Confiabilidade e Validade da Avaliação do Desempenho de Aprendizagem de *Machine Learning* na Educação Básica

Marcelo Fernando Rauber[1, 2], Abisague Belém Garcia[3], Christiane Gresse von Wangenheim[1], Adriano F. Borgatto[3], Ramon Mayor Martins[1], Jean C. R. Hauck[1]

[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 - Brazil

{marcelo.rauber@ifc.edu.br, abisague.garcia@prof.pmf.sc.gov.br, c.wangenheim@ufsc.br, adriano.borgatto@ufsc.br, ramon.mayor@posgrad.ufsc.br, jean.hauck@ufsc.br}

Abstract. *Observing the trend of teaching Machine Learning (ML) already in K-12, the need for assessing the learning also arises. In order to ensure a reliable and valid assessment, we present the evaluation of a rubric for the assessment of the learning of the application of ML concepts based on the learning outcomes of 108 middle and high school students. Both the reliability analysis (Omega coefficient of 0.646) and the analysis of the convergent validity of the construct through the polychoric correlation matrix indicate the possibility of two dimensions. Even indicating the need for a revision with a larger sample, these results can already support the application of the rubric.*

Resumo. *Observando a tendência de ensinar Machine Learning (ML) já na educação básica, surge também a necessidade de avaliação da aprendizagem. Com o objetivo de assegurar uma avaliação confiável e válida, apresentamos a avaliação de uma rubrica para a avaliação da aprendizagem da aplicação de conceitos de ML com base nos resultados da aprendizagem de 108 alunos do ensino fundamental e médio. Tanto a análise da confiabilidade (Coeficiente Ômega de 0,646) quanto a análise da validade convergente do construto por meio da matriz de correlação policórica indicam a possibilidade de duas dimensões. Mesmo indicando a necessidade de revisão com uma amostra maior, esses resultados já podem auxiliar na aplicação da rubrica.*

1. Introdução

O aprendizado de máquinas (*Machine Learning* - ML) tornou-se parte de nossa vida cotidiana, impactando profundamente nossa sociedade. O ML se concentra no desenvolvimento de sistemas que aprendem e evoluem a partir da sua própria experiência sem ter que ser explicitamente programados. Progressos recentes em ML foram alcançados especificamente por abordagens de aprendizagem profunda utilizando redes neurais, melhorando drasticamente o estado da arte em reconhecimento de imagem, detecção de objetos e reconhecimento de fala em muitos domínios (LeCun *et al*., 2015).

No entanto, a grande maioria da população não compreende a tecnologia por trás dela, que pode tornar o ML misterioso ou mesmo assustador, ofuscando seu potencial

impacto positivo na sociedade (Ho e Scadding, 2019). 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 (House of Lords, 2018). Portanto, torna-se importante introduzir conceitos e práticas básicas já na escola (Camada e Durães, 2020; Caruso e Cavalheiro, 2021), capacitando os estudantes a se tornarem mais do que apenas consumidores, mas também criadores de soluções inteligentes (Kandlhofer *et al*., 2016; Royal Society, 2017; Touretzky *et al*., 2019).

Como indicado pelas diretrizes curriculares (Touretzky *et al*., 2019), o ensino de Inteligência Artificial (IA) na educação básica também deve incluir o ML. Seguindo estas orientações, 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 (Marques *et al*., 2020), 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. Adotando metodologias ativas de aprendizado, focando no desenvolvimento centrado no ser humano de um modelo de ML, os estudantes devem aprender a preparar um conjunto de dados, treinar o modelo de ML e avaliar seu desempenho e predição de novas imagens (Lwakatare *et al*., 2019; Ramos *et al*., 2020). Tipicamente, ferramentas visuais, como o Google Teachable Machine (Google, 2020) são adotadas nesta fase educacional, não necessitando de qualquer programação. Isto permite que os estudantes executem um processo ML de forma interativa, utilizando um ciclo de treinamento-correção de alimentação, permitindo-lhes avaliar o estado atual do modelo e tomar as ações apropriadas (Gresse von Wangenheim *et al*., 2021).

Como parte do processo de aprendizado, é importante avaliar o aprendizado dos alunos fornecendo *feedback* tanto ao aluno quanto ao professor (Hattie e Timperley, 2007). Para um aprendizado efetivo, os estudantes precisam saber seu 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 (Sadler, 1989). Apesar dos muitos esforços para abordar a avaliação do ensino de computação na educação básica, focando no pensamento computacional, algoritmos e programação, e modelagem e simulação (Lye e Koh, 2014; Tang *et al*., 2019; Yasar *et al*., 2016), ainda faltam abordagens para a avaliação da aprendizagem de conceitos ML (Rauber e Gresse von Wangenheim, 2022). Os poucos existentes são relativamente simples baseados em quizzes ou autoavaliações. As exceções são Sakulkueakulsuk *et al*. (2018) e Gresse von Wangenheim *et al*. (2021) que propõem uma avaliação baseada no desempenho do modelo ML criado pelos estudantes voltado ao reconhecimento de imagens. E, embora Gresse von Wangenheim *et al*. (2021) apresentem uma avaliação inicial baseada em um painel de especialistas, nenhuma outra avaliação baseada em dados dos estudantes foi encontrada.

Com isso, este artigo apresenta os resultados de uma avaliação da confiabilidade e validade de uma avaliação baseada no desempenho com base em artefatos criados por estudantes. Para a avaliação é utilizada uma rubrica de pontuação para avaliar o aprendizado de aplicação de conceitos de ML com foco no reconhecimento de imagem com aprendizado supervisionado.

2. Metodologia de Pesquisa

Como resultado de pesquisas anteriores, foi desenvolvido sistematicamente um modelo de avaliação baseada no desempenho (Gresse von Wangenheim *et al*., 2021), incluindo uma rubrica, seguindo o método proposto por Moskal e Leyden (2000) e o projeto centrado em evidências (Mislevy *et al*., 2003). No contexto da pesquisa, essa rubrica foi revisada e foram ajustados os critérios ao curso "ML para todos!" (Gresse von Wangenheim *et al*., 2020) (Tabela 1).

Tabela 1 - Rubrica de pontuação

Critério		Níveis de Desempenho		
		Fraco - 0 pontos	**Aceitável - 1 ponto**	**Bom - 2 pontos**
	Gerenciamento de dados (LO5)			
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 (LO6)			
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 (LO7)			
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*., 2021)

O objetivo deste 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) (Basili *et al*., 1994), são derivadas as seguintes questões de análise:

QA1: Há evidência de consistência interna da rubrica?

QA2: Há evidência de validade convergente da rubrica?

Coleta dos dados. De acordo com o objetivo do estudo foram coletados artefatos em forma de resultados de aprendizagem criados pelos alunos ao longo do processo de desenvolvimento de um modelo de ML no contexto do curso "ML para Todos!". Utilizamos amostragem não-probabilidade em cada estudo de caso aplicando o método de amostragem de conveniência (Trochim e Donnelly, 2008), em que nossa amostra é composta por estudantes da educação básica matriculados no curso.

Análise dos dados. Reunimos os dados coletados em uma única amostra para análise de dados. Todos os artefatos coletados foram avaliados pelos autores adotando a rubrica (Tabela 1) alocando assim as pontuações referente ao nível de desempenho. Partes da avaliação foram automatizadas por meio de um script em Python, e, para determinar o critério de rotulação das imagens, foi inferido o rótulo a cada imagem utilizada pelo estudante por meio de um modelo de ML (Laydner, 2022). Como resultado foram identificadas as frequências de pontuações para cada critério da rubrica.

A confiabilidade refere-se à consistência ou estabilidade das pontuações dos critérios do instrumento de avaliação em um mesmo fator (Moskal e Leydens, 2000). 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 (Flora, 2020). 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, 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 (Lordelo et al., 2018).

Esta pesquisa foi aprovada pelo Comitê de Ética da Universidade Federal de Santa Catarina (No. 4.893.560).

3. Aplicação e coleta de dados

O curso "ML para Todos!" (Gresse von Wangenheim *et al*., 2020) foi projetado para ensinar conceitos básicos de ML com foco no reconhecimento de imagens para alunos de escolas de ensino fundamental e médio sem conhecimentos prévios de computação ou IA/ML. Os objetivos de aprendizado são definidos em alinhamento com as Diretrizes para ensino de IA (Touretzky *et al*., 2019) referentes à Grande Ideia 3 - Aprendizagem, alfabetização de IA (Long e Magerko, 2020) e um processo de ML centrado no ser humano (Amershi *et al*., 2019). O objetivo é levar os alunos ao nível de aplicação, desenvolvendo 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 preparação de dados, treinamento de modelo, avaliação de desempenho e previsão. A fim de permitir uma aplicação interdisciplinar, o modelo ML a ser desenvolvido enfoca a tarefa de reconhecimento de imagens de reciclagem, um tópico abordado na Educação Básica como parte das aulas de ciências (MINISTÉRIO DA EDUCAÇÃO, 2018), relacionado aos Objetivos de Desenvolvimento Sustentável das Nações Unidas (Pedro *et al.*, 2019). Os estudantes são orientados a desenvolver um modelo ML para reconhecimento de imagens usando o Google Teachable Machine (GTM) (Google, 2020). Usando um conjunto de imagens redimensionadas e não categorizadas, os alunos

são instruídos a preparam o conjunto de dados, organizando, limpando e rotulando as imagens em categorias de reciclagem. Em seguida, treinam o modelo, testam-no com novas imagens e analisam seu desempenho, considerando a precisão e a matriz de confusão apresentada pelo GTM. Os alunos também são orientados a ajustar o conjunto de dados e/ou parâmetros de treinamento a fim de melhorar o desempenho do modelo. Durante e após as atividades, os alunos são orientados a documentar os resultados do processo de ML, incluindo além do próprio modelo gerado na ferramenta GTM (arquivo .tm), relatórios online documentando a análise e interpretação do desempenho e resultados da predição (Figura 1). Esses artefatos criados ao longo do processo de ML são coletados como dados nesse estudo como base para a avaliação da aprendizagem do aluno com base no seu desempenho.

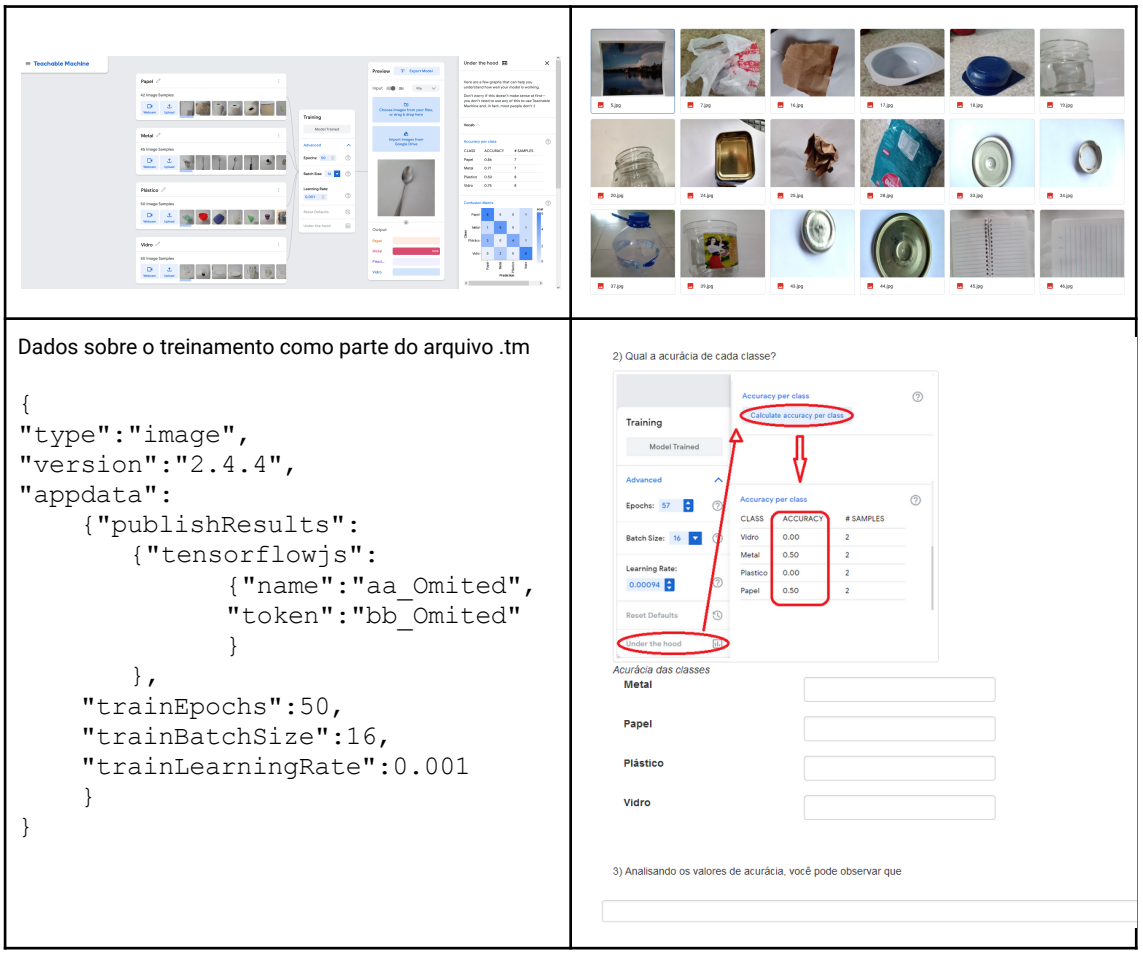

Figura 1. Exemplos de artefatos como parte da documentação dos resultados

O curso “ML para Todos!” foi aplicado em 4 casos de 2021 a 2022 com alunos dos anos finais do Ensino fundamental e Médio com idades entre 12 e 18 anos. Com exceção de uma aplicação (AP1) em uma escola na qual o curso foi aplicado como parte das aulas escolares, todas as aplicações foram conduzidas como atividades extracurriculares em forma de aulas remotas com instrutores. Como resultado, foram coletados artefatos criados de um total de 108 alunos.

4. Análise dos Dados

Com base na análise do desempenho dos artefatos criados pelos estudantes utilizando a rubrica (Tabela 1) foi levantada a distribuição de frequências de níveis de desempenho atingidos referente aos 11 critérios da rubrica, conforme apresentado na Tabela 2.

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
Total	**70**	**70**	**70**	**70**	**70**	**70**	**87**	**87**	**66**	**65**	**72**

Como nem todos os estudantes entregaram todos os artefatos, se obteve quantidades diferentes de artefatos em relação aos critérios. Podemos observar 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”.

Ao analisar as frequências de pontuações da Tabela 2, se observa a necessidade de agrupar alguns itens. A baixa frequência de respostas para cada nível de desempenho dos critérios gera imprecisão dos resultados em algumas análises estatísticas. Nos casos em que não houve pontuação em um nível de desempenho, foi eliminado esse nível de desempenho e os demais níveis de desempenho do critério foram recodificados para iniciarem sempre no primeiro nível (Fraco). Foram também eliminados para a análise artefatos incompletos criados pelo estudantes, pois o escore fica muito impreciso podendo comprometer a análise. Dessa forma, foram realizados os seguintes ajustes: a) foram eliminados as respostas (após inferência de acordo com a rúbrica) que atendam apenas 3 ou menos critérios; b) os critérios C6 e C7 foram recodificados, em que as pontuações inferidas com nível de desempenho de “Aceitável” foram transferidas para “Fraco” e de “Bom” para “Aceitável”; c) o critério C8 foi recodificado, manteve-se as pontuações inferidas no nível de desempenho “Fraco” e transferidas as inferências de “Bom” para “Aceitável”; d) o critério “C2-Relevância das imagens” foi eliminado por ter todos os escores no nível de desempenho “Bom”, com uma única exceção.

A Tabela 3 apresenta o conjunto de dados após a realização dos ajustes.

Tabela 3 - Distribuição de frequências de níveis de desempenho por critério da rubrica após ajuste dos dados.

	C1	C3	C4	C5	C6	C7	C8	C9	C10	C11
Fraco	22	3	3	17	60	3	31	19	12	9
Aceitável	16	28	44	6	10	66	38	10	8	32
Bom	32	39	23	47	-	-	-	37	45	19
Total	**70**	**70**	**70**	**70**	**70**	**69**	**69**	**66**	**65**	**60**

5. Resultados

5.1 Há evidência da consistência interna da rubrica?

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) (Brown, 2015). Como resultado foi obtido um valor Ômega Global de 0,646, pouco menor do que o valor mínimo de 0,7.

Ao analisar se a consistência interna aumenta eliminando um item (Tabela 4), se observa que o coeficiente aumenta eliminando vários itens. Isso inclui tanto os critérios "C5 - Limpeza dos dados", "C6 - Treinamento", "C8 - Interpretação dos testes", "C9 - Interpretação da acurácia", "C10 - Interpretação da matriz de confusão" e "C11 - Ajustes/Melhorias realizadas" o que pode ser um indício que o critério em questão não estaria associado ao mesmo traço latente. Destaca-se o critério "C6 - Treinamento", o qual a eliminação resultará no maior aumento indicando assim maior problema com esse item com relação a medir o mesmo traço latente.

Tabela 4 - Coeficiente Ômega - Análise com todos os critérios

Ômega	**Global**	**0,646**
Excluindo o critério	C1 - Quantidade de imagens	0,547
	C3 - Distribuição do conjunto de dados	0,488
	C4 - Rotulação das imagens	0,582
	C5 - Limpeza dos dados	0,694
	C6 - Treinamento	0,715
	C7 - Testes com novos objetos	0,500
	C8 - Interpretação dos testes	0,660
	C9 - Interpretação da acurácia	0,648
	C10 - Interpretação da matriz de confusão	0,670
	C11 - Ajustes / Melhorias realizadas	0,646

Observando esses critérios com consistência interna baixa e também motivado pelos resultados da análise da matriz de correlação (seção 5.2), não é possível concluir se a rubrica está medindo uma única ou mais dimensões.

Dessa forma, analisou-se também a hipótese de medição de duas dimensões pela rubrica, sendo uma dimensão formada pelos critérios de C1 a C5, os quais tratam do objetivo de aprendizagem "Gerenciamento de dados". E, uma segunda dimensão formada pelos critérios de C6 a C11, agrupando assim os critérios dos dois objetivos de aprendizagem: "Treinamento do modelo" e "Interpretação de desempenho".

Ao analisar somente a 1ª dimensão (Tabela 5) com os critérios C1, C3, C4, C5 resulta em um Ômega Global melhor com um valor de 0,721, acima do valor mínimo de consistência interna. Eliminando o critério "C5 - Limpeza dos dados" esse coeficiente aumentaria ainda para 0,853, indicando que o critério "C5 - Limpeza dos dados" pode não estar associado ao mesmo traço latente de "Gerenciamento de dados".

Da mesma maneira, ao analisar somente os critérios da 2ª dimensão (Tabela 6) compostos pelos critérios C6, C7, C8, C9, C10 e C11 resultou também em um valor o Ômega Global melhor de 0,732, também acima do valor mínimo de consistência

interna. Nesse caso, eliminando os critérios "C6 - Treinamento" e "C8 - Interpretação dos testes" o coeficiente Ômega aumentará para 0,726 e 0,728 respectivamente.

Tabela 5 - Coeficiente Ômega - Análise somente com os critérios de C1 a C5

Ômega	**Global (1ª Dimensão)**	**0,721**
Excluindo o critério	C1 - Quantidade de imagens	0,600
	C3 - Distribuição do conjunto de dados	0,449
	C4 - Rotulação das imagens	0,634
	C5 - Limpeza dos dados	0,853

Tabela 6 - Coeficiente Ômega - Análise somente com os critérios de C6 a C11

Ômega	**Global (2ª Dimensão)**	**0,732**
Excluindo o critério	C6 - Treinamento	0,779
	C7 - Testes com novos objetos	0,622
	C8 - Interpretação dos testes	0,766
	C9 - Interpretação da acurácia	0,570
	C10 - Interpretação da matriz de confusão	0,726
	C11 - Ajustes / Melhorias realizadas	0,666

5.2 Há evidência da validade convergente da rubrica?

Analisamos a validade convergente por meio do grau de correlação entre os critérios do instrumento. Para este propósito passamos a analisar a matriz de correlação policórica dos critérios da rubrica (Tabela 7). Nesta análise espera-se que os critérios que estejam medindo uma única dimensão apresentem correlações maiores ou iguais a 0,30 (DeVellis, 2017). Neste mesmo sentido, correlações (r) cujo valor em módulo não ultrapasse 0,5 ($0,30 \leq |r| < 0,50$) é considerada linear fraca, e até 0,7 ($0,50 \leq |r| < 0,70$) correlação moderada e acima ($0,70 \leq |r| < 0,90$) forte ou ($|r| \geq 0,90$) muito forte (Mukaka, 2012).

Tabela 7 - Matriz de correlação policórica

	C1	**C3**	**C4**	**C5**	**C6**	**C7**	**C8**	**C9**	**C10**	**C11**
C1 Quantidade de imagens	1									
C3 Distribuição do conjunto de dados	0,54	1								
C4 Rotulação das imagens	0,26	0,67	1							
C5 Limpeza dos dados	-0,19	-0,15	-0,13	1						
C6 Treinamento	-0,46	-0,22	0,07	0,34	1					
C7 Testes com novos objetos	0,35	0,58	0,26	0,03	-0,24	1				
C8 Interpretação dos testes	-0,19	0,18	-0,19	0,10	-0,27	0,51	1			
C9 Interpretação da acurácia	0,13	0,27	0,20	0,14	0,13	0,47	0,21	1		
C10 Interpretação da matriz de confusão	0,03	0,07	0,03	0,19	-0,46	0,01	0,27	0,51	1	
C11 Ajustes / Melhorias realizadas	0,08	0,12	0,17	0,37	0,43	0,17	-0,02	0,55	0,19	1

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 moderadas, em azul as correlações fracas em cinza as correlações próximas da condição de significância

estatística. O maior valor de 0,67 para a correlação foi alcançado para a associação entre a “C4 - Rotulação das imagens” e “C3 - Distribuição do conjunto de dados”, seguido de “C7 - Testes com novos objetos” e “C4 - Rotulação das imagens”. 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.

Analisando novamente a hipótese que a rubrica está medindo duas dimensões de C1 a C5 e outra de C6 a C11 (destacadas pelas bordas na Tabela 7), é esperado que contenham correlações mais altas dentro destas sub-matrizes. Referente a primeira dimensão podemos destacar a correlação moderada em dois pares de critérios “C3 - Distribuição do conjunto de dados” com “C1 - Quantidade de imagens”, e o par “C4 - Rotulação das imagens” com “C3 - Distribuição do conjunto de dados”. E na segunda dimensão há significância estatística com correlação moderada em 3 pares (C8xC7, C10xC9 e C11xC9) e fraca em outros 3 pares (C9xC7, C10xC6 e C11xC6).

6 Implicações e limitações

Os resultados 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. Há também um indicativo da rubrica medir duas dimensões distintas, uma voltada ao objetivo de aprendizagem “Gerenciamento de dados” com coeficiente Ômega de 0,721, e outra referente aos objetivos de aprendizagem “Treinamento do modelo” e "Interpretação de desempenho" com coeficiente Ômega de 0,732. Nestas análises os critérios “C5 - Limpeza dos dados” para a primeira dimensão e “C6 - Treinamento” e “C8 - Interpretação dos testes” merecem atenção, pois ao serem excluídos aumentam a consistência interna da rubrica de aprendizagem de ML.

As análises apontaram também a necessidade de revisão de alguns dos critérios. Entre eles o critério “C5 - Limpeza dos dados” que não apresenta relevância estatística nos indicadores de validade convergente por meio da análise da matriz de correlação policórica para a primeira dimensão. Um resultado análogo foi também observado referente ao critério “C8 - Interpretação dos testes” para a segunda dimensão. Já o critério “C6 - Treinamento” apresenta uma correlação fraca com os critérios “C10 - Interpretação da matriz de confusão” e “C11 - Ajustes/Melhorias realizadas”, e consequentemente a sua relacao aos demais critérios precisa ser analisada mais profundamente com base numa amostra maior.

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. Ao mesmo tempo 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 por Brennan e Resnick (2012), Avila *et al*. (2017) e Grover *et al*. (2015).

Ameaças à validade. A fim de minimizar impactos negativos a 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 rubrica, foi adotada uma metodologia sistemática seguindo a abordagem GQM (Basili *et al*., 1994). 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), inferindo a mesma rubrica. Somente os critérios C2, C5, C9 e C10 foram manualmente inferidos 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.

7. Conclusão

Em geral, os resultados desta avaliação mostram que a rubrica para 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. Tanto a análise da confiabilidade quanto a análise da validade convergente do construto por meio da matriz de correlação policórica indicam a possibilidade de que a rubrica está medindo duas dimensões, uma referente ao objetivo de aprendizagem de "Gerenciamento de dados" e outra referente as objetivos de aprendizagem de "Treinamento do modelo" e "Interpretação de desempenho".

Com base nesses resultados positivos está sendo implementado a integração da avaliação na ferramenta CodeMaster (Wangenheim *et al.*, 2018), de modo a fornecer ainda suporte automatizado que ajuda a garantir a consistência e a precisão dos resultados da avaliação, bem como a eliminar preconceitos. 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 outras para realizar 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

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.

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

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

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.

Brown T. A., (2015), *Confirmatory factor analysis for applied research*, Guilford publications.

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.

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.

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

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.

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

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.

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.

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.

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

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.

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

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.

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.

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

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.

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.

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.

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.

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.

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

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.

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

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

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

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.

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.

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

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

Sakulkueakulsuk B. *et al*., (2018), Kids making AI: Integrating machine learning, gamification, and social context in STEM education. Proc. of the Int. *Conference on Teaching, Assessment, and Learning for Engineering*, IEEE, 1005–1010.

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*.

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.

Trochim W. M. K. and Donnelly J. P., (2008), *The research methods knowledge base*, 3rd ed. Mason, Atomic Dog/Cengage Learning.

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.

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.