente descritas, na literatura médica, como relacionadas ao diagnóstico do TEA apresentariam um impacto positivo, negativo ou nulo ao serem mescladas a um teste diagnóstico. Esse cenário foi sugerido por Thabtah [13].

3. Materiais e métodos

Nesta seção está descrita a metodologia utilizada para realizar os experimentos que foram realizados a fim de validar a hipótese.

3.1 Modelo proposto

Este trabalho propõe a utilização de AM para selecionar três modelos de aprendizado sendo: um para adultos, um para adolescentes e um para crianças. A Figura 1 mostra o fluxo realizado durante o desenvolvimento deste 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 BDs foram para o pré-processamento (B), etapa responsável por verificar a existência de dados faltantes e ruídos. No caso dessas

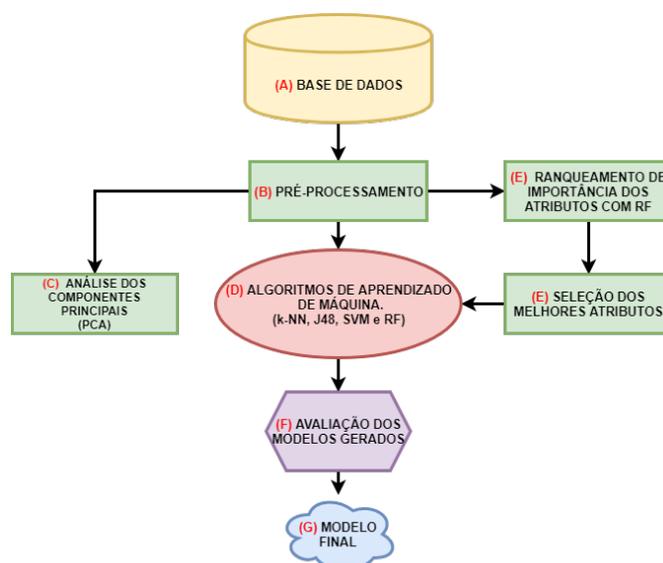


Figura 1. Fluxo de desenvolvimento seguido durante o trabalho

BDs não havia dados faltantes ou inconsistentes. Após o pré-processamento foi executada a PCA (C) para fornecer uma melhor visão da dimensionalidade 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).

3.2 Ferramentas utilizadas

Para o desenvolvimento deste projeto foi utilizada a Linguagem de Programação R³ 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 RStudio⁴ 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.

3.3 Bases de Dados

As bases de dados utilizadas durante o desenvolvimento deste trabalho foram doadas à comunidade científica pelo Dr. Fadi Thabtah da Universidade Huddersfield no Reino Unido, estando com acesso aberto e disponíveis para *download*⁵. 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 [12].

Essas bases podem ser consideradas recentes e ainda pouco exploradas em trabalhos 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 dos dados.

A Figura 2 expõe o número de amostras com e sem TEA e o total em cada Base de Dados.

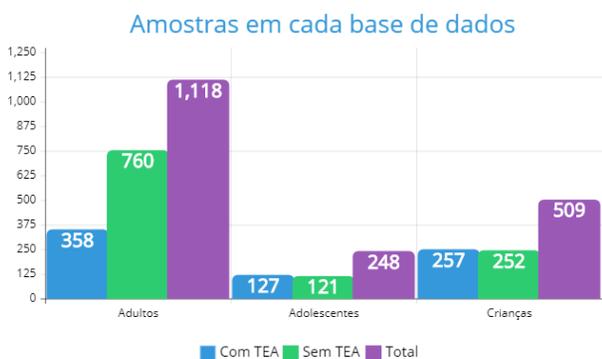


Figura 2. Número de amostras em cada Base de Dados

Todos os atributos presentes nas três Bases de Dados estão detalhados na Tabela 2 contendo o nome do atributo e o seu tipo.

³<<https://www.r-project.org/about.html>>

⁴<<https://www.rstudio.com/products/rstudio/>>

⁵<<https://archive.ics.uci.edu/ml/datasets/>>

Tabela 2. 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)

Pela Tabela 3 podemos ver a equivalência dos atributos A1 - A10, que representam as questões de 1 a 10 nos testes AQ-10 para Adultos, Adolescentes e Crianças [28]. 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 4 que contém 10 amostras de dados pertencentes à Base de Dados Adolescentes antes de iniciar o pré-processamento dos dados.

3.4 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 [49], bases de dados podem ser elementos bastante complexos por vários fatores como: a) Tamanho; b) Registros inconsistentes, faltantes e *Outliers*; c) Registros duplicados; d) Transformação de dados.

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 [49].

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;

Tabela 3. 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 em um grupo social, ele/a consegue facilmente seguir conversas de várias pessoas.	Quando está em um 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 trens, tipos de plantas, 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 4. Exemplo de 10 amostras da Base de Dados Adolescentes

Case.No	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	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

- **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;
- **Score.** O score faz referência à pontuação obtida pelo teste AQ realizado previamente e a partir desse valor o AQ-10 irá classificar o indivíduo como com ou sem TEA. Como o objetivo do trabalho é aplicar as técnicas de AM para substituir o modelo tradicional, o campo *Score* foi removido;
- **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.** Essa é 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 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.

3.5 Ranqueamento de Importância

O *Ranqueamento de Importância* pode ser descrito como um dos métodos de seleção de atributos. Esse pode ser utilizado para auxiliar na escolha dos atributos mais relevantes, do

ponto de vista computacional, de uma Base de Dados [50]. 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 utilizado por árvores de decisão [51]. O processo de construção de uma árvore vai da raiz até às 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. Esse algoritmo foi escolhido, pois o mesmo utiliza a métrica de *Mean Decrease Gini* (MDG) para a construção dos modelos e fornece a função de realizar o ranqueamento.

Os valores resultantes do ranqueamento podem ser vistos na Tabela 5. Para uma melhor visualização da variação da MDG, uma representação gráfica foi mostrada na Figura 3.

Tabela 5. 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

Embora o atributo *Ethnicity* (etnia) tenha apresentado um valor relevante na classificação, o mesmo foi removido por uma questão de desbalanceamento de dados. Dessa forma, o valor obtido por ele não deve ser considerado na análise, uma vez que os dados estão desbalanceados, o que favorece certas etnias.

Para a Base de Dados Adultos, o ranqueamento mostrou uma maior variação do MDG entre os atributos em relação às 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 MDG 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 MDG superior aos demais atributos.

3.6 Métricas de Avaliação

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 [13].

É chamada de *sensibilidade* a capacidade que um modelo tem de prever os Verdadeiros Positivos [52]. Nesse caso, a capacidade de que os testes têm de diagnosticar corretamente um indivíduo com TEA é calculada pela Equação 1.

$$\text{Sensibilidade} = \frac{VP}{VP + FN} \quad (1)$$

A *especificidade*, por sua vez, refere-se à capacidade do teste de prever corretamente os Verdadeiros Negativos [52]. Ou seja, diagnosticar corretamente quais indivíduos não possuem TEA. Seu valor é obtido pela Equação 2.

$$\text{Especificidade} = \frac{VN}{VN + FP} \quad (2)$$

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 [52]. Vide Equação 3.

$$\text{Acurácia} = \frac{VP + VN}{VP + VN + FP + FN} \quad (3)$$

A Figura 4 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 [52]. 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.

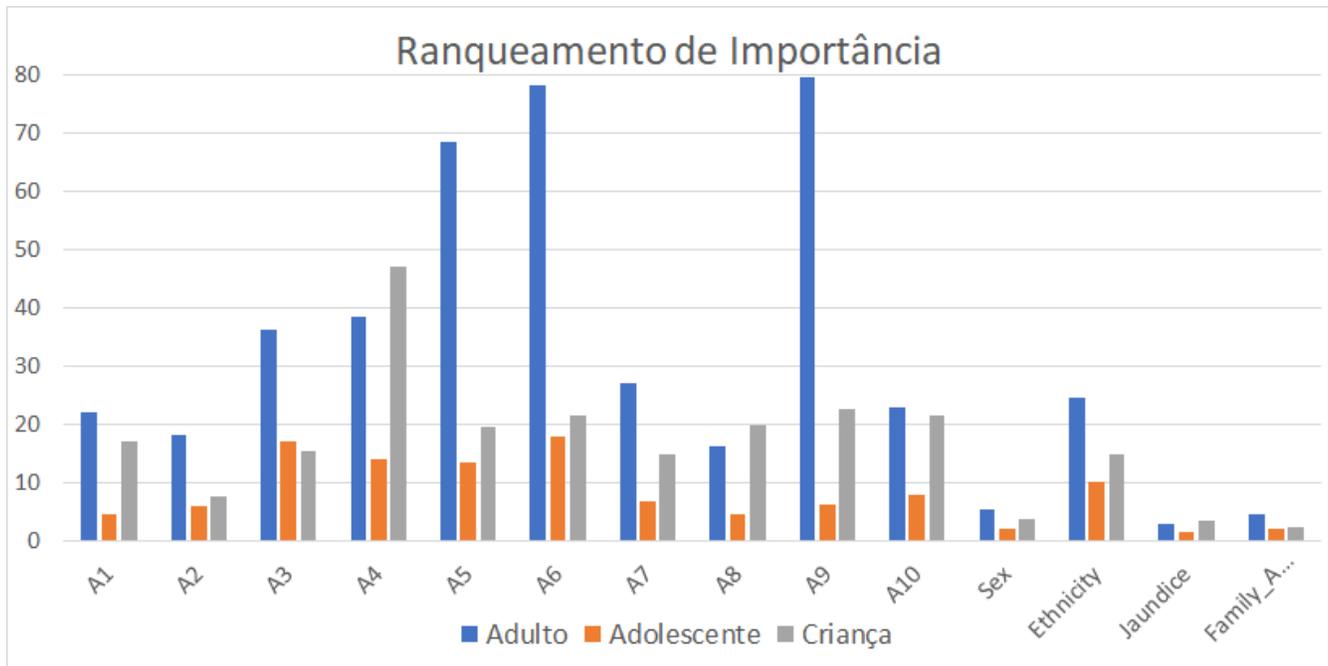


Figura 3. Variação da *Mean Decrease Gini* (MDG) gerada pela RF

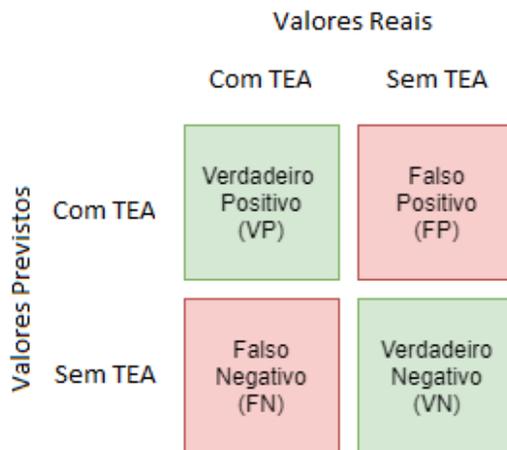


Figura 4. Possíveis classificações dos testes

3.7 Algoritmos e aplicação de AM

Para a realização do trabalho foram utilizados os seguintes algoritmos de AM.

- *Principal Component Analysis* (PCA) [53];
- *k-Nearest Neighbor* (k-NN) [54];
- Árvore de Decisão J48 [55];
- *Random Forest* (RF) [51];
- *Support Vector Machine* (SVM) [56]. Foram utilizadas 3 variações da SVM sendo elas:
 - SVM Linear;
 - SVM Polinomial;

– SVM Radial:

Foram utilizados diferentes tipos de SVM, pois inicialmente seu foco era a solução de problemas binários e, por meio dos hiperplanos, buscava a separação máxima entre as classes [56]. 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. 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 ser utilizados diferentes tipos de *kernels*, como por exemplo, linear, polinomial, radial e sigmoidal [57]. Além disso, é possível criar novos tipos ou fazer alterações nos modelos mais clássicos caso necessário.

Todos os algoritmos foram implementados com o pacote *CARET*⁶ (simplificação para *Classification And REgression Training*) e suas dependências, o qual foi escolhido por possibilitar a aplicação de vários tipos de algoritmos de AM.

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 em português), então um desses é utilizado para teste, enquanto os *k-1* restantes são usados para treinar

⁶<<https://cran.r-project.org/web/packages/caret/>>

o modelo. Assim, não é possível colocar os dados em uma disposição específica na Base de Dados [58]. Um exemplo de Validação Cruzada com 10 folhas pode ser visto na Figura 5.

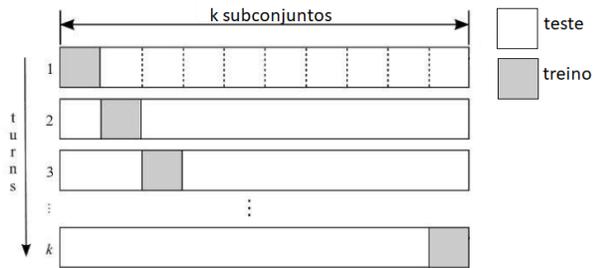


Figura 5. Exemplo de Validação Cruzada com 10 folhas

Foi aplicada a técnica de *Repeated Cross Validation* que possibilitou repetir o processo de CV quantas vezes forem solicitadas. Neste trabalho foram dez vezes. 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 seleção prévia das amostras das bases de dados que serão utilizadas no teste ou no treino [59]. Quando o processo de CV é repetido, oferece uma taxa de erro menor em relação à CV simples. Entretanto, o *Repeated Cross Validation* deve ser utilizado com cuidado, uma vez que o processo de se repetir a divisão de uma Base de Dados pode ser muito custoso, 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 ter sido repetido por 10 vezes não tornou a execução muito custosa.

4. Resultados e Discussões

Esta seção 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.

Importante salientar que cada base de dado foi explorada separadamente, uma vez que o teste foi desenvolvido para ser usado em diferentes faixas etárias. Para se criar um único modelo aplicável a todas as faixas etárias o ideal seria utilizar o Teste AQ em vez do AQ-10.

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 6.

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.

4.1 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 Linguagem de Programação R e neste trabalho o escolhido foi o *ggbiplot*⁷ 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 *Principal Component 1* (PC1) e *Principal Component 2* (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 em azul e vermelho denotam, 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 indicam que os atributos possuem 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. As Figuras 6, 7 e 8 mostram a aplicação do algoritmo nas bases de dados de Adolescentes, Crianças e Adultos.

4.2 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

⁷<<https://github.com/vqv/ggbiplot>>

Tabela 6. 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

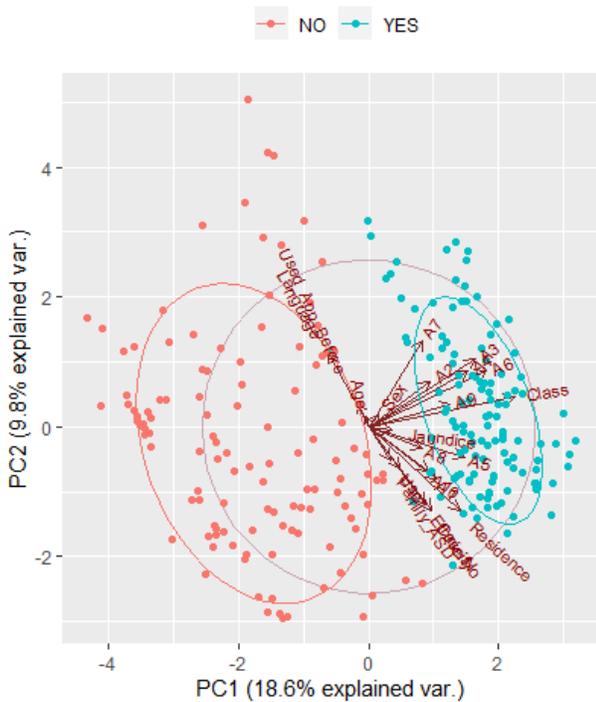


Figura 6. PCA_Adolescentes

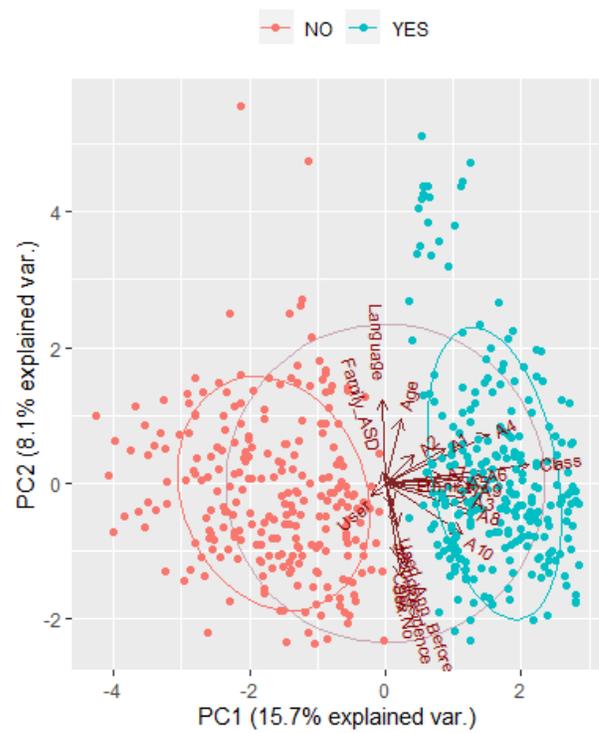


Figura 7. PCA_Crianças

menor dentre as três bases de dados selecionadas.

A Figura 9 mostra a Acurácia dos modelos gerados pelos algoritmos de AM quando aplicado 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 AQ-10, sem dados extras sugeridos por Abdeljaber [13]. Os melhores números 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.

4.3 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, respec-

tivamente). Contudo, essa diferença entre as classes não é o bastante para que a medida de Acurácia não possa mais ser empregada [29].

A Figura 10 mostra a Acurácia dos modelos preditivos gerados pelos algoritmos de AM que foram aplicados à 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 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

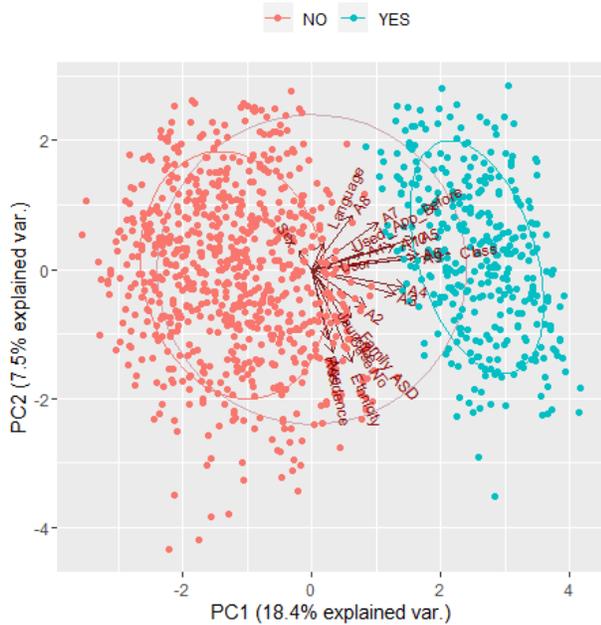


Figura 8. PCA_Adultos

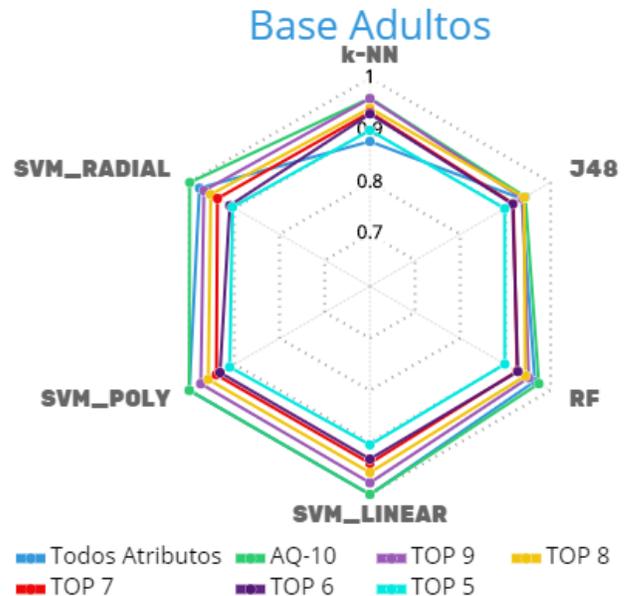


Figura 10. Resultados dos Experimentos na Base de Dados Adultos

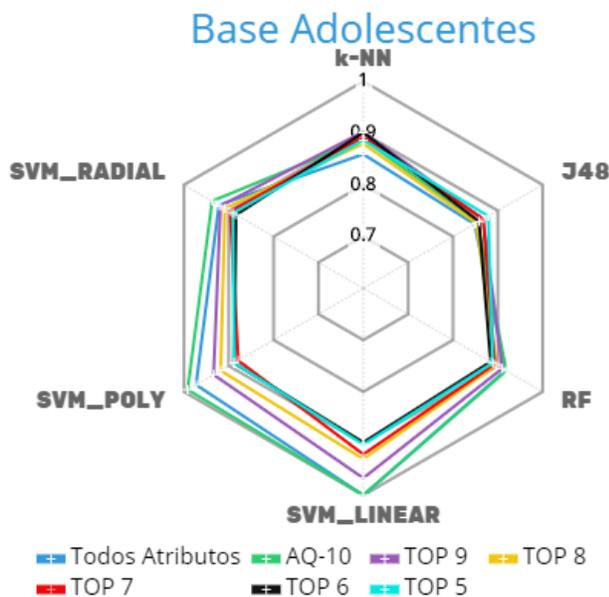


Figura 9. Resultados dos Experimentos na Base de Dados Adolescentes

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 à apenas 5 questões, o melhor modelo foi o gerado pela SVM Polykernel chegando a 0.91 de Acurácia.

4.4 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 11 mostra a Acurácia dos modelos preditivos gerados pelos algoritmos de AM aplicados à Base de Dados Crianças.

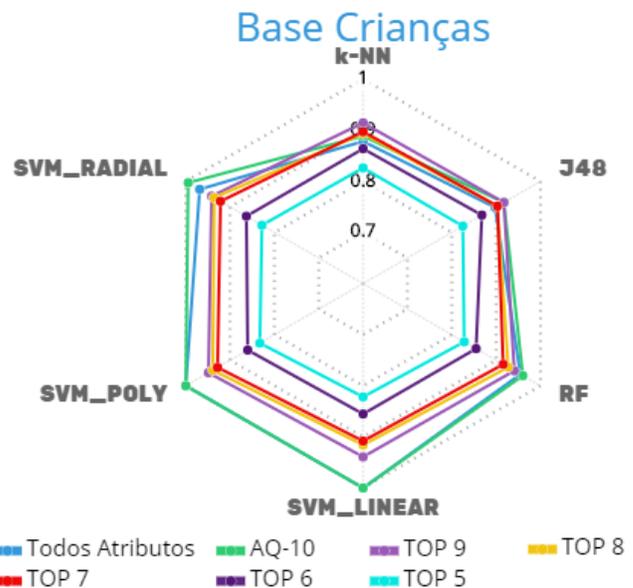


Figura 11. Resultados dos experimentos na Base de Dados Crianças

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

Diferentemente 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.

4.5 Comparação de Performance

Os valores detalhados dos experimentos realizados neste trabalho podem ser vistos na Tabela 7. Nela é possível observar desempenho de cada modelo de aprendizagem para cada cenário proposto e para cada Base de Dados. Os melhores resultados para cada base de dados e cenários propostos estão destacados em negrito.

Com os resultados pode-se observar que os demais atributos mencionados por Thabtah [12] não mostraram nenhum tipo de ganho quando se utilizou AM nessa Base de Dados.

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, obteve 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 *bootstrapping*. 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.

4.6 Questões Seleccionadas

A Tabela 8 apresenta as questões necessárias para cada modelo de aprendizado, utilizando aquele com melhor resultado descrito anteriormente.

Essas questões mostraram-se mais relevantes para os algoritmos de AM, uma vez que as mesmas foram seleccionadas

pelo ranqueamento de importância e 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 estudantes da área médica ou outros profissionais, como psicólogos, por exemplo.

4.7 Modelo Escolhido

Com base nos resultados descritos ao longo desta seção, foi possível verificar que as SVMs como o algoritmo de AM que geraram o melhor modelo de aprendizagem nas três BDs em praticamente todos os cenários.

Para o modelo obtido pela aplicação da SVM Linear na BD de adultos, composta por 1118 amostras e utilizando 5 preditores (variáveis), foi obtida uma acurácia de 0.9047. Na BD de adolescentes, composta por 248 amostras e também utilizando 5 preditores, foi alcançada uma acurácia de 0.9020. Já na BD de crianças, composta por 509 amostras, foi necessário utilizar 7 preditores para manter a acurácia acima de 0.9, a qual 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 [52] mencionada na Seção 2.2 e ilustrada pela Figura 4 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 [52].

A Tabela 9 mostra a Matriz de Confusão obtida pela SVM Linear nos cenários definidos como ideal. 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 9. Comparação das Matrizes de Confusão dos modelos escolhidos

		Sem TEA	Com TEA
Adultos	Sem TEA	63,6	5,2
	Com TEA	4,3	26,8
Adolescentes	Sem TEA	43,4	4,4
	Com TEA	5,4	46,8
Crianças	Sem TEA	45,4	5,1
	Com TEA	4,1	45,4

Com base nos resultados da Tabela 9 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%

Tabela 7. 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

Tabela 8. Questões necessárias para o modelo de aprendizagem

Base de dados	Questões Seleccionadas
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 em um 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.

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. Além disso, a Matriz de Confusão mostra que apesar de um maior desbalanceamento na BD de adultos, que apresenta o dobro de indivíduos sem TEA em relação ao

número de indivíduos com TEA, o percentual de FP e FN foi equivalente ao das demais BDs.

5. 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 de Adultos e Adolescentes. Na Base de Dados de 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 [12] 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.

Cabe ressaltar que este trabalho é uma ferramenta para auxílio no diagnóstico do Transtorno do Espectro Autista utilizando de Aprendizado de Máquina e não tem a pretensão de substituir a *expertise* de um profissional da medicina que é quem pode laudar se um indivíduo tem ou não TEA.

5.1 Trabalhos futuros

Como trabalhos futuros além de testar outros algoritmos de AM, seria de suma importância estudar a viabilidade de utili-

zar 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.

Ainda, 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 (*missclassification analysis*) para verificar qual o padrão dessas amostras classificadas incorretamente.

Agradecimentos

Os autores agradecem à CAPES pelo apoio financeiro recebido durante o mestrado, e ao Fadi Abdeljaber Thabtah - da Universidade de Huddersfield, Reino Unido - por disponibilizar as bases de dados utilizadas para o desenvolvimento da pesquisa.

Contribuição dos autores

Arthur Alexandre Artoni foi o responsável pelo design do modelo, implementação do protótipo, preparação e pré-processamento de dados, treinamento e testes de aprendizado de máquina em seu projeto de mestrado pelo Departamento de Computação da Universidade Estadual de Londrina, cuja dissertação foi defendida na área de Aprendizado de Máquina.

Cinthyan Renata Sachs Camerlengo de Barbosa foi a orientadora do referido projeto e responsável pelo desenvolvimento desta pesquisa e deste artigo pelo Departamento de Computação da Universidade Estadual de Londrina.

Marcelo Morandini da Escola de Arte, Ciências e Humanidades da Universidade de São Paulo fez a coorientação especificamente para desenvolvimento da escrita deste artigo.

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] SILVA, A.; GAIATO, M. B.; REVELES, L. T. *Mundo Singular: entenda o autismo*. Rio de Janeiro: Fontana, 2012.
- [3] FERREIRA, R. 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.
- [4] 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.

- [5] 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.
- [6] 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.
- [7] 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.
- [8] PENTEADO, F. A. O. et al. Software para auxílio ao diagnóstico de autismo. In: *VIII Congresso de extensão universitária da UNESP*. São Paulo: Universidade Estadual Paulista Júlio de Mesquita Filho, 2015. p. 1–4.
- [9] 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.
- [10] 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. 100, 2012.
- [11] XUE, M.; ZHU, C. A study and application on machine learning of artificial intelligence. In: *International Joint Conference on Artificial Intelligence*. Pasadena, California, USA: IEEE, 2009. p. 272–274.
- [12] 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.
- [13] ABDELJABER, F. *Detecting Autistic Traits using Computational Intelligence & Machine Learning Techniques*. Dissertação (Mestrado) — Department of Psychology, University of Huddersfield, January 2019.
- [14] GUEDES, N. P. 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.
- [15] LAMPREIA, A. R. S. *Percepções parentais sobre a perturbação do espectro do autismo: processo de diagnóstico, interferência e recursos*. Dissertação (Mestrado) — Faculdade de Psicologia, Universidade de Lisboa, 2015.
- [16] 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.
- [17] KANNER, L. et al. Autistic disturbances of affective contact. *Nervous child*, New York, v. 2, n. 3, p. 217–250, 1943.
- [18] 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.
- [19] 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.
- [20] 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.
- [21] ASSOCIATION, A. P. et al. *DSM-5: Manual diagnóstico e estatístico de transtornos mentais*. Porto Alegre: Artmed, 2014.
- [22] GROB, G. N. Origins of DSM: A study in appearance and reality. *American Journal of Psychiatry*, v. 148, n. 4, p. 421–431, 1991.
- [23] PORCIUNCULA, R. Investigação precoce do transtorno do espectro autista: Sinais que alertam para a intervenção. *ROTTA, NT; FILHO, CAB; BRIDI, FR. Neurologia e aprendizagem: abordagem multidisciplinar*. Porto Alegre: Artmed, p. 29–54, 2016.
- [24] ROTTA, N. T.; OHLWEILER, L.; RIESGO, R. S. *Transtornos da aprendizagem: abordagem neurobiológica e multidisciplinar*. Porto Alegre: Artmed, 190f., 2016.
- [25] 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.
- [26] 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.
- [27] BARON-COHEN, S. et al. The Autism-spectrum Quotient (AQ): Evidence from Asperger Syndrome/High-Functioning Autism, Males and Females, Scientists and Mathematicians. *Journal of autism and developmental disorders*, Springer, v. 31, n. 1, p. 5–17, 2001.
- [28] 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.
- [29] BISHOP, C. M. *Pattern recognition and machine learning*. New York, NY: Springer, 2006. (Information science and statistics).

- [30] MICHIE, D. et al. Machine Learning. *Neural and Statistical Classification*, Technometrics, v. 13, 1994.
- [31] ZHAO, Z. et al. Applying machine learning to identify autism with restricted kinematic features. *IEEE Access*, IEEE, v. 7, p. 157614–157622, 2019.
- [32] ALZUBI, R. et al. Hybrid feature selection method for autism spectrum disorder snps. In: *Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*. Manchester, UK: IEEE, 2017. p. 1–7.
- [33] LIU, W. et al. Automatic diagnosis of autism based on functional magnetic resonance imaging and elastic net. In: *5th Information Technology and Mechatronics Engineering Conference (ITOEC)*. Chongqing, China: IEEE, 2020. p. 104–108.
- [34] ELNAKIEB, Y. A. et al. Computer aided autism diagnosis using diffusion tensor imaging. *IEEE Access*, IEEE, Tainan, Taiwan, v. 8, p. 191298–191308, 2020.
- [35] HUANG, F. et al. Sparse low-rank constrained adaptive structure learning using multi-template for autism spectrum disorder diagnosis. In: *16th International Symposium on Biomedical Imaging (ISBI 2019)*. Venice: IEEE, 2019. p. 1555–1558.
- [36] ELNAKIEB, Y. et al. Autism spectrum disorder diagnosis framework using diffusion tensor imaging. In: *International Conference on Imaging Systems and Techniques (IST)*. Abu Dhabi, United Arab Emirates: IEEE, 2019. p. 1–5.
- [37] MOSTAFA, S.; TANG, L.; WU, F.-X. Diagnosis of autism spectrum disorder based on eigenvalues of brain networks. *IEEE Access*, IEEE, v. 7, p. 128474–128486, 2019.
- [38] HAPUTHANTHRI, D. et al. An EEG based Channel Optimized Classification Approach for Autism Spectrum Disorder. In: *Moratuwa Engineering Research Conference (MERCOn)*. Sri Lanka: IEEE, 2019. p. 123–128.
- [39] WU, C. et al. Machine Learning Based Autism Spectrum Disorder Detection from Videos. In: *2020 IEEE International Conference on E-health Networking, Application Services (HEALTHCOM)*. Virtual Conference: IEEE, 2021. p. 1–6.
- [40] SIDHU, G. Locally linear embedding and fmri feature selection in psychiatric classification. *IEEE Journal of Translational Engineering in Health and Medicine*, IEEE, v. 7, p. 1–11, 2019.
- [41] HASAN, C. Z. C.; JAILANI, R.; TAHIR, N. M. ANN and SVM Classifiers in Identifying Autism Spectrum Disorder Gait Based on Three-Dimensional Ground Reaction Forces. In: *TENCON 2018 - Region 10 Conference*. Jeju Island, Korea: IEEE, 2018. p. 2436–2440.
- [42] VIJAYALAKSHMI, K. et al. A hybrid recommender system using multiclassifier regression model for autism detection. In: *2020 International Conference on Smart Technologies in Computing, Electrical and Electronics (ICSTCEE)*. Virtual Conference: IEEE, 2020. p. 139–144.
- [43] AKTER, T. et al. Machine learning-based models for early stage detection of autism spectrum disorders. *IEEE Access*, IEEE, v. 7, p. 166509–166527, 2019.
- [44] ROOPA, B. S.; PRASAD, R. M. Concatenating framework in ASD analysis towards research progress. In: *1st International Conference on Advanced Technologies in Intelligent Control, Environment, Computing Communication Engineering (ICATIECE)*. Bangalore, India: IEEE, 2019. p. 269–271.
- [45] HUANG, Z.-A.; LIU, R.; TAN, K. C. Multi-task learning for efficient diagnosis of asd and adhd using resting-state fmri data. In: *2020 International Joint Conference on Neural Networks (IJCNN)*. Glasgow, United Kingdom: IEEE, 2020. p. 1–7.
- [46] ASLAM, A. R. et al. An 8.62 μ w Processor for Autism Spectrum Disorder Classification using Shallow Neural Network. In: *3rd International Conference on Artificial Intelligence Circuits and Systems (AICAS)*. Washington, DC, USA: IEEE, 2021. p. 1–4.
- [47] 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.
- [48] 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.
- [49] HAN, J.; PEI, J.; KAMBER, M. *Data mining: concepts and techniques*. Waltham, MA, USA: Elsevier, 2011.
- [50] GUYON, I.; ELISSEEFF, A. An introduction to variable and feature selection. *Journal of machine learning research*, v. 3, p. 1157–1182, 2003.
- [51] BREIMAN, L. Random Forests. *Machine learning*, Springer, v. 45, n. 1, p. 5–32, 2001.
- [52] ALPAYDIN, E. *Introduction to machine learning*. Cambridge, MA: MIT Press, 2009. 613 p.
- [53] RINGNÉR, M. What is principal component analysis? *Nature biotechnology*, Nature Publishing Group, v. 26, n. 3, p. 303, 2008.
- [54] CUNNINGHAM, P.; DELANY, S. J. K-nearest neighbour classifiers. *Multiple Classifier Systems*, Springer-Verlag, v. 34, n. 8, p. 1–17, 2007.
- [55] SALZBERG, S. L. C4.5: Programs for machine learning. *Machine Learning*, Springer, v. 16, n. 3, p. 235–240, 1994.
- [56] CORTES, C.; VAPNIK, V. Support-vector networks. *Machine Learning*, [S. l.], v. 20, n. 3, p. 273–297, Sep 1995.
- [57] 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.

- [58] 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.
- [59] 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.