Revista Brasileira de Informática na Educação – RBIE
Brazilian Journal of Computers in Education
(ISSN online: 2317-6121; print: 1414-5685)
http://br-ie.org/pub/index.php/rbie

Submission: 02/Jun/2023; 1st round notif.: 21/Sep/2023; New version: 11/Oct/2023; 2nd round notif.: 20/Oct/2023
Camera ready: 01/Nov/2023; Edition review: 18/Dec/2023; Available online: 18/Dec/2023; Published: 18/Dec/2023

Uma Análise Detalhada do Desempenho de Aprendizagem ensinando Machine Learning na Educação Básica aplicando a Teoria de Resposta ao Item

Title: A detailed analysis of the learning performance teaching Machine Learning in K-12 Education applying Item Response Theory

Título: Un análisis detallado del rendimiento en el aprendizaje de la enseñanza del Machine Learning en Educación Básica aplicando la Teoría de Respuesta al Ítem

Marcelo Fernando Rauber
Programa de Pós-Graduação em Ciência da Computação – Universidade Federal de Santa Catarina (UFSC) - Florianópolis - SC - Brasil, e, Instituto Federal Catarinense (IFC) - Camboriú - SC - Brasil
ORCID: 0000-0001-5653-7155
marcelo.rauber@ifc.edu.br

Christiane Gresse von Wangenheim
Programa de Pós-Graduação em Ciência da Computação – Universidade Federal de Santa Catarina (UFSC) - Florianópolis - SC - Brasil
ORCID: 0000-0002-6566-1606
c.wangenheim@ufsc.br

Adriano Ferreti Borgatto
Programa de Pós-Graduação em Métodos e Gestão em Avaliação - Universidade Federal de Santa Catarina (UFSC) - Florianópolis - SC - Brasil
ORCID: 0000-0001-6280-2525
adriano.borgatto@ufsc.br

Ramon Mayor Martins
Programa de Pós-Graduação em Ciência da Computação – Universidade Federal de Santa Catarina (UFSC) - Florianópolis - SC - Brasil
ORCID: 0000-0002-1952-0909
ramon.mayor@posgrad.ufsc.br

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 o projeto, desenvolvimento e implementação de um modelo de avaliação da aprendizagem de ML, com destaque para avaliação da validade e da confiabilidade da rubrica resultante. 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 da análise detalhada 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.*

Cite as: Rauber, M.F., Gresse von Wangenheim, C., Borgatto, A.F. & Martins, R.M. (2023). Uma análise detalhada do desempenho de aprendizagem de Machine Learning na Educação Básica aplicando a Teoria de Resposta ao Item. Revista Brasileira de Informática na Educação, 31, 1031-1056. https://doi.org/10.5753/rbie.2023.3442

Abstract

The current insertion of Machine Learning (ML) in everyday life demonstrates the importance of introducing the teaching of ML concepts already in middle and high school. Accompanying this trend arises the need to assess this learning. In this paper we present the design, development and implementation of an ML learning assessment model, with emphasis on the evaluation of the validity and reliability of a rubric for the performance-based assessment of learning outcomes of the application of ML concepts by middle and high school students. Adopting Item Response Theory we present a preliminary proposal of the construction of a scale for the level of student learning. The results of the detailed analysis show that it is possible to calibrate the parameters of the Item Response Theory with satisfactory indices of reliability and validity, which demonstrates the potential of using the rubric in order to help both students and teachers to promote the teaching of ML at this educational stage.
Keywords: *Learning Assessment, Machine Learning, Item Response Theory, IRT, Middle and High School*.

Resumen

La actual inserción del Machine Learning (ML) en la vida cotidiana demuestra la importancia de introducir la enseñanza de conceptos de ML desde la Educación Básica. Acompañando esta tendencia surge la necesidad de evaluar este aprendizaje. En este artículo se presenta el diseño, desarrollo e implementación de un modelo de evaluación del aprendizaje del ML, con énfasis en la evaluación de la validez y fiabilidad de la rúbrica resultante. Esta rúbrica tiene como objetivo evaluar el aprendizaje mediante el desempeño de los estudiantes con base en los resultados de aprendizaje de la aplicación de conceptos de ML por parte de estudiantes de los últimos años de Educación Básica y Educación Media. Adoptando la Teoría de Respuesta al Ítem presentamos una propuesta preliminar de construcción de una escala para el nivel de aprendizaje de los estudiantes. Los resultados del análisis detallado muestran que fue posible calibrar los parámetros de la Teoría de Respuesta al Ítem con índices satisfactorios de confiabilidad y validez, lo que demuestra el potencial del uso de la rúbrica para ayudar tanto a estudiantes como a investigadores y profesores a promover el desarrollo de la enseñanza del ML en la Educación Básica.
Palabras clave: *Evaluación del aprendizaje, Aprendizaje automático, Teoría de la Respuesta al Ítem, TRI, Educación 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 (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 da sua própria experiência sem ter que ser explicitamente programados, por meio da construção de um modelo matemático/estatístico baseado nos dados coletados (Mitchell, 1997). 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 (LeCun et al., 2015). 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* (Royal Society, 2017; UNESCO, 2022).

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 (Ho & 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). Então torna-se importante introduzir conceitos e práticas básicas já na escola (Camada & Durães, 2020; Caruso & 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 (Kandlhofer et al., 2016; Royal Society, 2017; UNESCO, 2022).

Com essa motivação já estão surgindo várias propostas propondo o ensino de IA/ML na Educação Básica, mas estão em estágios iniciais (Long & Magerko, 2020; Marques et al., 2020). 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 parágrafo II do artigo 32 da LDB (BRASIL, 1996) 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."* (Ministério da Educação, 2018)

Recentemente foram incluídas normas complementares à BNCC sobre Computação na Educação (Ministério da Educação, 2022) 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."* (Ministério da Educação, 2022)

O sequenciamento de atividades através do ciclo *Use-Modify-Create* proposto por Lee et al.

(2011), além de aumentar o engajamento e sentimento de criador, facilita o aprendizado dos processos educacionais inerentes ao pensamento computacional e outras atividades percebidas como complexas pelos estudantes (Lytle et al. 2019), comumente usado para a progressão do aprendizado de conceitos e práticas de computação, também pode ser adotado para o ensino de ML. O ciclo *Use-Modify-Create* proposto por Lee et al. (2011) reduz o estresse cognitivo através do gerenciamento/suporte adequado da intensidade da interatividade de elementos. Neste modelo, na primeira fase de *Use*, o aluno é estimulado a utilizar artefatos/códigos prontos, criados por outras pessoas. A seguir na fase *Modify*, o aluno deve alterar, completar ou acrescentar funcionalidades extras a artefatos/códigos existentes. Por fim, na fase *Create* o aluno é incentivado a produzir seus próprios artefatos, em um ciclo contínuo de teste, análise e refinamento.

Seguindo diretrizes curriculares abordando ML desde a Educação Básica (Long & Magerko, 2020; Touretzky et al., 2019), 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. Adotando o ciclo *Use-Modify-Create* (Lee et al., 2011), 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 (Amershi et al., 2019), 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). A utilização de ferramentas visuais, como o Google Teachable Machine (GTM) (Google, 2023) 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 (Gresse von Wangenheim et al., 2021).

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 (Huba & Freed, 2000). Avaliar e fornecer feedback adequado é importante tanto para o aluno quanto para o professor (Hattie & Timperley, 2007). Em um processo efetivo de aprendizado, é importante que os estudantes saibam 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). Nesse sentido, uma alternativa interessante são as rubricas (Morrison et al., 2019) e também são muito comuns (McMillan, 2018), já que podem ser usadas em questões dissertativas, avaliação de performance, desenvolvimento de produtos, portfólios, demonstrações e outros (McMillan, 2018; Morrison et al., 2019). Uma rubrica é um guia de pontuação que define critérios e seus diferentes níveis de desempenho (Morrison et al., 2019). O objetivo de se ter rubricas é comunicar um padrão de julgamento, permitindo assim aos alunos identificarem seus pontos fortes e fracos (Morrison et al., 2019).

Quando abordamos temas como pensamento computacional, algoritmos e programação, modelagem e simulação na educação básica, já há esforços consideráveis para abordar a avaliação (Alves et al., 2020a, 2021a; Lye & Koh, 2014; Tang et al., 2020; Yasar et al., 2016) inclusive para avaliar conceitos relacionados como design de interface (Alves et al., 2020b; Solecki et al., 2020) como também habilidades como a criatividade (Alves et al., 2020c, 2021b).

Entretanto, se observa ainda uma carência de abordagens para a avaliação da aprendizagem de conceitos ML de forma confiável e válida (Rauber & Gresse von Wangenheim, 2022). 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 (Hitron et al., 2019; Hsu et al., 2022). Como resultado, Hsu et al. (2022) relataram um valor Cronbach α de 0,883 para a confiabilidade de um questionário de autoavaliação com cinco itens usando uma escala Likert. Hitron et al. (2019) 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 (2021) 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. (2021) 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. (2021) ainda apresentam uma avaliação inicial da validade da rubrica com um painel de especialistas, continuado em Rauber et al. (2022) ao propor uma adaptação da rubrica de Gresse von Wangenheim et al. (2021) e apresentar resultados iniciais positivos relativos à validade e confiabilidade do instrumento com base uma série de estudos de casos realizados. Os resultados desse 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 (Rauber et al., 2022).

Na busca da validação da confiabilidade e da validade, uma alternativa à Teoria Clássica de Teste ou *Classical Test Theory*, é a Teoria de Resposta ao Item (TRI), muitas vezes abordada como moderna e superior, devido a criação e interpretação de uma escala (DeVellis, 2017). A TRI é *"uma coleção de modelos de medição que objetivam explicar conexões entre respostas observadas em item em uma escala e um construto subjacente"* (Cappelleri et al., 2014). 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 Avaliação da aprendizagem de conceitos de ML em nível *Use* na Educação Básica

Visando à avaliação baseada em desempenho do aprendizado de ML para classificação de imagens no nível de *Use* do ciclo *Use-Modify-Create* (Lee et al., 2011) nos anos finais do Ensino Fundamental e Médio, foi projetado, desenvolvido e implementado um modelo de avaliação seguindo a metodologia *Evidence-Centered Design* (Design Centrado em Evidências) (Mislevy et al., 2003; Seeratan & Mislevy, 2008) como resultado de um trabalho anterior (Gresse von Wangenheim et al., 2021).

2.1 Análise de domínio

O modelo de avaliação está inserido no contexto do curso "ML para Todos!" (Gresse von Wangenheim et al., 2020; Martins et al., 2023) que foi desenvolvido com base nas Diretrizes para Inteligência Artificial - Grande Ideia 3: Aprendizagem (Touretzky et al., 2019) e nas diretrizes curriculares sobre alfabetização em IA (Long & Magerko, 2020), bem como em um processo de ML centrado no ser humano (Amershi et al., 2019). Seu objetivo é promover a compreensão dos conceitos fundamentais de aprendizado de máquina com foco na classificação de imagens com redes neurais artificiais, incluindo preparação de dados, treinamento de modelos, avaliação de desempenho e previsão de modelos de ML no nível *Use*. O público-alvo são alunos de escolas públicas dos anos finais do Ensino Fundamental e Médio do Brasil com pelo menos 12 anos de idade.

Nessa etapa educacional, espera-se que os alunos sejam proficientes na língua portuguesa, tenham desenvolvido o raciocínio lógico e matemático e sejam capazes de usar computadores para tarefas rotineiras, como acessar a Internet (Ministério da Educação, 2018). No entanto, em sua maioria, o ensino de computação no Brasil apenas está disponível na forma de programas extracurriculares (Santos et al., 2018), e muitos alunos ainda não têm competências em computação e IA/ML.

Em relação à infraestrutura, mais da metade (66%) dos alunos de escolas urbanas têm pelo menos um computador ou tablet em casa, e 42% usam a tecnologia por mais de três horas por dia (CGI, 2019). Além disso, 98% dos alunos de escolas urbanas têm um *smartphone* para acessar a Internet (CGI, 2019).

Devido à falta de professores de informática (Ministério da Educação, 2020), o ensino de computação é normalmente introduzido de uma forma interdisciplinar, sendo lecionado por professores com formação em diversas áreas do conhecimento, como a história, as ciências, etc. Este fato pode dificultar a avaliação dos resultados da aprendizagem de computação e até mesmo produzir resultados pouco confiáveis. Além disso, dado que as turmas das escolas públicas do público alvo possuem frequentemente cerca de 30 alunos ou mais, a avaliação manual dos projetos dos alunos é trabalhosa e morosa.

2.2 Modelagem de domínio

Com base nos resultados da análise de domínio, 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, são apresentados na Tabela 1, usando o padrão de projeto *Principled Assessment Designs for Inquiry* (Seeratan & Mislevy, 2008), um framework conceitual onde são estabelecidos os elementos para avaliação em contextos educacionais e fornece uma estrutura para o design de avaliações que visa projetar avaliações autênticas e que reflitam as habilidades do mundo real que os alunos precisam desenvolver.

2.3 Framework de Avaliação

2.3.1 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 & Magerko, 2020) e um processo de ML centrado no ser humano (Amershi et al., 2019). O objetivo é introduzir uma compreensão dos conceitos básicos de ML com foco na classificação de imagens usando redes neurais artificiais. No nível *Use*, o objetivo geral de aprendizagem é apresentar aos alunos o desenvolvimento de um modelo de ML para classificação de imagens e obter uma compreensão básica de como funcionam o ML e as redes neurais. Isso abrange o conhecimento essencial e as habilidades apresentadas na Tabela 1.

Tabela 1: Modelação da avaliação das competências de ML no nível *Use*.

Elemento	Descrição
Fundamento	As redes neurais são uma técnica atual e fundamental em ML para o desenvolvimento de modelos de classificação de imagens.
Conhecimentos, competências e outros atributos essenciais	Compreensão dos conceitos básicos sobre redes neurais. 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.

Continua na próxima página.

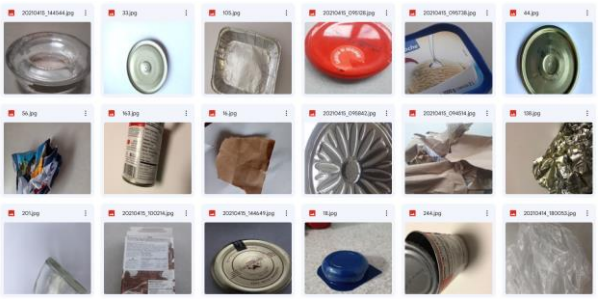
Tabela 1: Modelação da avaliação das competências de ML no nível *Use*. (*Continuação da página anterior.)*

Elemento	Descrição
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.
Potenciais produtos de trabalho (work products)	Arquivo Google Teachable Machine (.tm) incluindo o conjunto de dados (dataset) e os rótulos de categorias Relatório sobre a avaliação dos resultados dos testes com novos objetos Relatório sobre a avaliação do desempenho do modelo (tabela de acurácia e matriz de confusão) Relatório das melhorias efetuadas
Rubrica(s) potencial(ais)	Rubrica para aplicação de conceitos de ML para classificação de imagens - Nível *Use* (Gresse von Wangenheim et al., 2021)
Características	A tarefa deve permitir aos alunos limpar e rotular um conjunto de dados de imagens de lixo reciclável. A tarefa deve permitir que os alunos treinem o modelo ML. As tarefas devem permitir aos alunos analisar e interpretar o desempenho do modelo com base nos resultados da validação (acurácia das categorias e matriz de confusão) e no teste de novas imagens.
Características variáveis	Não são identificadas características variáveis no nível *Use*
Potenciais observações	Tamanho, distribuição e corretude dos rótulos do conjunto de dados Execução do treinamento do modelo Correta da análise e interpretação do desempenho (tabela de acurácia, matriz de confusão) Execução de ações de melhoria Correta análise e interpretação dos testes com novos dados

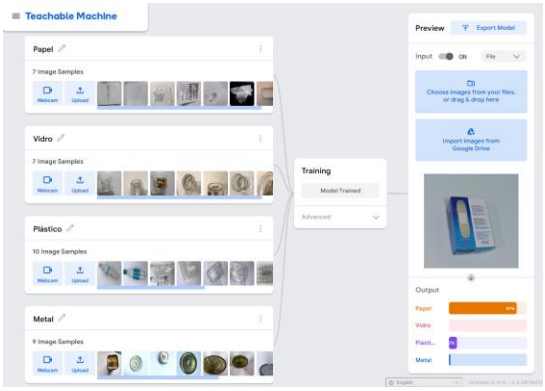
2.3.2 Modelo de Tarefa

O conjunto de produtos de trabalho a ser analisado está intrinsecamente inserido no curso "ML para Todos!" (Gresse von Wangenheim et al., 2020). Com duração de 8 horas, o curso ensina ML a alunos dos anos finais do Ensino Fundamental e Médio que têm habilidades básicas de computação e nenhuma experiência anterior com IA/ML. O curso pode ser implementado de duas maneiras: inserido no currículo das escolas como parte das aulas de ciências ou como atividade extracurricular. Ele também pode ser oferecido em diferentes modos, incluindo aprendizado remoto em ritmo próprio, cursos on-line remotos com instrutores ou aulas presenciais. O curso ensina conceitos básicos de ML e redes neurais e como desenvolver um modelo de ML predefinido para reconhecimento de imagens seguindo as etapas básicas de um processo de ML centrado no ser humano, incluindo preparação de dados, treinamento de modelos, avaliação de desempenho e previsão. A aplicação dos conceitos de ML é ensinada de forma interdisciplinar relacionada aos Objetivos de Desenvolvimento Sustentável (United Nations, 2015), concentrando-se na tarefa de classificar imagens de lixo reciclável. Após a motivação e a apresentação dos conceitos básicos de ML e redes neurais, os alunos iniciam o desenvolvimento do modelo de ML predefinido (Figura 1). Para criar o modelo de ML, os alunos são orientados passo a passo a usar um ambiente visual (Figura 1 - b), o Google Teachable Machine (GTM) (Google, 2023). Além disso, os alunos recebem um conjunto de 210 imagens redimensionadas e não categorizadas para preparar um conjunto de dados (Figura 1 - a). Os alunos precisam limpar o conjunto de dados e rotular as imagens com relação às categorias de reciclagem: metal, papel, plástico e vidro. Eles também são incentivados a expandir o conjunto de dados coletando imagens de lixo que tenham em mãos. Em seguida, eles são instruídos a treinar o modelo com o GTM, testar o modelo com novas imagens e interpretar o desempenho obtido pelo modelo, levando em conta seus testes, a acurácia do modelo e a matriz de confusão fornecida pelo GTM (Figura 1 - d). Durante o curso, os alunos também são instigados a ajustar o conjunto de dados e/ou alterar os

parâmetros de treinamento para melhorar o desempenho do modelo de ML (Figura 1 - c). O curso está disponível on-line gratuitamente em português brasileiro https://cursos.computacaonaescola.ufsc.br/.



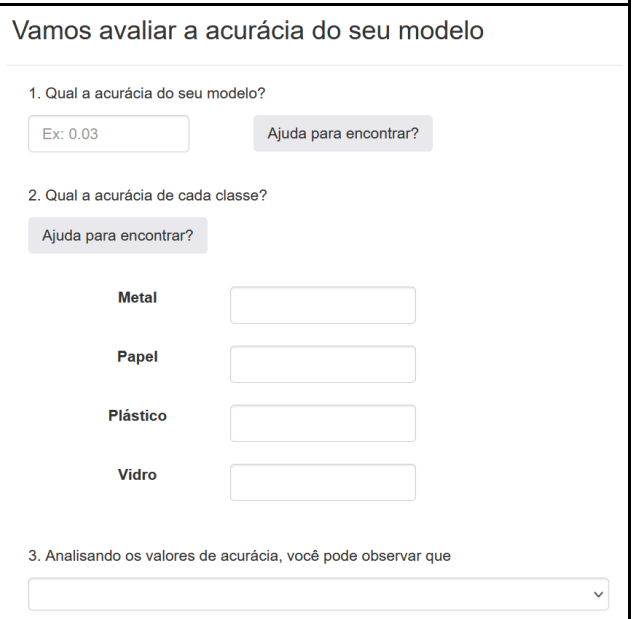
a) Preparar o conjunto de dados (dataset)



b) Treinamento e ajustes do modelo

```
{
"type":"image",
"version":"2.4.4",
"appdata":
  {"publishResults":
       {"tensorflowjs":
            {"name":"aa_Omited",
            "token":"bb_Omited"
            }
       },
  "trainEpochs":50,
  "trainBatchSize":16,
  "trainLearningRate":0.001
  }
}
```

c) Parâmetros de treinamento avaliados diretamente do arquivo Google Teachable Machine (.tm).



d) Exemplo de um relatório on-line para coleta de evidências sobre interpretação e análise da acurácia

Figura 1: Exemplos de produtos de trabalho criados pelos alunos como resultado do aprendizado.

2.3.3 Modelo de evidência

O modelo de evidência descreve como os detalhes das variáveis do modelo do aluno devem ser atualizados com base em um desempenho na forma de produtos de trabalho do aluno oriundos de suas tarefas (Mislevy et al., 2003). É composto por um modelo de avaliação e um modelo de medição. O modelo de avaliação explica como extrair variáveis observáveis relativas ao desempenho dos alunos a partir de produtos de trabalho de tarefas específicas e criar evidências que reflitam o nível de competência em informações dos alunos. As variáveis observáveis podem agrupar outras variáveis observáveis para descrever as características que estão sendo avaliadas (Mislevy et al., 2003). Portanto, neste artigo, os termos variáveis observáveis, itens e critérios de avaliação serão usados de forma intercambiável. Com foco na avaliação baseada no desempenho, o modelo de avaliação é apresentado como uma rubrica de pontuação (Tabela 2), adaptada a partir das propostas de Gresse von Wangenheim et al. (2021) e Rauber et al. (2022). Gresse von Wangenheim et al. (2021) avaliou a rubrica através de 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 (Rauber et al., 2022), apontou confiabilidade da rubrica com um valor de ômega global 0,646 e a validade convergente do construto com análise da matriz de correlação policórica. Nesta rubrica são definidas as variáveis observáveis a serem medidas para avaliar a capacidade de desenvolver um modelo de ML inferindo indiretamente a obtenção de competências de ML. Os níveis de desempenho foram definidos de acordo com os resultados de aprendizagem, especificando os critérios associados aos objetivos de aprendizagem e os indicadores que descrevem cada nível para avaliar o desempenho do aluno. Em termos do conceito que está sendo medido, níveis mais altos indicam uma compreensão maior. Os níveis de desempenho foram especificados em uma escala ordinal de 3 pontos, variando entre "Fraco", "Aceitável" e "Bom", em conformidade com o desempenho esperado para atingir o objetivo de aprendizado específico.

Tabela 2: Rubrica de pontuação – nível *Use*.

Critério / Variáveis Observáveis		**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 tem 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 consequente interpretação a respeito do modelo

Continua na próxima página.

Tabela 2: Rubrica de pontuação – nível *Use*.(*Continuação da página anterior*.)

Critério / Variáveis Observáveis		Níveis de Desempenho		
		Fraco - 0 pontos	**Aceitável - 1 ponto**	**Bom - 2 pontos**
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 consequente 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

Para o modelo de medição foi determinada uma pontuação geral que varia de 0 a 10 pontos, de acordo com o sistema de avaliação brasileiro. Assim, foi realizado um ajuste de escala, calculado a partir da conversão linear simples da média de pontos obtidos pelo aluno.

Uma análise da rubrica inicialmente proposta por Gresse von Wangenheim et al. (2021), realizada por um painel de especialistas, indicou uma concordância substancial de confiabilidade entre avaliadores, bem como validade de face em termos de correção, relevância, completude e clareza. Rauber et al. (Rauber et al., 2022) ao propor uma adaptação da rubrica de Gresse von Wangenheim et al. (2021), reportou resultados iniciais positivos relativos à validade e confiabilidade do instrumento com base em 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 (Rauber et al., 2022).

2.3.4 Implementação e entrega da Avaliação

Os produtos de trabalho elaborados pelos alunos durante o curso "ML para Todos!" 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 da performance e predição dos resultados como parte da execução do processo de ML.

3 Métodos

O objetivo do presente estudo é analisar de forma exploratória a rubrica (Tabela 2) 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 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)?

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

***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 (Trochim & Donnelly, 2008). Durante a aplicação do curso "ML para Todos!" foram coletados artefatos criados pelos alunos como resultados de aprendizagem.

Os dados foram coletados de 5 aplicações 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 (Tabela 3). 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. Também, no mesmo período, a aplicação (AP5) foi executada de forma autônoma e assíncrona pelos estudantes. Um total de 108 alunos entregaram, ainda que parcialmente, os artefatos criados ao longo do curso.

Tabela 3: Visão geral das aplicações.

Aplica-ção	Data	Instituição	Modo instrucional	Tipo de atividade	Idade (Anos)	Etapa educacional	No. de estudantes
AP1	Setembro 2021	Escola Básica Municipal. Dilma Lúcia dos Santos	Presencial	Como parte das aulas	15-16	Ensino Fundamental	12
AP2	Outubro 2021	Instituto Federal Catarinense (IFC)	Remoto no ritmo do instrutor	Extracurri-cular	15-17	Ensino Fundamental e Médio	10
AP3	Novembro 2021	Aberto a qualquer estudantes interessados, organizado pela Universidade Federal de Santa Catarina (UFSC)	"	"	12-18	"	35
AP4	Março 2022	Aberto a qualquer estudantes interessados, organizado pela Universidade Federal de Santa Catarina (UFSC)	"	"	14-18	"	40
AP5	–	Curso online	Remoto em ritmo próprio	"	≤ 18	"	11
Total							**108**

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 2) 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 (Laydner, 2022). 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 (Moskal & Leydens, 2000). Inicialmente, de acordo com a Teoria Clássica de Teste, 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 (Flora, 2020). A Teoria Clássica de Teste ainda fornece um conjunto de medidas estatísticas usadas para analisar e revisar itens, estimar suas características e fazer julgamentos

sobre a qualidade desses itens, tipicamente envolvendo medidas de dificuldade, discriminação e diferenciação (Bichi, 2016; Bennett & von Davier, 2017; Rust et al., 2020).

Após, investigamos a adequação e calibração à TRI da rubrica usando o modelo logístico de 2 parâmetros (Andrade et al., 2000; Paek & Cole, 2020), 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 (DeVellis, 2017), 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; Mukaka, 2012). Também foi realizada uma análise de dimensionalidade, com análise exploratória de matriz de correlação comparados com matrizes aleatórias paralelas e uma análise fatorial exploratória (Brown, 2015).

4 Preparação dos dados

Com base na análise do desempenho dos artefatos criados pelos estudantes utilizando a rubrica (Tabela 2), foram levantadas as frequências nos níveis de desempenho atingidos pelos estudantes ao longo do curso "ML para Todos!", conforme apresentado na Tabela 4. Alguns critérios da rubrica não puderam ser inferidos, pois alguns estudantes não entregaram alguns dos diferentes artefatos considerados (indicado por NA´s).

Tabela 4: 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
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

Como nem todos os estudantes entregaram todos os artefatos, há diferenças nas frequências em que os diferentes critérios foram avaliados. Assim, os critérios mais vezes avaliados foram "C7-Testes com novos objetos" e "C8-Interpretação dos testes" com 87 avaliações, enquanto somente 65 vezes foi avaliado o critério "C10-Interpretação da matriz de confusão".

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 5.

Tabela 5: 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 Teoria Clássica de Teste, 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.

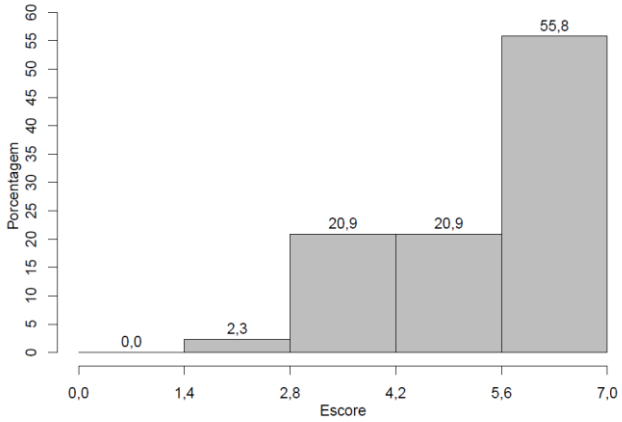


Figura 2: Desempenho dos estudantes.

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) (Brown, 2015). Como resultado foi obtido um valor aceitável Ômega Global de 0,781.

Ao se analisar se a consistência interna aumenta eliminando um item (Tabela 6), se observa que o coeficiente aumenta eliminando alguns itens (C1, C5 e C8), mesmo que não seja sensivelmente significante o aumento, poderia indicar necessidade de revisão destes itens.

Com relação a qualidade dos itens (Tabela 7) a análise de dificuldade (Dif), discriminação (Disc) e diferenciação (Bis) dos itens. A dificuldade dos itens foi analisada considerando o índice de dificuldade, que é dado em percentual, onde um valor alto indica que o item é fácil (Bennett & von Davier, 2017; Rust et al., 2020). Tipicamente o índice de dificuldade é classificado em intervalos, não havendo um consenso claro sobre esses limites. Assim definimos uma classificação para o índice de dificuldade: entre 0 a <15% muito difícil, de 15 a <35% difícil, de 35% a <65% médio, 65% a <85% fácil, e 85% ou mais muito fácil. De forma geral todos os itens apresentam índice de dificuldade adequado. Apenas o item C3 apresenta um índice de dificuldade muito fácil. A discriminação, analisada com o índice de discriminação, refere-se a capacidade de discriminação dos itens, calculada pela diferença entre as porcentagens do grupo de alunos com alto desempenho e o grupo com baixo desempenho (Bennett & von Davier, 2017; Rust et al., 2020). Variando entre -1 e 1, não há um consenso claro sobre a classificação em intervalos do índice de discriminação, mas índices negativos indicam a necessidade de revisão ou mesmo exclusão do item (Bichi, 2016; Bennett & von Davier, 2017; Rust et al., 2020). Assim, para o índice de discriminação assumimos a classificação de não discriminante se <0, baixa entre 0 a <0,1, moderada de 0,1 a <0.2, adequada de 0,2 a <0,3 e excelente se for 0.3 ou acima. De forma geral, todos os itens apresentam um bom índice de discriminação. Apenas o item C3 apresenta um índice de discriminação baixo, sendo que os demais variam de adequado a excelente. Para analisar a diferenciação, que também está relacionada a discriminação, utilizamos a correlação bisserial, que considera todas as respostas, diferentemente do índice de discriminação que

considera apenas os grupos extremos (Bichi, 2016). Variando entre -1 e 1, não há um consenso claro sobre a classificação em intervalos do bisserial, mas índices negativos indicam a necessidade de revisão ou mesmo exclusão do item (Bichi, 2016; Bennett & von Davier, 2017; Rust et al., 2020). Assumimos para a correlação bisserial classificação de inapropriado se <0, inadequado entre 0 a <0,1, moderada de 0,1 a <0.2, adequada de 0,2 a <0,3 e excelente se for 0.3 ou acima. De forma geral, todos os itens apresentam um bom índice bisserial, sendo a maioria excelentes.

Tabela 6: Coeficiente ômega ao excluir itens.

Item	Ômega
C1 - Quantidade de imagens	0,783
C3 - Distribuição do conjunto de dados	0,740
C5 - Limpeza dos dados	0,787
C8 - Interpretação dos testes	0,791
C9 - Interpretação da acurácia	0,669
C10 - Interpretação da matriz de confusão	0,736
C11 - Ajustes / Melhorias realizadas	0,752

Tabela 7: Qualidade dos itens segundo Teoria Clássica de Teste.

Item	Dif	Clas. Dif	Bis	Class. Bis	Disc	Clas. Disc
C1 - Quantidade de imagens	74,4	Fácil	0,106	Moderado	0,124	Moderada
C3 - Distribuição do conjunto de dados	97,7	Muito Fácil	0,454	Excelente	0,053	Baixa
C5 - Limpeza dos dados	79,1	Fácil	0,323	Excelente	0,241	Adequada
C8 - Interpretação dos testes	60,5	Médio	0,224	Adequado	0,151	Moderada
C9 - Interpretação da acurácia	69,8	Fácil	0,715	Excelente	0,491	Excelente
C10 - Interpretação da matriz de confusão	83,7	Fácil	0,564	Excelente	0,294	Adequada
C11 - Ajustes / Melhorias realizadas	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 itens calibrarem com a TRI num modelo de 2 parâmetros dicotomizados e qualidade desta calibração (Tabela 8) (Andrade et al., 2000; Paek & Cole, 2020). Ao analisar o resultado da calibração da TRI do modelo de 2 parâmetros, o parâmetro "a" indica o padrão de discriminação e está associado com a qualidade do item, enquanto o valor de "b" indica o índice de dificuldade do item.

Tabela 8: Resultado da calibração do modelo de 2 parâmetros dicotomizados.

Item	a			b	
	Valor	SE		Valor	SE
C1 - Quantidade de imagens	0,669	0,428		-1,761	1,092
C3 - Distribuição do conjunto de dados	0,821	0,516		-4,972	2,903
C5 - Limpeza dos dados	0,806	0,420		-1,881	0,920
C8 - Interpretação dos testes	0,732	0,410		-0,654	0,560
C9 - Interpretação da acurácia	1,345	0,490		-0,487	0,371
C10 - Interpretação da matriz de confusão	1,086	0,465		-1,856	0,693
C11 - Ajustes / Melhorias realizadas	0,944	0,438		-1,851	0,780

De forma geral, a calibração com a TRI demonstrou bons resultados. Mesmo não havendo um claro consenso sobre os limites de classificação do padrão de discriminação (parâmetro a), tipicamente, espera-se que cada item apresente um valor acima de 0,7 e abaixo de 5. Alguns autores também sugerem limites mais restritivos, em que valores razoavelmente bons variam de

0,8 a 2,5 (ex., De Ayala & Little, 2022). De uma forma geral os itens apresentam um bom padrão de discriminação. O que implica que os itens conseguem adequadamente distinguir entre alunos com bom e mau desempenho. Mesmo o item C1 que apresenta uma qualidade de discriminação ligeiramente baixa (0,669) mas muito próximo do valor adequado de 0,7, quanto C8 (0,732) ligeiramente acima, foram mantidos.

Também não há um claro consenso sobre a classificação dos limites a respeito do índice de dificuldade (parâmetro b), mas tipicamente, tanto a dificuldade do item quanto a habilidade do estudante espera-se que estejam localizados entre -5 e 5, mas há autores que sugerem limites mais restritivos variando entre -3 e 3 (ex. De Ayala & Little, 2022). Aqui, o nível de performance de todos os itens mostra um índice de dificuldade adequado, sendo que a maioria está inserida nos limites mais restritivos. O pior desempenho foi apresentado pelo item C3 (-4,972), mesmo assim, dentro do limite tipicamente utilizado.

Outra maneira de visualizar a relação entre a habilidade do aluno, índice de discriminação e índice dificuldade dos itens usando a TRI é através das curvas características dos itens (Figura 3). Elas indicam a probabilidade (P(Θ)) de selecionar a resposta adequada a cada item em função do nível de habilidade *theta* (Θ) do estudante. Foi mantida a escala padrão da TRI, em que a média é indicada pelo desempenho (*theta*) igual a zero e cada nível indica um desvio padrão.

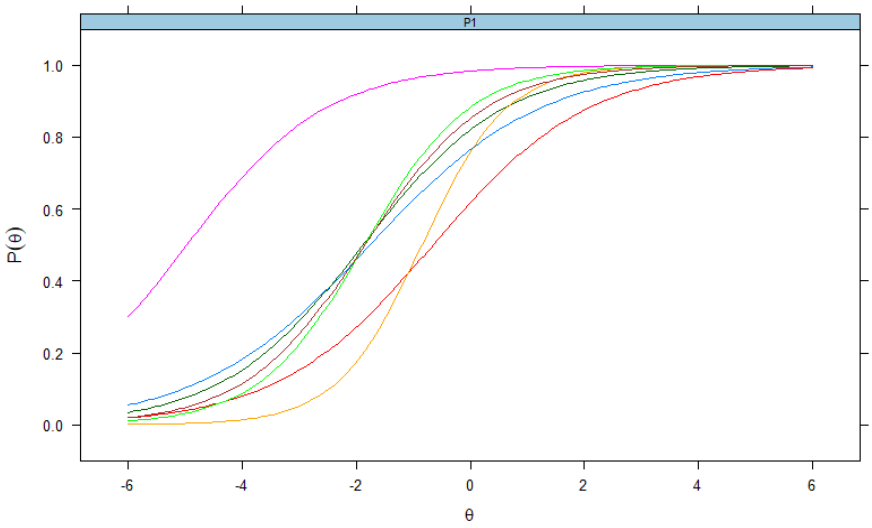


Figura 3: Gráfico de probabilidade de desempenho em função do score de todos os itens.

As curvas características dos itens colocam a habilidade do aluno e a dificuldade associada a cada categoria na mesma escala, permitindo que sejam relacionadas. Na Figura 3 o traço latente é apresentado variando de -6 (representando baixa habilidade) até 6 (alta habilidade). Assim, níveis de performance que são mais facilmente alcançados pelos alunos, estão representado com *theta* (Θ) negativo (eixo x), enquanto os mais difíceis com maior *theta* (Θ). Na Figura 3, a linha destacada das demais refere-se ao item "C3-Distribuição do conjunto de dados", mostrando que é mais facilmente atingido, mesmo por alunos com baixa habilidade.

Analisando a discriminação das curvas características dos itens, podemos observar que a inclinação das linhas no ponto médio indica um bom padrão de discriminação. Ela também representa adequadamente as habilidades dos alunos, uma vez que os baixos níveis de habilidade dos alunos (*theta*) tendem a apresentar uma pequena probabilidade de atingir o nível de desempenho "adequado". Ao mesmo tempo, os níveis crescentes de habilidade (*theta*) tendem a ter uma probabilidade maior de atingir o nível de desempenho "adequado". Isso reflete a boa aderência e os valores adequados de ajuste do modelo IRT usado para o traço latente.

A curva de informação do teste (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 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).

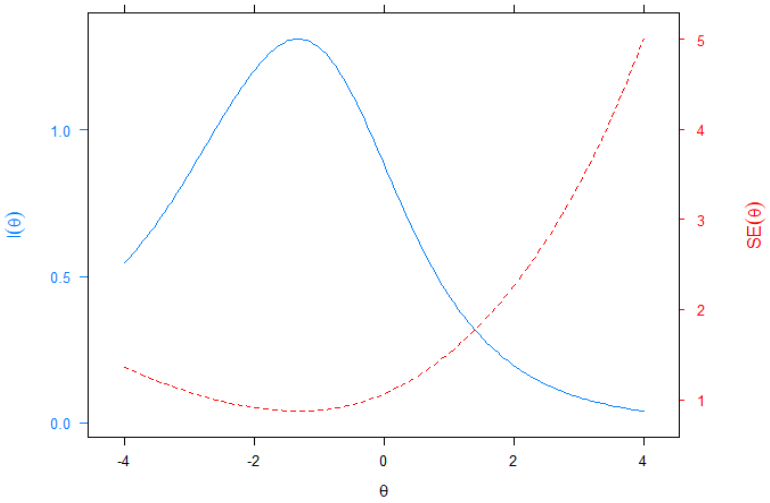


Figura 4: Informação do teste e erro padrão.

Na Tabela 9 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 (veja Tabela 8, coluna a-Padrão de discriminação). Como está sendo utilizado um modelo dicotomizado o item deve ser posicionado onde assume uma probabilidade maior que 50%.

Tabela 9: Probabilidade de posicionamento dos itens na escala (0,1).

Item	Níveis da Escala								
	-4	-3	-2	-1	0	1	2	3	4
C1 - Quantidade de imagens	0,18	0,30	0,46	*0,62*	0,76	0,86	0,93	0,96	0,98
C3 - Distribuição do conjunto de dados	*0,69*	0,83	0,92	0,96	0,98	0,99	1,00	1,00	1,00
C5 - Limpeza dos dados	0,15	0,29	0,48	*0,67*	0,82	0,91	0,96	0,98	0,99
C8 - Interpretação dos testes	0,08	0,15	0,27	0,44	*0,62*	0,77	0,87	0,94	0,97
C9 - Interpretação da acurácia	0,01	0,05	0,17	0,45	**0,76**	0,92	0,98	0,99	1,00
C10 - Interpretação da matriz de confusão	0,09	0,22	0,46	**0,72**	0,88	0,96	0,99	0,99	1,00
C11 - Ajustes / Melhorias realizadas	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 9, podemos inferir o grau de dificuldade de cada item, o que está representado na Figura 5.

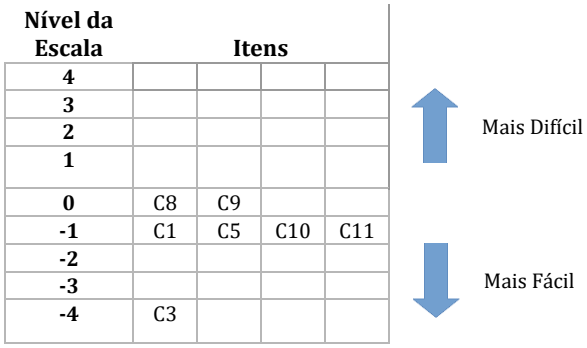


Figura 5: Posicionamento dos itens na escala (0,1).

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.

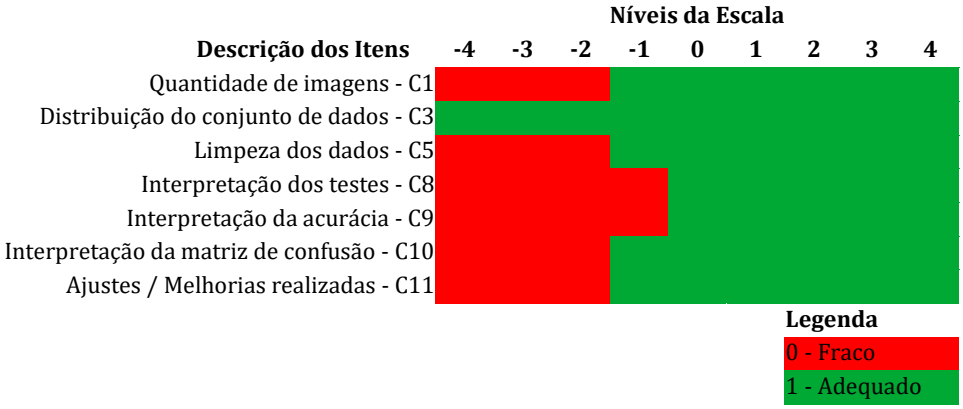


Figura 6: Representação gráfica do nível de desempenho dos estudantes.

Assim, partindo da representação gráfica na Figura 6 pode ser inferida uma interpretação a ser dada aos escores obtidos pelos estudantes (Tabela 10).

5.2 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 11). 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 uma correlação 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 10: 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 11: Matriz de correlação policórica.

Itens	C1	C3	C5	C8	C9	C10	C11
Quantidade de imagens - **C1**	1,000						
Distribuição do conjunto de dados - **C3**	0,565	1,000					
Limpeza dos dados - **C5**	-0,147	-0,515	1,000				
Interpretação dos testes - **C8**	-0,130	0,449	0,150	1,000			
Interpretação da acurácia - **C9**	0,544	0,442	0,132	0,415	1,000		
Interpretação da matriz de confusão - **C10**	-0,322	-0,457	0,483	0,306	0,418	1,000	
Ajustes / Melhorias realizadas - **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.

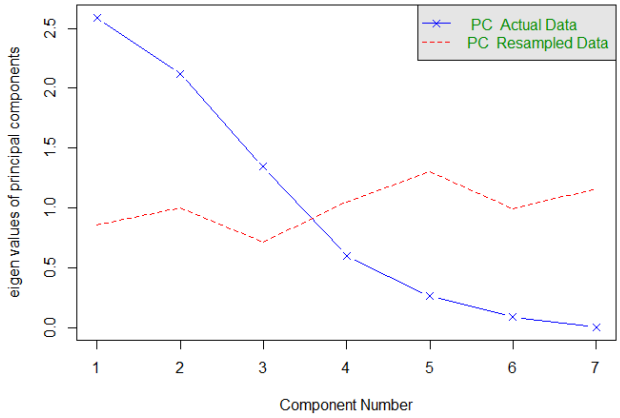


Figura 7: Matriz de correlação comparados com matrizes aleatórias paralelas

Partindo-se para uma análise fatorial exploratória (Tabela 12), 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 (Brown, 2015).

Tabela 12: 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 um 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 (Finch & West, 1997). 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. Adicionalmente, os resultados da análise dos itens da rubrica, seguindo a Teoria Clássica de Teste, apresentam índices adequados do grau de dificuldade, discriminação e diferenciação. Assim, demonstrando uma boa capacidade de diferenciar entre um bom e mau desempenho do estudante, uma importante característica em avaliações.

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 itens calibraram e a qualidade desta calibração chegou-se a valores adequados de qualidade de discriminação (parâmetro a) e também que a dificuldade dos itens (parâmetro b). Assim, de forma geral todos os itens estão adequados quanto a discriminação e dificuldade segundo a TRI. Mesmo adequados, resultados marginais foram encontrados para a "C1-Quantidade de imagens" com o menor padrão de discriminação e forma análoga, o item "C3-Distribuição do conjunto de dados" com o menor índice de dificuldade. Como abordado durante o curso com os alunos e discutido por Gresse von Wangenheim et al. (2021) e Martins et al. (2023), é importante que os alunos percebam que a acurácia do modelo desenvolvido melhora significativamente em função do número de imagens utilizadas para treinamento e alocadas em cada categoria, o que reflete a importância desses critérios. Talvez, a ferramenta GTM (Google, 2023) por utilizar uma rede neural pré-treinada com bons resultados, ofusque essa importância aos alunos, já que leva a bons resultados mesmo com poucas imagens.

Desta forma, foi possível identificar a dificuldade e discriminação dos itens em uma escala e a consequente 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 (Gresse von Wangenheim et al., 2021). Rauber et al. (2022) 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. (2022) 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 12) 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 8). 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 10).

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 (ex., Avila et al., 2017; Brennan & Resnick, 2012; Grover et al., 2015).

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 (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), 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 (parâmetro a) e também dificuldade dos itens (parâmetro 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 (Gresse von Wangenheim et al., 2018), 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 têm 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).

Artigo Premiado Estendido

Esta publicação é uma versão estendida de artigo premiado no II Simpósio Brasileiro de Educação em Computação (EduComp 2023), intitulado "Análise do desempenho de aprendizagem de Machine Learning na Educação Básica aplicando a Teoria de Resposta ao Item", DOI: 10.5753/educomp.2023.228159

Referências

Alves, N. da C., Gresse von Wangenheim, C., Hauck, J. C. R., & Borgatto, A. F. (2021). An Item Response Theory Analysis of Algorithms and Programming Concepts in App Inventor Projects. *Proc. of Brazilian Symposium on Computer Education*, Jataí, Goiás, Brazil. https://doi.org/10.5753/educomp.2021.14466 [GS Search]

Alves, N. da C., Gresse von Wangenheim, C., Hauck, J. C. R., & Borgatto, A. F. (2020). A Large-scale Evaluation of a Rubric for the Automatic Assessment of Algorithms and Programming Concepts. *Proc. of the 51st ACM Technical Symposium on Computer Science Education*, Portland, USA, Pages 556–562. https://doi.org/10.1145/3328778.3366840 [GS Search]

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. *Proc. of Simpósio Brasileiro de Informática na Educação*, Natal, Rio Grande do Norte, Brasil. https://doi.org/10.5753/cbie.sbie.2020.1563 [GS Search]

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. *Proc. of Simpósio Brasileiro de Informática na Educação*, Natal, Rio Grande do Norte, Brasil. https://doi.org/10.5753/cbie.sbie.2020.41 [GS Search]

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? *Proc. of Simpósio Brasileiro de Educação em Computação*, Jataí, Goiás. https://doi.org/10.5753/educomp.2021.14467 [GS Search]

Amershi, S., Begel, A., Bird, C., DeLine, R., Gall, H., Kamar, E., Nagappan, N., Nushi, B., & Zimmermann, T. (2019). Software Engineering for Machine Learning: A Case Study. *Proc. of IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice*, Montreal, Canada, 291–300. https://doi.org/10.1109/ICSE-SEIP.2019.00042 [GS Search]

Andrade, D. F., Tavares, H. R., & da Cunha Valle, R. (2000). *Teoria da Resposta ao Item: conceitos e aplicações*. São Paulo, SP, Brasil: ABE. [GS Search]

Avila, C., Cavalheiro, S., Bordini, A., Marques, M., Cardoso, M., & Feijo, G. (2017). Metodologias de Avaliação do Pensamento Computacional: Uma revisão sistemática. *Proc. of Simpósio Brasileiro de Informática na Educação*, Fortaleza, Ceará, Brasil, 28(1), 113. https://doi.org/10.5753/cbie.sbie.2017.113 [GS Search]

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

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

Bennett, R. E., von Davier, M. (2017). *Advancing human assessment: The methodological, psychological and policy contributions of ETS*. Springer Nature. https://doi.org/10.1007/978-3-319-58689-2 [GS Search]

Bichi, A. A. (2016). Classical Test Theory: An Introduction to Linear Modeling Approach to Test and Item Analysis. *International Journal for Social Studies*, 2(9), 27-33. [GS Search]

Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). The Guilford Press. [GS Search]

Brennan, K. & 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. [GS Search]

Camada, M. Y. & Durães, G. M. (2020). Ensino da Inteligência Artificial na Educação Básica: um novo horizonte para as pesquisas brasileiras. *Proc. of XXXI Brazilian Symposium on Informatics in Education*. Porto Alegre, Brasil, 1553–1562. https://doi.org/10.5753/cbie.sbie.2020.1553 [GS Search]

Cappelleri, J. C., Jason Lundy, J., & 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. https://doi.org/10.1016/j.clinthera.2014.04.006 [GS Search]

Caruso, A. L. M., & Cavalheiro, S. A. da C. (2021). Integração entre Pensamento Computacional e Inteligência Artificial: uma Revisão Sistemática de Literatura. *Proc. of XXXII Brazilian Symposium on Informatics in Education*, Porto Alegre, Brasil, 1051–1062. https://doi.org/10.5753/sbie.2021.218125 [GS Search]

CGI (2019). *TIC Educação 2019*. São Paulo, SP, Brasil: Cetic. https://www.cetic.br/pt/pesquisa/educacao/indicadores/

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

Finch, J. F. & West, SG (1997). The investigation of personality structure: statistical models. *Journal of Research in Personality*, 31(4), 439-485. https://doi.org/10.1006/jrpe.1997.2194 [GS Search]

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. https://doi.org/10.1177/2515245920951747 [GS Search]

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

Gresse von Wangenheim, C. G. von, Hauck, J. C. R., Demetrio, M. F., Pelle, R., Cruz Alves, N. da, Barbosa, H. & Azevedo, L. F. (2018). CodeMaster—Automatic Assessment and Grading of App Inventor and Snap! Programs. *Informatics in Education*, 17(1), 117–150. https://doi.org/10.15388/infedu.2018.08 [GS Search]

Gresse von Wangenheim, C., Alves, N. da C., Rauber, M. F., Hauck, J. C. R., & Yeter I. H. (2021a). A Proposal for Performance-based Assessment of the Learning of Machine Learning Concepts and Practices in K-12. *Informatics in Education*, 21(3), 479–500. https://doi.org/10.15388/infedu.2022.18 [GS Search]

Gresse von Wangenheim, C., Marques, L. S., & Hauck, J. C. R. (2020). Machine Learning for All – Introducing Machine Learning in K-12, *SocArXiv*, 1-10. https://doi.org/10.31235/osf.io/wj5ne [GS Search]

Grover, S., Pea, R., & 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*, Chicago, Illinois, USA, 15–20. [GS Search]

Hattie, J. & Timperley, H. (2007). The Power of Feedback. *Review of Educational Research*, 77(1), 81–112. https://doi.org/10.3102/003465430298487 [GS Search]

Hitron, T., Orlev, Y., Wald, I., Shamir, A., Erel, H., & 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*, Glasgow Scotland, Uk, 1–11. https://doi.org/10.1145/3290605.3300645 [GS Search]

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

House of Lords (2018). *AI in the UK: ready, willing and able*. London, UK: HL Paper 100. Retrieved 01/09/2022 from https://www.politico.eu/wp-content/uploads/2018/04/AI-in-the-UK-ReadyWillingAndAble-April-2018.pdf [GS Search]

Hsu, T.-C., Abelson, H., & van Brummelen, J. (2021). The Effects on Secondary School Students of Applying Experiential Learning to the Conversational AI Learning Curriculum. *The International Review of Research in Open and Distributed Learning*, 23(1), 82-103. https://doi.org/10.19173/irrodl.v22i4.5474 [GS Search]

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

Kandlhofer, M., Steinbauer, G., Hirschmugl-Gaisch, S., & Huber, P. (2016). Artificial intelligence and computer science in education: From kindergarten to university. *Proc. of the Frontiers in Education Conference*, Erie, PA, USA, 1–9. https://doi.org/10.1109/FIE.2016.7757570 [GS Search]

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436–444. https://doi.org/10.1038/nature14539 [GS Search]

Lee, I., Martin, F., Denner, J., Coulter, B., Allan, W., Erickson, J., Malyn-Smith, J., & Werner, L. (2011). Computational thinking for youth in practice. *ACM Inroads*, 2(1), 32–37. https://doi.org/10.1145/1929887.1929902 [GS Search]

Long, D., & Magerko, B. (2020). What is AI literacy? Competencies and design considerations. *Proc. of the Conference on Human Factors in Computing Systems*, Honolulu, HA, USA, 1–16. https://doi.org/10.1145/3313831.3376727 [GS Search]

Lordelo, L. M. K., Hongyu, K., Borja, P. C., & 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. https://doi.org/10.18607/ES201875266 [GS Search]

Lwakatare, L. E., Raj, A., Bosch, J., Olsson, H. H., & 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*, Montréal, Canada, 227–243. https://doi.org/10.1007/978-3-030-19034-7_14 [GS Search]

Lye, S. Y. & 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. https://doi.org/10.1016/j.chb.2014.09.012 [GS Search]

Lytle, N. et al. (2019). Use, modify, create: Comparing computational thinking lesson progressions for stem classes. *Proc. of the ACM Conference on Innovation and Technology in Computer Science Education*, Aberdeen, Scotland, UK, 395–401. https://doi.org/10.1145/3304221.3319786 [GS Search]

Marques, L. S., von Wangenheim, C. G., & Hauck, J. C. R. (2020). Ensino de Machine Learning na Educação Básica: um Mapeamento Sistemático do Estado da Arte. *Proc. of XXXI Simpósio Brasileiro de Informática na Educação*, Natal, Rio Grande do Norte, Brasil., 21–30. https://doi.org/10.5753/cbie.sbie.2020.21 [GS Search]

Martins, R. M., von Wangenheim, C. G., Rauber, M. F., & Hauck, J. C. (2023). Machine Learning for All!—Introducing Machine Learning in Middle and High School. *International Journal of Artificial Intelligence in Education*. 1-39. https://doi.org/10.1007/s40593-022-00325-y [GS Search]

McMillan, James H. (org.) (2013). *Sage handbook of research on classroom assessment*. Los Angeles, USA: Sage Publications. [GS Search]

Ministério da Educação (2018). *Base Nacional Comum Curricular*. Retrieved 01/05/2023 from http://basenacionalcomum.mec.gov.br/

Ministério da Educação (2020). *Census of Basic Education 2020*. Retrieved 01/05/2023 from https://download.inep.gov.br/publicacoes/institucionais/estatisticas_e_indicadores/notas_estatisticas_censo_escolar_2020.pdf

Ministério da Educação (2022). *Normas sobre Computação na Educação Básica – Complemento à Base Nacional Comum Curricular (BNCC)*. Parecer 02/2022 CNE/CEB/MEC. Retrieved 01/05/2023 from http://portal.mec.gov.br/component/content/article/323-secretarias-112877938/orgaos-vinculados-82187207/12992-diretrizes-para-a-educacao-basica?Itemid=164

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). A Brief Introduction to Evidence-Centered Design. *ETS Research Report Series*, 2003(1), i–29. https://doi.org/10.1002/j.2333-8504.2003.tb01908.x [GS Search]

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

Morrison, G. R., Ross, S. M., Morrison, J. R., & Kalman, H. K. (2019). *Designing effective instruction* (8h ed.). Hoboken, NJ, USA: Wiley. [GS Search]

Moskal, B. M., & Leydens, J. A. (2000). Scoring rubric development: Validity and reliability. *Practical assessment, research, and evaluation*, 7(1), 10. https://doi.org/10.7275/Q7RM-GG74 [GS Search]

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

Paek, I., & Cole, K. (2020). *Using R for Item Response Theory Model Applications*. New York, NY, USA: Routledge. https://doi.org/10.4324/9781351008167 [GS Search]

Ramos, G., Meek C., Simard P., Suh J., & Ghorashi S. (2020). Interactive machine teaching: a human-centered approach to building machine-learned models. *Human–Computer Interaction*, 35(5–6), 413–451. https://doi.org/10.1080/07370024.2020.1734931 [GS Search]

Rauber, M. F. & Gresse von Wangenheim, C. (2022). Assessing the Learning of Machine Learning in K-12: A Ten-Year Systematic Mapping. *Informatics in Education*. https://doi.org/10.15388/infedu.2023.11 [GS Search]

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*, Manaus, AM, Brasil. https://doi.org/10.5753/sbie.2022.224688 [GS Search]

Royal Society (2017). *Machine learning: the power and promise of computers that learn by example*. Retrieved 01/06/2022 from https://royalsociety.org/-/media/policy/projects/machine-learning/publications/machine-learning-report.pdf

Rust, J., Kosinski, M., & Stillwell, D. (2020). *Modern Psychometrics: The Science of Psychological Assessment* (4th ed.). Routledge. [GS Search]

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science*, 18(2), 119–144. https://doi.org/10.1007/BF00117714 [GS Search]

Santos, P. S., Araujo, L. G. J., & 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, 1–8. [GS Search]

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

Shamir G. & Levin I. (2021). Neural Network Construction Practices in Elementary School. *Künstliche Intelligenz*, 35(2), 181–189. https://doi.org/10.1007/s13218-021-00729-3 [GS Search]

Solecki, I., Porto, J. A., Alves, N. D. C., Gresse von Wangenheim, C., Hauck, J. C. R., & Borgatto, A. F. (2020). Automated Assessment of the Visual Design of Android Apps Developed with App Inventor. *Proc. of the 51st ACM Technical Symposium on Computer Science Education*, Portland, OR, USA, 51–57. https://doi.org/10.1145/3328778.3366868 [GS Search]

Tang, X., Yin, Y., Lin, Q., Hadad, R., & Zhai, X. (2020). Assessing computational thinking: A systematic review of empirical studies. *Computers & Education*, 148, 103798. https://doi.org/10.1016/j.compedu.2019.103798 [GS Search]

Touretzky, D., Gardner-McCune, C., Martin, F., & Seehorn D. (2019). Envisioning AI for K-12: What Should Every Child Know about AI? *Proc. of the AAAI Conference on Artificial Intelligence*, Honolulu, HA, USA. https://doi.org/10.1609/aaai.v33i01.33019795 [GS Search]

Trochim, W. M. K., & Donnelly, J. P. (2008). *The research methods knowledge base* (3rd ed.). Mason, OH, USA: Atomic Dog/Cengage Learning. [GS Search]

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

United Nations (2015). *The 17 Goals*. Department of Economic and Social Affairs, Sustainable Development. Retrieved 06/06/2022 from https://sdgs.un.org/goals

Yasar, O., Veronesi, P., Maliekal, J., Little, L., Vattana, S., & Yeter I. (2016). Computational Pedagogy: Fostering a New Method of Teaching. *Proc. of the Annual Conference & Exposition*, New Orleans, LA, USA. https://doi.org/10.18260/p.26550 [GS Search]