Robustness and Explainability of Machine Learning Algorithms for Autism Spectrum Disorder Screening

Saulo B. F. Lino[1], **Paulo T. Guerra**[1]

[1]Universidade Federal do Ceará (UFC) – Campus Quixadá
Av. José de Freitas Queiroz, 5003 – 63902-580 – Quixadá – CE – Brazil

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

Abstract. *This work evaluates the robustness and explainability of machine learning algorithms for Autism Spectrum Disorder (ASD) screening using AQ-10-CHILD data. Scenarios with artificial noise were analyzed to simulate form filling errors, and predictions were interpreted using SHAP values. The results indicate that Logistic Regression, SVM, and Multilayer Perceptron tend toward an additive logic. Random Forest, KNN, and Naive Bayes showed greater stability and a better ability to capture interactions among indicators. We thus argue that model selection should consider not only accuracy, but also resilience to noise and clinical interpretability in order to be suitable for real-world applications.*

Resumo. *Este trabalho avalia robustez e explicabilidade de algoritmos de aprendizado de máquina na triagem do Transtorno do Espectro Autista (TEA) com dados do AQ-10-CHILD. Foram analisados cenários com ruído artificial para simular erros de preenchimento, com interpretação das predições por valores SHAP. Os resultados indicam que Regressão Logística, SVM e Perceptron Multicamadas tendem a uma lógica aditiva. Florestas Aleatórias, KNN e Naive Bayes apresentaram maior estabilidade e melhor capacidade de capturar interações entre indicadores. Argumentamos assim que a escolha do modelo de predição deve considerar, além da acurácia, também a resiliência destes ao ruído e a sua interpretabilidade clínica, para que este seja adequado a cenários reais.*

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 sociocomunicativas, 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 ADI-R [Lord et al. 1994] e ADOS-R [Lord et al. 2000], que são demorados e complicados, exigindo 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 muito mais pessoas que estão no espectro permaneçam indetectadas [Fitzgerald 2017].

Para agilizar o encaminhamento de indivíduos que apresentam sinais dentro do espectro para uma avaliação mais aprofundada, foram desenvolvidos métodos de triagem autoadministrados ou administrados pelos pais ou cuidadores, principalmente com base em questionários, que em uma fase preliminar fornecem às pessoas o reconhecimento de possíveis características autistas. Como exemplos, temos AQ [Baron-Cohen et al. 2001] e QCHAT [Allison et al. 2012a].

Os questionários de triagem AQ10-CRIANCA e QCHAT10 representam versões simplificadas dos instrumentos AQ e QCHAT, projetados especificamente para otimizar o processo de triagem do TEA. Compostos por apenas 10 questões cada, esses métodos visam reduzir a carga de tempo e esforço necessários para a avaliação, tornando-os ferramentas mais acessíveis e eficientes. O AQ10-CRIANCA é adaptado para crianças de 4 a 11 anos [Allison et al. 2012a], enquanto o QCHAT10 é uma versão simplificada do QCHAT, direcionado para bebês entre 12 e 36 meses [Allison et al. 2012b].

Diversos estudos têm explorado o uso de aprendizado de máquina para acelerar a triagem do TEA e identificar os itens mais influentes dos questionários [Thabtah et al. 2019, Hossain et al. 2021, Artoni et al. 2022, Garg et al. 2022]. Contudo, estudos como [Lino et al. 2024] apontam para limitações dos algoritmos com respeito a sua tolerância a erros no processo de triagem. Modelos como Regressão Logística (RL), Máquina de Vetor de Suporte (SVM) e Perceptron Multicamadas (MLP) têm queda de acurácia de até 0,27 quando expostos a cenários que simulam erros na triagem, enquanto modelos como Floresta Aleatória (RF) têm menor variação [Lino et al. 2024]. Os autores argumentam que a eficácia da triagem automatizada deve depender não apenas da capacidade preditiva do modelo, mas de sua resiliência a equívocos comuns cometidos por pais ou cuidadores durante o preenchimento dos questionários.

Poucos estudos dedicam-se a investigar as razões por trás das predições, aspecto essencial para a confiança dos clínicos durante o processo de triagem. Tais lacunas evidenciam a necessidade de abordagens que combinem eficácia, eficiência, robustez a erros e interpretabilidade, sendo este o objetivo deste trabalho.

2. Robustez de Algoritmos de Aprendizado de Máquina em Dados do Transtorno do Espectro Autista

A escala AQ-10-CHILD é uma versão abreviada da escala *Autistic Quotient* (AQ) [Baron-Cohen et al. 2001], projetada para avaliar traços autistas em crianças de 4 a 11 anos. Esta escala é intuitiva em sua aplicação, permitindo que as respostas sejam preenchidas por pais ou cuidadores, tornando-a assim uma ferramenta versátil de triagem em um contexto de avaliação clínica [Allison et al. 2012a]. As questões do AQ-10-CHILD podem ser vistas na Tabela 1.

Nessa escala, consideram-se indicativas de traços autistas respostas positivas para A1, A5, A7 e A10, bem como negativas para A2, A3, A4, A6, A8 e A9. A escala recomenda investigação mais aprofundada quando se identificam 7 ou mais traços.[1]

[1]Embora ofereça uma triagem rápida, a escala não constitui instrumento de diagnóstico definitivo [Allison et al. 2012a]

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

Atributo	Descrição
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

Em [Thabtah et al. 2019], os autores apresentam uma base pública construída a partir de dados coletados pelo aplicativo *ASDTests*. Ela reúne respostas ao questionário de triagem e características dos participantes, totalizando 509 amostras de crianças, distribuídas em duas classes: 257 com TEA e 252 sem TEA.

Diversos trabalhos na literatura [Hossain et al. 2021, Artoni et al. 2022, Garg et al. 2022, Lino et al. 2024] mostram excelente desempenho dos modelos clássicos de aprendizado de máquina sobre essa base de dados (Tabela 2), com resultados superiores a 95% de acurácia.

Tabela 2. Tabela comparativa de desempenho dos algoritmos [Lino et al. 2024].

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

O excelente desempenho dos modelos pode contudo induzir a subestimação do problema real. Em [Lino et al. 2024], os autores argumentam que apesar da eficiência na predição, não está claro se os algoritmos estão efetivamente considerando o sentido das respostas às perguntas de A1–A10, para além de sua contagem.

Uma premissa que deve ser considerada para o problema real é que o formulário pode, inadvertidamente, ser respondido de forma errada e assim as bases de dados de triagem de autismo podem apresentar erros. Um algoritmo robusto deve ser capaz de indicar a necessidade de investigação de TEA, sempre que os elementos relevantes ao diagnóstico estejam presentes, ainda que porventura a pontuação total seja inferior a 7 [Lino et al. 2024].

Para avaliar a robustez dos algoritmos, [Lino et al. 2024] apresenta experimentos onde inserem erros artificiais na base AQ-10-CHILD, simulando cenários em que respostas equivocadas introduzem ruído. A Tabela 3 apresenta os resultados obtidos no trabalho.

Tabela 3. Acurácia dos algoritmos em cenários com erros [Lino et al. 2024].

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

Os resultados mostram que o cenário de acurácia (quase) perfeita apresentado na Tabela 2 passa a ter uma grande variação quando a hipótese de erros é admitida, o que pode incorrer em riscos às aplicações reais.

Nesse sentido, o presente trabalho busca, por meio de técnicas de Inteligência Artificial Explicável (XAI), investigar as razões por trás do comportamento desses modelos em cenários com ruídos (erros), oferecendo uma compreensão mais transparente e auditável do processo de triagem realizado pelos algoritmos de aprendizado de máquina.

3. Análise da Explicabilidade e Estabilidade da Decisão sob Ruído

O objetivo central do presente trabalho é investigar a explicabilidade dos modelos de aprendizado de máquina, em especial o comportamento dos algoritmos em cenários com inserção de erros artificiais.

O desenvolvimento experimental foi conduzido na linguagem Python, utilizando o ecossistema de bibliotecas Numpy [Harris et al. 2020], Pandas [Wes McKinney 2010] e Scikit-learn [Pedregosa et al. 2011] para manipulação e modelagem. A peça central desta etapa metodológica é a biblioteca SHAP [Lundberg and Lee 2017], empregada aqui para quantificar a contribuição marginal de cada atributo na classificação final. Todos os experimentos computacionais foram executados em ambiente Google Colab (CPU Intel Xeon @ 2.20GHz, 12GB RAM).

A etapa de coleta de dados manteve a fidelidade experimental presente no trabalho [Lino et al. 2024]. Para a análise detalhada de interpretação dos modelos, optou-se por restringir o escopo exclusivamente à utilização da base AQ10-CRIANCA, em especial por não apresentar desbalanceamento, reduzindo o risco de vieses que poderiam distorcer os valores SHAP e interferir na clareza da análise.

Diferentemente de [Lino et al. 2024], o foco analítico do presente trabalho desloca-se da avaliação de métricas de desempenho para a auditoria interpretativa. O objetivo central é utilizar a biblioteca SHAP para decompor a lógica de decisão dos modelos, investigando as hipóteses de causalidade e fragilidade levantadas anteriormente.

Ressalta-se também que todas as visualizações e cálculos de valores SHAP foram gerados exclusivamente a partir dos conjuntos de teste. Essa escolha metodológica é mandatória, pois o protocolo de injeção de ruído de [Lino et al. 2024] foi aplicado apenas a essa partição. Como os modelos foram treinados em dados íntegros para aprender o padrão "correto", a utilização dos dados de treinamento mascararia o fenômeno de degradação que desejamos observar. Portanto, para auditar verdadeiramente a reação dos algoritmos diante de informações conflitantes, a explicabilidade foi aferida estritamente sobre as amostras de teste, onde os erros artificiais foram introduzidos.

Para manter a relevância clínica e a viabilidade estatística, o escopo da análise de perturbação foi delimitado: dos 35 cenários de robustez de [Lino et al. 2024], foram selecionados apenas aqueles com injeção de 1 e 2 erros (aplicados a 50% do conjunto de teste). A hipótese motivadora foi que erros de preenchimento em uma ou duas questões representam as falhas humanas mais prováveis em ambientes reais de triagem. Além disso, a taxa de 50% garante um volume denso de amostras alteradas, permitindo significância estatística na comparação entre a predição original e a predição sob ruído.

3.1. Análise de Importância e Sensibilidade ao Erro

A investigação foi estruturada em quatro eixos principais de auditoria: Mapeamento de Importância Global; Sensibilidade por Questão (Erro Único); Sensibilidade Conjunta (2 Erros); e Validação da Hipótese Aditiva.

O mapeamento de importância global estabeleceu o ranking das variáveis preditoras por algoritmo com base na magnitude média dos valores SHAP absolutos. Esses rankings restringem-se às respostas comportamentais, fixando a "assinatura de decisão" de cada modelo e permitindo confrontar se as variáveis mais influentes correspondem aos pontos críticos de vulnerabilidade observados nos testes.

As análises de sensibilidade por questão quantificam o impacto de erros isolados (Erro Único) e combinados (2 Erros): para cada pergunta A1–A10, registrou-se o total de amostras perturbadas e a taxa de inversão de classe; e, para pares de erros em respostas, mapeou-se o potencial de interação destes quanto à desestabilização do modelo.

Por fim, a validação da hipótese aditiva verificou se a vulnerabilidade decorre de um mecanismo de decisão por pontuação cumulativa, analisando a variação na soma bruta das respostas (0 a 10) nas instâncias com inversão de classe e a prevalência de casos limítrofes próximos à fronteira de decisão.

É fundamental ressaltar uma distinção metodológica nas unidades de medida dos valores SHAP, decorrente da escolha dos *Explainers* adequados a cada arquitetura. O modelo RF foi interpretado via *TreeExplainer* (cálculo exato dos valores Shapley) e os modelos KNN e NB via *KernelExplainer* sobre a função de probabilidade. Nesses casos, os valores SHAP expressam a contribuição direta para a probabilidade da classe alvo (Espaço de Probabilidade 0–1). Já os modelos lineares (RL, SVM) foram interpretados via *LinearExplainer*, e o MLP utilizou *KernelExplainer* configurado para operar em *log-odds*, de modo a linearizar a interpretação antes da saturação da ativação (Espaço Log-Odds).

Consequentemente, a análise comparativa entre os modelos foca na importância relativa (ranking) das variáveis e na direção do impacto, e não na magnitude absoluta dos valores, visto que operam em escalas matemáticas distintas. A Tabela 4 compila os resultados do mapeamento de importância global para cada modelo.

3.2. Convergência na Explicabilidade: MLP, SVM e RL

A Tabela 4 apresenta o ranking de importância das variáveis para a RL nos 10 conjuntos de teste. Observa-se que, embora haja uma alternância na liderança entre as perguntas A4, A8 e A9 dependendo da amostragem, o grupo das cinco variáveis mais influentes (Top-5) permanece estável, composto consistentemente por A4, A8, A9, A2 e A7. Inversamente, a pergunta A3 figura sistematicamente como a menos relevante na maioria dos cenários.

Os resultados para o SVM seguem um comportamento análogo ao modelo logístico, reafirmando a relevância do mesmo quinteto de variáveis. Contudo, o SVM demonstra uma leve predileção pela pergunta A8, que assume a primeira posição em 50% dos datasets, superando a frequência de liderança da variável A4 neste classificador.

O modelo MLP apresentou a maior convergência entre todos os modelos: a resposta A4 é a preditora mais forte, ocupando o primeiro lugar em 8 dos 10 *datasets*. A alta

RL	Ordem de Importância (Decrescente)									
	1º	2º	3º	4º	5º	6º	7º	8º	9º	10º
Dataset 1	A8	A9	A2	A4	A7	A1	A6	A5	A10	A3
Dataset 2	A9	A2	A4	A8	A7	A1	A6	A10	A5	A3
Dataset 3	A2	A4	A8	A9	A7	A1	A10	A6	A5	A3
Dataset 4	A4	A8	A2	A7	A9	A5	A6	A1	A10	A3
Dataset 5	A4	A8	A2	A7	A9	A10	A6	A1	A5	A3
Dataset 6	A8	A2	A4	A7	A9	A1	A5	A10	A6	A3
Dataset 7	A4	A9	A8	A2	A7	A1	A6	A10	A5	A3
Dataset 8	A9	A8	A2	A4	A7	A1	A6	A5	A10	A3
Dataset 9	A4	A8	A9	A2	A7	A1	A10	A6	A3	A5
Dataset 10	A4	A8	A9	A7	A2	A10	A1	A6	A5	A3

KNN	Ordem de Importância (Decrescente)									
	1º	2º	3º	4º	5º	6º	7º	8º	9º	10º
Dataset 1	A4	A9	A8	A1	A5	A10	A6	A7	A3	A2
Dataset 2	A4	A9	A1	A8	A10	A6	A3	A5	A7	A2
Dataset 3	A4	A9	A6	A10	A8	A1	A5	A7	A3	A2
Dataset 4	A4	A5	A10	A9	A8	A6	A1	A3	A7	A2
Dataset 5	A4	A8	A5	A10	A1	A3	A9	A7	A6	A2
Dataset 6	A4	A5	A1	A10	A8	A9	A7	A3	A6	A2
Dataset 7	A4	A8	A10	A9	A1	A6	A5	A3	A7	A2
Dataset 8	A4	A9	A5	A8	A1	A6	A10	A7	A3	A2
Dataset 9	A4	A8	A5	A9	A3	A10	A7	A1	A6	A2
Dataset 10	A4	A6	A9	A8	A1	A3	A10	A5	A7	A2

SVM	Ordem de Importância (Decrescente)									
	1º	2º	3º	4º	5º	6º	7º	8º	9º	10º
Dataset 1	A8	A9	A2	A4	A7	A1	A6	A10	A5	A3
Dataset 2	A4	A2	A8	A9	A7	A1	A6	A10	A5	A3
Dataset 3	A2	A8	A9	A4	A7	A10	A1	A6	A5	A3
Dataset 4	A8	A9	A4	A2	A7	A5	A1	A6	A10	A3
Dataset 5	A4	A9	A8	A2	A7	A6	A10	A1	A5	A3
Dataset 6	A8	A9	A7	A4	A2	A1	A5	A10	A6	A3
Dataset 7	A9	A8	A4	A2	A7	A1	A6	A10	A5	A3
Dataset 8	A8	A9	A2	A7	A4	A6	A1	A5	A10	A3
Dataset 9	A4	A8	A9	A2	A7	A1	A10	A6	A3	A5
Dataset 10	A8	A9	A4	A7	A2	A1	A10	A6	A5	A3

NB	Ordem de Importância (Decrescente)									
	1º	2º	3º	4º	5º	6º	7º	8º	9º	10º
Dataset 1	A4	A6	A3	A5	A10	A9	A1	A8	A7	A2
Dataset 2	A6	A3	A4	A5	A10	A1	A9	A8	A7	A2
Dataset 3	A6	A3	A4	A10	A5	A9	A1	A8	A7	A2
Dataset 4	A5	A6	A3	A4	A10	A9	A1	A8	A7	A2
Dataset 5	A6	A10	A4	A3	A5	A1	A9	A8	A7	A2
Dataset 6	A5	A3	A6	A4	A10	A1	A9	A8	A7	A2
Dataset 7	A6	A4	A10	A5	A3	A1	A9	A8	A7	A2
Dataset 8	A6	A3	A4	A5	A10	A1	A9	A8	A7	A2
Dataset 9	A3	A6	A5	A10	A4	A8	A1	A9	A7	A2
Dataset 10	A3	A6	A4	A10	A5	A1	A9	A8	A7	A2

MLP	Ordem de Importância (Decrescente)									
	1º	2º	3º	4º	5º	6º	7º	8º	9º	10º
Dataset 1	A9	A4	A8	A6	A2	A7	A1	A5	A10	A3
Dataset 2	A4	A8	A2	A1	A9	A6	A7	A10	A5	A3
Dataset 3	A4	A8	A9	A6	A7	A2	A10	A1	A5	A3
Dataset 4	A4	A8	A9	A7	A1	A3	A6	A2	A10	A5
Dataset 5	A4	A6	A8	A9	A7	A2	A1	A5	A10	A3
Dataset 6	A4	A8	A9	A2	A7	A5	A1	A6	A10	A3
Dataset 7	A8	A4	A7	A9	A2	A6	A10	A1	A5	A3
Dataset 8	A4	A8	A9	A2	A7	A1	A6	A10	A5	A3
Dataset 9	A4	A8	A9	A2	A7	A6	A1	A3	A5	A10
Dataset 10	A4	A8	A9	A2	A7	A6	A1	A5	A10	A3

RF	Ordem de Importância (Decrescente)									
	1º	2º	3º	4º	5º	6º	7º	8º	9º	10º
Dataset 1	A4	A9	A8	A10	A1	A7	A5	A6	A3	A2
Dataset 2	A4	A9	A1	A10	A8	A5	A6	A7	A3	A2
Dataset 3	A4	A9	A1	A7	A3	A5	A10	A8	A6	A2
Dataset 4	A4	A9	A10	A5	A8	A1	A6	A7	A3	A2
Dataset 5	A4	A8	A10	A5	A9	A7	A1	A6	A3	A2
Dataset 6	A4	A10	A9	A1	A5	A8	A7	A6	A3	A2
Dataset 7	A4	A10	A8	A9	A6	A1	A5	A7	A3	A2
Dataset 8	A4	A9	A6	A1	A10	A8	A5	A7	A3	A2
Dataset 9	A4	A8	A9	A7	A5	A10	A6	A1	A3	A2
Dataset 10	A4	A9	A6	A1	A10	A8	A7	A5	A3	A2

Tabela 4. Ranking por modelo de importância das respostas do AQ10-CRIANCA.

similaridade deste ranking com os anteriores reforça a hipótese de que o MLP convergiu para uma fronteira de decisão linear, priorizando os mesmos padrões detectados pela RL e SVM.

A Figura 1, lado esquerdo, apresenta o mapa de calor da sensibilidade unitária, com a taxa média de inversão de classe quando uma única resposta é alterada nos modelos RL, MLP e SVM. Em nossos experimentos, verificamos resultados idênticos, o que corrobora a análise de robustez de [Lino et al. 2024] e indica equivalência funcional entre os modelos.

Um fenômeno notável observa-se ao confrontar este mapa com os rankings de importância SHAP: as variáveis de maior relevância global (como A4 ou A8) não resultaram, necessariamente, em maiores taxas de erro. A análise da mecânica de decisão revelou que a vulnerabilidade destes modelos é regida por uma lógica estritamente aditiva, e não semântica. A inversão de classe ocorre exclusivamente quando a alteração da resposta é capaz de cruzar o limiar de corte.

A Figura 1, lado direito, ilustra o impacto da alteração simultânea de dois atributos. A análise não sugere a existência de combinações semânticas críticas baseadas na importância das variáveis, parecendo indicar que a distribuição de erros segue a mesma regra aritmética observada anteriormente, apenas ampliando o intervalo de vulnerabilidade.

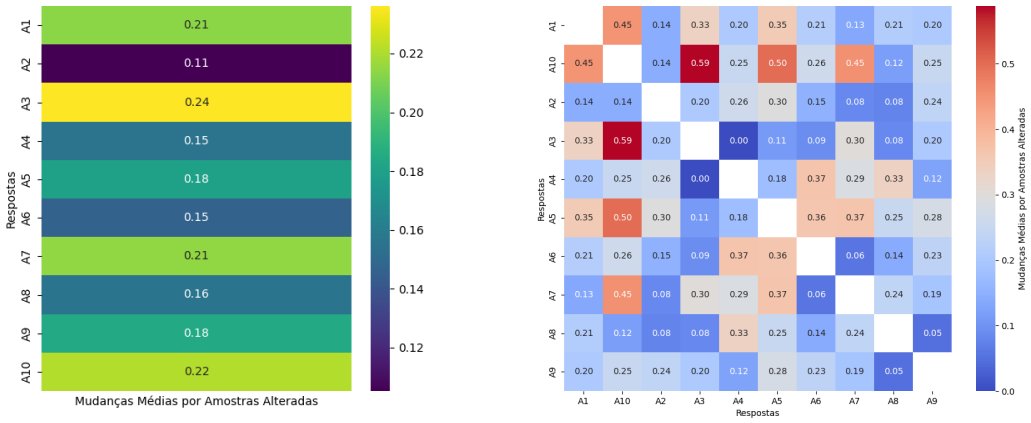


Figura 1. Mapa de calor de sensibilidade a erro único (esquerda) e duplo (direita) para os modelos RL, MLP e SVM (média de 10 bases)

4. Estabilidade na Explicabilidade: KNN, NB e RF

Para o KNN, a Tabela 4 mostra hegemonia absoluta da variável A4. Do mesmo modo, constata-se que A2 aparece invariavelmente como a de menor contribuição (10ª) em todas as amostras. Contudo, variáveis como A9 e A5 apresentam alta volatilidade posicional. A variável A9, por exemplo, figura como a 2ª mais importante na maioria dos cenários (datasets 1, 2, 3 e 8), mas cai abruptamente para a 7ª posição no dataset 5.

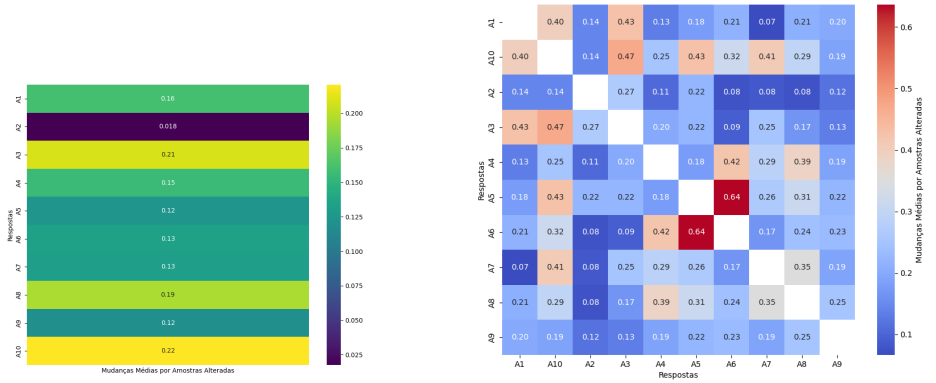
Para o NB, a tabela mostra uma hierarquia notavelmente estável. Os itens A3, A4, A5, A6 e A10 dominam consistentemente as cinco primeiras posições em praticamente todos os conjuntos de teste. Aqui, a variável A2 também figura unanimemente na última posição. Note-se, contudo, que há uma alternância na definição do item mais importante.

Diferente das variações observadas no KNN e NB, o RF demonstrou uma hegemonia absoluta na seleção do atributo principal. Em todos os 10 conjuntos de dados avaliados (100%), a variável A4 foi isolada como o preditor de maior ganho de informação (valor SHAP mais alto). Essa consistência sugere que, para a estrutura de árvores de decisão, a característica A4 é o "nó raiz" mais eficiente para separar as classes, superando qualquer outro sintoma em poder discriminativo primário.

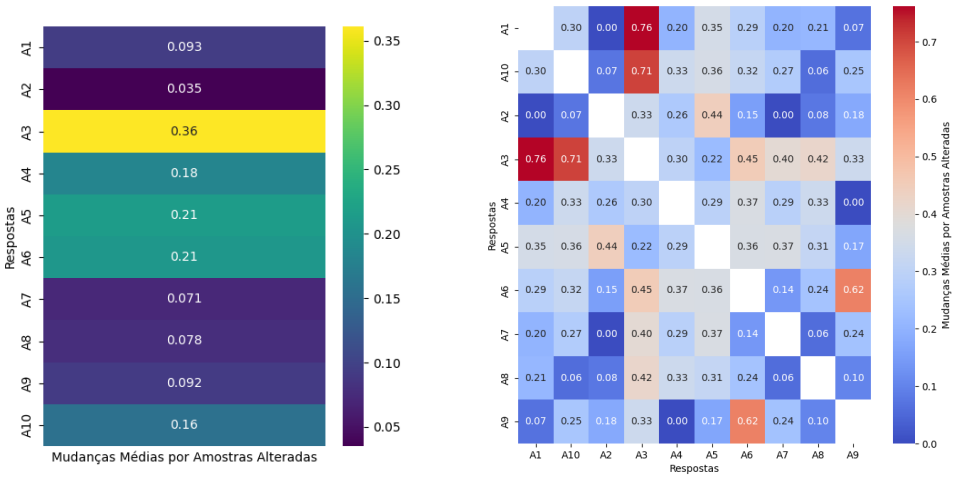
Para o RF, observa-se ainda uma reorganização nas variáveis secundárias em comparação ao NB. Enquanto o NB valorizava fortemente a variável A3, o RF tende a rebaixá-la para a penúltima posição (9ª), promovendo as variáveis A9 e A10 para o topo do ranking. Por fim, confirma-se o padrão de irrelevância da variável A2, que permaneceu estagnada na 10ª posição em todas as iterações.

Quanto à sensibilidade por questão, a Figura 2 evidencia uma diferença arquitetural do modelo KNN em relação aos anteriores. Enquanto RL e SVM exibem vulnerabilidade aritmética, o KNN apresenta robustez em captar parte da complexidade dos dados. A variável A2, por exemplo, tem taxa de inversão de classe ínfima (0.018), sugerindo irrelevância para a separação dos *clusters*, ao passo que A10 (0.22) e A3 (0.21) concentram a discriminação, indicando que o KNN não trata o questionário como lista plana, mas constrói a fronteira de decisão pela densidade de informação de cada pergunta.

KNN



NB



RF

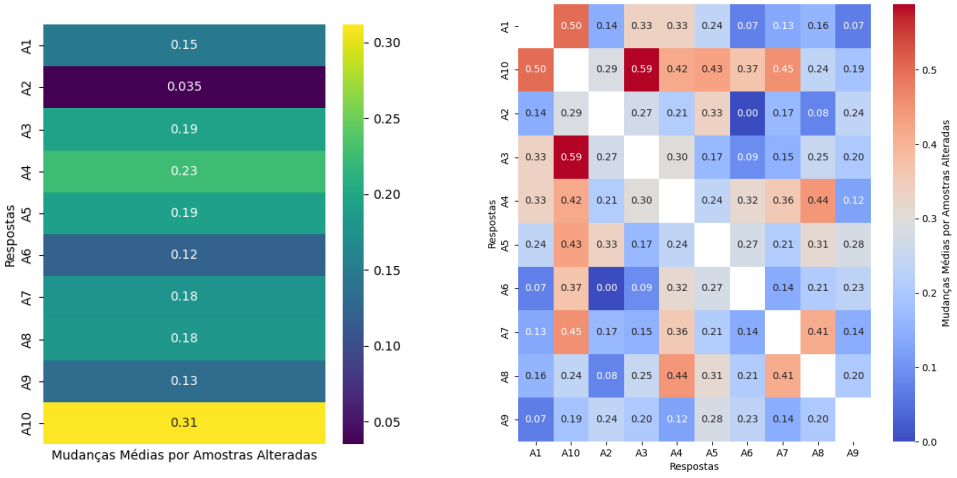


Figura 2. Mapa de calor de sensibilidade a erro único (esquerda) e duplo (direita) para os modelos KNN, NB e RF (média de 10 bases)

No caso do modelo NB, a Figura 2 mostra a relação entre importância e robustez: o mapa de calor mostra correlação direta, pois as variáveis do topo do ranking (A3, A4, A5, A6) coincidem com as maiores taxas de alteração de predição sob ruído. A análise também inclui erros duplos, revelando acumulação de evidências, já que as taxas de alteração disparam quando o erro incide simultaneamente em duas variáveis de alto peso informacional.

Para o RF, a Figura 2 mostra um resultado contraintuitivo em relação à tabela de importância SHAP: embora A4 seja a mais importante globalmente, A10 é a mais sensível a alterações isoladas. A mudança em A10 provoca variação média de probabilidade de 0.31, superando A4 (0.23), o que sugere que, nas árvores de decisão, A10 atua como um divisor em casos de incerteza, enquanto A4 fornece uma base mais estável.

Na análise do erro duplo, o RF demonstra um comportamento de decisão mais sofisticado: ele não apenas consulta o valor das variáveis, mas assimila como elas se relacionam. A alta sensibilidade ao par (A3,A10), por exemplo, sugere que o modelo identificou um núcleo essencial para o diagnóstico.

Em síntese, os resultados indicam que KNN, NB e, sobretudo, RF superam a lógica puramente aditiva dos modelos RL, SVM e MLP, com maior capacidade de preservar a decisão sob ruído e de refletir interações entre sintomas.

5. Conclusão

A análise de explicabilidade das predições sob cenários de ruído permitiu investigar a lógica interna de decisão dos classificadores, revelando uma dicotomia fundamental entre os modelos de predição. Ao examinar o comportamento dos algoritmos de RL, SVM e MLP, confirmou-se a hipótese de que estas arquiteturas convergem para um mecanismo de aditividade simples. Essa característica torna tais classificadores vulneráveis a erros pontuais: se a soma das respostas de um paciente estiver próxima ao limiar de corte, a alteração de uma única variável (por erro de preenchimento ou imprecisão) é matematicamente suficiente para cruzar a fronteira de decisão e inverter o diagnóstico.

Em contrapartida, a análise do RF revelou uma arquitetura de decisão substancialmente mais robusta e alinhada à realidade clínica, demonstrando operar através de uma lógica condicional e hierarquizada. O classificador parece assimilar que a relevância de uma resposta nuclear pode aumentar ou diminuir dependendo da presença de outras respostas, capturando interações não-lineares que escapam à detecção dos modelos aditivos.

Embora a análise de robustez e de explicabilidade tenha indicado o RF como o modelo mais adequado para este domínio, vale destacar que o NB e o KNN demonstraram tolerância aos cenários de falha que pode ser explorada caso a implementação do RF não seja possível.

Como trabalhos futuros, é fundamental ampliar o escopo da investigação avaliando classes de algoritmos que não foram cobertas neste estudo, como *frameworks* baseados em *Gradient Boosting* (XGBoost, LightGBM e CatBoost) e *Deep Learning*. Propõe-se também expandir esta análise para outros instrumentos de triagem para o TEA além da base AQ-10-CHILD, avaliando a estabilidade das atribuições de SHAP e a robustez dos modelos em cenários clínicos variados.

6. Declaração sobre uso de Inteligência Artificial

Em conformidade com o Código de Conduta para autores da SBC, declaramos o uso de ferramentas de Inteligencia Artificial Generativa (Gemini e GPT/Prism) para revisão linguística e sintetização de trechos do texto.

Referências

Allison, C., Auyeung, B., and Baron-Cohen, S. (2012a). 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., Sastry, B., Green, J., Argyle, E., and Brayne, C. (2012b). Short report: a short strategy for identifying children at high risk of autism spectrum conditions in primary care. *Molecular Autism*, 3(1):1–7.

APA, A. P. A. (2013). Diagnostic and statistical manual of mental disorders. *The 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.

Fitzgerald, M. (2017). The clinical gestalts of autism: over 40 years of clinical experience with autism. In *Autism-paradigms, recent research and clinical applications*. 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.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., and Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, 585(7825):357–362.

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.

Lino, S., Cruz, L., and Guerra, P. T. (2024). Análise da robustez de algoritmos de aprendizado de máquina em dados do transtorno do espectro autista. In *Anais do XXXIX Simpósio Brasileiro de Bancos de Dados*, pages 53–65, Porto Alegre, RS, Brasil. SBC.

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. *Advances in neural information processing systems*, 30.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

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.

Wes McKinney (2010). Data Structures for Statistical Computing in Python. In Stéfan van der Walt and Jarrod Millman, editors, *Proceedings of the 9th Python in Science Conference*, pages 56 – 61.