Análise do desempenho de aprendizagem de *Machine Learning* na Educação Básica aplicando a Teoria de Resposta ao Item

Marcelo Fernando Rauber[1, 2], Christiane Gresse von Wangenheim[1], Adriano F. Borgatto[3], Ramon Mayor Martins[1]

marcelo.rauber@ifc.edu.br, c.wangenheim@ufsc.br, adriano.borgatto@ufsc.br, ramon.mayor@posgrad.ufsc.br

[1] Programa de Pós-Graduação em Ciência da Computação – Universidade Federal de Santa Catarina (UFSC) - Florianópolis - SC - Brasil.
[2] Instituto Federal Catarinense (IFC) - Camboriú - SC - Brasil.
[3] Programa de Pós-Graduação em Métodos e Gestão em Avaliação - Universidade Federal de Santa Catarina (UFSC) - Florianópolis - SC - Brasil.

RESUMO

A atual inserção de *Machine Learning* (ML) no dia-a-dia demonstra a importância de introduzir o ensino de conceitos de ML desde a Educação Básica. Acompanhando esta tendência surge a necessidade de avaliar essa aprendizagem. Neste artigo apresentamos a avaliação da validade e da confiabilidade de uma rubrica. Essa rubrica visa avaliar a aprendizagem pelo desempenho do aluno com base nos resultados da aprendizagem da aplicação de conceitos de ML por alunos dos anos finais do Ensino Fundamental e do Ensino Médio. Adotando a Teoria de Resposta ao Item apresentamos uma proposta preliminar da construção de uma escala para o nível de aprendizagem dos estudantes. Os resultados mostram que foi possível calibrar os parâmetros da Teoria de Resposta ao Item com índices satisfatórios de confiabilidade e validade, o que demonstra o potencial de utilização da rubrica de modo a auxiliar tanto alunos quanto pesquisadores e professores a promover o desenvolvimento do ensino de ML na Educação Básica.

PALAVRAS-CHAVE

Avaliação da aprendizagem, *Machine Learning*, Teoria de Resposta ao Item, TRI, Educação Básica

1 INTRODUÇÃO

Um conjunto diverso de tecnologias de Inteligência Artificial (IA) está sendo empregado atualmente tanto em ambientes corporativos quanto impactando profundamente nossa vida diária, nossa cultura, diversidade, educação, conhecimento científico, comunicação e informação [56]. Uma das principais técnicas de IA é o aprendizado de máquinas ou *Machine Learning* (ML). O ML se concentra no desenvolvimento de sistemas que aprendem e evoluem a partir da sua própria experiência sem ter que ser explicitamente programados, por meio da construção de um modelo matemático/estatístico baseado nos dados coletados [41]. Progressos recentes em ML foram alcançados especificamente por abordagens de aprendizagem profunda utilizando redes neurais, melhorando drasticamente o estado da arte na visão computacional por meio de reconhecimento de imagem [33]. Atualmente, também podemos encontrar aplicações de ML em chatbots, sistemas de recomendação, assistentes pessoais com processamento de linguagem natural, reconhecimento de padrões para encontrar atividades financeiras não usuais e o uso sensores para coleta de dados (como pressão ou temperatura) integrando um ambiente de internet das coisas ou *internet of things* [49, 56].

Mesmo com a atual inserção de sistemas inteligentes de forma ampla, ainda há uma grande parcela da população que não compreende a tecnologia por trás do ML, o que pode torná-lo misterioso ou mesmo assustador, ofuscando seu potencial impacto positivo na sociedade [27]. Assim, para desmistificar o que é ML, como funciona e quais são seus impactos e limitações, há uma necessidade crescente de compreensão pública do ML [28]. Então torna-se importante introduzir conceitos e práticas básicas já na escola [13, 15], despertando os estudantes a serem mais do que meros consumidores de aplicações de ML, mas que também passem a ser criadores de soluções inteligentes e eticamente corretas [31, 49, 56].

Com essa motivação já estão surgindo várias propostas propondo o ensino de IA/ML na Educação Básica, mas estão em estágios iniciais [38, 34]. No Brasil, apesar da Lei de Diretrizes e Bases da Educação não abordar explicitamente os conhecimentos associados ao ML, vale destacar sua inata inserção quando foram definidos os objetivos da formação básica do cidadão, no

Fica permitido ao(s) autor(es) ou a terceiros a reprodução ou distribuição, em parte ou no todo, do material extraído dessa obra, de forma verbatim, adaptada ou remixada, bem como a criação ou produção a partir do conteúdo dessa obra, para fins não comerciais, desde que sejam atribuídos os devidos créditos à criação original, sob os termos da licença CC BY-NC 4.0.
EduComp'23, Abril 24-29, 2023, Recife, Pernambuco, Brasil (On-line)
© 2023 Copyright mantido pelo(s) autor(es). Direitos de publicação licenciados à Sociedade Brasileira de Computação (SBC).

parágrafo II do artigo 32 da LDB [9] tanto no Ensino Fundamental quanto seu aprofundamento no Ensino Médio: "*a compreensão do ambiente natural e social, do sistema político, da tecnologia, das artes e dos valores em que se fundamenta a sociedade*". De forma análoga a Base Nacional Comum Curricular (BNCC) não aborda diretamente o ML, mas fica claro em vários trechos da mesma que novos conhecimentos devem ser abordados, como por exemplo:

> *"É preciso garantir aos jovens aprendizagens para atuar em uma sociedade em constante mudança, prepará-los para profissões que ainda não existem, para usar tecnologias que ainda não foram inventadas e para resolver problemas que ainda não conhecemos. Certamente, grande parte das futuras profissões envolverá, direta ou indiretamente, computação e tecnologias digitais."* [39]

Recentemente foram incluídas normas complementares à BNCC sobre Computação na Educação [60] onde foram incluídas a habilidades relacionadas a IA e ML para o Ensino Médio:

> *"(EM13CO10) Conhecer os fundamentos da Inteligência Artificial, comparando-a com a inteligência humana, analisando suas potencialidades, riscos e limites."*
> *"(EM13CO12) Produzir, analisar, gerir e compartilhar informações a partir de dados, utilizando princípios de ciência de dados."* [60]

Seguindo diretrizes curriculares abordando ML desde a Educação Básica [54, 34], o ensino de ML neste estágio educacional deve incluir uma compreensão dos conceitos básicos de ML, tais como algoritmos de aprendizagem e fundamentos de redes neurais, assim como limitações e considerações éticas relacionadas ao ML. Espera-se que os estudantes não somente obtenham uma compreensão desses conceitos mas aprendam também a aplicá-los criando modelos de ML. Ao adotar metodologias ativas no processo de aprendizado, focando no desenvolvimento centrado no ser humano de um modelo de ML [7], os estudantes devem aprender a preparar um conjunto de dados, treinar o modelo de ML e avaliar seu desempenho e predição de novas imagens [36, 46]. A utilização de ferramentas visuais, como o Google Teachable Machine (GTM) [20] é tipicamente adotada nesta etapa educacional, não necessitando de qualquer programação. Desta forma os estudantes podem executar um processo ML de forma interativa, utilizando um ciclo de treinamento, *feedback* e correção, permitindo-lhes avaliar o desempenho do modelo de ML e tomar as ações apropriadas [22].

A avaliação de aprendizagem é uma etapa importante do processo de aprendizado. Avaliar é resultado de uma experiência educacional, compreendendo os processos de coletar e analisar informações de fontes diversas a fim de entender profundamente o que os estudantes sabem, entendem e podem realizar com seus conhecimentos [30]. Avaliar e fornecer *feedback* adequado é importante tanto para o aluno quanto para o professor [25]. Em um processo efetivo de aprendizado, é importante que os estudantes saibam seu o nível de desempenho em uma tarefa, como seu próprio desempenho se relaciona ao bom desempenho e o que fazer para fechar a lacuna entre eles [50]. Nesse sentido, uma alternativa interessante são as rubricas [58] e também são muito comuns [59], já que podem ser usadas em questões dissertativas, avaliação de performance, desenvolvimento de produtos, portfólios, demonstrações e outros [58, 59]. Uma rubrica é um guia de pontuação que define critérios e seus diferentes níveis de desempenho [59]. O objetivo de se ter rubricas é comunicar um padrão de julgamento, permitindo assim aos alunos identificarem seus pontos fortes e fracos [59].

Quando abordamos temas como pensamento computacional, algoritmos e programação, modelagem e simulação na educação básica, já há esforços consideráveis para abordar a avaliação [37, 53, 57, 2, 1] inclusive para avaliar conceitos relacionados como design de interface [52, 3] como também habilidades como a criatividade [5, 4].

Entretanto, se observa ainda uma carência de abordagens para a avaliação da aprendizagem de conceitos ML de forma confiável e válida [47]. As poucas propostas de avaliação de aprendizagem de ML existentes para Educação Básica são relativamente simples, baseados em quizzes ou autoavaliações. Análises da confiabilidade através da consistência interna foram relatadas por [26, 29]. Como resultado, Hsu *et al*. [29] relataram um valor Cronbach α de 0,883 para a confiabilidade de um questionário de auto-avaliação com cinco itens usando uma escala Likert. Hitron *et al*. [26] também relataram uma alta confiabilidade (Kappa = 92%) da codificação realizada pelos pesquisadores ao rotular manualmente os itens de resposta curta do ensaio. Com o objetivo de avaliar a validade do conteúdo, Shamir e Levin [51] não informaram resultados específicos, mas mencionaram que alunos e professores revisaram as perguntas analisando a capacidade de leitura e compreensão do item. Gresse von Wangenheim *et al*. [22] sugerem uma rubrica para a avaliação baseada no desempenho do modelo ML criado pelos estudantes a partir de atividades voltadas ao reconhecimento de imagens. Gresse von Wangenheim *et al*. [22] ainda apresentam uma avaliação inicial da validade da rubrica com um um painel de especialistas, continuado em Rauber *et al*. [48] ao propor uma adaptação da rubrica de Gresse von Wangenheim *et al*. [22] e apresentar resultados iniciais positivos relativos à validade e confiabilidade do

instrumento com base uma série de estudos de casos realizados. Os resultados deste estudo apontam a confiabilidade da rubrica com um valor de ômega global 0,646 e a validade convergente do construto por meio da matriz de correlação policórica.

Na busca da validação da confiabilidade e da validade, uma alternativa à Teoria Clássica de Teste ou *Classical Test Theory* (CTT), é a Teoria de Resposta ao Item (TRI), muitas vezes abordada como moderna e superior, devido a criação e interpretação de uma escala [17]. A TRI é "*uma coleção de modelos de medição que objetivam explicar conexões entre respostas observadas em ítem em uma escala e um construto subjacente*" [14]. Porém, atualmente ainda não existem pesquisas que avaliam a validade de modelos de avaliação de aprendizagem de conceitos de ML com a TRI.

Neste contexto, este artigo apresenta os resultados de uma avaliação de dimensionalidade e da calibração dos parâmetros da TRI de uma avaliação baseada no desempenho com base em artefatos criados por estudantes como resultado da aprendizagem. E, a consequente definição inicial de uma escala para o nível de aprendizagem dos estudantes.

2 MÉTODOS

Integrado ao curso "ML para todos!" [23] e como resultado de pesquisas anteriores, foi desenvolvido sistematicamente um modelo de avaliação baseada no desempenho que inclui uma rubrica [22], seguindo o método proposto por Moskal e Leyden [42] e o projeto centrado em evidências [40]. Para o contexto desta pesquisa, essa rubrica foi revisada e ajustada (Tabela 1).

Tabela 1: Rubrica de pontuação

Critério		Níveis de Desempenho		
		Fraco - 0 pontos	Aceitável - 1 ponto	Bom - 2 pontos
	Gerenciamento de dados			
C1	Quantidade de imagens	Menos de 20 imagens por categoria	21 - 35 imagens por categoria	Mais de 36 imagens por categoria
C2	Relevância das imagens	Muitas imagens não estão relacionadas a tarefa (irrelevantes) e/ou ao menos uma imagem contém conteúdo não ético (violência, nudez, etc)	Ao menos uma imagem é irrelevante mas não contém imagens não éticas.	Todas as imagens são relacionadas a tarefa de ML e éticas.
C3	Distribuição do conjunto de dados	A quantidade de imagens em cada categoria varia muito. Mais de 10% de variação em ao menos uma categoria (relativo ao total).	A quantidade de imagens entre as categorias têm entre 3% e 10% de variação.	Todas as categorias têm a mesma quantidade de imagens (menos de 3% de variação).
C4	Rotulação das imagens	Menos de 20% das imagens foram rotuladas corretamente	Entre 20% e 95% das imagens foram rotuladas corretamente	Mais de 95% das imagens foram rotuladas corretamente
C5	Limpeza dos dados	Há várias imagens confusas (fora de foco, vários objetos na mesma imagem, etc.)	Há uma imagem confusa	Nenhuma imagem confusa foi incluída no conjunto de dados
	Treinamento do modelo			
C6	Treinamento	O modelo não foi treinado	O modelo foi treinado usando os parâmetros padrões.	O modelo foi treinado com parâmetros ajustados (ex. épocas, batch size, taxa de aprendizado)
	Interpretação de desempenho			
C7	Testes com novos objetos	Nenhum objeto testado	1-3 objetos testados	Mais de 3 objetos testados
C8	Interpretação dos testes	Interpretação errada	(Não aplicável)	Correta interpretação
C9	Interpretação da acurácia	Categorias com baixa acurácia não são identificadas corretamente e interpretação incorreta em relação ao modelo	Categorias corretamente identificadas com baixa acurácia, mas interpretação incorreta em relação ao modelo	Categorias corretamente identificadas com baixa acurácia e a conseqüente interpretação a respeito do modelo
C10	Interpretação da matriz de confusão	As classificações errôneas não são identificadas corretamente e a interpretação a respeito do modelo é incorreta	As classificações errôneas foram corretamente identificadas, mas a interpretação a respeito do modelo é incorreta	Identificação correta de erros de classificação e a conseqüente interpretação com respeito ao modelo
C11	Ajustes / Melhorias realizadas	Nenhuma nova iteração de desenvolvimento foi relatada	Uma nova iteração com mudanças no conjunto de dados e/ou parâmetros de treinamento foi relatada	Várias iterações com mudanças no conjunto de dados e/ou parâmetros de treinamento foram relatadas

(Adaptado de Gresse von Wangenheim et al. [22])

O objetivo do presente estudo é analisar de forma exploratória a rubrica a fim de estimar a sua confiabilidade e validade de construto para a avaliação da aprendizagem dos conceitos de ML a partir da perspectiva dos pesquisadores no contexto da Educação Básica. Seguindo

a abordagem Goal Question Metric (GQM) [10], são analisadas as seguintes questões:
QA1: Há evidências da confiabilidade da rubrica por meio da TRI?
QA2: Há evidências de validade da rubrica por meio da carga fatorial da análise fatorial (dimensionalidade)?

Research design. A pesquisa foi realizada de forma exploratória com base em uma série de estudos de casos, aplicando o curso "ML para Todos!" na prática.

Coleta dos dados. A amostra é composta por estudantes da educação básica matriculados no curso, onde foi utilizada uma amostragem não-probabilística em cada estudo de caso aplicando o método de amostragem de conveniência [55]. Durante a aplicação do curso "ML para Todos!" foram coletados artefatos criados pelos alunos como resultados de aprendizagem.

Análise dos dados. Todos os dados coletados foram reunidos em uma única amostra para análise. Os autores avaliaram os artefatos coletados seguindo a rubrica (Tabela 1) indicando o nível de desempenho referente a cada critério. Algumas partes desse processo foram automatizadas utilizando um script em Python, por exemplo, para o critério de rotulação das imagens à inferência foi realizada por meio de um modelo ML [32]. Como resultado foram somados os quantitativos em cada nível dos critérios da rubrica, e em seguida se realizou a investigação das evidências da confiabilidade e da validade.

A confiabilidade refere-se à consistência ou estabilidade das pontuações dos critérios do instrumento de avaliação em um mesmo fator [42]. Inicialmente, de acordo com a TCT, se realizou a análise do desempenho de acertos e também a consistência interna foi analisada usando o coeficiente Ômega. Ao contrário do coeficiente alfa comumente utilizado, o coeficiente ômega trabalha com as cargas fatoriais, o que torna os cálculos mais estáveis, com nível de confiabilidade maior e de forma independente do número de itens do instrumento [19]. Também foram avaliados o grau de dificuldade, bisserial e discriminação dos itens [14]. Após, investigamos a adequação e calibração à TRI da rubrica usando o modelo logístico de 2 parâmetros [16, 44], utilizando a escala padrão do modelo (0,1), onde zero indica a média do grupo e 1 equivale ao desvio padrão.

A validade de construto, por outro lado, refere-se à capacidade que os critérios do instrumento conseguem medir o traço latente que o mesmo se propõe a medir [17], envolvendo a validade convergente que é obtida pelo grau de correlação entre os critérios do instrumento. Assim, foi analisada a matriz de correlação policórica, que melhor se adapta a itens categóricos [35, 43]. Também foi realizada uma análise de dimensionalidade, com análise exploratória de matriz de correlação comparados com matrizes aleatórias paralelas e uma análise fatorial exploratória [11].

3 APLICAÇÃO E COLETA DE DADOS

Com foco em alunos de escolas de ensino fundamental e médio que não tenham conhecimentos prévios de programação ou IA/ML, o curso "ML para Todos!" [23] busca promover a construção de conhecimentos a respeito de conceitos básicos de ML com foco no reconhecimento de imagens. Os objetivos de aprendizagem estão alinhados a um processo de ML centrado no ser humano [7] chegando no nível de aplicação da taxonomia de Bloom [6], seguindo as as diretrizes para ensino de IA referentes à "Grande Ideia 3 - Aprendizagem" [54] e alfabetização de IA [34]. O curso ensina a desenvolver um modelo pré-definido de ML para o reconhecimento de imagens seguindo os passos básicos de um processo de ML centrado no ser humano incluindo a preparação de dados, treinamento de modelo, avaliação de desempenho e previsão. Visando uma aplicação interdisciplinar, propõe a construção de um modelo de ML para reconhecimento de imagens abordando o tema de reciclagem de lixo, um tópico abordado na Educação Básica como parte das aulas de ciências [39], e também relacionado aos Objetivos de Desenvolvimento Sustentável das Nações Unidas [45]. Para construção do modelo de ML os estudantes são orientados a usar um ambiente visual, o GTM [20]. Após a motivação e apresentação de conceitos básicos de ML e redes neurais os alunos iniciam o desenvolvimento do modelo de ML. Além das instruções para construção do modelo é fornecido aos alunos um conjunto de imagens redimensionadas e não categorizadas, sendo os alunos então instruídos a prepararem o conjunto de dados, organizando, limpando e rotulando as imagens nas categorias de reciclagem: papel, vidro, metal e plástico. Na sequência são orientados a treinar o modelo no GTM, testar o modelo com novas imagens e interpretar o desempenho alcançado pelo modelo, levando em consideração seus testes, a precisão do modelo e a matriz de confusão fornecidas pelo GTM. Durante o curso os alunos também são instigados a ajustar o conjunto de dados e/ou alterar os parâmetros de treinamento e desta forma tentar melhorar o desempenho do modelo de ML. Os alunos também são orientados a documentar os resultados do processo de ML durante e após as atividades, o que inclui além envio do modelo gerado na ferramenta GTM (arquivo .tm), relatórios preenchidos de forma online que documentam a análise e interpretação do desempenho e resultados da predição (Figura 1). Todos estes artefatos resultantes do processo de criação do modelo de ML são coletados como base para

a avaliação da aprendizagem do aluno com base no seu desempenho.

Foram realizados 4 casos de aplicação do curso "ML para todos!" nos anos de 2021 a 2022 com alunos dos anos finais do Ensino Fundamental e Médio com idades entre 12 e 18 anos. As aulas foram remotas on-line com instrutores, sendo consideradas como atividades extracurriculares para os alunos. A escolha deste formato se deu por dois principais motivos: a simultaneidade de oferta com pandemia mundial da Covid-19 e a intencionalidade de disponibilizar o treinamento a uma ampla e distribuída parcela da população alvo. A exceção foi de uma aplicação em uma escola municipal, na qual o curso foi aplicado de forma híbrida como parte das aulas escolares. Um total de 108 alunos entregaram, ainda que parcialmente, os artefatos criados ao longo do curso. Esta pesquisa foi aprovada pelo Comitê de Ética da Universidade Federal de Santa Catarina (No. 4.893.560).

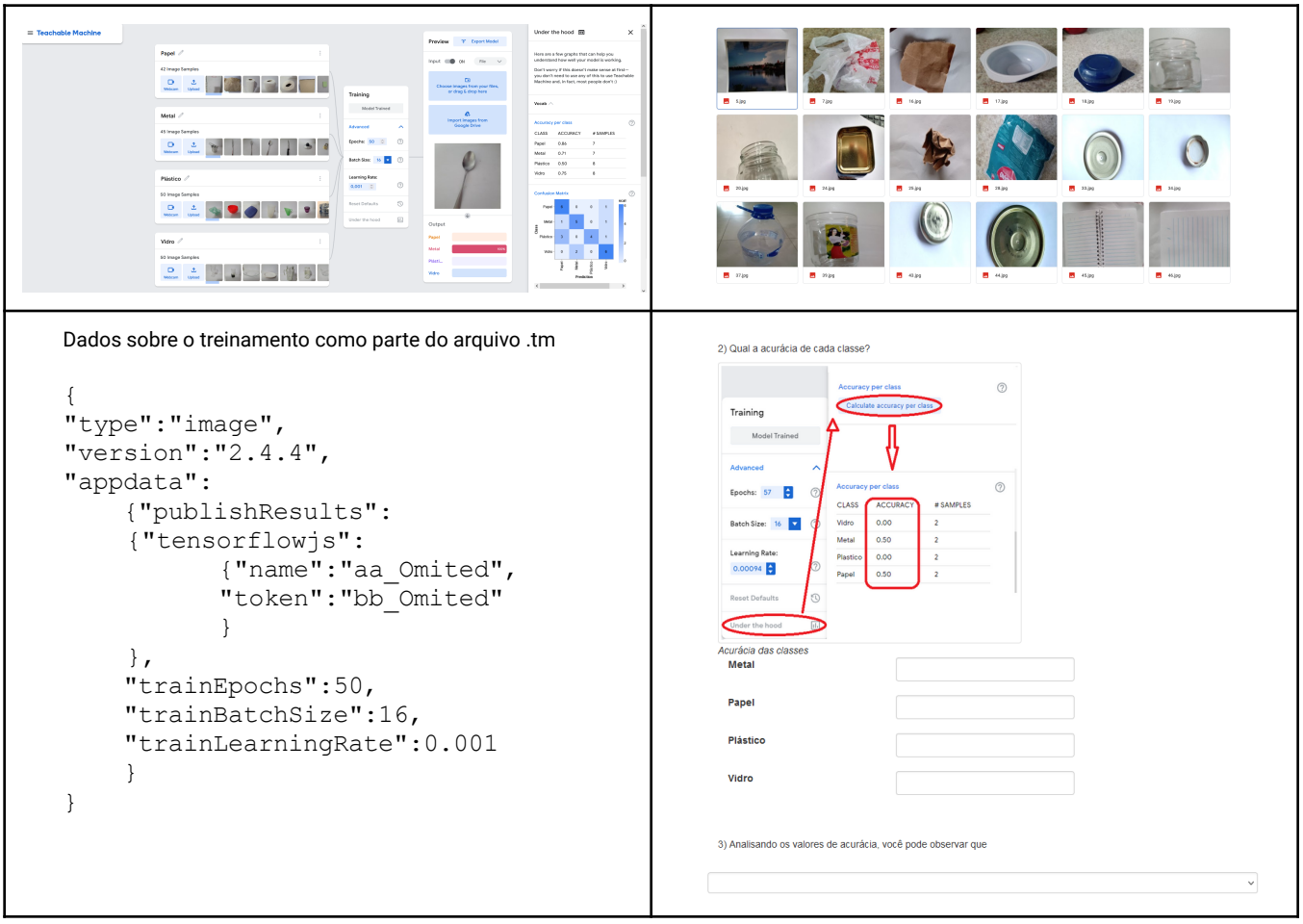

Figura 1: Exemplos de artefatos criados pelos alunos como resultado da aprendizagem

4 ANÁLISE DOS DADOS

Com base na análise do desempenho dos artefatos criados pelos estudantes utilizando a rubrica (Tabela 1), foram levantadas as frequências nos níveis de desempenho atingidos pelos estudantes ao longo do curso "ML para Todos!", conforme apresentado na Tabela 2. Alguns critérios da rubrica não puderam ser inferidos, pois alguns estudantes não entregaram alguns dos diferentes artefatos considerados (indicado por NA´s).

Como nem todos os estudantes entregaram todos os artefatos, há diferenças nas somas dos dados considerados nas frequências nos níveis de desempenho dos diferentes critérios. Observa-se que as frequências variam de 87 referente aos critérios "C7-Testes com novos objetos" e "C8-Interpretação dos testes", até somente 65 em relação ao critério "C10-Interpretação da matriz de confusão".

Tabela 2: Distribuição de frequências de níveis de desempenho por critério da rubrica

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11
Fraco	22	1	3	3	17	0	0	34	19	12	13
Aceitável	16	0	28	44	6	60	3	0	10	8	37
Bom	32	69	39	23	47	10	84	53	37	45	22
Soma dos Entregues	70	70	70	70	70	70	87	87	66	65	72
NA´s	38	38	38	38	38	38	21	21	42	43	36
Total	108	108	108	108	108	108	108	108	108	108	108

Diante do grande número de NA´s, e da potencial imprecisão ser inserida ao manter-se esses dados, optou-se por desconsiderar os estudantes com respostas com NA´s. Também, devido ao atual tamanho limitado da amostra, optou-se por uma análise dicotomizada, agrupando os níveis "1-Aceitável" e "2-Bom" dos critérios da rubrica para um novo nível de desempenho chamado "1-Adequado". O resultado é apresentado na Tabela 3.

Tabela 3: Distribuição de frequências de níveis de desempenho por critério da rubrica sem NA´s e dicotomizada

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11
Fraco	11	0	1	1	9	0	0	17	13	7	8
Adequado	32	43	42	42	34	43	43	26	30	36	35
Total	43	43	43	43	43	43	43	43	43	43	43

Ao analisar as frequências de pontuações da amostra apresentadas na Tabela 3, se observa a necessidade de eliminar alguns itens que não apresentam variação. Desta forma, os itens indicados por C2, C6 e C7 foram eliminados, já que nesta amostra não apresentaram nenhuma resposta na categoria 0-fraco. O item C4 também foi eliminado, pois nesta amostra os itens C3 e C4 apresentam os mesmos quantitativos e suas categorias.

5 RESULTADOS

Partindo-se da CTT, o desempenho dos respondentes é visualizado na Figura 2. O número médio de acertos dos 7 itens considerados foi de 5,5 com desvio padrão de 1,4.

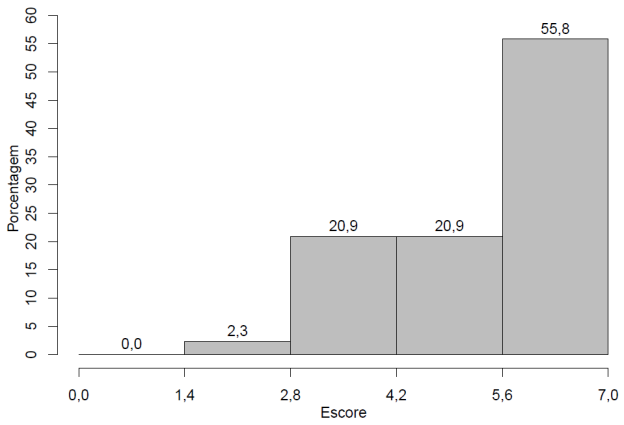

Figura 2: Desempenho dos estudantes

A confiabilidade medindo a consistência interna da rubrica ML foi analisada por meio do coeficiente Ômega. De acordo com a literatura, Ômega > 0,70 indica confiabilidade do conjunto de fatores (um valor entre 0,7 e 0,8 é aceitável, de 0,8 até 0,9 são bons, e maiores ou iguais a 0,9 são excelentes) [11]. Como resultado foi obtido um valor aceitável Ômega Global de 0,781.

Ao se analisar se a consistência interna aumenta eliminando um item (Tabela 4), se observa que o coeficiente aumenta eliminando alguns itens (C1, C5 e C8), o que pode indicar necessidade de revisão destes ítens.

Tabela 4: Coeficiente ômega ao excluir itens

Item	Ômega
C1	0,783
C3	0,740
C5	0,787
C8	0,791
C9	0,669
C10	0,736
C11	0,752

Com relação a qualidade dos itens (Tabela 5) a análise de dificuldade (Dif), bisserial (Bis) e discriminação (Disc) dos itens, apenas para o item C3 é indicada a revisão por apresentar uma dificuldade e discriminação baixa.

Tabela 5: Qualidade dos itens segundo TCT

Item	Dif	Clas. Dif	Bis	Class. Bis	Disc	Clas. Disc
C1	74,4	Fácil	0,106	Moderado	0,124	Moderada
C3	97,7	Muito Fácil	0,454	Excelente	0,053	Baixa
C5	79,1	Fácil	0,323	Excelente	0,241	Adequada
C8	60,5	Médio	0,224	Adequado	0,151	Moderada
C9	69,8	Fácil	0,715	Excelente	0,491	Excelente
C10	83,7	Fácil	0,564	Excelente	0,294	Adequada
C11	81,4	Fácil	0,429	Excelente	0,312	Excelente

5.1 Há evidências da confiabilidade da rubrica por meio da TRI?

Apesar de normalmente iniciar-se com a análise da carga fatorial associada à avaliação, dado a pequena amostra atualmente disponível, optou-se por iniciar verificando se os ítens calibrarem com a TRI num modelo de 2 parâmetros dicotomizados e qualidade desta calibração (Tabela 6) [16, 44]. Ao analisar o resultado da calibração da TRI do modelo de 2 parametros, o parametro "a" indica o padrão de discriminação e está associado com a qualidade do do ítem, enquanto o valor de "b" indica o índice de dificuldade do item.

Tabela 6: Resultado da calibração do modelo de 2 parâmetros dicotomizados

Item	a		b	
	Valor	SE	Valor	SE
C1	0,669	0,428	-1,761	1,092
C3	0,821	0,516	-4,972	2,903
C5	0,806	0,420	-1,881	0,920
C8	0,732	0,410	-0,654	0,560
C9	1,345	0,490	-0,487	0,371
C10	1,086	0,465	-1,856	0,693
C11	0,944	0,438	-1,851	0,780

O item C1 apresenta uma qualidade de discriminação (valor de a) ligeiramente baixa, mas muito próxima do valor adequado de 0,7, portanfo foi mantido. Também o item C3 apresenta um índice de dificuldade (parâmetro b) ligeiramente alto, próximo a -5, mas foi mantido. De uma forma geral, os demais itens parecem estar adequados como também pode ser observado na Figura 3, onde a inclinação das linhas em seu ponto médio indica o seu padrão de discriminação.

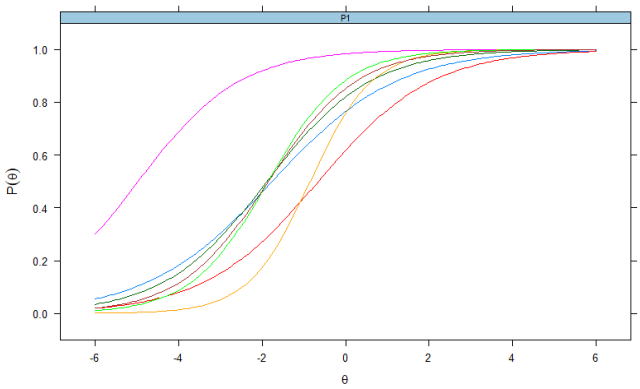

Figura 3: Gráfico de probabilidade de desempenho em função do score de todos os itens

A Figura 4 apresenta o gráfico que mostra onde estão distribuídas as dificuldades dos itens em relação a média do grupo, onde em conjunto com a Figura 3, se percebe que na amostra utilizada a proficiência (theta) necessária para ter uma probabilidade de acerto dos itens aponta que os itens são ligeiramente fáceis, estando abaixo da média de acertos do grupo (theta igual a zero).

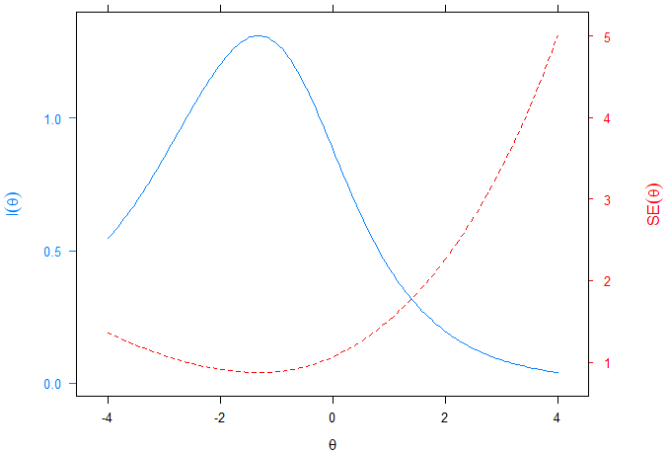

Figura 4: Informação do teste e erro padrão

Escala. Foi mantida a escala padrão da TRI, em que a média é indicada pelo desempenho igual a zero e cada nível indica um desvio padrão.

Na Tabela 7 é indicada a probabilidade de um item estar posicionado em determinado nível da escala. Pela cor de fundo podemos observar a indicação de onde devem ser posicionados os itens dentro desta escala, marcado em verde quando o padrão de discriminação for superior a 1,0. Como está sendo utilizado um modelo dicotomizado o item deve ser posicionado onde assume uma probabilidade maior que 50%.

Tabela 7: Probabilidade de posicionamento dos itens na escala (0,1)

Item	Níveis da Escala								
	-4	-3	-2	-1	0	1	2	3	4
C1	0,18	0,30	0,46	0,62	0,76	0,86	0,93	0,96	0,98
C3	0,69	0,83	0,92	0,96	0,98	0,99	1,00	1,00	1,00
C5	0,15	0,29	0,48	0,67	0,82	0,91	0,96	0,98	0,99
C8	0,08	0,15	0,27	0,44	0,62	0,77	0,87	0,94	0,97
C9	0,01	0,05	0,17	0,45	0,76	0,92	0,98	0,99	1,00
C10	0,09	0,22	0,46	0,72	0,88	0,96	0,99	0,99	1,00
C11	0,12	0,25	0,46	0,69	0,85	0,94	0,97	0,99	1,00

A partir da probabilidade apresentada na da Tabela 7, podemos inferir o grau de dificuldade de cada item, o que está representado na Figura 5.

Nível da Escala	Itens			
4				
3				
2				
1				
0	C8	C9		
-1	C1	C5	C10	C11
-2				
-3				
-4	C3			

Mais Difícil

Mais Fácil

Figura 5: Posicionamento dos itens na escala (0,1)

De acordo com o modelo dicotômico utilizado, a Figura 6 destaca a interpretação dos níveis da escala, na qual se evidencia o eventual desempenho de um estudante diante da escala da rubrica proposta.

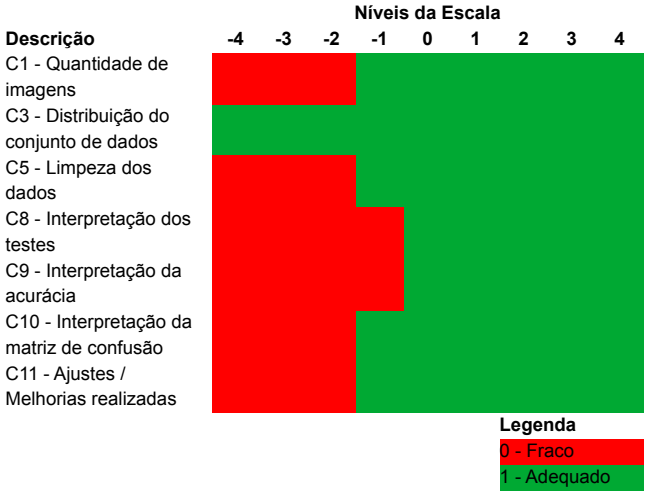

Figura 6: Representação gráfica do nível de desempenho dos estudantes

Assim, partindo da representação gráfica na Figura 6 pode ser inferida uma interpretação a ser dada aos escores obtidos pelos estudantes (Tabela 8).

5.1 Há evidências de validade da rubrica por meio da carga fatorial da análise fatorial (dimensionalidade)?

Foi analisada a validade convergente primeiramente por meio do grau de correlação entre os critérios do instrumento. Para este propósito foi analisada a matriz de correlação policórica dos critérios da rubrica (Tabela 9). Nesta análise espera-se que os critérios que estejam medindo uma única dimensão apresentem correlações maiores ou iguais a 0,30 [17]. Neste mesmo sentido, correlações (r) cujo valor em módulo não ultrapasse 0,5 (0,30 ≤ | r | < 0,50) é considerada uma correlação linear fraca, e até 0,7 (0,50 ≤ | r | < 0,70) correlação moderada e acima (0,70 ≤ | r | < 0,90) forte ou (| r | ≥ 0,90) muito forte [43].

Tabela 8: Interpretação do desempenho dos estudantes de acordo com a escala

Nível da escala	Análise descritiva
Abaixo de -2	As imagens utilizadas e rotuladas em cada categoria estão adequadamente distribuídas no modelo criado.
-1 até 0	A quantidade de imagens, sua distribuição, a limpeza dos dados, interpretação da matriz de confusão e ajustes e melhorias estão adequados ao modelo criado. No entanto, a interpretação dos testes e da acurácia está fraca.
Acima 0	A quantidade de imagens, sua distribuição, a limpeza dos dados, interpretação da matriz de confusão, interpretação de testes, interpretação da acurácia e ajustes e melhorias estão adequados ao modelo criado.

Tabela 9: Matriz de correlação policórica

Itens	C1	C3	C5	C8	C9	C10	C11
C1	1,000						
C3	0,565	1,000					
C5	-0,147	-0,515	1,000				
C8	-0,130	0,449	0,150	1,000			
C9	0,544	0,442	0,132	0,415	1,000		
C10	-0,322	-0,457	0,483	0,306	0,418	1,000	
C11	0,140	-0,497	0,408	-0,149	0,364	0,540	1,000

Se observa na matriz de correlação policórica para a rubrica que há vários pares de critérios que apresentam correlação acima de 0,3, o que indica relação estatística na associação entre o par. Destacadas em verde estão as correlações em condição de significância estatística. O maior valor de 0,56 para a correlação foi alcançado para a associação entre os critérios C1xC3. Também podemos observar correlações negativas, que indicam que há uma relação inversamente proporcional entre o par, isto é, quando um critério da rubrica aumenta o outro diminui, algo que não é esperado nesta análise.

Já a análise exploratória de matriz de correlação comparados com matrizes aleatórias paralelas (Figura 7) indica a existência de 3 fracas dimensões ou traços latentes na amostra, que estão representadas pelos x´s em azul acima da linha pontilhada vermelha.

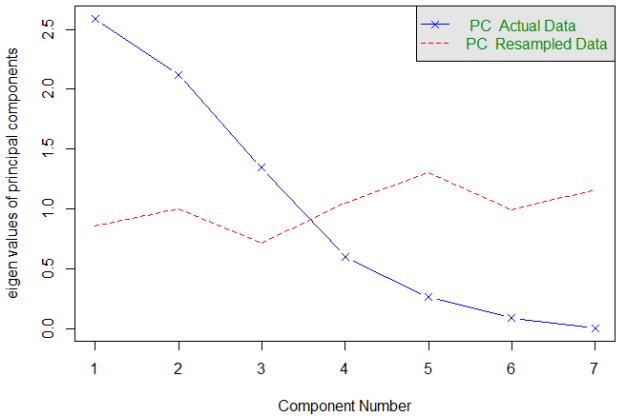

Figura 7: Matriz de correlação comparados com matrizes aleatórias paralelas

Partindo-se para uma análise fatorial exploratória (Tabela 10), cujas estatísticas são pouco sensíveis ao tamanho da amostra, consideramos os valores que avaliam a qualidade do ajuste do modelo quanto ao erro quadrático médio de aproximação (RMSEA), Índice de ajuste comparativo (CFI) e Índice de Tucker– Lewis (TLI). O ajuste é considerado adequado quando o RMSEA < 0,05, o TLI > 0,90 e CFI > 0,90 [11].

Tabela 10: Análise fatorial exploratória

Dimensões	Variação explicada pela dimensão	RMSEA	TLI	CFI
Uma	0,37	0,072	0,764	0,842
Duas	0,66	0	1,24	1
Três	0,76	0	1,797	1

Com relação a variabilidade do fator, todas as 3 propostas de dimensões testadas ficaram adequadas com uma valor acima de 20%. O RMSEA para uma dimensão, apesar de não ser o ideal abaixo de 0,05, está em uma condição adequada na faixa entre 0,06 e 0,08 [18]. É interessante observar que para duas e três dimensões, apesar de haver convergência do modelo, o RMSEA ficou em zero, o CFI em 1 e o TLI acima de 1, resultados estes de que não eram esperados, possivelmente oriundos da não convergência dos modelos.

6 DISCUSSÃO

A quantidade de respostas consideradas, isto é, as de artefatos de 43 alunos depois de serem eliminados os respondentes com respostas parciais (NA´s) é pequena para análises utilizando a TRI. Mesmo assim, os resultados da amostra indicam que a rubrica para avaliar o desempenho da aprendizagem de ML atingiu níveis mínimos de consistência interna e há indícios significativos de validade convergente. Como resultado foi obtido um valor aceitável de Ômega Global de 0,781.

A matriz de correlação comparada com matrizes aleatórias indica um modelo com 3 dimensões fracas. Ao mesmo tempo, os dados da análise fatorial exploratória indicam melhores parâmetros para o modelo com uma dimensão. A análise fatorial também é fortemente influenciada pelo tamanho da amostra.

Ao considerar a TRI num modelo de 2 parâmetros dicotomizados, apesar das exclusões devido a dados nulos (NA´s) e invariabilidade em itens da amostra, os ítens calibraram e a qualidade desta calibração chegou-se a valores adequados de qualidade de discriminação (parametro a) e também que a dificuldade dos ítens (parametro b). Desta forma, foi possível identificar a dificuldade e discriminação dos ítens em uma escala e a conseguente definição da interpretação a ser dada ao desempenho dos estudantes.

Ainda assim, a distribuição normal da dificuldade dos itens mostrou-se inferior à média da amostra. Desta forma uma sugestão de melhoria da rubrica pode ser a inclusão de itens com uma maior dificuldade, ampliando assim o espectro de cobertura dos itens em relação a dificuldade.

Os resultados da presente pesquisa estão de acordo com as análises preliminares realizadas que indicaram substancial concordância *inter-rater* de um painel de especialistas quanto rubrica utilizada na avaliação da aprendizagem, bem como a validade em termos de corretude, relevância, completude e clareza [22]. Rauber *et al*. [48] ao analisar a confiabilidade do instrumento aponta um coeficiente ômega global de 0,646 ante 0,781 apontado neste estudo, o que se mostra melhor. Ainda Rauber *et al*. [48] ao analisar a validade do instrumento e considerar a matriz de correlação policórica discute a possibilidade da existência de duas dimensões, análogo ao encontrado neste trabalho, que indicou a possibilidade de 3 dimensões fracas. No entanto, ao ser realizada a análise dos indicadores de qualidade de ajuste do modelo (Tabela 10) comparando exploratoriamente a existência de uma, duas ou três dimensões, se percebe a indicação de uma única dimensão. Fato este também corroborado pela qualidade da calibração dos parâmetros da TRI (Tabela 6). De posse desses resultados, foi possível a criar uma escala, posicionar os critérios da rubrica nesta escala (Figura 5) e interpretar o resultado do desempenho dos estudantes (Tabela 8).

Também é de suma importância aumentar o tamanho da amostra e sua variabilidade para poder incluir os itens excluídos por invariabilidade e possibilitar futuras análises politômicas.

De forma geral, os resultados da análise mostram que a rubrica de ML está muito próxima de ser um instrumento confiável e válido, podendo ser aplicada para avaliar a aprendizagem de ML voltada a classificação de imagens com GTM na Educação Básica. Contudo, observando as questões identificadas é importante ressaltar que os resultados da rubrica devem ser revisados pelo instrutor. A rubrica também representa apenas uma alternativa para medir a aprendizagem de ML do estudante e que deve ser completada por outros métodos de avaliação, tais como entrevistas, revisões por pares, apresentações, etc., como sugerido por exemplo também no contexto da aprendizagem de pensamento computacional [12, 8, 24].

Ameaças à validade. A fim de minimizar impactos de validade nesse estudo, identificamos ameaças potenciais e aplicamos estratégias de mitigação. A fim de mitigar as ameaças relacionadas ao projeto do estudo e definição da análise, foi adotada uma metodologia sistemática seguindo a abordagem GQM [10]. Outra questão refere-se à qualidade dos dados agrupados em uma única amostra. Isso foi possível pela padronização dos dados, todos coletados da mesma maneira em de aplicações do curso "ML para Todos!". Outro risco se refere à validade das pontuações alocadas com base nos dados coletados. Como nosso estudo se limita às avaliações utilizando a rubrica de ML, este risco é minimizado, pois as análises foram realizadas de forma (semi-) automatizada (utilizando um script Python), a partir da mesma rubrica. Somente os critérios C2, C5, C9 e C10 foram manualmente analisados pelos autores. Neste caso, a avaliação foi feita por um pesquisador e revisada por um segundo pesquisador para reduzir o risco de erros na pontuação. Outro risco é o agrupamento de dados de vários contextos. Entretanto, como o objetivo é analisar a validade da rubrica de forma independente do contexto, isto não é considerado um problema aqui. Outra ameaça à validade externa está associada ao tamanho da amostra e à diversidade dos dados utilizados. Nossa análise é baseada em uma amostra de 108 alunos. Isto é considerado um tamanho de amostra suficiente para uma pesquisa exploratória, porém levando em consideração os resultados das análises deve ser aumentado no futuro para revisar os resultados obtidos, incluindo possivelmente análises politômicas dos itens.

7 CONCLUSÃO

Em geral, os resultados desta avaliação mostram que a rubrica para a avaliação da aprendizagem de ML está próxima de representar um instrumento com confiabilidade e validade aceitáveis que poderá ser usado para a avaliação da construção de modelos de ML para classificação de imagens usando GTM, como parte da educação em computação nas escolas.

Foi possível calibrar os itens com a TRI num modelo de 2 parâmetros dicotomizados, com qualidade de discriminação (parametro a) e também dificuldade dos ítens (parametro b) dentro das condições de aceitabilidade. A análise de dimensionalidade deve ser tomada de forma exploratória, dado o tamanho da amostra. Apesar do indicativo de 3 dimensões fracas, os valores encontrados referentes a qualidade do ajuste a essas dimensões, se mostram mais consistentes com o modelo de uma única dimensão.

Com base nesses resultados positivos, está sendo implementada a integração da avaliação na ferramenta CodeMaster [21], de modo a fornecer suporte completamente automatizado que ajuda a garantir a consistência, rapidez e a precisão dos resultados da avaliação, bem como a eliminar viés. Os atuais resultados e a implementação proposta tem o potencial de auxiliar em um processo de avaliação adequado tanto aos estudantes quanto à avaliação da sua aprendizagem. Além disso, também poderá reduzir a carga de trabalho dos professores e deixá-los livres para dedicar mais tempo a outras atividades com os alunos, bem como para realizar outras avaliações complementares sobre fatores que não são facilmente automatizados, como a criatividade.

AGRADECIMENTOS

Gostaríamos de agradecer a todos os alunos que participaram do curso.

O presente trabalho foi realizado com apoio do Conselho Nacional de Desenvolvimento Científico e Tecnológico – Brasil (CNPq).

REFERÊNCIAS

[1] Alves, N. da C., Gresse von Wangenheim, C., Hauck, J. C. R., Borgatto, A. F., (2021a), An Item Response Theory Analysis of Algorithms and Programming Concepts in App Inventor Projects. In: *Anais do Simpósio Brasileiro de Educação em Computação*, Jataí, Goiás.

[2] Alves, N. da C., Gresse von Wangenheim, C., Hauck, J. C. R., Borgatto, A. F., (2020a), A Large-scale Evaluation of a Rubric for the Automatic Assessment of Algorithms and Programming Concepts. In: Proc. of the 51st ACM Technical Symposium on Computer Science Education, Portland/USA, Pages 556–562.

[3] Alves, N. da C., Solecki, I., Gresse von Wangenheim, C., Borgatto, A. F. Hauck, J. C. R., Ferreira, M. N. F., (2020b), Análise do Nível de Dificuldade dos Conceitos de Design de Interface de Usuário usando a Teoria de Resposta ao Item. In: *Anais do Simpósio Brasileiro de Informática na Educação*, Natal, Brasil.

[4] Alves, N. da C., Gresse von Wangenheim, C., Alberto, M., Martins-Pacheco, L. H., (2020c), Uma Proposta de Avaliação da Originalidade do Produto no Ensino de Algoritmos e Programação na Educação Básica. In: Anais do Simpósio Brasileiro de Informática na Educação, Natal, Brasil.

[5] Alves, N. da C., Gresse von Wangenheim, C., Martins-Pacheco, L. H., Borgatto, A. F., (2021b), Existem concordância e confiabilidade na avaliação da criatividade de resultados tangíveis da aprendizagem de computação na Educação Básica? In: Anais do Simpósio Brasileiro de Educação em Computação, Jataí, Goiás.

[6] Anderson L. W. and Krathwohl D. R., (2001), *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*, Longman.

[7] Amershi S. et al. (2019), Software Engineering for Machine Learning: A Case Study. Proc. of the *41st International Conference on Software Engineering: Software Engineering in Practice*, IEEE, 291–300.

[8] Avila C. et al. (2017), Metodologias de Avaliação do Pensamento Computacional: uma revisão sistemática. *Anais do Simpósio Brasileiro de Informática na Educação*, 113.

[9] BRASIL, (1996), LEI Nº 9.394, de 20 de dezembro de 1996. Estabelece as diretrizes e bases da educação nacional. Retrieved 01/09/2022 from http://www.planalto.gov.br/ccivil_03/leis/l9394.htm

[10] Basili V. R., Caldiera G., and Rombach H. D., (1994), Goal Question Metric Paradigm. In *Encyclopedia of Software Engineering*, Wiley.

[11] Brown T. A., (2015), Confirmatory factor analysis for applied research, Second edition. The Guilford Press.

[12] Brennan K. e Resnick M., (2012), New frameworks for studying and assessing the development of computational thinking. *Proc. of the Annual Meeting of the American Educational Research Association, Vancouver, Canada*, 25.

[13] Camada M. Y. e Durães G. M., (2020), Ensino da Inteligência Artificial na Educação Básica: um novo horizonte para as pesquisas brasileiras. *Anais do XXXI Simpósio Brasileiro de Informática na Educação, SBC,* 1553–1562.

[14] Cappelleri J. C., Jason Lundy J., and Hays R. D., (2014), *Overview of Classical Test Theory and Item Response Theory for the Quantitative Assessment of Items in Developing Patient-Reported Outcomes Measures*. Clinical Therapeutics, 36(5), 648–662.

[15] Caruso A. L. M. e Cavalheiro S. A. da C., (2021), Integração entre Pensamento Computacional e Inteligência Artificial: uma Revisão Sistemática de Literatura. *Anais do XXXII Simpósio Brasileiro de Informática na Educação*, SBC, 1051–1062.

[16] de Andrade D. F., Tavares H. R., and da Cunha Valle R., (2000), Teoria da Resposta ao Item: conceitos e aplicações. *ABE, Sao Paulo*.

[17] DeVellis R. F., (2017), *Scale development: theory and applications*, 4th ed. SAGE.

[18] Finch, J. F. & West, SG (1997). The investigation of personality structure: statistical models. *Journal of Research in Personality*, 31(4), 439-485.

[19] Flora D. B., (2020), Your Coefficient Alpha Is Probably Wrong, but Which Coefficient Omega Is Right? A Tutorial on Using R to Obtain Better Reliability Estimates. *Advances in Methods and Practices in Psychological Science*, 3(4), 484–501.

[20] Google, (2020), Google Teachable Machine. Retrieved 01/06/2022 from https://teachablemachine.withgoogle.com/,

[21] Gresse von Wangenheim C. *et al*., (2018), CodeMaster - Automatic Assessment and Grading of App Inventor and Snap! Programs. *Informatics in Education*, 17(1), 117–150.

[22] Gresse von Wangenheim C., Alves N. da C., Rauber M. F., Hauck J. C. R., and Yeter I. H., (2021), A Proposal for Performance-based Assessment of the Learning of Machine Learning Concepts and Practices in K-12. *Informatics in Education,* online.

[23] Gresse von Wangenheim C., Marques L. S., and Hauck J. C. R., (2020), Machine Learning for All – Introducing Machine Learning in K-12, SocArXiv, 1-10.

[24] Grover S., Pea R., and Cooper S., (2015), "Systems of Assessments" for deeper learning of computational thinking in K-12. *Proc. of the Annual Meeting of the American Educational Research Association*, 15–20.

[25] Hattie J. and Timperley H., (2007), The Power of Feedback. *Review of Educational Research*, 77(1), 81–112.

[26] Hitron T., Orlev Y., Wald I., Shamir A., Erel H., and Zuckerman O., (2019), *Can Children Understand Machine Learning Concepts?: The Effect of Uncovering Black Boxes*, *Proc. of the 2019 CHI Conference on Human Factors in Computing Systems*, ACM, 1–11.

[27] Ho J. W. and Scadding M., (2019), Classroom Activities for Teaching Artificial Intelligence to Primary School Students. *Proc. of the Int. Conference on Computational Thinking*, 157-159.

[28] House of Lords, (2018), AI in the UK: ready, willing and able, HL Paper 100.

[29] Hsu T.-C., Abelson H., and Van Brummelen J., (2021), The Effects of Applying Experiential Learning into the Conversational AI Learning Platform on Secondary School Students. *Preview Version. Accepted by the IRRODL Special Issue "AI e-Learning and Online Curriculum".*

[30] Huba, M. E., and Freed, J. E. (2000). *Learner-centered assessment on college campuses: Shifting the focus from teaching to learning*. Allyn & Bacon.

[31] Kandlhofer M., Steinbauer G., Hirschmugl-Gaisch S., and Huber P., (2016), Artificial intelligence and computer science in education: From kindergarten to university. Proc. of the *Frontiers in Education Conference, IEEE*, 1–9.

[32] Laydner M., (2022), Automação da Avaliação de Aprendizagem de Machine Learning para classificação de Imagens no Ensino Fundamental. Trabalho de Conclusão de Curso. (Graduação em Sistemas de Informação) – Universidade Federal de Santa Catarina.

[33] LeCun Y., Bengio Y., and Hinton G., (2015), Deep learning. *Nature*, 521(7553), 436–444.

[34] Long D. and Magerko B., (2020), What is AI literacy? Competencies and design considerations. *Proc. of the Conference on Human Factors in Computing Systems*, ACM, 1–16.

[35] Lordelo L. M. K., Hongyu K., Borja P. C., e Porsani M. J., (2018), Análise Fatorial por Meio da Matriz de Correlação de Pearson e Policórica no Campo das Cisternas. *E&S Engineering and Science*, **7**(1), 58–70.

[36] Lwakatare L. E., Raj A., Bosch J., Olsson H. H., and Crnkovic I., (2019), A taxonomy of software engineering challenges for machine learning systems: An empirical investigation. Proc. of the *Int. Conference on Agile Software Development*, Springer, 227–243.

[37] Lye S. Y. and Koh J. H. L., (2014), Review on teaching and learning of computational thinking through programming: What is next for K-12? *Computers in Human Behavior*, 41, 51–61.

[38] Marques L. S., von Wangenheim C. G., e Rossa Hauck J. C., (2020), Ensino de Machine Learning na Educação Básica: um Mapeamento Sistemático do Estado da Arte. *Anais do XXXI Simpósio Brasileiro de Informática na Educação*, SBC, 21–30.

[39] Ministério da Educação, (2018), Base Nacional Comum Curricular. Retrieved 01/06/2022 from http://basenacionalcomum.mec.gov.br/

[40] Mislevy R. J., Almond R. G., and Lukas J. F., (2003), A Brief Introduction to Evidence-Centered Design. *ETS Research Report Series*, 2003(1), i–29.

[41] Mitchell T. M., (1997), Machine Learning, New York:McGraw-Hill.

[42] Moskal B. M. and Leydens J. A., (2000), Scoring rubric development: Validity and reliability. *Practical assessment, research, and evaluation*, 7(1), 10.

[43] Mukaka M. M., (2012), A guide to appropriate use of correlation coefficient in medical research. *Malawi Medical journal*, 24(3), 69–71.

[44] Paek I. and Cole K., (2019), *Using R for Item Response Theory Model Applications*, 0 ed. Routledge.

[45] Pedro F., Subosa M., Rivas A., and Valverde P., (2019), Artificial intelligence in education: Challenges and opportunities for sustainable development. UNESCO.

[46] Ramos G., Meek C., Simard P., Suh J., and Ghorashi S., (2020), Interactive machine teaching: a human-centered approach to building machine-learned models. *Human–Computer Interaction*, 35(5–6), 413–451.

[47] Rauber M. F. and Gresse von Wangenheim C., (2022), Assessing the Learning of Machine Learning in K-12: A Ten-Year Systematic Mapping. *Informatics in Education*, online.

[48] Rauber M. F., Garcia A. B. Gresse von Wangenheim C., Borgatto, A.F, Martins, R. M. Hauck, J. C.. (2022), Confiabilidade e Validade da Avaliação do Desempenho de Aprendizagem de Machine Learning na Educação Básica. Proc. of *XXXIII Simpósio Brasileiro de Informática na Educação*, online.

[49] Royal Society, (2017), *Machine learning: the power and promise of computers that learn by example.* Retrieved 01/06/2022 from royalsociety.org/machine-learning.

[50] Sadler D. R., (1989), Formative assessment and the design of instructional systems. *Instructional Science*, 18(2), 119–144.

[51] Shamir G. and Levin I., (2021), Neural Network Construction Practices in Elementary School. Künstliche Intelligenz, 35(2), 181–189.

[52] SOLECKI, I.; PORTO, J. A.; ALVES, N. d. C., GRESSE VON WANGENHEIM, C., HAUCK, J. C. R., BORGATTO, A. F. Automated Assessment of the Visual Design of Android Apps Developed with App Inventor. In: *Proc. of the 51st ACM Technical Symposium on Computer Science Education*, Portland, USA, 2020, Pages 51–57.

[53] Tang D., Utsumi Y., and Lao N., (2019), PIC: A Personal Image Classification Webtool for High School Students. *Proc. of the 2019 IJCAI EduAI Workshop. IJCAI*.

[54] Touretzky D., Gardner-McCune C., Martin F., and Seehorn D., (2019), Envisioning AI for K-12: What Should Every Child Know about AI? Proc. of the *AAAI Conference on Artificial Intelligence*, 9795–9799.
[55] Trochim W. M. K. and Donnelly J. P., (2008), *The research methods knowledge base*, 3rd ed. Mason, Atomic Dog/Cengage Learning.
[56] UNESCO, (2022), K-12 AI curricula: a mapping of government-endorsed AI curricula. Retrieved 06/06/2022 from https://unesdoc.unesco.org/ark:/48223/pf0000380602
[57] Yasar O., Veronesi P., Maliekal J., Little L., Vattana S., and Yeter I., (2016), Computational Pedagogy: Fostering a New Method of Teaching. Proc. of the *Annual Conference & Exposition Proceedings*, ASEE, 26550.
[58] Morrison, Gary R. et al. (2019), *Designing effective instruction*. Eighth editioned. Hoboken, NJ: Wiley.
[59] McMillan, James H. (org.) (2013), *Sage handbook of research on classroom assessment*. Los Angeles: Sage Publications.
[60] Ministério da Educação (2022), *Normas sobre Computação na Educação Básica – Complemento à Base Nacional Comum Curricular (BNCC)*. Parecer 02/2022 CNE/CEB/MEC. Retrieved 20/12/2022 from http://portal.mec.gov.br/component/content/article/323-secretarias-112877938/orgaos-vinculados-82187207/12992-diretrizes-para-a-educacao-basica?Itemid=164