CBIE 2023
CONGRESSO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO
Uma escola para o futuro: Tecnologia e conectividade a serviço da educação

Uma proposta para avaliação do desempenho de aprendizagem de conceitos e práticas de *Machine Learning* em nível *Create* na Educação Básica

Marcelo Fernando Rauber[1, 2], Christiane Gresse von Wangenheim[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.

marcelo.rauber@ifc.edu.br, c.wangenheim@ufsc.br

Abstract. *There is a trend to include Machine Learning (ML) teaching already in K-12 by taking students to create their own intelligent solutions. In this context, we propose a model for assessing the students' learning based on a scoring rubric, which was evaluated by experts. The results provide a first indication of the model's adequacy regarding internal consistency and content validity in terms of correctness, relevance, completeness, and clarity. The specialists were also unanimous in pointing out the adequacy and applicability of the assessment model in the context of K-12 in order to support assessment in the context of teaching ML.*

Resumo. *Há uma tendência de incluir o ensino de Machine Learning (ML) já na Educação Básica, levando os alunos a criar suas próprias soluções inteligentes. Nesse contexto, propomos um modelo para avaliar o aprendizado dos alunos com base em uma rubrica de pontuação, que foi avaliada por especialistas. Os resultados fornecem uma primeira indicação da adequação do modelo em relação à consistência interna e à validade do conteúdo em termos de correção, relevância, integridade e clareza. Os especialistas também foram unânimes em apontar a adequação e a aplicabilidade do modelo de avaliação no contexto da Educação Básica, a fim de apoiar a avaliação no contexto do ensino de ML.*

1. Introdução

A Inteligência Artificial (IA) impacta nossas vidas cotidianas empregando diferentes tecnologias (Unesco, 2022). 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 de dados coletados e sua própria experiência, sem serem explicitamente programados, por meio da construção de modelos matemáticos/estatísticos (Mitchell, 1997). Recentemente, abordagens de aprendizagem profunda utilizando redes neurais, melhoraram drasticamente na visão computacional por meio de reconhecimento de imagens (LeCun *et al.*, 2015; Royal Society, 2017).

Porém, uma parcela significativa da população em geral ainda não compreende a tecnologia por trás do ML (Ho e Scadding, 2019), apontando a necessidade de esclarecer o que é ML, como funciona e quais seus impactos e limitações. Então torna-se importante introduzir conceitos e práticas básicas já na escola (Camada e Durães, 2020; Caruso e Cavalheiro, 2021), 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 (Royal Society, 2017; Touretzky *et al.*, 2019). Isto também é abordado indiretamente pela Base Nacional Comum Curricular (BNCC) que indica que os jovens devem ser preparados com novos conhecimentos, inerentes uma sociedade e profissões em constante mudança, incluindo a computação e tecnologias digitais, inclusive habilidades relacionadas a IA e ML para o Ensino Médio (MEC, 2022).

Para o ensino de computação, e também de ML, pode ser adotado o ciclo *Use-Modify-Create* (Lee *et al.*, 2011), que leva os alunos a uma coerente progressão nos níveis de dificuldade inerentes às atividades. Iniciando pelo nível *Use*, alunos sem experiência podem ser ensinados a inspecionar e manipular modelos de ML predefinidos, depois a *modificar* esses modelos até que, no nível *Create*, são incentivados a desenvolver seus próprios projetos de ML. Assim, desenvolver atividades em nível *Create* envolve atividades cognitivas complexas de nível superior da taxonomia de Bloom (Anderson e Krathwohl, 2001). Pois abordar uma temática em nível *Create* leva a soluções *open-ended*, em quais não há um produto final pré-definido, permitindo aos alunos explorar possibilidades infinitas e não se limitar a uma única resposta ou solução. Assim, os alunos podem criar novos projetos ou derivar novos projetos a partir de soluções existentes. Permite ainda, que desenvolvam novas funcionalidades, em um ciclo interativo de desenvolvimento envolvendo testes, análise e refinamento. Isso aumenta consideravelmente o desafio inerente à aplicação de seus conhecimentos e habilidades.

A avaliação de aprendizagem é uma etapa importante do processo de aprendizado. E mesmo já existindo esforços para definir avaliações para a aprendizagem de pensamento computacional, algoritmos e programação, na educação básica (Da Cruz Alves *et al*., 2019; Alves *et al*., 2020), se observa ainda uma carência de abordagens para a avaliação da aprendizagem de conceitos ML de forma confiável e válida (Rauber e Gresse von Wangenheim, 2022). As poucas propostas existentes para Educação Básica são relativamente simples, baseados em quizzes ou autoavaliações, muitas vezes visando avaliar uma unidade instrucional e não o aprendizado dos alunos. Poucos trabalhos analisam a confiabilidade ou validade dessas avaliações (Rauber e Gresse von Wangenheim, 2022). Uma exceção é a uma rubrica em nível *Use* inicialmente proposta por Gresse von Wangenheim *et al.* (2021). Esta foi avaliada por um painel de especialistas, e indicou uma concordância substancial de confiabilidade entre avaliadores, bem como validade de face em termos de correção, relevância, completude e clareza. Uma análise estatística da versão atualizada da rubrica, indica que os itens da rubrica têm um grande poder de discriminação e diferenciação (Rauber *et al.*, 2023). Em termos de confiabilidade, obteve-se boa consistência interna com um ômega global de 0,834 (Rauber *et al.*, 2023). Porém esta rubrica avalia somente a aprendizagem de ML no nível *Use*.

Assim, este artigo propõe um modelo de avaliação baseado em desempenho da aprendizagem de ML para classificação de imagens com aprendizado supervisionado no nível de *Create* para os anos finais do Ensino Fundamental e Médio.

2. Metodologia de Pesquisa

O modelo de avaliação foi projetado, desenvolvido e implementado seguindo a

metodologia de Design Centrado em Evidências (ECD) (Mislevy *et al.*, 2003; Seeratan e Mislevy, 2008) e Moskal e Leyden (2000) incluindo as seguintes fases:

Análise e Modelagem de domínio: Partindo-se de revisões do estado da arte (Marques *et al.*, 2020; Rauber and Gresse von Wangenheim, 2022), o domínio de ML foi analisado, incluindo suas características, propostas de diretrizes, o público envolvido e o curso "Apps inteligentes para Todos!" (Almeida, 2022), no qual o modelo de avaliação é adotado. As principais características da avaliação foram elencadas, seguindo o *Principled Assessment Designs for Inquiry* (Seeratan e Mislevy, 2008).

Desenvolvimento do framework conceitual: Esta etapa inclui a definição de competências do estudante, modelo de tarefa e modelo de evidência seguindo Mislevy *et al.* (2003) e Seeratan e Mislevy (2008). Como parte de uma proposta inicial do modelo, foi definida uma rubrica de pontuação para identificar os critérios com os quais o resultado de aprendizagem dos alunos é medido. Ela representa um esquema descritivo de pontuação (Brookhart, 1999) para avaliações baseadas no desempenho de artefatos de ML criados como resultados de aprendizagem em nível *Create*.

Avaliação da consistência interna e da validade do conteúdo. A proposta inicial do modelo de avaliação foi avaliado por um painel de especialistas. Os especialistas foram convidados a avaliar dois exemplos de produtos de trabalhos reais produzidos por alunos, sendo um bom e outro ruim, utilizando a rubrica de pontuação. Em seguida, os especialistas forneceram um *feedback* referente a rubrica, ao responder de um questionário com itens voltados a aplicabilidade, corretude, relevância, completude e clareza. Com base nos dados coletados foi avaliada a consistência interna analisando a confiabilidade entre avaliadores usando o coeficiente kappa de Fleiss (Fleiss *et al.*, 2003; Gamer *et al.*, 2019), que está relacionada à questão de que a pontuação de um aluno pode ser diferente entre avaliadores diferentes. A validade de conteúdo foi analisada com base nas respostas dos especialistas, avaliando até que ponto os critérios refletem as variáveis do construto e determinando se a medida é bem construída (Moskal e Leydens, 2000; Rubio *et al.*, 2003). A validade do conteúdo foi analisada por meio de estatísticas descritivas e do índice de validade do conteúdo proposto por Lawshe (1975). Os resultados foram interpretados e discutidos no respectivo contexto educacional.

3. Avaliação da aprendizagem de ML em nível *Create* na Educação Básica

O presente modelo de avaliação foi desenvolvido para o contexto do curso "Apps inteligentes para Todos!" (Almeida, 2022), que visa levar o aluno ao nível *Create,* criando o seu próprio modelo de ML para uma necessidade identificada e implantá-lo em um aplicativo móvel voltado ao ensino de ML nos anos finais do Ensino Fundamental e Médio.

3.1 Análise e Modelagem de domínio

Propostas de diretrizes curriculares Touretzky *et al.* (2019) e Long e Magerko (2020), sugerem que o ensino de ML no Ensino Fundamental e Médio, deve abranger uma compreensão dos conceitos básicos de ML, como algoritmos de aprendizagem e fundamentos de redes neurais, bem como limitações e considerações éticas. Espera-se que os alunos vão além da mera compreensão dos conceitos de ML, mas que se tornem

criadores de modelos de ML. Tipicamente são adotadas metodologias ativas de aprendizagem com foco no desenvolvimento centrado no ser humano de um modelo de ML (Amershi *et al.*, 2019), a fim de ensinar aos alunos como preparar um conjunto de dados, treinar um modelo de ML, avaliar seu desempenho e usá-lo para a classificação de novas imagens (Ramos *et al.*, 2020). Portanto, normalmente são usadas ferramentas visuais sem codificação, como o Google Teachable Machine (GTM; Google, 2023). Isso permite que os alunos executem um processo de ML de forma interativa, por meio de um ciclo de treinamento, *feedback* e correção, permitindo que eles avaliem o desempenho do modelo de ML e, assim, façam alterações no modelo com o objetivo de aprimorá-lo (Gresse von Wangenheim *et al.*, 2021).

Na faixa etária alvo do curso "Apps inteligentes para Todos!", de alunos entre 12 e 18 anos de idade, é esperado que tenham proficiência na língua portuguesa, com raciocínio lógico e matemático desenvolvidos e capazes de utilizar computadores para tarefas corriqueiras, como navegar na Internet (MEC, 2018). Porém, mesmo com a recente inclusão de normas complementares à BNCC sobre Computação na Educação (MEC, 2022) para a maioria dos estudantes, o ensino de computação apenas está disponível por meio de cursos extracurriculares (Santos *et al.*, 2018). Ainda assim, 42% dos alunos de escolas urbanas utilizam recursos tecnológicos por mais de 3 horas diárias, e 98% têm acesso a um a um *smartphone* com acesso a Internet. É notável a carência de formação de professores de informática (MEC, 2020), o que leva o ensino de computação a ser introduzido de forma interdisciplinar, sendo lecionado por educadores de outras áreas de formação, com turmas comumente com 30 alunos ou mais, o que torna a avaliação manual dos projetos trabalhosa e morosa.

A partir dessa análise foram estabelecidas as principais características e fundamentos para o desenvolvimento do modelo de avaliação que leva em conta os artefatos computacionais no contexto da educação computacional de ML em nível *Create*, o que inclui os conhecimentos e competências esperados dos alunos (Tabela 1).

Tabela 1: Modelagem da avaliação das competências de ML no nível *Create*.

Elemento	**Descrição**
Conhecimentos, competências e outros atributos essenciais	Compreensão dos conceitos básicos sobre redes neurais. Habilidade para identificar e descrever uma nova aplicação de um modelo de ML, usando classificação de imagens. Habilidade para especificar e analisar riscos e requisitos de desempenho para um novo modelo de ML. Habilidade de recolher, limpar e rotular dados para a formação de um modelo de ML. Habilidade de treinar um modelo ML para classificação de imagens utilizando uma ferramenta visual. Habilidade para analisar, interpretar o desempenho, e, melhorar o modelo de ML treinado. Habilidade de testar o modelo de ML com novas imagens para previsão.
Conhecimentos, competências e atributos adicionais	Habilidade e maturidade para compreender instruções em português do Brasil. Habilidade de utilizar um computador (operações básicas) e de acessar à Internet através de um navegador (browser). Habilidade de realizar login em páginas da Internet com uso de dados pessoais.

3.2 Framework de Avaliação

3.2.1 Competências do estudante

As competências esperadas dos alunos foram definidas com base nas Diretrizes para Ensino de IA - Grande Ideia 3: Aprendizagem (Touretzky *et al.*, 2019), diretrizes curriculares para alfabetização em IA (Long e Magerko, 2020) e um processo de ML centrado no ser humano (Amershi *et al.*, 2019). No curso "Apps inteligentes para Todos!" (Almeida, 2022), com relação ao modelo de ML, o objetivo geral de

aprendizagem é permitir aos alunos criar o seu próprio modelo de ML para classificação de imagens, desenvolvendo assim conhecimentos e competências de ML (Tabela 1).

3.2.2 Modelo de Tarefa

Para dar continuidade ao curso "ML para Todos!" (Gresse von Wangenheim *et al.*, 2020), o curso "Apps inteligentes para Todos!" (Almeida, 2022) é voltado ao nível *Create*, onde o aluno desenvolve seu próprio modelo de ML de classificação de imagens de temática livre, e sua implementação com App Inventor. A duração total é estimada em 25 horas, das quais 11 horas são destinadas ao desenvolvimento do modelo de ML em nível *Create*. O curso adota o processo de *Design Thinking*, que leva o aluno a identificar um problema em relação a sua vida cotidiana ou comunidade, ajudando-o a desenvolver uma solução de ML útil e usável, prototipando-o e implementando-o em um aplicativo móvel. Os alunos são instruídos sobre o processo de ML, e preparam os seus próprios conjuntos de imagens. O Google Teachable Machine (Google, 2023) é usado para treinar e avaliar o desempenho do modelo de ML, que ao final é exportado na Google *Cloud*. Em seguida, os alunos começam a criação de aplicativos esboçando a interface do aplicativo. Eles programam e testam o aplicativo no nível do *wireframe* e projetando o design visual usando o App Inventor com a extensão TMIC para implantar o modelo de ML desenvolvido (Oliveira, 2022).

3.2.3 Modelo de evidência

O modelo de evidência detalha como as variáveis do modelo de aluno devem ser atualizadas, tendo como base o desempenho apresentado pelos alunos em seus produtos de trabalho, incluindo o modelo de avaliação e um modelo de medição (Mislevy *et al.*, 2003). O modelo de avaliação explica como atualizar as variáveis observáveis relativas ao desempenho do aluno a partir dos produtos de trabalho criados ao longo das tarefas específicas do curso, gerando assim evidências que refletem seu nível de competência. Neste artigo, os termos variáveis observáveis, itens e critérios de avaliação serão usados de forma intercambiável. O modelo de avaliação é apresentado como uma rubrica de pontuação (Tabela 2), adaptada a partir da proposta de rubrica desenvolvida para o nível *Use* por Gresse von Wangenheim *et al.* (2021) e Rauber *et al.* (2023). Foram definidas variáveis observáveis a serem medidas para avaliar a capacidade dos alunos de criar um modelo de ML, inferindo indiretamente as competências de ML em nível *Create*. Os níveis de desempenho foram definidos de acordo com resultados de aprendizagem. Em termos do conceito medido em cada variável observável, níveis de desempenho mais altos indicam uma compreensão maior. Os níveis de desempenho foram especificados em uma escala ordinal de zero a três pontos, variando entre "Não entregue", "Fraco", "Aceitável" e "Bom", em conformidade com o desempenho esperado para atingir o objetivo de aprendizado específico.

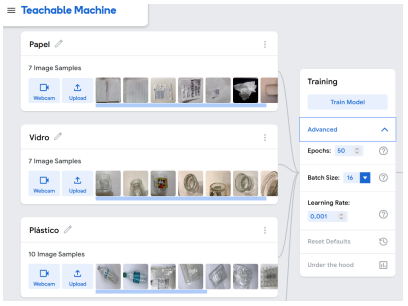
O modelo de medição foi definido em forma de uma pontuação geral que varia de zero a dez pontos, de acordo com o sistema de avaliação brasileiro, que é calculado a partir da soma média de pontos das variáveis observáveis multiplicada por dez.

Tabela 2: Rubrica de avaliação de desempenho de aprendizagem de ML – nível *Create*.

Critério / Variáveis observáveis	**Níveis de Desempenho**			
	Não entregue - 0 pontos	**Fraco - 1 ponto**	**Aceitável - 2 pontos**	**Bom - 3 pontos**
Análise de requisitos de ML				
Objetivo do modelo de ML especificado	Não descreveu	Descrição incompleta ou incorreta	-	Descrição completa e correta
Especificação de riscos e requisitos de desempenho	Não descreveu	Descrição incompleta ou incorreta	-	Descrição completa e correta
Análise dos risco de erro em relação ao modelo proposto	Não descreveu	Incorreta identificação do nível de risco	-	Correta identificação do nível de risco
Acurácia correta em relação ao risco de erro no modelo proposto	Não descreveu	Descrição incompleta ou incorreta	-	Descrição completa e correta (Mínimo de 75% para risco médio e baixo, mínimo de 95% para alto)
Gerenciamento de dados				
Quantidade de imagens	Não enviou informações (do arquivo TM)	Menos de 20 imagens por categoria	21 - 35 imagens por categoria	Mais de 36 imagens por categoria
Distribuição do conjunto de dados	Não enviou informações (do arquivo TM)	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)
Imagens com conteúdo ético	Não enviou informações (do arquivo TM)	Ao menos 3 imagens contém conteúdo não ético (violência, nudez, armas)	Ao menos uma imagem não contém imagens não éticas.	Todas as imagens são éticas.
Treinamento do modelo				
Treinamento	Não enviou informações (do arquivo TM)	O modelo não foi treinado	O modelo foi treinado usando os parâmetros padrões (Epochs: 50, batch size: 16, Learning rate: 0,001)	O modelo foi treinado com parâmetros ajustados
Interpretação de desempenho				
Análise de acurácia por categoria	Não enviou as informações para viabilizar a análise	Categorias com baixa acurácia não identificadas	-	Todas as categorias com baixa acurácia identificadas corretamente
Interpretação da acurácia	Não enviou as informações para viabilizar a análise	Interpretação incorreta em relação ao modelo	-	Correta interpretação a respeito do modelo
Análise da matriz de confusão	Não enviou as informações para viabilizar a análise	Mais de dois erros na identificação de classificações errôneas	Até dois erros na identificação de classificações errôneas	Identificação correta de erros de classificação
Interpretação da matriz de confusão	Não enviou as informações para viabilizar a análise	Interpretação a respeito do modelo é incorreta	-	Correta interpretação em respeito ao modelo
Ajustes / Melhorias realizadas	Não enviou as informações para viabilizar a análise	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
Testes com novos objetos	Não enviou as informações para viabilizar a análise	Nenhum objeto testado	1-5 objetos testados	Mais de 5 objetos testados
Análise dos resultados de testes	Não enviou as informações para viabilizar a análise	Indicação errada da quantidade de erros nos testes	-	Indicação correta da quantidade de erros nos testes
Interpretação dos testes	Não enviou as informações para viabilizar a análise	Interpretação errada	-	Correta interpretação
O modelo alcançou a mínima acurácia planejada?	Não enviou as informações para viabilizar a análise	Não	Ficou próximo	Sim

3.3 Implementação e aplicação da Avaliação

Os produtos de trabalho elaborados pelos alunos durante o curso “Apps inteligentes para Todos!” (Almeida, 2022) são coletados como resultados de aprendizagem e avaliados usando o modelo de evidência. Isso inclui o modelo de ML desenvolvido contido no arquivo GTM (.tm), bem como os relatórios on-line preenchidos pelos alunos que documentam a análise e a interpretação do desempenho e predição dos resultados como parte da execução do processo de ML (Figura 1).



ANÁLISE DO MODELO DE ML

Objetivo do modelo de ML		
Tarefa	Classificar a categoria de resíduos orgânicos por meio de imagens capturadas de um aplicativo Android.	
Tipo da tarefa	*Single-label* classificação de imagens	
Categorias	2 categorias de resíduos orgânicos	"Pode ir na composteira" e "não pode ir na composteira"
Contexto de uso	O modelo é utilizado como exemplo no contexto de ensino de na Educação Básica.	
Conjunto de dados		
Fonte de dados	Coleta própria. Captura com a câmera de celular por voluntários do grupo Computação na Escola (CnE).	

Figura 1: Exemplos de resultados de aprendizagem.

4. Avaliação da rubrica

Com o objetivo de avaliar a rubrica em termos de consistência interna e validade de conteúdo foi realizado um estudo de caso por meio de um painel de especialistas, no mês de maio de 2023. Foram selecionados especialistas de forma sistemática que participam da iniciativa “Computação na Escola” da Universidade Federal de Santa Catarina (UFSC), com experiência na área acadêmica e/ou que tenham experiência no ensino de ML. Todas as participações foram voluntárias. Foram convidados 6 especialistas, dos quais todos responderam. A maioria dos participantes possui experiência e conhecimentos em computação (Figura 2). Todos os especialistas já trabalharam com a criação de modelos de ML. Apenas um especialista tem menos de dois anos de experiência prática em ensino de computação e não publicou algum artigo voltado ao ensino de computação na Educação Básica.

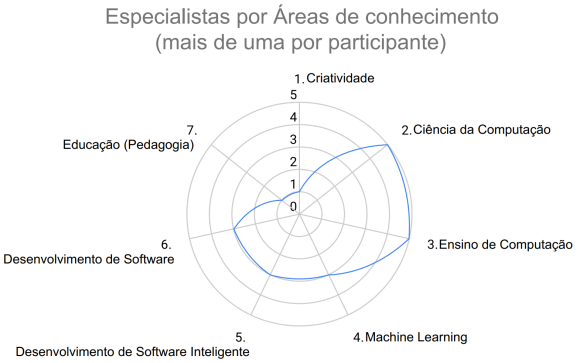


Figura 2: Áreas de conhecimento do painel de especialistas.

4.1 Quais são as evidências de consistência interna da rubrica de avaliação baseada em desempenho em nível *Create*?

Com o objetivo de avaliar se a rubrica criada permite uma avaliação confiável (Moskal e Leydens, 2000), foi analisada a confiabilidade entre as respostas das avaliações dos

especialistas (*inter-rater*) em relação aos dois exemplos de resultados de aprendizagem de ML em nível *Create*. Foi utilizado Fleiss Kappa (Fleiss *et al.*, 2003), uma medida que estende Cohen's Kappa para avaliação do nível de concordância entre dois ou mais avaliadores. O valor resultante do Fleiss Kappa varia entre -1 e 1, onde valores maiores indicam maior concordância entre os avaliadores. Tipicamente, valores negativos indicam concordância pobre, valores entre 0,01-0,20 indicam concordância leve, 0,21-0,40 concordância razoável, 0,41-0,60 concordância moderada, 0,61-0,80 concordância substancial e 0,81-0,99 concordância quase perfeita (Landis e Koch, 1977).

Analisando os dezessete itens da rubrica utilizando a avaliação dos seis especialistas, foi obtido um valor de Fleiss Kappa = 0,534, indicando concordância moderada, confirmada pelo p-value (p=0) o que sugere uma evidência estatisticamente significante de que existe uma concordância entre os avaliadores além do que seria esperado pelo acaso. Também foram calculados os valores individuais de Fleiss Kappa para cada nível de desempenho separadamente e comparados a todas as categorias juntas (Tabela 3).

Tabela 3: Fleiss Kappa por nível de desempenho.

Nível de Desempenho	Não entregue - 0 pontos	Fraco - 1 ponto	Aceitável - 2 pontos	Bom - 3 pontos
Fleiss Kappa	0,441	0,455	0,593	0,674

Um nível de concordância substancial pode ser observado no nível de "Bom", ao passo que os níveis "Aceitável", "Fraco" e "Não Entregue" apresentam apenas uma concordância moderada. Isso implica que parece ser mais fácil reconhecer um alto nível de desempenho, ante distinguir entre os níveis de baixa a intermediário desempenho.

4.2 Quais são as evidências de validade de conteúdo da rubrica de avaliação baseada em desempenho em nível *Create*?

A maioria dos participantes considerou os critérios e níveis de desempenho da rubrica como corretos (100%), completos (66,7%) e claros (66,7%).

Corretude: Com relação a corretude, todos os especialistas consideraram que os níveis de desempenho critérios apresentados e descritos na rubrica estão corretos.

Relevância: Para avaliar a relevância, os especialistas foram indagados a classificar cada um dos critérios da rubrica em uma escala ordinal de 3 pontos (Likert), variando de irrelevante a essencial. A grande maioria dos critérios foi considerada essencial, com alguns especialistas considerando alguns critérios como desejáveis (Figura 3). Nenhum critério foi considerado irrelevante.

Visando analisar a validade de conteúdo, foi calculado o índice de validade de conteúdo (*content validity ratio*), que é dado pelo cálculo CVR=(Ne-N/2)/(N/2), em que Ne é o número de especialistas que marcam essencial e N é o número total de especialistas (Lawshe, 1975). Tipicamente, é esperado que o índice de validade de conteúdo fique acima de 0,49 (Lawshe, 1975). Se observa que a maioria dos critérios está adequada com base nos índices de validade de conteúdo à direita da Figura 3. Porém, alguns critérios apresentaram um índice de validade de conteúdo abaixo do esperado, que foram considerados desejáveis mas não essenciais. Consequentemente, há o indicativo que esses critérios podem ser reconsiderados, sendo importante avaliar a validade de critério e de construto da rubrica.

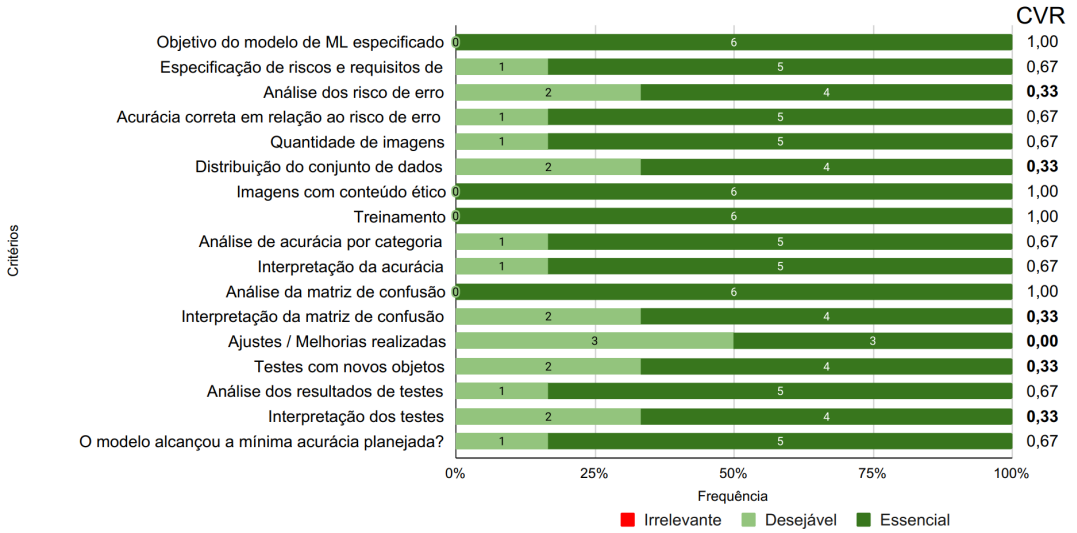


Figura 3: Relevância de cada critério.

Completude: A grande maioria dos especialistas concorda que o modelo de avaliação abrange todos os aspectos necessários. Um especialista sugeriu a inclusão de um critério para tratar da variância do conjunto das imagens, abordando aspectos como distorção de perspectiva, luminosidade e distância dos objetos de interesse, já que estes têm influência nos resultados. Outro especialista sugeriu a inclusão de um critério para avaliação da descrição do conjunto de dados, avaliando a fonte dos dados utilizados para treinar o modelo de ML de classificação de imagens e como se deu sua rotulação.

Clareza: Se observou que os níveis de desempenho "Não entregue" e "Fraco" foram confundidos algumas vezes, levando a avaliações subjetivas, quando a entrega do aluno não atingiu níveis mínimos dos esperados. No caso, o aluno ao invés de descrever ou apresentar os artefatos gerados colocou descrições como "Não sei" ou "Não me lembro". Esse também pode ser o motivo da falta de uma maior concordância entre os avaliadores. Um dos especialistas aponta que o critério "Análise dos risco de erro em relação ao modelo proposto" não apresenta quais são os possíveis níveis de risco para identificar se o nível alocado é o correto, sugerindo que estes sejam apresentados. Dois especialistas relatam que em relação aos critérios de "Análise de requisitos de ML" os termos "completa" e "correta" lhe pareceram ambíguos, sugerindo que haja alguma indicação do que representam. Dois especialistas ainda sugerem melhorar a descrição na rubrica, visando esclarecer qual a diferença entre critérios que envolvem análise e interpretação.

De forma unânime, todos os especialistas concordam na adequação e aplicabilidade da rubrica no contexto da Educação Básica, junto com outras formas de avaliação, como entrevistas ou observações, visando um entendimento compreensivo do desempenho dos estudantes.

4.3 Ameaças à validade

A fim de minimizar os impactos negativos à validade neste estudo, desenvolvemos sistematicamente o modelo de avaliação adotando a metodologia de Design Centrado em Evidências (Mislevy *et al.*, 2003; Seeratan e Mislevy, 2008) com base em uma

análise e modelagem de contexto e realizamos uma avaliação inicial por meio de um painel de especialistas. A fim de mitigar a ameaça relacionada à diversidade e tamanho da amostra, o painel abrange profissionais experientes nas áreas de interesse, bem como do público alvo, que inclui professores da educação básica. Em termos de tamanho, seis especialistas são considerados suficientes para obter resultados iniciais (Lawshe, 1975). Para reduzir as ameaças relacionadas à análise de dados, a avaliação estatística segue Lawshe (1975), Moskal e Leydens (2000) e Rubio *et al.* (2003). Para minimizar o impacto do viés devido a subjetividade do *feedback* dos especialistas, seguimos a metodologia proposta por (Lawshe, 1975). Dada a natureza inicial do modelo de avaliação desenvolvido, ressalta-se a importância de estudos em maior escala para confirmar os resultados e analisar questões em aberto.

5 Discussão e Conclusão

Neste artigo apresentamos uma proposta de um modelo para a avaliação baseada em desempenho do aprendizado de ML para classificação de imagens no nível *Create* no contexto dos anos finais do Ensino Fundamental e Médio.

Os resultados da avaliação de consistência interna sugerem que a rubrica está adequada, com um valor de Fleiss Kappa de 0,534, o que indica uma concordância moderada e estatisticamente significante entre as respostas dos especialistas. Complementarmente, a análise do Fleiss Kappa de cada nível de desempenho indica uma maior concordância entre os especialistas ao avaliar produtos de trabalho que indicam alto desempenho dos estudantes.

A avaliação da validade de conteúdo também corrobora com a adequação da rubrica. Os especialistas foram unânimes em apontar a corretude dos critérios e níveis de desempenho. Os especialistas apontam a maior parte dos critérios utilizados na rubrica como essenciais, ao mesmo tempo que nenhum critério chegou a ser considerado irrelevante. É interessante observar que os critérios de "gerenciamento de dados", "treinamento do modelo" e "interpretação do desempenho", já tiveram sua validade de critério e de construto avaliadas anteriormente em nível *Use*, obtendo valores adequados (Gresse von Wangenheim *et al.*, 2021; Rauber *et al.*, 2023). Assim, corroborando a indicação de que todos os critérios devem ser mantidos. A grande maioria dos especialistas confirma que a rubrica é completa e clara, sendo apresentadas poucas sugestões de melhorias.

Com base nas respostas dos especialistas obteve-se uma primeira indicação da confiabilidade e validade de conteúdo do modelo de avaliação em termos de corretude, relevância, completude, clareza e aplicabilidade. Com base nesses primeiros *feedbacks* positivos, estamos automatizando o modelo de avaliação como parte da ferramenta CodeMaster (Gresse von Wangenheim *et al.*, 2018), para minimizar o esforço e assegurar a consistência e a precisão dos resultados da avaliação, bem como a eliminar preconceitos. Obviamente, na prática educacional, esse tipo de avaliação pode ser complementado por outros tipos de avaliação, como observações ou entrevistas.

Agradecimentos

Gostaríamos de agradecer a todos os especialistas que participaram da avaliação.

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

Referências

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

Almeida, B. C. S. (2022). *Desenvolvimento de um Curso Ensinando a Criação de Apps Inteligentes para a Classificação de Imagens com Machine Learning e Design Thinking*. TCC. (Graduação em Sistemas de Informação) – UFSC.

Amershi, S. *et al.*(2019). Software Engineering for Machine Learning: A Case Study. *Proc. of 41st Int.Conf. on Software Engineering: Software Engineering in Practice*, Montreal, Canada.

Alves, N. da C., Gresse von Wangenheim, C., Alberto, M., and Martins-Pacheco, L. H. (2020), Uma Proposta de Avaliação da Originalidade do Produto no Ensino de Algoritmos e Programação na Educação Básica. *Proc. of XXXI Simpósio Brasileiro de Informática na Educação*, SBC.

Brookhart, S. M. (1999). *The Art and Science of Classroom Assessment.* ASHE-ERIC Higher Education Report, 27(1).

Camada M. Y. and Durães G. M., (2020), Ensino da Inteligência Artificial na Educação Básica: um novo horizonte para as pesquisas brasileiras. *Proc. of XXXI Simpósio Brasileiro de Informática na Educação, SBC*.

Caruso A. L. M. and Cavalheiro S. A. da C., (2021), Integração entre Pensamento Computacional e Inteligência Artificial: uma Revisão Sistemática de Literatura. *Proc. of XXXII Simpósio Brasileiro de Informática na Educação*, SBC.

CGI (2019). *TIC Educação 2019*. São Paulo, SP, Brasil: Cetic.

Da Cruz Alves, N., Gresse Von Wangenheim, C., and Hauck, J. C. R. (2019). Approaches to Assess Computational Thinking Competences Based on Code Analysis in K-12 Education: A Systematic Mapping Study. *Informatics in Education*, 18(1).

Fleiss, J. L., Levin, B., and Paik, M. C. (2003). *Statistical methods for rates and proportions (3rd ed)*. John Wiley & Sons, Inc.

Google (2023). Google Teachable Machine. Retrieved 01/06/2023 from https://teachablemachine.withgoogle.com/.

Gamer, M., Lemon, J., and Singh, I. F. P. (2019). irr: Various Coefficients of Interrater Reliability and Agreement (0.84.1). https://cran.r-project.org/web/packages/irr/index.html

Gresse von Wangenheim, C. G. von, Hauck, J. C. R., Demetrio, M. F., Pelle, R., Cruz Alves, N. da, Barbosa, H. and Azevedo, L. F. (2018). CodeMaster—Automatic Assessment and Grading of App Inventor and Snap! Programs. *Informatics in Education*, 17(1).

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,* 21(3).

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.

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

Landis, J. R., and Koch, G. G. (1977). The measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1).

Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel Psychology*, 28(4).

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

Lee, I., Martin, F., Denner, J., Coulter, B., Allan, W., Erickson, J., Malyn-Smith, J., and Werner, L. (2011). Computational thinking for youth in practice. *ACM Inroads*, 2(1), 32–37.

Long, D., Magerko, B. (2020). What is AI literacy? Competencies and design considerations. *Proc. of the Conf. on Human Factors in Computing Systems*, Honolulu, HI, USA, 1–16.

Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel psychology*, 28(4), 563–575.

Marques, L. S., Gresse von Wangenheim, C., and Hauck, J. C. (2020). Teaching machine learning in school: A systematic mapping of the state of the art. *Informatics in Education*, 19(2), 283-321.

MEC (2018), *Base Nacional Comum Curricular. Ministry of Education.* Brazil.

MEC (2020), *Census of Basic Education 2020. Ministry of Education.* Brazil.

MEC (2022), *Normas sobre Computação na Educação Básica – Complemento à Base Nacional Comum Curricular (BNCC).* Parecer 02/2022 CNE/CEB/MEC.

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

Mitchell, T. M. (1997), *Machine Learning.* New York, NY, USA: McGraw-Hill.

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

Oliveira, F. P. (2022), *TMIC - Uma extensão do App Inventor para a implantação de modelos de ML voltados a classificação de imagens treinados no Teachable Machine*. TCC. (Graduação em Sistemas de Informação) – UFSC.

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).

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*, 22(2), 295-328.

Rauber, M. F., Gresse von Wangenheim, C., Barbetta, P. A., Borgatto, A. F., Martins, R. M. and Hauck, J. R. (2023). Reliability and Validity of an Automated Model for Assessing the Learning of Machine Learning in Middle and High School: Experiences from the "ML for All!" course. *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.

Rubio, D. M., Berg-Weger, M., Tebb, S. S., Lee, E. S., and Rauch, S. (2003). *Objectifying content validity: Conducting a content validity study in social work research*. Social Work Research, 27(2).

Santos, P. S., Araujo, L. G. J., and Bittencourt, R. A. (2018). A mapping study of computational thinking and programming in brazilian k-12 education. *Proc. of Frontiers in Education Conference*, San Jose, CA, USA.

Seeratan, K. L., and Mislevy, R. J. (2008). *Design patterns for assessing internal knowledge representations (PADI Technical Report 22)*. Menlo Park, USA: SRI International.

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*, Honolulu, HI, USA, 33(01).

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