

# Aplicação de Aprendizado de Máquina para Auxílio no Diagnóstico do Transtorno do Espectro Autista em Adultos

**Arthur Alexandre Artoni**  
Universidade Estadual de  
Londrina - UEL  
arthurartoni@uel.br

**Bruno Prece**  
Universidade Estadual de  
Londrina - UEL  
bprece@gmail.com

**Gustavo Scaranti**  
Universidade Estadual de  
Londrina - UEL  
gustavoscaranti@hotmail.com

**Sylvio Barbon Junior**  
Universidade Estadual de  
Londrina - UEL  
barbon@uel.br

**Cinthyan R. S. C. de Barbosa**  
Universidade Estadual de  
Londrina - UEL  
cinthyan@uel.br

## ABSTRACT

Autism Spectrum Disorder (ASD), also known simply as autism, is a psychiatric disorder that manifests itself before the age of three and is characterized by symptoms which affect the cognitive abilities and social interaction of individuals. There are several techniques used by experts in the individual's screening to detect the presence of this disorder. One of these techniques is the AQ-10 test, which consists of the application of 10 questions and from them, it is possible to generate a score and detect whether the individual has ASD or not. Through a database containing samples of AQ-10 test applications, in adults, this work aims at the application of machine learning algorithms using classification techniques to demonstrate possible solutions and alternatives to perform or assist in the diagnosis of ASD.

## Author Keywords

Autism Spectrum Disorder (ASD), Machine Learning, diagnosis.

## RESUMO

O Transtorno do Espectro Autista (TEA), também conhecido simplesmente como autismo, é um distúrbio psiquiátrico que se manifesta antes dos três anos de idade e é caracterizado por sintomas que afetam as habilidades cognitivas e a interação social dos indivíduos. Há diversas técnicas utilizadas por especialistas na triagem dos indivíduos a fim de detectar a presença desse transtorno. Uma destas técnicas é o teste AQ-10 que consiste na aplicação de 10 questões e a partir delas é possível gerar uma pontuação e detectar se o indivíduo possui ou não TEA. Através de uma base de dados contendo amostras de aplicações do teste AQ-10, em indivíduos em idade adulta, este trabalho tem como objetivo a aplicação de algoritmos de aprendizado de máquina

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM or the author must be honored. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. *UbiComp '13*, Sept 9-12, 2013, Zurich, Switzerland. Copyright 2013 ACM 978-1-4503-1770-2/13/09...\$15.00.

utilizando técnicas de classificação para demonstrar possíveis soluções e alternativas para realizar ou auxiliar no diagnóstico do TEA.

## Palavras-Chave

Transtorno do Espectro Autista, Aprendizado de Máquina, diagnóstico.

## ACM Classification Keywords

I.2.4 Knowledge Representation Formalisms and Methods: Miscellaneous

## INTRODUÇÃO

Atualmente um dos grandes desafios dos profissionais da área da saúde é o diagnóstico de doenças como o Transtorno do Espectro Autista (TEA). O TEA consiste em um transtorno global do desenvolvimento, que afeta a parte neurológica do paciente, manifestando-se na infância, geralmente até os três anos de idade, não possuindo cura, ou seja, estará presente durante toda a vida desses indivíduos. O TEA afeta principalmente o comportamento social, o desenvolvimento da função comunicativa e a percepção do indivíduo [23].

Para um preciso diagnóstico do TEA, é necessário que os profissionais da área médica consigam realizar uma precisa análise comportamental do indivíduo, em todas as fases da vida. Infelizmente, vários fatores podem retardar o diagnóstico correto e por consequência, o início do tratamento do TEA, dentre eles: demora na detecção dos primeiros sintomas, variabilidade sintomática, falta de treinamento dos profissionais da área de saúde e dificuldade de acesso aos serviços de saúde [29] [18].

Normalmente o diagnóstico do TEA é feito por um profissional especialista no assunto, através da utilização de escalas, ou seja, métodos elaborados por especialistas em diagnóstico para identificar o TEA. No geral, escalas são formadas por questões objetivas que são respondidas por um especialista da área, baseando-se na observação do indivíduo e em entrevistas, que podem ser realizadas com o próprio indivíduo ou responsáveis. Existem diversas escalas, como CARS (*Childhood Autism Rating Scale*), MCHAT (*Modified Checklist for Autism in Toddlers*), dentre outras.

Além do uso das escalas é possível utilizar técnicas de *machine learning* para realizar o diagnóstico do TEA [4][26]. O termo *machine learning*, ou em português, aprendizado de máquina, consiste em uma técnica na qual o computador simula ou realiza um estudo do comportamento dos dados. Ele funciona buscando um novo conhecimento ou habilidade e organiza de uma forma na qual seja possível realizar um melhoramento progressivo do seu próprio rendimento. O Aprendizado de Máquina consiste em uma técnica de Inteligência Artificial, aplicada em vários ramos do conhecimento como: Sistemas Especialistas, Ferramentas de Suporte a Decisão Médica, Processamento de Linguagem Natural, Reconhecimento de Padrões etc [28][27][13].

Segundo Thabtah [25], tipicamente o diagnóstico do TEA, utilizando *machine learning*, é encarado como um problema de Classificação no qual um modelo é construído se baseando em dados previamente classificados de uma base de dados. O uso de *machine learning* pode oferecer métodos e 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.

A base de dados utilizada neste trabalho foi doada para pesquisas por Thabtah [25]. Essa base é formada por indivíduos adultos, a base de dados possui 10 perguntas pertencentes ao teste *Autism Quocient-10* (AQ-10) para adultos e outras características, que segundo Thabtah [25], teriam benefício ao serem utilizadas em conjunto com os dados do teste. No total conjunto de dados conta ao todo com 21 atributos e 704 amostras.

## DEFINIÇÕES BÁSICAS

### Entendendo O Transtorno do Espectro Autista

Atualmente, se busca definir de forma mais precisa o Transtorno do Espectro Autista (TEA) e existem várias definições de autismo atualmente e a mais aceita, segundo Lampreia [17], é que o TEA pode ser considerado como uma síndrome neuropsiquiátrica, que causa déficits comportamentais, emocionais, comunicativos e em especial de na capacidade do indivíduo se relacionar com outras pessoas [1] [6].

O termo Autismo foi utilizado pela primeira vez por Bleuler em 1911, ao se referir a indivíduos que teriam perdido o contato com a realidade. Em 1943, o termo Autismo foi novamente utilizado por Kanner para descrever o comportamento de 11 crianças por ele estudadas. Em 1944, Asperger descreveu várias características semelhantes em indivíduos com dificuldade de comunicação e interação social, mas que possuíam inteligência normal [15] [6].

Inicialmente o autismo era considerado uma doença, porém, atualmente o TEA é considerado uma perturbação do desenvolvimento neurológico, que afeta a maneira como seu portador compreende o mundo ao seu redor. Pode-se classificar o TEA de várias formas. Para Lampreia [17], o autismo pode ser dividido em três níveis, sendo o nível 1 o mais leve e o nível 3 o mais severo.

De acordo com Didehbani et al [8], o primeiro nível tende a ter dificuldades em processar sinais sociais e assim acaba se

tornando oprimido e ansioso em interações sociais, principalmente com pessoas desconhecidas e dificuldades de expor pensamentos e emoções. O segundo nível apresenta graves prejuízos na sua comunicação verbal e não verbal e uma extrema dificuldade de aceitar mudanças. O terceiro nível apresenta um comprometimento funcional mais grave em interações sociais, precisando de um apoio para sua comunicação ser funcional. O correto diagnóstico dos níveis acima é de extrema importância para indicar o melhor tratamento.

O TEA é uma patologia relativamente comum, embora as fontes de dados variem quanto a sua incidência (para Gomes et al [12], estima que um a cada 88 nascidos vivos apresenta TEA, enquanto para Skafidas et al[24] afeta 1 a cada 150) pode-se dizer que ela afeta aproximadamente 1% dos nascimentos no planeta, sendo o sexo masculino quatro vezes mais propenso segundo [29]. De acordo com Gomes et al [12], no ano de 2010 no Brasil existiam cerca de 500 mil pessoas com autismo.

Segundo [29], quanto mais cedo o TEA é diagnosticado e o tratamento iniciado, melhores poderão ser os resultados.

“Devido à plasticidade cerebral, a precocidade do início da intervenção desempenha papel importante, potencializando os efeitos positivos da mesma. Ademais, estudos indicam que os ganhos decorrentes da intervenção precoce podem reduzir consideravelmente os gastos dos familiares no tratamento das crianças com TEA, bem como os dos sistemas de saúde pública, quando se analisa os resultados em longo prazo” [29].

Mesmo ciente da grande importância do diagnóstico precoce do autismo, segundo [9], recentes pesquisas realizadas pelo *WORLD BANK* em 2015, mostram que 86,5% dos diagnosticados com TEA encontram-se na América do Norte, Europa e Japão. Entretanto, essas regiões possuem apenas 10% das crianças que vivem no planeta. Desta forma é possível notar uma dificuldade de diagnóstico do autismo em locais economicamente mais pobres, como é o caso do Brasil. Uma das principais causas de atraso no diagnóstico do TEA é a falta de profissionais treinados/habilitados para reconhecer as primeiras manifestações. Na maior parte das vezes, os pais, não os profissionais da saúde, acabam sendo os primeiros a suspeitarem [29].

O problema, relatado acima é comum em toda a América Latina, segundo [30]. O autismo em crianças latinas é diagnosticado cerca de dois anos e meio anos mais tarde do que em crianças não latinas. A dificuldade de crianças latinas de encontrar o diagnóstico e o tratamento acaba por refletir em casos mais severos de autismo, decorrência dessa demora.

### Diagnóstico Tradicional do TEA

O TEA não é uma doença simples de ser detectada, uma vez que não existem exames laboratoriais para sua confirmação, portanto o diagnóstico deve ser feito por um profissional da saúde especializado em autismo. Para isso, é realizada uma análise comportamental do paciente, entrevista com os pais

e/ou responsáveis, aplicação de uma escala diagnóstica[10].

Uma escala diagnóstica consiste em métodos ou manuais, elaborados para identificar o TEA. Existem diversas escalas como DSM-V, CARS, M-CHAT dentre outras, onde cada uma delas possui suas particularidades. O Manual Estatístico e Diagnóstico da Associação Americana de Psiquiatria, o DSM (*Diagnostic and Statistical Manual of Mental Disorders*) é um dos mais conhecidos. Esse manual serve para o diagnóstico de vários outros tipos de transtornos mentais. Atualmente a versão mais recente desse manual foi lançada em 2013 e recebe o nome de DSM-V [29]. A escala CARS (*Childhood Autism Rating Scale*) foi criada por Schopler e Reichler em 1971 [19][22]. A escala M-CHAT (*Modified Checklist for Autism in Toddlers*) foi uma elaboração da *American healthcare system* modificando a escala CHAT a fim de deixar os resultados mais exatos [14]. Todas essas escalas mencionadas, foram criadas para serem utilizadas por profissionais da área médica.

Além disso, existem escalas diagnósticas autoaplicáveis, ou seja, escalas criadas para serem aplicadas por pessoas sem experiência na área. O *Autism Quocient* (AQ) é um teste criado para realizar o diagnóstico do TEA. Foi proposto por Baron et al [3], com o objetivo de ser a primeira escala de diagnóstico de autismo, que poderia ser aplicada sem a necessidade de um especialista. Ela é composta com 50 questões objetivas, subdivididas em 5 grupos diferentes de 10 questões cada, sendo eles: habilidades sociais, falta de concentração, atenção aos detalhes, comunicação e imaginação.

O AQ-10 é uma adaptação do AQ, no qual as questões foram reduzidas apenas para as 10 mais relevantes. Existem três versões do teste AQ-10 para crianças, adolescentes e adultos e cada uma dessas versões conta com as 10 questões mais impactantes para o diagnóstico do TEA em cada fase da vida [2].

### Inteligência Artificial e Aprendizado de Máquina

No mundo atual, imerso em tecnologia, é natural que essa tecnologia seja utilizada em benefício da qualidade de vida das pessoas. Uma das maneiras de aplicar a computação para a resolução dos problemas do dia a dia se dá através do uso de Inteligência Artificial (IA). A IA consiste na criação de um programa que seja capaz de apresentar características de inteligência, ou seja, capacidade de aprender a executar uma tarefa [28].

Uma das técnicas de IA é o *machine learning* ou em português, aprendizado de máquina. É um conjunto de técnicas onde computador simula ou realiza um estudo do comportamento dos dados para modelar um novo conhecimento ou habilidade e a melhor maneira possível de representá-lo, a fim de realizar um melhoramento progressivo do seu próprio rendimento. Atualmente é amplamente aplicada em vários ramos do conhecimento como: Sistemas Especialistas, Ferramentas de Suporte a Decisão Médica, Processamento de Linguagem Natural e Reconhecimento de Padrões [28][13].

Existem várias técnicas de aprendizado de máquina, sendo que os mais utilizados são o aprendizado supervisionado e o não supervisionado. O aprendizado não supervisionado consiste em um modelo no qual nenhum tipo de resposta

Atributo	Tipo
Idade	Inteiro
Gênero	Texto
Etnia	Texto
Nascido com icterícia	Booleano (Sim ou Não)
Membro da família	Booleano (Sim ou Não)
Distúrbio Difuso de Desenvolvimento	Booleano (Sim ou Não)
Quem está completando o teste?	Texto
País de residência	Texto
Usou o aplicativo de triagem antes	Booleano (Sim ou Não)
Tipo de método de triagem	Inteiro (0,1,2,3)
Questões [1-10]	Binário (0, 1)
Resultado da triagem	Inteiro
Classe	Booleano (Sim ou Não)

Tabela 1. Atributos presentes na base de dados

esperada é dada para a resolução do problema e seu objetivo é descobrir por si mesmo a melhor maneira de reconhecer os padrões de dados [28].

O aprendizado supervisionado, por sua vez, funciona de maneira oposta ao não supervisionado. Antes da aplicação do algoritmo um rótulo ou classe é aplicado aos dados, no caso do reconhecimento de TEA a classe poderia ser Possui TEA, *sim* ou *não*. Esse aprendizado funciona basicamente com duas fases. A primeira consiste no treinamento, onde uma parte dos dados da base é usada para criar um modelo. Após o treino, outra parte dos dados que não foi utilizada no treinamento é usada para avaliar o desempenho. Com essas etapas é possível aos algoritmos modelarem uma solução para identificar se o indivíduo possui ou não o TEA com base nos exemplos fornecidos [28].

## MATERIAIS E MÉTODOS

### Obtenção de Dados

Os dados utilizados para os experimentos desenvolvidos neste trabalho foram obtidos do Repositório *UCI Machine Learning Repository*<sup>1</sup>, disponibilizando uma base de dados com dados de adultos. A base de dados é composta por 10 perguntas pertencentes à AQ-10 para adultos. O conjunto de dados conta ao todo com 21 características, apresentadas pela Tabela 1 e contém 704 amostra [25].

### Algoritmos utilizados

Os seguintes algoritmos foram utilizados:

- **Random Forest (RF)**: é um algoritmo proposto por [5] baseado no modelo de agregação de ideias para resolução genérica de problemas complexos de regressão e classificação. Seu funcionamento consiste na combinação de múltiplas árvores de decisão do tipo *Classification and Regression Tree* (CART) que são criadas utilizando amostras originadas da separação da base de dados inicial em diversos subconjuntos através da técnica de *bootstrap aggregati*

*ng* (*bagging*), escolhendo randomicamente as amostras que irão compor cada subconjunto. Além da sua utilização na geração de modelos de aprendizado, a RF

<sup>1</sup><https://archive.ics.uci.edu/ml/datasets/Autism+Screening+Adult>

possui um método de ranqueamento de importância das características do problema. O ranqueamento é realizado durante a construção do modelo de aprendizado, onde os atributos mais relevantes tendem a alocar os nós mais relevantes nos níveis mais próximos à raiz da árvore [11].

- **Support Vector Machine (SVM):** a SVM é um algoritmo que faz uso de vetores para reconhecer padrões. Ele funciona traçando hiperplanos que tem como objetivo separar as classes e assim conseguir solucionar o problema. Trata-se de um algoritmo muito poderoso para predição, pois permite a alteração de núcleos, o que muda drasticamente o resultado dependendo do tipo de problema [7]. A SVM pode ser utilizada com quatro tipos de núcleos: linear, polinomial, radial e sigmoidal;
- **Decision Tree J48:** é uma implementação em Linguagem de Programação Java do modelo C4.5 criado por Ross Quinlan, atualmente sendo uma das árvores de decisão mais usadas para problemas de classificação [21];
- **Principal Component Analysis (PCA):** a PCA, ou em português, Análise dos Componentes Principais é um algoritmo que usa um modelo matemático para a transformação dos dados. Ela converte um conjunto de observações de variáveis, que podem estar correlacionadas em um conjunto de valores de variáveis linearmente não correlacionadas chamadas de componentes principais. A PCA é bem útil para a observação dos espaços de dados podendo indicar algumas características do problema, como por exemplo, se o problema seria linearmente separável [20];

### Fluxo de Desenvolvimento

As etapas de desenvolvimento deste trabalho seguem o modelo mostrado na Figura 1, iniciando-se pela fase de pré-processamento dos dados, na qual dados ruidosos foram removidos da base de dados. Na sequência foi realizada uma Análise dos Componentes Principais (PCA) utilizando a linguagem de programação R<sup>2</sup>, a fim de se obter uma visão inicial do problema. Na sequência, a *Random Forest* foi utilizada para o ranqueamento da importância dos atributos. Após esse ranqueamento, os atributos mais relevantes foram selecionados. Por fim, os algoritmos de aprendizado de máquina foram aplicados e os resultados foram obtidos e analisados.

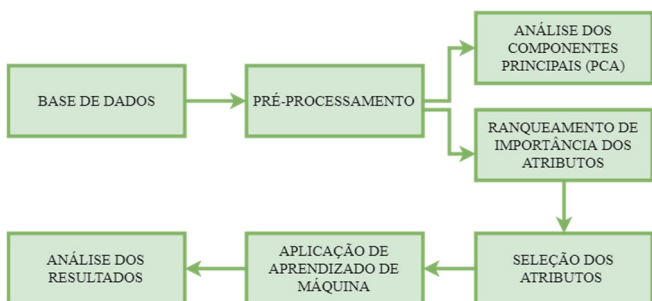


Figura 1. Etapas de desenvolvimento

<sup>2</sup><https://www.r-project.org/>

### Análise dos Componentes Principais

Inicialmente foi realizada uma Análise dos Componentes Principais utilizando a linguagem de programação R a fim de obter uma análise da correlação entre os atributos. Analisando os resultados dos Componentes Principais foi possível observar indícios de que o diagnóstico do TEA, através da base de dados escolhida, seria um problema linearmente separável.

### Ranqueamento de Importância

Após a análise da PCA a *Random Forest* foi utilizada para o ranqueamento da importância dos atributos da base de dados. A Tabela 2 apresenta o resultado da aplicação da *Random Forest* como seletor de atributos da base de dados. Nota-se que há perguntas do teste AQ-10 que apresentam grandes relevâncias em relação à classificação das amostras, além disso, os atributos de dados pessoais e outras informações complementares apresentado na Tabela 2 foram omitidos. A baixa relevância dessas características se dá principalmente pela não influência de questões étnicas, faixa etária e distúrbios de desenvolvimento em membros da família em relação ao autismo do indivíduo em que o teste está sendo aplicado.

Atributo	Gini	Atributo	Gini
Questão 9	60.69	Questão 1	16.27
Questão 6	43.97	Questão 7	15.42
Questão 5	34.35	Questão 10	13.23
Questão 4	23.99	Questão 2	11.69
Questão 3	18.42	Questão 8	11.16

Tabela 2. Ranqueamento de Importância

A Figura 2 deixa mais clara a variação de importância de cada questão referente ao teste AQ-10 para Adultos.

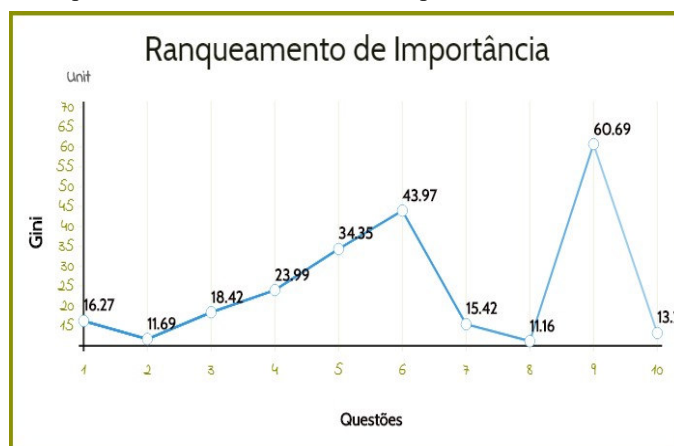


Figura 2. Variação da importância das questões

### Seleção dos Atributos

Com base nos resultados obtidos no ranqueamento de importância realizado é possível deduzir que os demais atributos, além das 10 questões, não apresentem nenhum ganho para o modelo de aprendizado. Além disso, como uma das questões conseguiu um Gini de mais de 60 também é possível questionar se todas as questões são realmente necessárias para um modelo de aprendizado de máquina. À partir disso, quatro cenários foram propostos:

- utilizar toda a base de dados;

- utilizar apenas as 10 questões;
- utilizar as três questões com maior índice Gini;
- utilizar as seis questões com maior índice Gini.

### Aplicação dos Algoritmos de Aprendizado de Máquina

Conforme apresentado na Tabela 1, o conjunto de dados possui uma coluna referente à classificação dada ao indivíduo, indicando se ele possui ou não TEA. Esse elemento binário da base de dados possibilita que sejam utilizados algoritmos que utilizam aprendizado supervisionado para aplicar técnicas de classificação, a fim de, identificar se o indivíduo possui o TEA.

Foram escolhidos três algoritmos distintos para serem aplicados à base de dados. A *Random Forest*, *J48* e a *SVM-Linear*. A SVM foi utilizada com núcleo linear, devido as análises realizadas no início deste trabalho durante a PCA. Como a mesma sugeria que o problema poderia ser linearmente separável, o núcleo linear foi escolhido. De todos os núcleos de SVMs o Linear é o mais simples, uma vez que os problemas lineares têm um grau de complexidade menor que os polinomiais. Por se tratar de um núcleo mais simples, na maior parte das vezes ele acaba tendo um tempo de execução reduzido e um baixo custo computacional [7].

Todos os algoritmos foram treinados utilizando a técnica de *Cross Validation* com 10 *fold*s (dobras). Trata-se de uma técnica matemática utilizada para não induzir nenhum resultado pela disposição dos dados na base de dados utilizada. Ele funciona dividindo os dados em k subconjuntos exclusivos e com tamanhos iguais e a partir disto, um subconjunto é utilizado para teste e os k-1 restantes são utilizados para treinar o modelo. Dessa forma, não é possível colocar os dados em uma disposição específica na base de dados, a fim de melhorar o desempenho [16].

## RESULTADOS

### Classificação dos indivíduos quanto presença de TEA

Nas Tabelas 3 e 4 estão dispostos os resultados obtidos pela classificação dos indivíduos por métodos de classificação, onde foram aplicados três algoritmos: RF, J48 e SVM-linear.

A Tabela 3 mostra os resultados dos dois primeiros cenários propostos, a base completa e apenas as 10 perguntas.

	Base Completa	Apenas as 10 Perguntas
<i>Random Forest</i>	649 - 92.18 %	680 - 96.59 %
<i>J48</i>	652 - 92.61 %	652 - 92.61 %
<i>SVM(Linear)</i>	704 - 100 %	704 - 100%

Tabela 3. Classificação dos indivíduos

A Tabela 4 mostra os resultados dos dois últimos cenários propostos, as 3 questões com maior Gini e as 6 questões com maior Gini respectivamente.

Nota-se na Tabela 3 que as características não relacionadas ao teste AQ-10, não agregam benefícios na classificação dos indivíduos, pelo contrário, acaba interferindo de modo negativo durante a classificação, como por exemplo, no caso

	(Q9,Q6,Q5)	(Q9,Q6,Q5,Q4,Q3,Q1)
<i>Random Forest</i>	624 - 88.63 %	649 - 92.18 %
<i>J48</i>	622 - 88.35 %	646 - 91.76 %
<i>SVM(Linear)</i>	624 - 88.63 %	651 - 92.47%

Tabela 4. Classificação dos indivíduos

do modelo obtido pela *Random Forest*, pode ser visto uma melhora significativa entre os cenários 1 e 2, porém nos obtidos na J48 e SVM os resultados foram basicamente os mesmos em ambos os cenários. Também foi possível observar um melhor rendimento da SVM nesses cenários, onde o mesmo conseguiu classificar corretamente 100% das amostras. Já a *Random Forest* obteve melhor resultado contra a J48 em todos os cenários onde foram analisadas somente as perguntas do teste AQ-10.

A fim de verificar a possibilidade de reduzir a quantidade de perguntas do teste, foram mantidas apenas as três perguntas com maiores índices de Gini, como pode ser visto na Tabela 4, sendo este, o terceiro cenário. Todos os algoritmos foram semelhantes em seus desempenhos, classificando corretamente cerca de 88% dos indivíduos. Esse resultado aproximado nos três algoritmos se deve principalmente à pouca quantidade de perguntas para realizar a classificação, uma vez que a menor quantidade de atributos acaba por diminuir a efetividade dos modelos mais robustos como a RF e a SVM. Em especial no caso da RF, a menor quantidade de atributos representa uma diminuição na variabilidade dos conjuntos gerados pelo método *bagging*, que é o grande diferencial da RF. Devido a isso, nesse cenário o *bagging* acabou não sendo tão eficiente e as árvores geradas pelo modelo acabam ficando mais semelhantes entre si, aproximando o resultado da RF e J48.

No quarto cenário foi utilizado o dobro de perguntas, ou seja, as seis perguntas com maiores índices de Gini para os testes, como pode ser visto na Tabela 4. O aumento na quantidade de informações comparado ao cenário três, possibilitou uma melhora na classificação dos indivíduos, aumentando a quantidade de informações para a geração de um modelo de aprendizado. Além disso, pode-se notar um melhor desempenho da SVM em relação à *Random Forest*, onde havia obtido a mesma performance no cenário com três atributos. Contudo, mesmo dobrando o número de perguntas, no melhor dos cenários o ganho foi inferior a 4%.

Com a aplicação dos algoritmos, pode-se verificar que a linearidade do problema permite que sejam gerados classificadores com pelo menos três algoritmos distintos e foram obtidos bons resultados, até mesmo com uma quantidade limitada de informação, como a que foi utilizada no cenário 3. Nota-se que a SVM com núcleo linear, além de permitir um alto poder de generalização, também se sobressai em praticamente todos os testes realizados, não sofrendo interferência nem mesmo com as informações irrelevantes das amostras, como por exemplo, que afetaram negativamente a *Random Forest*.

## CONCLUSÕES

Durante este trabalho ficou evidente a necessidade de ferramentas e modelos computacionais que possam auxiliar

e ajudar os profissionais da saúde a diagnosticar um problema tão comum e de difícil identificação como é o caso do TEA. Lembrando que nem todos os profissionais da área da saúde têm o conhecimento ou recursos necessários para realizar esse diagnóstico com algum tipo de apoio. Assim, este trabalho será também muito útil aos aprendizes de áreas médicas como psiquiatria e neurologia.

Após realizar várias análises e experimentos, no decorrer deste trabalho pode-se concluir que para essa base de dados, as características físicas e sociais além de não exerceram nenhum tipo de ganho de informação, também acabaram diminuindo a capacidade preditiva dos modelos gerados.

Sobre o AQ-10 teste para adultos, foi possível obter a importância de cada pergunta pertencente ao teste utilizando a *Random Forest* para o ranqueamento da importância das características. Também foi possível comparar a performance do teste usando das perguntas mais importantes, mesmo utilizando apenas as top 3 perguntas, na classificação todos os modelos apresentaram rendimento superior a 88%. Esses resultados podem ser considerados bons em ambos os cenários, haja vista que foram utilizadas apenas 3 das 10 questões. Isso levanta uma questão interessante sobre a análise da importância das perguntas e a necessidade do profissional da saúde de aprender que durante um teste, algumas perguntas podem ter mais relevância do que outras.

Além disso, este modelo de aprendizado pode ser aplicado na construção de ferramentas computacionais voltadas para o ensino de profissionais da área médica ou também como ferramentas de auxílio ao diagnóstico.

Para estudos futuros espera-se aplicar o mesmo modelo desenvolvido neste trabalho a outras duas bases de dados semelhantes a esse, com dados dos testes AQ-10 também para crianças e adolescentes, a fim de verificar se eles apresentam as mesmas características que foram encontradas na base de dados deste trabalho.

## REFERÊNCIAS

- Alencar, C. N., do, K. C. F., Matos, and dos Santos, A. C. *Rel085 - cartilha educativa como importante ferramenta para a detecção precoce de sinais clínicos de autismo*, (2015).
- Allison, C., Auyeung, B., and 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* 51, 2 (2012), 202–212.
- Baron-Cohen, S., Wheelwright, S., Skinner, R., Martin, J., and Clubley, E. The autism-spectrum quotient (aq): Evidence from asperger syndrome/high-functioning autism, males and females, scientists and mathematicians. *Journal of autism and developmental disorders* 31, 1 (2001), 5–17.
- Bone, D., Goodwin, M. S., Black, M. P., Lee, C.-C., Audhkhazi, K., and Narayanan, S. Applying machine learning to facilitate autism diagnostics: pitfalls and promises. *Journal of autism and developmental disorders* 45, 5 (2015), 1121–1136.
- Breiman, L. Random forests. *Machine learning* 45, 1 (2001), 5–32.
- Brito, M. M. V. A. *A contribuição do PECS no desenvolvimento da comunicação de uma aluna com perturbações do espectro do autismo*. Master’s thesis, 2015.
- Cortes, C., and Vapnik, V. Support-vector networks. *Machine Learning* 20, 3 (Sep 1995), 273–297.
- Didehbani, N., Allen, T., Kandalaft, M., Krawczyk, D., and Chapman, S. Virtual reality social cognition training for children with high functioning autism. *Computers in Human Behavior* 62 (2016), 703–711.
- Durkin, M. S., Elsabbagh, M., Barbaro, J., Gladstone, M., Happe, F., Hoekstra, R. A., Lee, L.-C., Rattazzi, A., Stapel-Wax, J., Stone, W. L., Tager-Flusberg, H., Thurm, A., Tomlinson, M., and Shih, A. Autism screening and diagnosis in low resource settings: Challenges and opportunities to enhance research and services worldwide. *Autism Research* 8, 5 (2015), 473–476.
- Ferreira, R. d. S. Autism testing: Uma ferramenta móvel no auxílio ao pré-diagnóstico do autismo. *Conferência Internacional sobre Informática na Educação*, (2010).
- Genuer, R., Poggi, J.-M., and Tuleau-Malot, C. Variable selection using random forests. *Pattern Recognition Letters* 31, 14 (2010), 2225–2236.
- Gomes, P., Lima, L. H., Bueno, M. K., Araújo, L. A., and Souza, N. M. Autism in brazil: a systematic review of family challenges and coping strategies. *Jornal de pediatria* 91, 2 (2015), 111–121.
- Ishak, W. H. W., and Siraj, F. Artificial intelligence in medical application: An exploration. *Health Informatics Europe Journal* 16 (2002).
- Kleinman, J. M., Robins, D. L., Ventola, P. E., Pandey, J., Boorstein, H. C., Esser, E. L., Wilson, L. B., Rosenthal, M. A., Sutura, S., Verbalis, A. D., Barton, M., Hodgson, S., Green, J., Dumont-Mathieu, T., Volkmar, F., Chawarska, K., Klin, A., and Fein, D. 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* 38, 5 (2008), 827–839.
- Klin, A. Autismo e síndrome de asperger: uma visão geral autism and asperger syndrome: an overview. *Revista Brasileira de Psiquiatria* 28, Supl I (2006), S3–11.
- Kohavi, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *IJCAI*, vol. 14, Montreal, Canada (1995), 1137–1145.

17. Lampreia, A. R. d. S. *Percepções parentais sobre a perturbação do espectro do autismo: processo de diagnóstico, interferência e recursos*. PhD thesis, 2015.
18. Penteadó, F. A. d. O., Salvadeo, D. H. P., Aporta, A. P., Garcia, R. V. B., and Manoni, N. d. V. Software para auxílio ao diagnóstico de autismo. In *Congresso de extensão universitária da UNESP*, Universidade Estadual Paulista (UNESP) (2015), 1–4.
19. Rellini, E., Tortolani, D., Trillo, S., Carbone, S., and Montecchi, F. 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* 34, 6 (2004), 703–708.
20. Ringnér, M. What is principal component analysis? *Nature biotechnology* 26, 3 (2008), 303.
21. Salzberg, S. L. C4. 5: Programs for machine learning by j. ross quinlan. morgan kaufmann publishers, inc., 1993. *Machine Learning* 16, 3 (1994), 235–240.
22. Schopler, E., Reichler, R. J., DeVellis, R. F., and Daly, K. Toward objective classification of childhood autism: Childhood autism rating scale (cars). *Journal of autism and developmental disorders* 10, 1 (1980), 91–103.
23. Silva, A., Gaiato, M. B., and Reveles, L. T. *Mundo singular. Entenda o Autismo*. Rio de Janeiro: Editora Fontana (2012).
24. Skafidas, E., Testa, R., Zantomio, D., Chana, G., Everall, I. P., and Pantelis, C. Predicting the diagnosis of autism spectrum disorder using gene pathway analysis. *Molecular psychiatry* 19, 4 (2014), 504–510.
25. Thabtah, F. Autism spectrum disorder screening: machine learning adaptation and dsm-5 fulfillment. In *International Conference on Medical and Health Informatics*, ACM (2017), 1–6.
26. Wall, D., Kosmicki, J., Deluca, T., Harstad, E., and Fusaro, V. Use of machine learning to shorten observation-based screening and diagnosis of autism. *Translational psychiatry* 2(4), (2012), e100.
27. Wilson, R., and Obimbo, C. Self-organizing feature maps for user-to-root and remote-to-local network intrusion detection on the kdd cup 1999 dataset. In *Internet Security (WorldCIS), 2011 World Congress on*, IEEE (2011), 42–47.
28. Xue, M., and Zhu, C. A study and application on machine learning of artificial intelligence. In *Artificial Intelligence, 2009. JCAI'09. International Joint Conference on*, IEEE (2009), 272–274.
29. Zanon, R. B., Backes, B., and Bosa, C. A. Identificação dos primeiros sintomas do autismo pelos pais. *Psicologia: Teoria e Pesquisa* 30, 1 (2014), 25–33.
30. Zuckerman, K. E., Sinche, B., Mejia, A., Cobian, M., Becker, T., and Nicolaidis, C. Latino parents' perspectives on barriers to autism diagnosis. *Academic pediatrics* 14, 3 (2014), 301–308.