

# Análise da Robustez de Algoritmos de Aprendizado de Máquina em Dados do Transtorno do Espectro Autista

Saulo B. F. Lino<sup>1</sup>, Lívia A. Cruz<sup>1</sup>, Paulo T. Guerra<sup>1</sup>

<sup>1</sup>Universidade Federal do Ceará (UFC)  
Campus de Quixadá – Quixadá, CE – Brasil

saulobruno@alu.ufc.br, livia.almada@ufc.br, paulodetarso@ufc.br

**Abstract.** *Autism Spectrum Disorder (ASD) is a neurological condition that affects communication, social interaction, behavior, and learning. Screening methods such as AQ and Q-CHAT have been developed to speed up the identification of autistic signs. The present work analyzes the performance of machine learning algorithms in ASD screening, such as SVM, MLP, Logistic Regression, Naive Bayes, Random Forest, and KNN, and the robustness of these models in the face of possible errors in the data. The algorithms are evaluated on datasets with samples based on personal characteristics and simplified questions from the AQ and Q-CHAT instruments. The experiments show good performance of the SVM, MLP, and Logistic Regression methods, but with a significant reduction in their accuracy in scenarios with errors.*

**Resumo.** *O Transtorno do Espectro Autista (TEA) é uma condição neurológica que afeta a comunicação, interação social, comportamento e aprendizado. Métodos de triagem como AQ e Q-CHAT foram desenvolvidos para agilizar a identificação de sinais autistas. O presente trabalho analisa o desempenho de algoritmos de aprendizado de máquina na triagem do TEA, tais como SVM, MLP, Regressão Logística, Naive Bayes, Floresta Aleatória e KNN, e a robustez destes modelos diante de possíveis erros nos dados. Os algoritmos são avaliados em conjuntos de dados com amostras baseadas em características pessoais e questões simplificadas dos instrumentos AQ e Q-CHAT. Os experimentos apontam um bom desempenho obtido pelos métodos SVM, MLP e Regressão Logística, porém com significativa redução da acurácia em cenários com erros.*

## 1. Introdução

O Transtorno do Espectro Autista (TEA) é um distúrbio do desenvolvimento neurológico que afeta a forma como as pessoas se comunicam, interagem com os outros, se comportam e aprendem [Thabtah et al. 2019]. Os sinais desse transtorno surgem ainda na infância e persistem ao longo da vida, sendo caracterizado por prejuízos no desenvolvimento de habilidades socio-comunicativas, habilidades cognitivas e por comportamentos e interesses repetitivos ou restritos [APA 2013].

O processo de diagnóstico oficial de indivíduos com autismo é longo, exigindo recursos clínicos e métodos de diagnóstico tais como *Autism Diagnostic Interview Revised* (ADI-R) [Lord et al. 1994] e *Autism Diagnostic Observation Schedule Revised* (ADOS-R) [Lord et al. 2000]. Para agilizar o encaminhamento de indivíduos que apresentam sinais dentro do espectro para uma avaliação mais aprofundada, foram desen-

volvidos métodos de triagem autoadministrados ou administrados pelos pais ou cuidadores, principalmente com base em questionários, como o *Autistic Quotient* (AQ) [Baron-Cohen et al. 2001] e *Quantitative Checklist for Autism in Toddlers* (Q-CHAT) [Allison et al. 2012].

A qualidade da classificação resultante para os indivíduos submetidos a essa triagem depende principalmente de três fatores cruciais: (1) a concepção dos itens no método de triagem, (2) a experiência e o conhecimento do usuário encarregado da triagem e, mais importante, (3) as regras criadas manualmente que estão associadas à função de pontuação [Thabtah and Peebles 2020].

Este estudo analisa o desempenho de algoritmos de aprendizado de máquina na tarefa de triagem do TEA e investiga a robustez dos modelos ao introduzir erros artificiais nos dados. A introdução dos erros busca simular possíveis equívocos cometidos pelos pais ou cuidadores ao responder os questionários, dado que certos comportamentos podem passar despercebidos até mesmo pelas pessoas mais próximas ao paciente. Essa estratégia permite avaliar como tais modelos de aprendizado reagem diante de situações de erro na coleta de dados.

## 2. Escalas Diagnósticas para o TEA

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 porque esses exames não existem ou por serem muito complexos e demorados [Ferreira 2010].

Infelizmente, o processo de diagnóstico de indivíduos com autismo é longo, exigindo recursos clínicos e métodos de diagnóstico tais como *Autism Diagnostic Interview Revised* (ADI-R) [Lord et al. 1994] e *Autism Diagnostic Observation Schedule Revised* (ADOS-R) [Lord et al. 2000]. A aplicação desses métodos exige que profissionais de saúde realizem uma avaliação clínica abrangente, levando em consideração diversos domínios, incluindo comportamento excessivo, comunicação, autocuidado e habilidades sociais [Thabtah et al. 2018]. Por conta disso, acredita-se que muitas mais pessoas que estão no espectro permanecem indetectadas [Fitzgerald 2017].

Assim, as ferramentas de triagem, como o *Autism Quotient* (AQ) e *Quantitative Checklist for Autism in Toddlers* (Q-CHAT), surgiram com o objetivo principal de identificar indivíduos que apresentam possíveis sinais de TEA e encaminhá-los para avaliações diagnósticas detalhadas realizadas por profissionais de saúde qualificados [Baron-Cohen et al. 2001, Kleinman et al. 2008]. O *Autism Quotient - 10* e o *Quantitative Checklist for Autism in Toddlers - 10* (Q-CHAT-10) são versões mais concisas destas escalas e de mais rápida aplicação [Allison et al. 2012, Kleinman et al. 2008].

### 2.1. *Autism Quotient - 10* (AQ - 10)

A escala *Autism Quotient* (AQ) é uma ferramenta de triagem que permite avaliar a presença de traços autistas em crianças, adolescentes e adultos. Elaborada por [Baron-Cohen et al. 2001], a escala AQ é um questionário autoadministrado que aborda uma série de comportamentos sociais, interesses e características de comunicação relacionados ao TEA. Com um total de 50 itens, os indivíduos preenchem o questionário

e, com base em suas respostas, obtêm uma pontuação que pode indicar a presença de características autistas [Baron-Cohen et al. 2001].

A escala *Autistic Quotient - 10* (AQ - 10) é uma versão abreviada da escala AQ, projetada para avaliar traços autistas em crianças, adolescentes e adultos. Composta por apenas 10 itens. A AQ-10 oferece uma abordagem mais rápida e simplificada para identificar possíveis traços autistas. Esta escala é flexível em sua aplicação, permitindo que os itens sejam preenchidos por pais, cuidadores ou pelos próprios indivíduos, tornando-a uma ferramenta versátil de triagem em um contexto clínico [Allison et al. 2012]. A Tabela 1 apresenta a síntese das questões do AQ-10.

**Tabela 1. Síntese do questionário AQ-10 aplicado a crianças de 4 a 11 anos.**

<b>Atributo</b>	<b>Descrição</b>
A1	Ele/ela costuma notar pequenos sons quando outros não notam
A2	Ele/ela geralmente se concentra mais na imagem como um todo, em vez dos pequenos detalhes
A3	Em um grupo social, ele/ela pode acompanhar facilmente várias conversas diferentes
A4	Ele/ela acha fácil alternar entre diferentes atividades
A5	Ele/ela não sabe como manter uma conversa com seus pares
A6	Ele/ela é bom em conversa social
A7	Quando lhe é contada uma história, ele/ela acha difícil entender as intenções ou sentimentos dos personagens
A8	Quando ele/ela estava na pré-escola, ele/ela costumava gostar de brincar de faz de conta com outras crianças
A9	Ele/ela acha fácil entender o que alguém está pensando ou sentindo apenas olhando para o rosto deles
A10	Ele/ela acha difícil fazer novos amigos

Nessa escala, respostas positivas para as questões A1, A5, A7 e A10 ou negativas para as perguntas A2, A3, A4, A6, A8 e A9 são consideradas indicativas de traços autistas. A escala sugere a necessidade de uma investigação mais aprofundada quando 7 ou mais traços são identificados. Embora a AQ-10 ofereça uma maneira rápida de identificar possíveis traços autistas, é importante ressaltar que ela é uma ferramenta de triagem e não um instrumento de diagnóstico definitivo [Allison et al. 2012].

## **2.2. Quantitative-Checklist for Autism in Toddlers - 10 (Q-CHAT-10)**

O *Quantitative-Checklist for Autism in Toddlers* (Q-CHAT) é um instrumento de triagem desenvolvido para identificar sinais precoces de autismo em crianças com idades entre 18 e 24 meses. Ele é preenchido pelos pais ou cuidadores e consiste em 25 itens que avaliam comportamentos associados ao espectro do autismo, como interação social, comunicação e comportamentos repetitivos [Allison et al. 2012].

O *Quantitative-Checklist for Autism in Toddlers - 10* (Q-CHAT-10) é uma versão abreviada do Q-CHAT, com apenas 10 itens. Ele é um instrumento de triagem desenvolvido para identificar sinais precoces de autismo em crianças pequenas com idades entre

18 e 24 meses [Allison et al. 2008]. O Q-CHAT-10 é preenchido pelos pais ou cuidadores e avalia comportamentos associados ao autismo, como interação social, comunicação e comportamentos repetitivos. A Tabela 2 apresenta a síntese das questões do Q-CHAT-10.

**Tabela 2. Síntese do questionário Q-CHAT-10 aplicado a crianças de 18 e 24 meses.**

<b>Atributo</b>	<b>Descrição</b>
A1	Seu filho olha para você quando você chama pelo nome dele/dela?
A2	Quão fácil é para você conseguir contato visual com seu filho?
A3	Seu filho aponta para indicar que ele/ela quer algo? (por exemplo, um brinquedo fora de alcance)
A4	Seu filho aponta para compartilhar interesse com você? (por exemplo, apontando para uma visão interessante)
A5	Seu filho finge? (por exemplo, cuidar de bonecas, falar em um telefone de brinquedo)
A6	Seu filho segue para onde você está olhando?
A7	Se você ou outra pessoa da família estiver visivelmente chateada, seu filho mostra sinais de querer confortá-los? (por exemplo, acariciar o cabelo, abraçá-los)
A8	Você descreveria as primeiras palavras do seu filho como muito ou bastante comuns?
A9	Seu filho usa gestos simples? (por exemplo, acenar adeus)
A10	Seu filho fica olhando para o nada sem um propósito aparente?

Nessa escala, respostas positivas para as questões A1 a A9 ou resposta negativa para a pergunta A10 são consideradas indicativas de traços autistas. A escala sugere a necessidade de uma investigação mais aprofundada quando 4 ou mais traços são identificados. Assim como o AQ-10, o Q-CHAT-10 não é um instrumento diagnóstico definitivo. Ele serve como uma ferramenta de triagem eficaz, permitindo a identificação precoce de sinais de autismo e facilitando encaminhamentos para avaliações diagnósticas mais abrangentes, se necessário. Profissionais de saúde qualificados devem conduzir essas avaliações para confirmar ou descartar a presença de TEA.

### **3. Análise do desempenho de algoritmos de aprendizado de máquina em bases de dados de triagem de autismo**

#### **3.1. Conjunto de dados**

Os algoritmos são avaliados em 2 conjuntos de dados coletados usando um aplicativo móvel chamado *ASDTests* [Thabtah 2017], cujas amostras são baseadas em características pessoais dos pacientes e nas questões dos métodos de triagem Q-CHAT-10 e AQ-Criança-10, versões simplificadas dos instrumentos AQ e Q-CHAT. Este aplicativo contém quatro questionários baseados nos métodos de triagem Q-CHAT-10, AQ-Criança-10, AQ-Adolescente-10 e AQ-Adulto-10, que atendem a diferentes públicos-alvo (crianças entre 18 e 24 meses, crianças entre 4 e 11 anos, adolescentes e adultos, respectivamente).

Foram utilizadas duas bases de dados, BASE-AQ-10 e BASE-Q-CHAT-10, que contêm características pessoais dos pacientes e suas respostas aos questionários de triagem AQ-Criança-10 e Q-CHAT-10, respectivamente. A base BASE-AQ-10 contém 509 amostras de crianças entre 4 e 11 anos, sendo dividida em duas classes: 257 possuem TEA e 252 não possuem TEA. Já a base BASE-Q-CHAT-10 contém 1054 amostras de crianças entre 12 e 36 meses, sendo dividida em duas classes: 728 possuem TEA e 326 não possuem TEA. Ambas as bases de dados estão disponíveis na plataforma Kaggle<sup>1</sup>.

Os dois conjuntos de dados têm dez atributos binários que representam as respostas às perguntas do método de triagem (A1 a A10), bem como outros atributos relativos a gênero, etnia, icterícia, TEA na família, residência, idade e pontuação/resultados da triagem.

As respostas às perguntas dos métodos de triagem (A1 até A10) foram previamente convertidas em 0 ou 1 antes da disponibilização da base, com base na opção selecionada em uma escala para cada resposta pelo usuário que completou o questionário. Respostas resultaram em atribuição de um ponto quando positivas para as perguntas 1, 5, 7 e 10 e negativas para as perguntas 2, 3, 4, 6, 8 e 9 na BASE-AQ-10. Já na base BASE-Q-CHAT-10, exceto para a questão 10, respostas negativas resultaram em atribuição de um ponto.

O valor da classe foi atribuído durante o processo de coleta de dados pelo aplicativo com base nas pontuações das respostas às perguntas dos métodos de triagem. O valor da classe “Não” foi atribuído quando a pontuação final do método AQ-Criança-10 era menor que 7. Caso contrário, era atribuído “Sim”, o que indicava que a pessoa deveria proceder com um diagnóstico mais aprofundado. Para o método Q-CHAT-10, a pontuação de corte era menor que 4. Portanto, nesse caso, se a pontuação total foi maior ou igual a 4, foi indicado que a criança precisaria prosseguir com um diagnóstico mais aprofundado.

### 3.2. Preparação dos dados

A segunda etapa realizada foi o pré-processamento dos dados. Nesta etapa, diversas tarefas foram executadas para preparar os conjuntos de dados para o treinamento e teste dos modelos. Aplicamos engenharia de atributos para transformar os dados originais em formatos mais apropriados.

Nesta etapa, foram realizados os seguintes procedimentos para o conjunto de dados BASE-AQ-10: os atributos que são meta-informações (“Why taken the screening”, “Used\_App\_Before”, “Screening Type”, “Residence”, “Language” e “User”) foram removidos; o atributo “Score” foi removido; o atributo categórico “Ethnicity” foi codificado numericamente; os atributos categóricos “Sex”, “Family\_ASD”, “Jaundice” e o rótulo “Class” tiveram seus valores codificados em 0 e 1; normalizamos os conjuntos de dados com a *Standard Scaler* para o experimento com KNN.

Já para o conjunto de dados BASE-Q-CHAT-10, foram realizados os seguintes procedimentos: os atributos que são meta-informações foram removidos; o atributo

---

<sup>1</sup>Bases de dados disponíveis em <https://www.kaggle.com/datasets/basmarg/autism-screening-child-two-version> (BASE-AQ-10) e <https://www.kaggle.com/datasets/fabdelja/autism-screening-for-toddlers> (BASE-Q-CHAT-10)

“Qchat-10-Score” foi removido; o atributo categórico “Ethnicity” foi codificado utilizando a técnica *One-Hot-Encoding* ou codificação numérica, dependendo do algoritmo que estava sendo executado; os atributos categóricos “Sex”, “Family\_mem\_with\_ASD”, “Jaundice” e o rótulo “Class/ASD Traits” tiveram seus valores codificados em 0 e 1; normalizamos o conjunto de dados com a técnica *MinMax Scaler* para o experimento com método KNN.

### 3.3. Execução dos algoritmos de aprendizado de máquina

Os algoritmos de aprendizado de máquina avaliados na classificação de triagem do autismo foram: *Support Vector Machine* com *kernel* linear (SVM), *Multilayer Perceptron* (MLP), Regressão Logística (RL), *Naive Bayes* (NB), Floresta Aleatória (RF) e *K-Nearest Neighbors* (KNN).

Para treinamento do modelo, foi utilizado 70% do conjunto de dados, reservado o restante para os testes. Os hiperparâmetros foram otimizados por meio da aplicação de uma *GridSearch* com validação cruzada *k-fold* utilizando *k* igual a 5. A acurácia foi usada como métrica para seleção dos melhores hiperparâmetros.

A avaliação dos modelos utilizou as métricas acurácia, precisão, revocação e *F1-measure* em dez cenários de testes. Os resultados dos algoritmos para cada uma das bases de dados são sintetizados nas Tabelas 3 e 4.

**Tabela 3. Tabela comparativa de desempenho dos algoritmos (BASE-Q-CHAT-10).**

Métrica	RF	NB	RL	MLP	KNN	SVM
Acurácia	0.96 ± 0.01	0.95 ± 0.01	<b>1.00 ± 0.00</b>	<b>1.00 ± 0.00</b>	0.96 ± 0.01	<b>1.00 ± 0.00</b>
Precisão	0.96 ± 0.02	0.96 ± 0.01	<b>1.00 ± 0.00</b>	<b>1.00 ± 0.00</b>	0.98 ± 0.01	<b>1.00 ± 0.00</b>
Revocação	0.99 ± 0.01	0.97 ± 0.01	<b>1.00 ± 0.00</b>	<b>1.00 ± 0.00</b>	0.96 ± 0.01	<b>1.00 ± 0.00</b>
<i>F1-measure</i>	0.97 ± 0.01	0.97 ± 0.01	<b>1.00 ± 0.00</b>	<b>1.00 ± 0.00</b>	0.97 ± 0.01	<b>1.00 ± 0.00</b>

**Tabela 4. Tabela comparativa de desempenho dos algoritmos (BASE-AQ-10).**

Métrica	RF	NB	RL	MLP	KNN	SVM
Acurácia	0.95 ± 0.01	0.91 ± 0.02	<b>1.00 ± 0.00</b>	<b>1.00 ± 0.00</b>	0.93 ± 0.02	<b>1.00 ± 0.00</b>
Precisão	0.95 ± 0.02	0.93 ± 0.03	<b>1.00 ± 0.00</b>	<b>1.00 ± 0.00</b>	0.92 ± 0.03	<b>1.00 ± 0.00</b>
Revocação	0.95 ± 0.02	0.88 ± 0.04	<b>1.00 ± 0.00</b>	<b>1.00 ± 0.00</b>	0.95 ± 0.02	<b>1.00 ± 0.00</b>
<i>F1-measure</i>	0.95 ± 0.01	0.91 ± 0.03	<b>1.00 ± 0.00</b>	<b>1.00 ± 0.00</b>	0.93 ± 0.02	<b>1.00 ± 0.00</b>

Podemos observar que os modelos de Regressão Logística, SVM com *kernel* linear e MLP obtiveram pontuações perfeitas em todas as métricas para ambas as bases de dados. Além disso, é possível perceber que os demais modelos, até mesmo *Naive Bayes*, que se trata de um modelo mais básico, obtiveram resultados excelentes, com desempenho acima de 0.95 em todas as métricas, para ambas as bases de dados. Esses resultados estão coerentes com aqueles apresentados em [Hossain et al. 2021, Artoni et al. 2022, Garg et al. 2022].

Os algoritmos, de certo modo, estão sendo bem sucedidos em replicar as regras de pontuação atribuídas aos métodos de triagem Q-CHAT-10 e AQ-Criança-10, de forma que se a soma dos valores dos atributos “A1” até “A10” for maior ou igual a 4 para

BASE-Q-CHAT-10 ou maior ou igual ou maior ou igual a 7 para a BASE-AQ-10, os modelos indicam que o paciente necessita de uma investigação mais aprofundada sobre TEA. Contudo, não está claro quando os algoritmos estão efetivamente considerando os valores dos atributos “A1” até “A10” individualmente, para além de sua soma.

Devido a respostas inadvertidamente respondidas de forma errada, as bases de dados de triagem de autismo podem apresentar erros em elementos principais para diagnósticos. Nesse sentido, um algoritmo robusto deve ser capaz de indicar a necessidade de investigação de TEA, sempre que os elementos relevantes ao diagnóstico estivessem presentes, ainda que por ventura pontuação total seja inferior a 7.

Investigaremos assim a robustez dos modelos introduzindo erros artificiais no conjunto de dados, simulando possíveis equívocos cometidos pelos pais ou cuidadores ao responder os questionários, dado que certos comportamentos podem passar despercebidos até mesmo pelas pessoas mais próximas ao paciente. Essa avaliação nos ajudará a entender como os modelos reagem diante de situações de erro na coleta de dados.

#### **4. Análise da robustez de algoritmos de aprendizado de máquina em bases de dados de triagem de autismo com erros**

Para avaliar a robustez dos algoritmos de aprendizado de máquina selecionados, inserimos erros artificiais nas bases de dados AQ-10 e Q-CHAT-10, simulando um cenário comum a área da saúde, onde as bases de dados obtidas podem apresentar erros devido a respostas equivocadas fornecidas pelos pacientes.

As respostas presentes nas novas bases de dados foram deliberadamente alteradas de 0 para 1 ou de 1 para 0 a fim de inserir erros artificiais. Variamos a quantidade de erros por amostra (i.e., a quantidade de respostas alteradas de 0 para 1 ou de 1 para 0 para cada amostra) e a porcentagem de amostras do conjunto de dados com erros a fim de contemplar uma ampla gama de cenários de erros na coleta de dados.

Para os experimentos, foram gerados dez cenários de testes aleatórios distintos, onde calculamos a média e desvio padrão do desempenho dos algoritmos de acordo com as métricas acurácia, precisão, revocação e *F1-measure*. As Tabelas 5 e 6 apresentam os resultados de acurácia para as bases BASE-Q-CHAT-10 e BASE-AQ-10, respectivamente. Por limitações de espaço, omitimos as tabelas com os resultados de precisão, revocação e *F1-measure*.

##### **4.1. Análise da acurácia para a base de dados Q-CHAT-10**

Observando os resultados da Tabela 5, vemos que para a base BASE-Q-CHAT-10 os algoritmos RL, MLP e SVM apresentam os melhores resultados gerais de acurácia entre os casos analisados. Nos cenários onde 1 erro foi introduzido, esses algoritmos obtiveram resultados de acurácia entre 0.99 e 0.95, com até 0.01 de variação no desvio padrão. É interessante observar contudo que os algoritmos RF, NB e KNN apesar de apresentarem resultados gerais piores, entre 0.95 e 0.93, obtiveram uma menor diferença de variação, apenas 0.02 entre valores mínimos e máximos.

Nos cenários onde 2 erros são introduzidos, os melhores resultados são alcançados pelo SVM, ainda que RL e MLP tenham obtidos resultados percentualmente próximos. Observamos contudo uma maior variação de resultados nesses algoritmos, de 0.99 a 0.92.

**Tabela 5. Acurácia dos algoritmos em cenários com erros (BASE-Q-CHAT-10).**

Erros	Alter.	RF	NB	RL	MLP	KNN	SVM
1	5%	0.95 ± 0.01	0.94 ± 0.01	<b>0.99 ± 0.00</b>	<b>0.99 ± 0.00</b>	0.95 ± 0.00	<b>0.99 ± 0.00</b>
	10%	0.96 ± 0.01	0.94 ± 0.01	<b>0.99 ± 0.00</b>	0.99 ± 0.01	0.96 ± 0.01	<b>0.99 ± 0.00</b>
	15%	0.95 ± 0.01	0.93 ± 0.01	<b>0.98 ± 0.00</b>	<b>0.98 ± 0.00</b>	0.95 ± 0.01	0.98 ± 0.01
	20%	0.95 ± 0.01	0.95 ± 0.01	0.98 ± 0.01	<b>0.98 ± 0.00</b>	0.95 ± 0.01	0.98 ± 0.01
	30%	0.94 ± 0.01	0.94 ± 0.01	0.96 ± 0.01	0.96 ± 0.01	0.94 ± 0.01	<b>0.97 ± 0.01</b>
	40%	0.94 ± 0.01	0.94 ± 0.01	<b>0.96 ± 0.01</b>	0.95 ± 0.01	0.94 ± 0.01	<b>0.96 ± 0.01</b>
	50%	0.93 ± 0.01	0.94 ± 0.01	<b>0.95 ± 0.01</b>	<b>0.95 ± 0.01</b>	0.93 ± 0.01	<b>0.95 ± 0.01</b>
2	5%	0.95 ± 0.01	0.94 ± 0.01	<b>0.99 ± 0.00</b>	<b>0.99 ± 0.00</b>	0.95 ± 0.01	<b>0.99 ± 0.00</b>
	10%	0.95 ± 0.01	0.94 ± 0.01	0.98 ± 0.01	0.98 ± 0.01	0.95 ± 0.01	<b>0.99 ± 0.01</b>
	15%	0.94 ± 0.01	0.92 ± 0.01	0.97 ± 0.01	0.98 ± 0.01	0.94 ± 0.01	<b>0.98 ± 0.00</b>
	20%	0.94 ± 0.01	0.93 ± 0.01	0.97 ± 0.01	0.97 ± 0.01	0.94 ± 0.01	<b>0.98 ± 0.01</b>
	30%	0.92 ± 0.01	0.93 ± 0.01	0.95 ± 0.01	0.95 ± 0.01	0.92 ± 0.01	<b>0.96 ± 0.02</b>
	40%	0.92 ± 0.01	0.92 ± 0.01	0.94 ± 0.02	0.94 ± 0.01	0.92 ± 0.02	<b>0.95 ± 0.01</b>
	50%	0.90 ± 0.01	0.92 ± 0.01	<b>0.93 ± 0.01</b>	0.92 ± 0.01	0.91 ± 0.01	<b>0.93 ± 0.01</b>
3	5%	0.94 ± 0.01	0.94 ± 0.01	0.98 ± 0.01	0.98 ± 0.01	0.95 ± 0.01	<b>0.99 ± 0.00</b>
	10%	0.94 ± 0.01	0.93 ± 0.01	0.97 ± 0.01	0.97 ± 0.01	0.95 ± 0.01	<b>0.98 ± 0.01</b>
	15%	0.93 ± 0.01	0.92 ± 0.01	0.96 ± 0.01	0.96 ± 0.01	0.93 ± 0.01	<b>0.97 ± 0.01</b>
	20%	0.93 ± 0.01	0.93 ± 0.01	<b>0.95 ± 0.01</b>	<b>0.95 ± 0.01</b>	0.92 ± 0.02	<b>0.95 ± 0.01</b>
	30%	0.90 ± 0.02	0.91 ± 0.01	<b>0.92 ± 0.02</b>	<b>0.92 ± 0.02</b>	0.90 ± 0.01	<b>0.92 ± 0.02</b>
	40%	0.89 ± 0.02	0.90 ± 0.02	<b>0.91 ± 0.02</b>	<b>0.91 ± 0.02</b>	0.89 ± 0.02	0.90 ± 0.01
	50%	0.87 ± 0.01	<b>0.89 ± 0.01</b>	0.89 ± 0.02	0.89 ± 0.02	0.88 ± 0.01	0.89 ± 0.02
4	5%	0.94 ± 0.01	0.93 ± 0.01	<b>0.98 ± 0.01</b>	<b>0.98 ± 0.01</b>	0.94 ± 0.01	<b>0.98 ± 0.01</b>
	10%	0.93 ± 0.01	0.93 ± 0.02	0.96 ± 0.01	<b>0.97 ± 0.01</b>	0.94 ± 0.01	<b>0.97 ± 0.01</b>
	15%	0.92 ± 0.01	0.91 ± 0.01	<b>0.95 ± 0.01</b>	<b>0.95 ± 0.01</b>	0.92 ± 0.01	<b>0.95 ± 0.01</b>
	20%	0.91 ± 0.01	0.90 ± 0.01	<b>0.93 ± 0.01</b>	<b>0.93 ± 0.01</b>	0.91 ± 0.01	<b>0.93 ± 0.01</b>
	30%	0.87 ± 0.02	0.88 ± 0.01	0.89 ± 0.02	0.89 ± 0.02	0.88 ± 0.01	<b>0.89 ± 0.01</b>
	40%	0.86 ± 0.01	0.86 ± 0.01	<b>0.87 ± 0.02</b>	0.86 ± 0.02	0.86 ± 0.02	<b>0.87 ± 0.02</b>
	50%	0.83 ± 0.01	<b>0.84 ± 0.01</b>	<b>0.84 ± 0.01</b>	<b>0.84 ± 0.01</b>	0.83 ± 0.02	<b>0.84 ± 0.01</b>
5	5%	0.94 ± 0.01	0.93 ± 0.01	0.97 ± 0.01	0.97 ± 0.01	0.94 ± 0.01	<b>0.98 ± 0.01</b>
	10%	0.93 ± 0.01	0.92 ± 0.02	0.95 ± 0.01	<b>0.96 ± 0.01</b>	0.93 ± 0.01	<b>0.96 ± 0.01</b>
	15%	0.91 ± 0.02	0.90 ± 0.01	0.93 ± 0.01	<b>0.94 ± 0.01</b>	0.90 ± 0.01	0.93 ± 0.01
	20%	0.90 ± 0.01	0.89 ± 0.01	0.91 ± 0.01	<b>0.92 ± 0.01</b>	0.89 ± 0.02	0.91 ± 0.01
	30%	0.84 ± 0.02	0.84 ± 0.01	<b>0.86 ± 0.02</b>	<b>0.86 ± 0.02</b>	0.85 ± 0.01	<b>0.86 ± 0.02</b>
	40%	0.82 ± 0.02	0.82 ± 0.01	<b>0.83 ± 0.02</b>	<b>0.83 ± 0.02</b>	0.82 ± 0.01	<b>0.83 ± 0.02</b>
	50%	0.80 ± 0.01	0.80 ± 0.02	<b>0.81 ± 0.01</b>	0.80 ± 0.02	0.79 ± 0.01	<b>0.81 ± 0.01</b>



**Tabela 6. Acurácia dos algoritmos em cenários com erros (BASE-AQ-10).**

Erros	Alter.	RF	NB	RL	MLP	KNN	SVM
1	5%	0.95 ± 0.01	0.93 ± 0.03	<b>0.99 ± 0.01</b>	<b>0.99 ± 0.01</b>	0.93 ± 0.02	<b>0.99 ± 0.01</b>
	10%	0.96 ± 0.01	0.92 ± 0.02	<b>0.98 ± 0.01</b>	<b>0.98 ± 0.01</b>	0.92 ± 0.01	<b>0.98 ± 0.01</b>
	15%	0.95 ± 0.01	0.89 ± 0.02	<b>0.97 ± 0.01</b>	<b>0.97 ± 0.01</b>	0.91 ± 0.02	0.97 ± 0.02
	20%	<b>0.95 ± 0.01</b>	0.89 ± 0.02	0.94 ± 0.02	0.94 ± 0.02	0.90 ± 0.03	<b>0.95 ± 0.01</b>
	30%	0.94 ± 0.01	0.91 ± 0.03	<b>0.95 ± 0.01</b>	0.95 ± 0.02	0.91 ± 0.03	0.94 ± 0.02
	40%	<b>0.94 ± 0.01</b>	0.89 ± 0.03	0.93 ± 0.01	0.93 ± 0.01	0.90 ± 0.02	0.93 ± 0.02
	50%	<b>0.93 ± 0.01</b>	0.89 ± 0.02	0.93 ± 0.02	0.93 ± 0.02	0.90 ± 0.02	0.92 ± 0.02
2	5%	0.95 ± 0.01	0.92 ± 0.03	0.98 ± 0.01	0.98 ± 0.01	0.92 ± 0.02	<b>0.99 ± 0.01</b>
	10%	0.95 ± 0.01	0.92 ± 0.02	0.97 ± 0.01	0.97 ± 0.01	0.92 ± 0.03	<b>0.98 ± 0.01</b>
	15%	0.94 ± 0.01	0.89 ± 0.02	0.94 ± 0.02	0.96 ± 0.01	0.90 ± 0.03	<b>0.97 ± 0.01</b>
	20%	<b>0.94 ± 0.01</b>	0.87 ± 0.02	0.92 ± 0.02	0.93 ± 0.02	0.88 ± 0.02	<b>0.94 ± 0.01</b>
	30%	<b>0.92 ± 0.01</b>	0.88 ± 0.02	0.91 ± 0.02	0.91 ± 0.02	0.87 ± 0.02	0.91 ± 0.02
	40%	<b>0.92 ± 0.01</b>	0.87 ± 0.03	0.88 ± 0.02	0.88 ± 0.02	0.85 ± 0.02	0.89 ± 0.02
	50%	<b>0.90 ± 0.01</b>	0.87 ± 0.02	0.88 ± 0.01	0.88 ± 0.03	0.85 ± 0.03	0.88 ± 0.02
3	5%	0.94 ± 0.01	0.91 ± 0.03	0.97 ± 0.02	<b>0.98 ± 0.01</b>	0.91 ± 0.02	<b>0.98 ± 0.01</b>
	10%	0.94 ± 0.01	0.91 ± 0.02	0.95 ± 0.02	<b>0.96 ± 0.02</b>	0.89 ± 0.03	<b>0.96 ± 0.02</b>
	15%	<b>0.93 ± 0.01</b>	0.88 ± 0.02	0.92 ± 0.02	0.93 ± 0.02	0.88 ± 0.02	0.93 ± 0.02
	20%	<b>0.93 ± 0.01</b>	0.86 ± 0.03	0.89 ± 0.02	0.89 ± 0.02	0.85 ± 0.01	0.90 ± 0.02
	30%	<b>0.90 ± 0.02</b>	0.85 ± 0.02	0.86 ± 0.02	0.86 ± 0.02	0.85 ± 0.03	0.85 ± 0.02
	5%	<b>0.89 ± 0.02</b>	0.82 ± 0.02	0.83 ± 0.02	0.82 ± 0.03	0.81 ± 0.04	0.83 ± 0.02
	5%	<b>0.87 ± 0.01</b>	0.81 ± 0.03	0.82 ± 0.02	0.81 ± 0.03	0.79 ± 0.02	0.81 ± 0.01
4	5%	0.94 ± 0.01	0.91 ± 0.03	0.96 ± 0.01	0.97 ± 0.01	0.91 ± 0.02	<b>0.98 ± 0.01</b>
	10%	0.93 ± 0.01	0.91 ± 0.02	0.93 ± 0.02	<b>0.95 ± 0.01</b>	0.89 ± 0.02	0.95 ± 0.02
	15%	<b>0.92 ± 0.01</b>	0.87 ± 0.03	0.90 ± 0.03	0.92 ± 0.02	0.86 ± 0.02	0.90 ± 0.02
	20%	<b>0.91 ± 0.01</b>	0.86 ± 0.01	0.86 ± 0.03	0.88 ± 0.02	0.83 ± 0.02	0.87 ± 0.02
	30%	<b>0.87 ± 0.02</b>	0.82 ± 0.01	0.82 ± 0.02	0.82 ± 0.02	0.80 ± 0.03	0.82 ± 0.02
	40%	<b>0.86 ± 0.01</b>	0.78 ± 0.03	0.78 ± 0.02	0.77 ± 0.02	0.77 ± 0.03	0.78 ± 0.02
	50%	<b>0.83 ± 0.01</b>	0.75 ± 0.02	0.75 ± 0.03	0.75 ± 0.02	0.72 ± 0.04	0.75 ± 0.03
5	5%	0.94 ± 0.01	0.90 ± 0.03	0.96 ± 0.02	<b>0.97 ± 0.01</b>	0.90 ± 0.02	<b>0.97 ± 0.01</b>
	10%	0.93 ± 0.01	0.90 ± 0.02	0.92 ± 0.02	<b>0.94 ± 0.01</b>	0.88 ± 0.01	0.93 ± 0.01
	15%	<b>0.91 ± 0.02</b>	0.86 ± 0.02	0.88 ± 0.01	0.90 ± 0.01	0.85 ± 0.02	0.89 ± 0.01
	20%	<b>0.90 ± 0.01</b>	0.83 ± 0.03	0.84 ± 0.03	0.86 ± 0.02	0.82 ± 0.02	0.84 ± 0.02
	30%	<b>0.84 ± 0.02</b>	0.78 ± 0.02	0.78 ± 0.01	0.79 ± 0.01	0.76 ± 0.03	0.78 ± 0.02
	40%	<b>0.82 ± 0.02</b>	0.73 ± 0.02	0.73 ± 0.02	0.73 ± 0.03	0.73 ± 0.04	0.73 ± 0.02
	50%	<b>0.80 ± 0.01</b>	0.69 ± 0.02	0.68 ± 0.02	0.70 ± 0.02	0.68 ± 0.03	0.69 ± 0.02

Novamente, RF, NB e KNN apresentaram uma menor variação, sendo 0.04 para RF e KNN, e apenas 0.02 para NB.

Nos cenários onde 5 erros são introduzidos, os melhores resultados são novamente alcançados pelos algoritmos RL, MLP e SVM, agora com variações que chegaram a 0.17. Diferente dos cenários anteriores, as variações de RF, NB e KNN também foram expressivas, sendo 0.14, 0.13 e 0.15, respectivamente.

#### 4.2. Análise da acurácia para a base de dados AQ-10

Observando agora os resultados da Tabela 6, vemos que para a base BASE-AQ-10 os algoritmos MLP e SVM apresentam os melhores resultados gerais de acurácia entre os casos onde até 10% das amostras são alteradas. Para cenários onde 15% ou mais amostras tem erros introduzidos, o algoritmo RF obtém os melhores resultados gerais.

Nos cenários onde 1 erro foi introduzido, os algoritmos MLP e SVM obtiveram resultados de acurácia entre 0.99 e 0.92, com até 0.02 de variação no desvio padrão. Enquanto isso, RF obteve resultados entre 0.95 e 0.93, uma variação de apenas 0.02, sendo a menor variação entre os algoritmos analisados.

Nos cenários onde 2 erros são introduzidos, os melhores resultados foram alcançados pelos algoritmos SVM, para erros em até 20% das amostras e RF, para erros em 20% ou mais das amostras. O algoritmo SVM obteve resultados de acurácia entre 0.99 e 0.88, uma variação de 0.11, enquanto RF obteve resultados entre 0.95 e 0.90, variando apenas 0.05, sendo novamente o algoritmo de menor variação entre os métodos analisados.

Nos cenários onde 5 erros são introduzidos, os melhores resultados foram alcançados pelos algoritmos MLP e SVM, para erros em até 5% das amostras, por MLP, para erros em 10% das amostras, e RF para erros em 15% ou mais das amostras. O algoritmo RF teve resultados de acurácia entre 0.94 e 0.80, uma variação de 0.14, a menor entre os algoritmos analisados, enquanto MLP teve resultados entre 0.97 e 0.70, variando 0.27, a segunda maior entre eles.

#### 4.3. Análise da precisão, revocação e *F1-measure*

Com respeito a precisão, os experimentos na base BASE-Q-CHAT-10 mostram um bom desempenho dos algoritmos RL, MLP e SVM no cenários com inserção de até 3 erros, com resultados entre 0.99 e 0.91 e desvio padrão de até 0.01. Para cenários com 4 ou 5 erros em 30% ou mais das amostras, os melhores resultado de precisão são alcançados pelo algoritmo NB, variando entre 0.91 e 0.84, com desvio padrão de 0.01. No geral, os resultados de precisão foram semelhante também para a base BASE-AQ-10, exceto pelo algoritmo RF possuir os melhores resultados em todos os cenários onde 30% ou mais amostras tiveram erros inseridos.

Quanto ao revocação, os experimentos na base BASE-Q-CHAT-10 também mostram um bom desempenho dos algoritmos RL, MLP e SVM no cenários com inserção de até 3 erros, com resultados entre 1.00 e 0.93 e desvio padrão de até 0.02. Para cenários com 3 a 5 erros em 30% ou mais das amostras, o algoritmo RF também aparece entre os melhores resultados, variando entre 0.95 e 0.93, com desvio padrão de até 0.02. Para a base BASE-AQ-10, o algoritmo RF apresentou os melhores resultados em todos os cenários analisados.

Com relação a *F1-measure*, o bom desempenho dos algoritmos RL, MLP e SVM se repete para a base BASE-Q-CHAT-10 e para a base BASE-AQ-10 com 2 erros em até 15% das amostras. Por sua vez, o algoritmo RF apresenta os melhores resultados de *F1-measure* em praticamente todos os outros cenários relacionados aos testes com a base BASE-AQ-10.

Por fim, é importante observar que a diferença de resultado observada entre as duas bases pode ser consequência dos limiares utilizados pelos métodos AQ-criança-10 e Q-CHAT-10 para indicar a necessidade de investigação aprofundada. No primeiro caso, é necessário obter 7 pontos ou mais para esse indicativo, enquanto no segundo caso são 4 pontos ou mais. Desse modo, o método de inserção de erros tem probabilidades diferentes de resultar em inconsistências em cada base, o que pode ter relação direta com a variação do desempenho dos algoritmos.

## 5. Conclusões

O presente trabalho analisa o desempenho de diversos algoritmos de aprendizado de máquina na triagem do Transtorno do Espectro Autista (TEA). Por meio da análise de métricas como acurácia, revocação, precisão e *F1-measure*, foi possível avaliar a eficácia desses modelos em identificar a necessidade de aprofundar a investigação diagnóstica, seguindo as escalas AQ-10 e Q-CHAT-10.

Os resultados destacaram modelos como regressão logística, SVM com kernel linear, floresta aleatória e *multilayer perceptron* como os mais eficazes para essa tarefa específica. No entanto, é importante ressaltar que a sensibilidade dos modelos regressão logística, SVM e *multilayer perceptron* foi afetada significativamente quando as bases possuíam erros, demonstrando a necessidade de considerar este efeito quando da aplicação em cenários realistas.

Como trabalhos futuros, sugere-se realizar a análise dos algoritmos em outras bases de dados, com maior número de amostras ou com o conjunto completo de perguntas do AQ e Q-CHAT. Outro aspecto relevante para trabalhos futuros é incluir a análise de técnicas de explicabilidade, como LIME [Ribeiro et al. 2016] e SHAP [Lundberg and Lee 2017], aos modelos mais robustos gerados pelos diferentes algoritmos de classificação. Isso permitiria uma análise mais aprofundada das decisões dos modelos e destacar os principais fatores que influenciam as classificações.

## Referências

- Allison, C., Auyeung, B., and Baron-Cohen, S. (2012). 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):202–212.
- Allison, C., Baron-Cohen, S., Wheelwright, S., Charman, T., Richler, J., Pasco, G., and Brayne, C. (2008). The q-chat (quantitative checklist for autism in toddlers): a normally distributed quantitative measure of autistic traits at 18–24 months of age: preliminary report. *Journal of autism and developmental disorders*, 38:1414–1425.
- APA, A. P. A. (2013). *Diagnostic and Statistical Manual of Mental Disorders*. American Psychiatric Association.

- Artoni, A. A., Barbosa, C., and Morandini, M. (2022). Autism spectrum disorder diagnosis assistance using machine learning. *Revista de Informática Teórica e Aplicada*, 29(3):36–53.
- Baron-Cohen, S., Wheelwright, S., Skinner, R., Martin, J., and Clubley, E. (2001). The autism-spectrum quotient (aq): Evidence from asperger syndrome/high-functioning autism, males and females, scientists and mathematicians. *Journal of autism and developmental disorders*, 31:5–17.
- Ferreira, R. d. S. (2010). 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*, volume 13, pages 178–187.
- Fitzgerald, M. (2017). The clinical gestalts of autism: Over 40 years of clinical experience with autism. In Fitzgerald, M. and Yip, J., editors, *Autism*, chapter 2. IntechOpen.
- Garg, A., Parashar, A., Barman, D., Jain, S., Singhal, D., Masud, M., and Abouhawwash, M. (2022). Autism spectrum disorder prediction by an explainable deep learning approach. *Computers, Materials & Continua*, 71(1):1459–1471.
- Hossain, M. D., Kabir, M. A., Anwar, A., and Islam, M. Z. (2021). Detecting autism spectrum disorder using machine learning techniques: An experimental analysis on toddler, child, adolescent and adult datasets. *Health Information Science and Systems*, 9:1–13.
- Kleinman, J. M., Robins, D. L., Ventola, P. E., Pandey, J., Boorstein, H. C., Esser, E. L., Wilson, L. B., Rosenthal, M. A., Sutera, S., Verbalis, A. D., Barton, M., Hodgson, S., Green, J., Dumont-Mathieu, T., Volkmar, F., Chawarska, K., Klin, A., and Fein, D. (2008). 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):827–839.
- Lord, C., Risi, S., Lambrecht, L., Cook, E. H., Leventhal, B. L., DiLavore, P. C., Pickles, A., and Rutter, M. (2000). The autism diagnostic observation schedule—generic: A standard measure of social and communication deficits associated with the spectrum of autism. *Journal of autism and developmental disorders*, 30:205–223.
- Lord, C., Rutter, M., and Le Couteur, A. (1994). Autism diagnostic interview-revised: a revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. *Journal of autism and developmental disorders*, 24(5):659–685.
- Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.
- Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). ”why should i trust you?” explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.
- Thabtah, F. (2017). ASDTests: A mobile app for ASD screening. Disponível em: <https://www.asdtests.com/>. Acesso em: 10 de maio de 2024.

- Thabtah, F., Abdelhamid, N., and Peebles, D. (2019). A machine learning autism classification based on logistic regression analysis. *Health Information Science and Systems*, 7:1–11.
- Thabtah, F., Kamalov, F., and Rajab, K. (2018). A new computational intelligence approach to detect autistic features for autism screening. *International journal of medical informatics*, 117:112–124.
- Thabtah, F. and Peebles, D. (2020). A new machine learning model based on induction of rules for autism detection. *Health informatics journal*, 26(1):264–286.