

Classificação Automática de Movimentos Pedagógicos como Apoio à Análise e à Avaliação da Prática Docente: Um Estudo com a Base TalkMoves

Alex Almeida do Amaral¹, Huliane Medeiros da Silva², Geiser Chalco Chalco²

¹Programa de Pós-Graduação em Ciência da Computação (PPGCC)

Universidade Federal Rural do Semi-Árido (UFERSA)

Universidade do Estado do Rio Grande do Norte (UERN) – Mossoró – RN – Brasil

²Departamento de Engenharias e Tecnologia (DETEC)

Universidade Federal Rural do Semi-Árido (UFERSA) – Pau dos Ferros – RN – Brasil

alex.amaral@alunos.ufersa.edu.br, {huliane, geiser}@ufersa.edu.br

Abstract. *Systematic evaluation of teaching practice through classroom observation depends on trained raters, which limits its large-scale application. This study investigates to what extent classical machine learning techniques can support the automatic identification of pedagogical moves, understood as discursive actions used to guide classroom participation and reasoning. English-language mathematics lesson transcripts from the TalkMoves dataset were used. Different classification algorithms and model combination strategies were compared using the same textual representation. The best individual model achieved approximately 65% accuracy, while the best combination reached 65.5% accuracy and a 65.2% F1-score. The results indicate that these techniques can support the screening of classroom excerpts and the production of formative feedback, but their performance is not sufficient to replace human evaluation or produce autonomous judgments about teaching practice. The study provides a reproducible and low-cost baseline for future research on the automated analysis of classroom discourse.*

Keywords: *teaching practice analysis, teacher evaluation, classroom discourse, artificial intelligence in education, machine learning, TalkMoves.*

Resumo. *A avaliação sistemática da prática docente por meio da observação de aulas depende de avaliadores treinados, o que limita sua aplicação em larga escala. Este trabalho investiga em que medida técnicas clássicas de aprendizado de máquina podem apoiar a identificação automática de movimentos pedagógicos, entendidos como ações discursivas empregadas para conduzir a participação e o raciocínio em sala de aula. Foram utilizadas transcrições em inglês de aulas de matemática da base TalkMoves. Diferentes algoritmos de classificação e estratégias de combinação de modelos foram comparados a partir de uma mesma representação textual. O melhor modelo individual alcançou aproximadamente 65% de acurácia, enquanto a melhor combinação obteve 65,5% de acurácia e 65,2% de F1-score. Os resultados indicam que essas técnicas podem apoiar a triagem de trechos e a elaboração de feedback formativo, mas não apresentam desempenho suficiente para substituir a avaliação humana ou produzir julgamentos autônomos sobre a prática docente. O estudo*

oferece um baseline reproduzível e de baixo custo para futuras pesquisas sobre a análise automatizada do discurso em sala de aula.

Palavras-chave: *análise da prática docente, avaliação docente, discurso de sala de aula, inteligência artificial na educação, aprendizado de máquina, Talk-Moves.*

1. Introdução

Avaliar a prática docente é uma das tarefas mais sensíveis e menos automatizadas dos sistemas educacionais. Os modelos consolidados de observação, como o *Framework for Teaching* [Danielson 2007], o *Classroom Assessment Scoring System* [Pianta et al. 2008] e o *Stallings Observation System* [Stallings 1977], dependem de avaliadores presenciais e exigem treinamento extenso, o que torna o processo caro, pouco escalável e suscetível a variações entre observadores [Chanchal and Zualkernan 2022, Almubarak et al. 2024]. Esse gargalo tem consequência educacional direta: sem instrumentos escaláveis, o professor raramente recebe *feedback* sistemático sobre o próprio discurso, justamente o aspecto da aula que a pesquisa associa de forma mais robusta à aprendizagem. Uma síntese de mais de 800 meta-análises situa o *feedback* docente e a qualidade do discurso entre os fatores de maior impacto sobre o desempenho dos estudantes [Hattie 2009]. O conceito de *accountable talk* descreve movimentos pedagógicos (como verificação de entendimento, reformulação e solicitação de raciocínio) diretamente ligados à participação e à aprendizagem [Michaels et al. 2008].

Foi a partir dessa categorização que se construiu a base *TalkMoves* [Suresh et al. 2022], com 567 transcrições de aulas de matemática do nível K–12 (etapa que, nos Estados Unidos, abrange da educação infantil ao fim do ensino médio e corresponde, de forma aproximada, à educação básica brasileira), cada turno anotado segundo o movimento exercido por professor ou aluno. A digitalização do registro de aulas tornou viável tratar esse discurso como dado textual e analisá-lo automaticamente [Alic et al. 2022, Demszky and Hill 2023], abrindo espaço para que ferramentas computacionais apoiem a observação de aulas e a formação docente em escala. O problema que motiva este trabalho é, portanto, educacional antes de ser computacional: como apoiar, de forma escalável e transparente, a análise das práticas discursivas que o professor mobiliza em aula. Os algoritmos de classificação são, aqui, um meio para investigar esse problema, e não o seu fim.

Duas lacunas orientam o recorte. A primeira é de método: a literatura recente sobre a *TalkMoves* concentra-se em modelos contextuais profundos e em *Large Language Models* [Suresh et al. 2021, Cao et al. 2025, Long et al. 2024, Tran et al. 2024], deixando pouco documentado o desempenho de *baselines* clássicos, interpretáveis e de baixo custo, que sirvam de referência e de instrumento auditável para o professor. A segunda é prática: em contextos de pesquisa com recursos limitados, comuns em programas de pós-graduação brasileiros, é desejável dispor de *pipelines* reproduzíveis e explicáveis, inspecionáveis por educadores antes de qualquer migração para arquiteturas mais pesadas [Gunasekara and Saarela 2025, Türkmen 2025].

Questão de pesquisa. Em que medida técnicas clássicas de aprendizado de máquina podem apoiar a identificação automática de movimentos pedagógicos em transcrições de sala de aula, de modo a subsidiar processos de análise e avaliação da prática docente?

Objetivo geral. Investigar a viabilidade da classificação automática de movimentos pedagógicos presentes em interações discursivas de sala de aula por meio de técnicas de aprendizado de máquina, visando apoiar processos de análise e avaliação da prática docente. Como objetivos específicos: (a) investigar a viabilidade da identificação automática de práticas discursivas docentes em transcrições; (b) comparar diferentes abordagens de aprendizado de máquina para a classificação de movimentos pedagógicos; (c) avaliar o impacto da combinação de classificadores sobre a estabilidade e a qualidade da classificação; (d) analisar a estrutura dos dados discursivos resultantes da representação adotada; (e) analisar a influência dos componentes latentes da representação textual sobre as decisões dos modelos; e (f) discutir o potencial dessas evidências para a avaliação e a formação docente.

Contribuições. O trabalho oferece quatro contribuições articuladas em torno do problema educacional: (i) um *baseline* reproduzível e de baixo custo para a classificação de movimentos pedagógicos na *TalkMoves*, documentando um patamar interpretável contra o qual modelos mais pesados podem ser comparados; (ii) a comparação sistemática entre famílias de algoritmos e estratégias de combinação (modelos individuais, comitês homogêneos, heterogêneos e *stacking*); (iii) a análise da influência dos componentes latentes por meio da importância de *Random Forest* e dos valores SHAP, oferecendo uma primeira camada de inspeção das decisões dos modelos; e (iv) a discussão das implicações educacionais do desempenho obtido, em particular o que um resultado em torno de 65% significa, e o que não significa, para o apoio à observação de aulas, ao *feedback* e à formação docente. O fio que une essas contribuições é o deslocamento do eixo do trabalho, de uma comparação de algoritmos para o estudo, com apoio computacional, da prática discursiva docente.

As Seções seguintes apresentam o referencial teórico (Seção 2), os trabalhos relacionados (Seção 3), a metodologia (Seção 4), os resultados e a discussão (Seção 5), a análise de importância das variáveis (Seção 6), as limitações e ameaças à validade (Seção 7) e a conclusão (Seção 8).

2. Referencial Teórico

2.1. Avaliação da prática docente e observação de aulas

A prática docente compreende o conjunto de decisões pedagógicas tomadas pelo professor em sala de aula, desde a escolha das estratégias até a forma de interagir com os estudantes. Essas decisões envolvem a integração entre o conhecimento do conteúdo e o repertório pedagógico e estão relacionadas entre si, respondendo também às características do contexto da turma [Shulman 1987, Grossman et al. 2009]. Avaliar essa prática não é avaliar um ato pontual, e há evidência robusta de que ela afeta a aprendizagem [Hattie 2009, Danielson 2007]. Os instrumentos consolidados operam por observação humana: o *Framework for Teaching* [Danielson 2007] organiza a atuação em quatro domínios; o *Teacher Code* [Cohen 2011] registra comportamentos específicos; o CLASS [Pianta et al. 2008] acrescenta eixos de qualidade; e o *Stallings Observation System* [Stallings 1977] mantém uma taxonomia detalhada de atividades. Apesar das diferenças, compartilham limitações estruturais: alto custo de treinamento, baixa escalabilidade e variação entre avaliadores [Ninin 2009, Chanchal and Zulkernan 2022]. Trabalhos recentes reafirmam esse gargalo e já exploram alternativas apoiadas por inteligência artificial para a avaliação e a observação da docência [Almubarak et al. 2024].

São essas limitações que abrem espaço para abordagens automatizadas, que tratam a transcrição da aula como dado computacional.

No centro dessa avaliação está o discurso da aula, que se organiza em movimentos pedagógicos: as ações com que o professor conduz a participação e o raciocínio dos estudantes, como verificar o entendimento, reformular uma contribuição (*revoicing*) ou pedir que o aluno justifique o que afirmou. O termo tem origem no modelo de *accountable talk* [Michaels et al. 2008], que descreve a fala como meio de construção coletiva do conhecimento. Reconhecer esses movimentos é importante para a avaliação docente porque eles estimulam a elaboração e a explicitação do pensamento dos estudantes. Além disso, o *feedback* e a qualidade das interações em sala de aula estão entre os fatores associados à aprendizagem [Hattie 2009]. Classificar o discurso do professor segundo esses movimentos é uma forma de tornar observável, e analisável em escala, aquilo que liga a condução da aula ao que os alunos aprendem.

2.2. Análise computacional do discurso e IA na Educação

A análise do discurso de sala de aula como dado textual situa-se na confluência de *Learning Analytics*, *Educational Data Mining* e *Artificial Intelligence in Education* (AIED), com forte presença de classificação supervisionada e mineração de texto, inclusive na pesquisa brasileira em Informática na Educação [Oliveira et al. 2024]. A base *TalkMoves* [Suresh et al. 2022], anotada segundo o modelo de *accountable talk* [Michaels et al. 2008], e a base *NCTE Transcripts* [Demszky and Hill 2023], com 1.660 observações de aulas e escores de observação, transformaram a fala docente em objeto mensurável. Sobre esses recursos, estudos classificam tipos de questionamento [Alic et al. 2022], modelam movimentos pedagógicos em tutoria [Cao et al. 2025] e avaliam a qualidade de discussões [Tran et al. 2024]. A automação da observação tem sido proposta como uma forma de reduzir custos e ampliar a cobertura do *feedback*. A aplicação *TalkMoves*, por exemplo, fornece ao professor uma análise automatizada dos movimentos presentes em seu discurso [Suresh et al. 2021]. Outras abordagens investigam a observação automatizada de aulas a partir de dados de áudio, inclusive em ambientes com recursos computacionais limitados [Chanchal and Zualkernan 2022]. Trabalhos recentes com *Large Language Models* [Long et al. 2024, Tran et al. 2024] ganham consistência na codificação de diálogos, mas levantam preocupações de custo, dependência de serviços remotos e baixa transparência, o que valoriza *baselines* interpretáveis como ponto de comparação.

2.3. Explicabilidade e fundamentos de aprendizado de máquina

Em educação, a decisão do modelo precisa ser inspecionável: apontar a presença de um movimento sem permitir entender em que se baseou é pouco útil para formação e avaliação. Revisões recentes sobre *Explainable AI* (XAI) na pesquisa educacional [Türkmen 2025, Gunasekara and Saarela 2025] mostram que métodos como SHAP [Lundberg and Lee 2017] tornam transparentes decisões de modelos opacos. Quanto aos algoritmos, foram considerados perfis de inferência distintos para comparar paradigmas: *k-NN* [Cover and Hart 1967], baseado em vizinhança; Árvores de Decisão [Breiman et al. 1984], baseadas em regras; *Naive Bayes* [Mitchell 1997], probabilístico; *SVM* [Cortes and Vapnik 1995], de margem máxima com *kernels*; e *MLP* [Rumelhart et al. 1986], neural. Para a análise não supervisionada usou-se o

k-Means [MacQueen 1967]. Os comitês combinam modelos para reduzir variância [Dietterich 2000], sendo homogêneos (mesma família) ou heterogêneos (famílias distintas), com integração por votação ou por um meta-classificador treinado sobre as predições-base (*stacking*) [Wolpert 1992].

3. Trabalhos Relacionados

Os primeiros estudos sobre a classificação automática de movimentos pedagógicos investigaram transcrições de aulas reais de matemática e obtiveram resultados próximos aos níveis de concordância observados entre anotadores humanos [Suresh et al. 2019]. Posteriormente, a base *TalkMoves* consolidou esse esforço ao disponibilizar transcrições anotadas de aulas de matemática [Suresh et al. 2022]. A aplicação derivada dessa iniciativa levou a análise automatizada ao professor e reportou *F1-score* de aproximadamente 79% para a classificação dos movimentos docentes com modelos contextuais [Suresh et al. 2021]. Esse resultado constitui uma referência comparativa importante, embora tenha sido obtido a partir de uma formulação experimental diferente da adotada neste trabalho.

Estudos posteriores ampliaram as possibilidades de análise computacional do discurso em sala de aula. Entre essas iniciativas estão a classificação de perguntas de *funneling* e *focusing* [Alic et al. 2022], a investigação da relação entre movimentos dialógicos identificados automaticamente e escores de observação docente [Demszky and Hill 2023] e a extensão da análise de movimentos pedagógicos para contextos de tutoria [Cao et al. 2025]. Trabalhos mais recentes também avaliam o uso de *Large Language Models* na análise de diálogos educacionais [Long et al. 2024, Tran et al. 2024]. Embora esses modelos apresentem maior capacidade de representação contextual, seu emprego envolve desafios relacionados ao custo computacional e à transparência das decisões. No contexto brasileiro, também se observa a aplicação de técnicas de aprendizado de máquina a problemas educacionais, acompanhada de preocupações relacionadas ao viés e à justiça algorítmica [Oliveira et al. 2024].

Esses estudos indicam um espaço ainda pouco explorado: o desenvolvimento de *baselines* clássicos sobre a base *TalkMoves*, fundamentados em representações textuais esparsas, como o *TF-IDF*, e em estratégias de combinação de classificadores. Investigar esse espaço é relevante por permitir o estabelecimento de uma referência reproduzível e de baixo custo para a classificação dos movimentos pedagógicos, além de favorecer o uso da base em contextos com restrições de *hardware*. Esse é o recorte adotado neste trabalho.

4. Metodologia

A metodologia adota um desenho comparativo: a mesma base, com o mesmo pré-processamento, é submetida a diferentes famílias de algoritmos e a três camadas de combinação, isolando o efeito de cada etapa do *pipeline* (Figura 1). A escolha por algoritmos clássicos com representação esparsa (*TF-IDF*), em vez de *embeddings* contextuais profundos, foi deliberada: (a) define um *baseline* reproduzível e de baixo custo para a *TalkMoves*; (b) permite examinar a influência dos componentes latentes por meio da importância de atributos e dos valores SHAP, aspecto relevante em contextos educacionais; e (c) estabelece a referência necessária antes de migrar para arquiteturas mais pesadas.

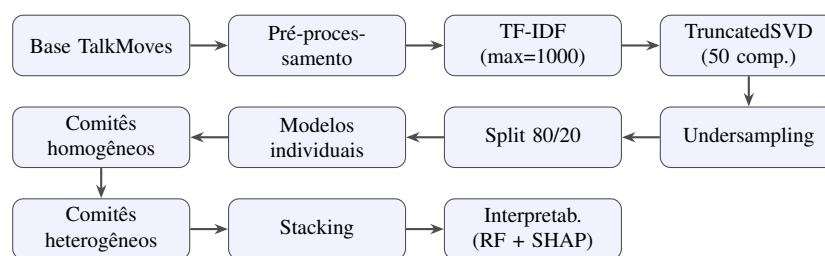


Figura 1. Visão geral do *pipeline* metodológico, do dado bruto à interpretação dos modelos.

4.1. Base de dados e pré-processamento

A *TalkMoves* [Suresh et al. 2022] reúne 567 transcrições de aulas de matemática do nível K–12, todas em inglês, totalizando 174.186 falas de professores e 59.874 de alunos (cerca de 234 mil turnos brutos, aproximadamente 1,8 milhão de palavras e 15.830 termos únicos). Cada turno tem sete atributos, entre eles o *Type of Move* (alvo da classificação), com categorias como *checking understanding*, *revoicing* e *press for reasoning*, definidas pelo modelo de *accountable talk* [Michaels et al. 2008]. Dois turnos ilustram a dificuldade: “*Can you explain how you got that answer?*” é *Checking Understanding*, enquanto “*Is what you said connected to what João mentioned earlier?*” é *Revoicing*; estruturas e vocabulário semelhantes, funções pedagógicas distintas, o que indica que a tarefa exige sensibilidade a sinais lexicais combinados.

Todo o pré-processamento e a vetorização foram realizados sobre o texto original em inglês, sem tradução das falas. O pré-processamento converteu as falas para minúsculas, removeu pontuação e marcou os falantes (T/S). Foram excluídos turnos com conteúdo entre colchetes ([End of Class]), apenas *timestamps*, vazios ou com *Speaker* ausente/número inválido, além dos rótulos *None* (não pedagógicos); restaram 91.891 sentenças válidas. A distinção entre as cerca de 234 mil instâncias brutas e as 91.891 trabalhadas é importante para a leitura dos resultados. A vetorização usou *TF-IDF* [Sparck Jones 1972, Manning et al. 2008] via *TfidfVectorizer* do *scikit-learn*, com *max_features=1000*: testes preliminares com 500, 1000 e 2000 *features* mostraram queda perceptível abaixo de 1000 e ganho apenas marginal acima, com maior custo. Aplicou-se *undersampling* sobre a classe majoritária, opção justificada pelo desbalanceamento residual e por preservar o sinal real de cada classe sem geração sintética (como faria o *SMOTE*), preferível em discurso natural anotado. A representação esparsa foi reduzida por *TruncatedSVD* [Deerwester et al. 1990, Halko et al. 2011], adequado a matrizes esparsas, com 50 componentes: a variância explicada acumulada satura acima desse valor, e 50 componentes mantêm o custo dos modelos não lineares (sobretudo o *SVM RBF*) viável para validar múltiplas configurações de comitês, atenuando ainda o risco de *overfitting*.

4.2. Divisão, métricas, algoritmos e comitês

Os dados foram particionados em 80% para treino e 20% para teste, com *random_state* fixo para reprodutibilidade; a divisão única foi adotada por compatibilidade com o custo do espaço de comitês avaliado, e a validação cruzada com teste de significância (Wilcoxon, McNemar) fica como trabalho futuro (Seção 7). As métricas foram acurácia, precisão, *recall* e *F1-score (macro)*, com desvio padrão sobre repetições

com diferentes sementes. O experimento contemplou cinco supervisionados (k -NN, AD, NB, *MLP*, *SVM*) e o k -Means, cobrindo paradigmas distintos; os hiperparâmetros foram ajustados por busca exploratória ($k \in \{3, 5, 7\}$; profundidades 5/10/15; camadas ocultas 50/100/150; *kernels* linear/polinomial/RBF). A combinação teve três camadas: cinco comitês *homogêneos* (5, 10 e 20 membros, votação majoritária); quatro comitês *heterogêneos* a partir dos três melhores grupos, $X=MLP$, $Y=SVM$ (*RBF*), $Z=k$ -NN (XY, XZ, YZ, XYZ, mesmos tamanhos); e um meta-classificador de Regressão Logística em *stacking* [Wolpert 1992] sobre os dois melhores heterogêneos. A nomenclatura X/Y/Z é mantida em todo o texto. Tudo foi implementado em *scikit-learn*. A análise opera em três níveis: descritivo (histogramas, *boxplots*, correlação), preditivo (métricas) e interpretativo, com importância de *Random Forest* (usado apenas como ferramenta de interpretação, não como modelo competitivo) e valores SHAP [Lundberg and Lee 2017].

5. Resultados e Discussão

5.1. Análise descritiva da base

Após remover o rótulo *None*, a distribuição entre os movimentos pedagógicos tornou-se mais equilibrada (Figura 2), embora persista desequilíbrio residual. Os 50 componentes do *TruncatedSVD* (Figura 4) apresentam perfis distintos: Comp_1 a Comp_3 são os mais informativos (alta variabilidade, *outliers*, compressão), enquanto Comp_4, Comp_5, Comp_8–Comp_10 concentram-se em torno de zero, com dominância de ruído. Optou-se por não remover *outliers*: em dados textuais vetorizados, eles podem representar falas semanticamente ricas, e sua exclusão comprometeria a representatividade. O mapa de correlação (Figura 3) mostra correlações fora da diagonal entre $-0,07$ e $+0,07$; trata-se de propriedade construtiva do *TruncatedSVD*, não de indicador de qualidade, garantindo apenas a ausência de redundância linear entre as 50 dimensões. A adequação à classificação é avaliada empiricamente a seguir.

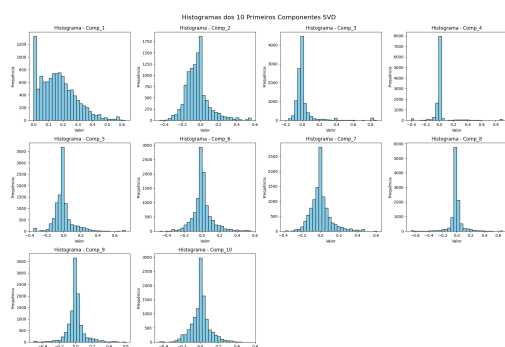


Figura 2. Frequência das classes após limpeza.

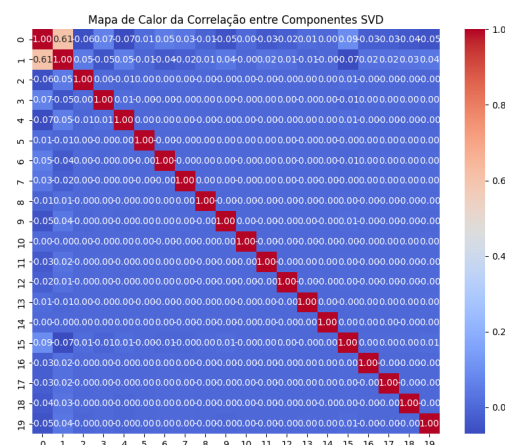


Figura 3. Correlação entre componentes SVD.

5.2. Classificadores individuais

O *SVM* com *kernel RBF* obteve o melhor desempenho (acurácia de 64,9%), seguido pelo *MLP* com 100 e 50 neurônios (63,3% e 63,2%), modelos também mais estáveis (menor desvio padrão) e com bom equilíbrio entre métricas (Tabela 1). Esses três modelos foram selecionados como base dos comitês heterogêneos.

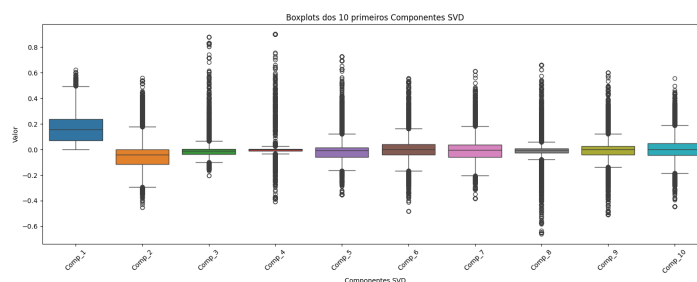


Figura 4. Boxplots dos 10 primeiros componentes SVD.

Tabela 1. Desempenho dos classificadores individuais (melhor modelo em destaque)

Modelo	Acurácia (%)	Precisão (%)	Recall (%)	F1-Score (%)	Desvio Padrão
SVM (kernel RBF)	64,90	65,37	64,90	64,65	0,0091
MLP (100 neurônios)	63,29	63,23	63,29	63,03	0,0128
MLP (50 neurônios)	63,20	62,91	63,20	62,80	0,0105
MLP (150 neurônios)	62,93	62,65	62,93	62,56	0,0091
SVM (linear)	62,10	62,57	62,10	61,66	0,0111
SVM (polinomial)	56,62	61,56	56,62	56,56	0,0118
<i>k</i> -NN (<i>k</i> =7)	54,19	57,20	54,19	55,04	0,0123
<i>k</i> -NN (<i>k</i> =3)	53,22	57,32	53,22	54,22	0,0154
<i>k</i> -NN (<i>k</i> =5)	52,58	56,05	52,58	53,55	0,0117
Árvore (profundidade 10)	52,85	54,01	52,85	53,12	0,0117
Árvore (profundidade 15)	51,38	50,98	51,38	51,04	0,0080
Naive Bayes	48,53	49,57	48,53	47,99	0,0099
Árvore (profundidade 5)	46,69	49,24	46,69	46,05	0,0097

5.3. Agrupamento não supervisionado (*k*-Means)

Avaliou-se o *k*-Means para verificar se os vetores reduzidos formam agrupamentos correspondentes às classes. O número de *clusters* foi explorado em torno do número de categorias pedagógicas presentes: privilegiou-se `n_clusters=7`, por corresponder à quantidade de movimentos do professor anotados na *TalkMoves*, testando-se os vizinhos 5 e 8 para verificar a sensibilidade. Mesmo variando inicialização e reinicializações, os índices permaneceram baixos (Tabela 2): melhor ARI de 0,083 e melhor *Silhouette* de 0,080. As categorias pedagógicas não formam aglomerados separáveis por distância euclidiana no espaço *TF-IDF+SVD*, o que é coerente com o dado: em discurso natural, turnos com funções distintas compartilham quase todo o vocabulário, e a informação que os distingue está mais na estrutura sintática e dialógica do que na distância vetorial. O resultado não invalida a supervisão; apenas evidencia o limite de métodos baseados em centroide para esse tipo de dado, sugerindo *clustering* por densidade ou sobre *embeddings* contextuais como trabalho futuro.

5.4. Comitês homogêneos e heterogêneos

Nos comitês homogêneos, o *MLP* liderou (acurácia $\approx 63,7\%$), seguido do *SVM* (62,7%); *k*-NN, AD e NB ficaram em 53,2%, 51,6% e 49,2%, em linha com seus equivalentes individuais. Para cada família, os tamanhos 5, 10 e 20 produziram métricas idênticas (Tabela 3 apresenta um valor por família). O motivo é a configuração: treinados sobre o mesmo conjunto, com a mesma representação e sem perturbação dos dados (sem

Tabela 2. Resultados do *k-Means* (melhores valores por métrica em destaque)

Configuração	n_clusters	init	n_init	ARI	Silhouette
C1	7	random	20	0,0826	0,0574
C2	7	random	10	0,0582	0,0447
C3	5	random	20	0,0469	0,0666
C4	8	k-means++	20	0,0170	0,0798
C5	8	random	20	0,0183	0,0721

bagging, *bootstrap* ou variação de *seed* nos modelos determinísticos), a votação majoritária colapsa para a decisão única do classificador. O achado é correto, mas mostra que, nessa configuração, o comitê homogêneo equivale a repetir o mesmo modelo; introduzir *bagging* e subespaços aleatórios é direção natural para versões futuras.

Tabela 3. Comitês homogêneos (métricas idênticas para 5/10/20 membros; melhor família em destaque)

Comitê (família)	Acurácia (%)	Precisão (%)	Recall (%)	F1-Score (%)
<i>k-NN</i>	53,19	54,65	53,19	53,10
<i>AD</i>	51,63	52,45	51,63	51,36
<i>NB</i>	49,23	50,36	49,23	48,85
<i>MLP</i>	63,68	63,42	63,68	63,34
<i>SVM</i>	62,73	63,14	62,73	62,54

Nos comitês heterogêneos (Tabela 4), os melhores foram o XY (*MLP* + *SVM RBF*) de tamanho 10 ou 20, com **65,1%** de acurácia e **64,7%** de *F1-score*, e o XYZ (*MLP* + *SVM RBF* + *k-NN*) de tamanho 20, com **65,0%** e **64,7%**. O ganho sobre o melhor individual (*SVM RBF*, 64,9%) é marginal e próximo do desvio padrão das execuções. A estabilidade entre os tamanhos 10 e 20 reforça que o diferencial não é o número de membros, mas a diversidade entre famílias: o XY combina modelos com fronteiras de decisão estruturalmente distintas.

Tabela 4. Comitês heterogêneos (melhores resultados em destaque)

Comitê	Tamanho	Acurácia (%)	Precisão (%)	Recall (%)	F1-Score (%)
XY	5	64,85	64,74	64,85	64,55
XY	10	65,06	64,91	65,07	64,73
XY	20	65,06	64,91	65,07	64,73
XZ	5	64,26	64,19	64,27	63,93
XZ	10	61,93	62,98	61,94	62,36
XZ	20	61,93	62,98	61,94	62,36
YZ	5	64,17	64,25	64,18	63,77
YZ	10	63,01	63,24	63,01	62,67
YZ	20	63,01	63,24	63,01	62,67
XYZ	5	64,94	64,84	64,94	64,58
XYZ	10	64,72	64,64	64,73	64,37
XYZ	20	65,03	64,97	65,04	64,66

5.5. Meta-classificador (stacking) e leitura educacional

Sobre as predições dos dois melhores heterogêneos, o *stacking* por Regressão Logística [Wolpert 1992] alcançou o melhor desempenho geral, com *F1-score* de 65,2% na configuração XYZ (Tabela 5). O ganho absoluto sobre o melhor individual é de cerca de 0,5 ponto percentual: modesto, mas obtido com a menor variância entre execuções, o que sustenta o uso da técnica onde a estabilidade importa tanto quanto a acurácia. Tratar esse ganho como estatisticamente significativo, porém, exigiria um teste pareado não conduzido nesta versão, indicado como prioridade futura.

Tabela 5. Meta-classificadores baseados nos comitês XY e XYZ (melhor configuração em destaque)

Comitê	Acurácia (%)	Precisão (%)	Recall (%)	F1-Score (%)
XY (<i>MLP</i> + <i>SVM RBF</i>)	65,28	65,05	65,28	64,95
XYZ (<i>MLP</i> + <i>SVM RBF</i> + <i>k-NN</i>)	65,46	65,27	65,46	65,17

Para a Informática na Educação, o número isolado importa menos do que sua leitura. Uma acurácia em torno de 65% sobre um conjunto multiclasse com classes lexicalmente sobrepostas está bem acima do acaso, mas abaixo dos $\approx 79\%$ de *F1-score* de modelos contextuais para movimentos do professor [Suresh et al. 2021]. A distância tem significado pedagógico: o *baseline* clássico não substitui o avaliador humano nem entrega um juízo final sobre a aula, e tomá-lo como nota de desempenho docente seria um equívoco; mas um modelo que acerta cerca de dois terços dos movimentos pode atuar como camada de triagem e apoio, sinalizando trechos para inspeção, estimando a frequência de movimentos, como a solicitação de raciocínio, ou compondo relatórios que sejam posteriormente revisados pelo professor. É nesse papel de instrumento de apoio, e não de árbitro, que reside a contribuição. A análise de importância apresentada na Seção 6 oferece uma primeira camada de inspeção das decisões dos modelos, embora sua tradução em sinais textuais diretamente compreensíveis por educadores ainda dependa do mapeamento dos componentes latentes para o vocabulário original. A confusão entre classes próximas, também evidenciada pelo baixo desempenho do *k-Means*, indica que os movimentos pedagogicamente mais sutis são os mais difíceis de automatizar com representações clássicas. Esse resultado reforça a necessidade de representações sensíveis ao contexto que preservem a possibilidade de inspeção das decisões dos modelos.

6. Análise da Importância das Variáveis

Em contextos educacionais, examinar a contribuição dos atributos é condição para que o resultado seja inspecionável por educadores [Türkmen 2025, Gunasekara and Saarela 2025]. Aplicaram-se duas técnicas: a importância nativa do *Random Forest* (média da redução de impureza, usado apenas como ferramenta de interpretação) e o método SHAP [Lundberg and Lee 2017]. No *Random Forest* (Figura 5), *comp_5* foi a mais importante, seguida de *comp_1* e *comp_4*, com distribuição relativamente uniforme, indício de captura de interações não lineares. Nos gráficos SHAP (Figura 6), o *MLP* concentra impacto em *comp_1*, *comp_2* e *comp_5*, com ampla dispersão típica de padrões não lineares; já o *SVM* mantém *comp_1* e *comp_2*, com *comp_3* ganhando destaque e dispersão mais concentrada, sugerindo relação mais linear. As diferenças refletem a natureza dos modelos e têm implicação direta na combinação: ao integrar modelos que privilegiam aspectos distintos do mesmo

espaço, os comitês heterogêneos obtêm visão complementar do dado, o que sustenta os ganhos modestos observados. Convém esclarecer o que esses componentes representam para um leitor não técnico: cada componente corresponde a uma dimensão latente que sintetiza padrões de coocorrência entre diferentes termos do vocabulário, e não a uma palavra ou expressão isolada. Assim, a maior importância atribuída a um componente indica que determinados padrões lexicais latentes contribuíram mais fortemente para as decisões do modelo. Entretanto, a análise realizada não permite identificar diretamente quais marcadores discursivos concretos estão associados a cada dimensão. A recuperação do vocabulário correspondente aos componentes permanece, portanto, como trabalho futuro, conforme discutido na Seção 8. Do ponto de vista educacional, essa análise constitui uma etapa inicial para examinar os fatores que influenciam as decisões dos modelos. Entretanto, sua tradução em explicações diretamente compreensíveis por professores depende do mapeamento futuro dos componentes latentes para termos e marcadores discursivos concretos.

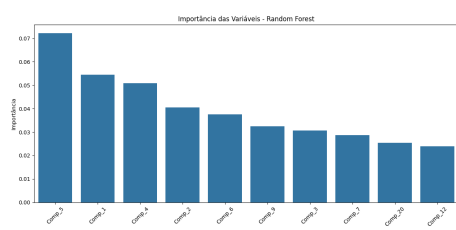


Figura 5. Importância (Random Forest).

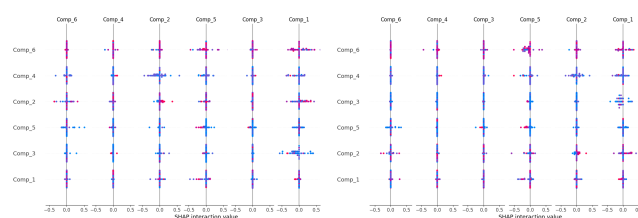


Figura 6. SHAP para os modelos MLP (esq.) e SVM (dir.).

7. Limitações e Ameaças à Validade

Limitações. A avaliação usou divisão única treino/teste (80/20), sensível à amostragem e sem teste de significância; o ganho do *stacking* sobre o melhor individual (~0,5 pp) está dentro da variabilidade entre execuções. A representação *TF-IDF + SVD* é clássica e não captura semântica contextual, o que se reflete no *k-Means* e na confusão entre classes próximas. O pré-processamento descartou os turnos *None*, e a base resultante mantém desbalanceamento residual, apenas atenuado pelo *undersampling*. Os comitês homogêneos foram montados sem perturbação dos dados, limitando seu ganho. Por fim, e do ponto de vista educacional decisivo, o estudo avalia a viabilidade técnica da classificação, mas não inclui validação de uso com professores nem evidência de que a ferramenta melhore a observação de aulas ou a formação docente.

Ameaças à validade. *Interna:* a ausência de validação cruzada e de teste pareado é a principal ameaça, pois parte das diferenças pode decorrer da partição específica; a busca de hiperparâmetros, não exaustiva, também pode favorecer modelos de forma não controlada. *Externa:* a *TalkMoves* cobre matemática K–12 em inglês, anotada por *accountable talk*; os resultados não se transferem automaticamente para outras disciplinas, níveis, idiomas ou esquemas de anotação, e a generalização ao contexto brasileiro exigiria dados em português. *De construção:* acertar o rótulo de *Type of Move* é *proxy* da qualidade discursiva, não sua medida direta, e a anotação carrega decisões e possíveis vieses dos anotadores originais. *De conclusão:* como as diferenças entre os melhores modelos são da ordem do desvio padrão e não foram testadas, conclusões sobre superioridade

devem ser tratadas como indicativas, e as métricas *macro* podem mascarar desempenho fraco em classes minoritárias, justamente as pedagogicamente mais relevantes.

8. Conclusão

Este estudo investigou em que medida o aprendizado de máquina clássico pode apoiar a identificação automática de movimentos pedagógicos no discurso de sala de aula, tomando a *TalkMoves* como objeto e tratando os algoritmos como meio para estudar um problema de avaliação da prática docente. Caracterizou-se a viabilidade da identificação automática (objetivo a), comparando famílias de algoritmos (b) e estratégias de combinação (c) sobre a mesma representação, analisando a estrutura dos dados (d), analisando a influência dos componentes latentes da representação textual (e) e discutindo as implicações para a avaliação e a formação docente (f). O *SVM RBF* e o *MLP* (100) lideraram entre os individuais (64,9% e 63,3%); a combinação *MLP* + *SVM* chegou a 65,1% de acurácia e 64,7% de *F1-score*, e o *stacking XYZ* a 65,5% e 65,2%, com menor variância. Em resposta direta à questão de pesquisa, os resultados indicam que técnicas clássicas apoiam a identificação de movimentos pedagógicos de forma parcial, porém útil: o acerto em torno de dois terços, somado à transparência do *pipeline*, sustenta o uso desses modelos como apoio interpretável à análise da prática docente, e não como substitutos do avaliador humano. Cumpre-se, com isso, o objetivo geral de investigar a viabilidade dessa classificação como subsídio à avaliação e à formação docente.

A contribuição para a Informática na Educação é dupla. Metodologicamente, documenta-se um *baseline* reproduzível e interpretável para a análise de práticas discursivas docentes na *TalkMoves*, com a interpretabilidade tratada como requisito, e não como acessório. Em termos aplicados, esse *baseline* é posicionado como instrumento auditável de apoio à observação de aulas e ao *feedback* ao professor, e não como árbitro da qualidade docente. O estudo evidencia, ainda, que os movimentos mais sutis são os mais difíceis de automatizar com representações clássicas, resultado relevante para a área. Persistem limitações que circunscrevem esses achados: a avaliação repousa em uma única partição treino/teste, sem teste de significância; a representação *TF-IDF* + *SVD* não capta semântica contextual; e o trabalho verifica a viabilidade técnica, sem validação de uso com professores (Seção 7). A agenda futura segue dessas limitações: (i) validação cruzada estratificada com teste de Wilcoxon pareado; (ii) comparação com *embeddings* contextuais, preservando a interpretabilidade; (iii) *F1-score* por classe, com atenção às minoritárias; (iv) *bagging* e subespaços aleatórios nos comitês; (v) um estudo de uso com professores, para verificar se a ferramenta apoia de fato a observação de aulas e a formação continuada; e (vi) projetar os componentes latentes do *TruncatedSVD* de volta ao vocabulário original, com o objetivo de identificar os termos e marcadores discursivos associados às dimensões mais influentes e tornar as explicações mais compreensíveis para educadores.

Considerações Éticas, Anonimização e Uso de IA

A pesquisa utiliza exclusivamente a base pública *TalkMoves* [Suresh et al. 2022], disponibilizada de forma anonimizada para fins de pesquisa. Não houve coleta direta de dados pelos autores, e não são apresentados elementos que permitam identificar professores, estudantes ou instituições participantes da base original. Eventuais identificadores presentes nos exemplos seguem o processo de anonimização adotado pelos responsáveis

pela base. Os resultados são propostos como apoio à observação de aulas e ao *feedback* formativo, e não como instrumento autônomo de julgamento da prática docente.

Declaração de Uso de Inteligência Artificial. Em conformidade com as diretrizes do SBIE 2026, declara-se que ferramentas de IA generativa foram utilizadas exclusivamente como apoio à escrita e à organização do manuscrito (revisão textual, aprimoramento da redação, padronização terminológica e estruturação de seções). A concepção da pesquisa, o desenho experimental, a execução dos experimentos, a coleta e o tratamento dos dados, a análise e a interpretação dos resultados e a redação das conclusões são de responsabilidade integral dos autores, que revisaram e validaram todo o conteúdo. Nenhum dado, resultado, métrica ou referência foi gerado por IA sem verificação humana.

Referências

- Alic, S., Demszky, D., Mancenido, Z., Liu, J., Hill, H., and Jurafsky, D. (2022). Computationally identifying funneling and focusing questions in classroom discourse. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 224–233. Association for Computational Linguistics (ACL).
- Almubarak, A., Alhalabi, W., Albidewi, I., and Alharbi, E. (2024). An analytical approach for an ai-based teacher performance evaluation system in saudi arabia's schools. *Discover Applied Sciences*, 6(406). Published online: 29 July 2024.
- Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification and Regression Trees*. Wadsworth, Belmont, CA.
- Cao, J., Suresh, A., Jacobs, J., Clevenger, C., Howard, A., Brown, C., Milne, B., Fischhaber, T., Sumner, T., and Martin, J. H. (2025). Enhancing talk moves analysis in mathematics tutoring through classroom teaching discourse. In *Proceedings of the 31st International Conference on Computational Linguistics (COLING 2025)*, pages 7710–7725, Abu Dhabi. Association for Computational Linguistics (ACL).
- Chanchal, A. and Zuolkernan, I. (2022). Exploring semi-supervised learning for audio-based automated classroom observations. In *Proceedings of the 19th International Conference on Cognition and Exploratory Learning in the Digital Age, CELDA 2022*, pages 241–250, Portugal. IADIS Press. Publisher Copyright: © 2022 Proceedings of the 19th International Conference on Cognition and Exploratory Learning in the Digital Age, CELDA 2022. All rights reserved.
- Cohen, D. K. (2011). *Teaching and Its Predicaments*. Harvard University Press, Cambridge, MA.
- Cortes, C. and Vapnik, V. (1995). Support vector networks. *Machine Learning*, 20(3):273–297.
- Cover, T. and Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27.
- Danielson, C. (2007). *The Framework for Teaching*. ASCD.
- Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

- Demszky, D. and Hill, H. (2023). The NCTE transcripts: A dataset of elementary math classroom transcripts. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 528–538, Toronto, Canada. Association for Computational Linguistics (ACL).
- Dietterich, T. G. (2000). Ensemble methods in machine learning. In *International Workshop on Multiple Classifier Systems*, pages 1–15.
- Grossman, P., Compton, C., Igra, D., Ronfeldt, M., Shahan, E., and Williamson, P. W. (2009). Teaching practice: A cross-professional perspective. *Teachers College Record*, 111(9):2055–2100.
- Gunasekara, S. and Saarela, M. (2025). Explainable ai in education: Techniques and qualitative assessment. *Applied Sciences*, 15(3):1239.
- Halko, N., Martinsson, P.-G., and Tropp, J. A. (2011). Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, 53(2):217–288.
- Hattie, J. (2009). Visible learning: A synthesis of over 800 meta-analyses relating to achievement. *Routledge*.
- Long, Y., Luo, H., and Zhang, Y. (2024). Evaluating large language models in analysing classroom dialogue. *npj Science of Learning*, 9(1):60.
- Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, pages 4765–4774.
- MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297.
- Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.
- Michaels, S., O’Connor, C., and Resnick, L. B. (2008). Deliberative discourse idealized and realized: Accountable talk in the classroom and in civic life. *Studies in Philosophy and Education*, 27(4):283–297.
- Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill.
- Ninin, M. O. G. a. (2009). A atividade de observação nas práticas de orientação a professores: uma perspectiva crítica. *DELTA: Documentação de Estudos em Lingüística Teórica e Aplicada*, 25(2):253–282.
- Oliveira, M., Cabral, L. d. S., and Mello, R. F. (2024). Análise de modelos de aprendizado de máquina para a predição do desempenho de alunos com enfoque na detecção de vies algorítmico. In *Anais do XXXV Simpósio Brasileiro de Informática na Educação (SBIE 2024)*, Rio de Janeiro. Sociedade Brasileira de Computação (SBC).
- Pianta, R. C., La Paro, K. M., and Hamre, B. K. (2008). *Classroom Assessment Scoring System (CLASS)*. Paul H. Brookes Publishing Co.
- Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, 323(6088):533–536.

- Shulman, L. S. (1987). Knowledge and teaching: Foundations of the new reform. *Harvard Educational Review*, 57(1):1–22.
- Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–21.
- Stallings, J. A. (1977). *The Stallings Classroom Observation System*. Stanford University Press.
- Suresh, A., Jacobs, J., Harty, C., Perkoff, M., Martin, J. H., and Sumner, T. (2022). The talkmoves dataset: K–12 mathematics lesson transcripts annotated for teacher and student discursive moves. In *Proceedings of the 13th Language Resources and Evaluation Conference (LREC)*, pages 4654–4662, Marseille. European Language Resources Association.
- Suresh, A., Jacobs, J., Lai, V., Tan, C., Ward, W., Martin, J. H., and Sumner, T. (2021). Using transformers to provide teachers with personalized feedback on their classroom discourse: The talkmoves application. In *Proceedings of the AAAI 2021 Spring Symposium on Artificial Intelligence for K-12 Education*. arXiv:2105.07949.
- Suresh, A., Sumner, T., Jacobs, J., Foland, B., and Ward, W. (2019). Automating analysis and feedback to improve mathematics teachers’ classroom discourse. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9721–9728.
- Tran, N., Pierce, B., Litman, D., Correnti, R., and Matsumura, L. C. (2024). Analyzing large language models for classroom discussion assessment. In *Proceedings of the 17th International Conference on Educational Data Mining (EDM 2024)*, pages 500–508. International Educational Data Mining Society.
- Türkmen, G. (2025). The review of studies on explainable artificial intelligence in educational research. *Journal of Educational Computing Research*.
- Wolpert, D. H. (1992). Stacked generalization. *Neural Networks*, 5(2):241–259.