Identificação de Características Relevantes para Construir Modelos Preditivos de Desempenhos de Alunos

Douglas H. Silva[1], Lucas M. Melo[2], Reginaldo G. Leão Jr.[2]

[1]Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG)
Departamento de Computação – Campus Divinópolis (DECOMDV)
Caixa Postal 400 — 35.503-822 — Divinópolis — MG — Brasil

[2]Instituto Federal de Minas Gerais (IFMG) – Campus Arcos – Arcos, MG – Brasil

douglassilva@cefetmg.br, lucas.morais.melo@educacao.mg.gov.br,
reginaldo.junior@ifmg.edu.br

Abstract. *Educational data mining (EDM) applies the process of data mining with machine learning algorithms in the educational field. This process is applied in educational institutions to assist managers in making decisions about school problems, such as truancy. However, EDM depends on the quality and quantity of available data to run the algorithms. In this context, this work analyzes the identification of relevant features defined by a feature selection algorithm, for designing EDM models with machine learning algorithms. The dataset used contains school and socioeconomic information on Portuguese high school students. The algorithm applied to select the features was Recursive Feature Elimination, and the machine learning algorithms applied were Decision Tree, Gradient Tree Boosting, and Support Vector Machines. Analysis of the results demonstrated that school and socioeconomic information are relevant to predicting student performance.*

Resumo. *A mineração de dados educacionais (MDE) aplica o processo de mineração de dados com algoritmos de aprendizagem de máquina no campo educacional. Esse processo é aplicado em instituições de ensino para auxiliar os gestores nas tomadas de decisões sobre problemas escolares, como a evasão escolar. Contudo, a MDE depende da qualidade e da quantidade de dados disponíveis para executar os algoritmos. Neste contexto, este trabalho analisa a identificação de características relevantes definidas por um algoritmo de seleção de características, para construir modelos de MDE com algoritmos de aprendizagem de máquina. O conjunto de dados utilizados contém informações escolares e socioeconômicas de alunos portugueses do ensino médio. O algoritmo aplicado para selecionar as características foi o Recursive Feature Elimination e os algoritmos de aprendizagem de máquina aplicados foram Decision Tree, Gradient Tree Boosting e Support Vector Machines. A análise dos resultados demonstrou que as informações escolares e socioeconômicas são relevantes para predizer os desempenhos dos alunos.*

1. Introdução

Nos últimos anos, o avanço do uso de tecnologias digitais da informação e comunicação, e o crescimento do acesso a recursos computacionais por instituições escolares, geraram um grande volume de dados educacionais. Deste modo, trabalhar com grande volume de dados é uma tarefa humanamente inviável. Por conseguinte, a aplicação de algoritmos de

aprendizagem de máquina em dados educacionais com técnicas de mineração de dados, objetivando variadas finalidades, é factível (Kampff 2010; F. D. Santos 2016; F. P. Santos 2017). Na literatura, alguns estudos aplicaram esses algoritmos para solucionar problemas envolvendo dados educacionais, com as seguintes finalidades: aprendizagem personalizada e aprendizagem adaptativa para alunos (Costa et al. 2021); sistemas de recomendação para *e-learning* (Khanal et al. 2020); predição do desempenho final de alunos (Romero et al. 2013); realidade aumentada ou virtual (Yufeia et al. 2020), e; mineração de dados educacionais (Lemay, Baek e Doleck 2021).

A mineração de dados é profundamente dependente da qualidade e da quantidade de dados disponíveis para executar os algoritmos. Uma das dificuldades estaria centrada em encontrar uma fonte segura para selecionar o maior número possível de dados relevantes e investigar um problema de interesse (Manhães e Cruz 2020). Neste contexto, o trabalho de Souza (2021) julgou que os dados mais relevantes para predizer o desempenho de alunos de duas escolas portuguesas, usando um conjunto de dados públicos, eram as notas e a quantidade de faltas dos alunos. Essa análise foi baseada na seleção de características do algoritmo de aprendizagem de máquina Árvore de Decisão (*Decision Tree*). A partir dessa análise, este trabalho questionou se a seleção de características aplicada no conjunto de dados público foi relevante para projetar modelos de outras abordagens de classificação e formulou o seguinte pressuposto: para cada abordagem de modelo de classificação, existe um subconjunto de atributos mais influente no conjunto de dados público, capaz de predizer os desempenhos dos alunos com mais exatidão.

O principal objetivo deste trabalho foi a verificação do pressuposto formulado, aplicando seleção de características e abordagens de aprendizagem de máquina supervisionada. A seleção de características aplicada foi o RFE (*Recursive Feature Elimination*), um método da abordagem *wrapper*. Já as abordagens de aprendizagem de máquina aplicadas foram *Decision Tree*, *Gradient Tree Boosting* e *Support Vector Machines*. A principal contribuição da metodologia proposta foi o reconhecimento da relação entre as abordagens de aprendizagem de cada modelo e a identificação de características relevantes para predizer desempenhos de alunos em um conjunto de dados.

A aplicação de algoritmos de aprendizagem de máquina em dados educacionais justifica-se pela possibilidade de sua ação preditiva em diferentes contextos escolares, dentre os quais podem ser citados a evasão escolar, um problema que atinge principalmente a etapa do ensino médio no Brasil, e o desempenho escolar dos alunos. Pois, tendo em mãos resultados provenientes de modelos preditivos, gestores escolares podem redirecionar ações e buscar alternativas em prol da minimização de possíveis problemas que afligem suas instituições de ensino (Alamri et al. 2021; Albreiki, Zaki e Alashwal 2021; Dabhade et al. 2021; Hussain e Khan 2021; Kabathova e Drlik 2021; Sekeroglu, Dimililer e Tuncal 2019; Vijayalakshmi e Venkatachalapathy 2019; Yakubu e Abubakar 2022).

O trabalho está estruturado da seguinte forma. Na seção 2, é apresentado o referencial teórico com os conceitos de seleção de características e aprendizagem de máquina supervisionada. A seção 3, apresenta a metodologia aplicada para identificar as características relevantes e criar cada modelo de classificação. Na seção 4, são apresentados os resultados da seleção de características e os desempenhos dos modelos de classificação. A seção 5, apresenta uma discussão das implicações dos resultados

obtidos em comparação com outro resultado da literatura. Por fim, a seção 6 apresenta as considerações finais e os trabalhos futuros.

2. Referencial Teórico

Esta seção apresenta os conceitos de seleção de características e aprendizagem de máquina supervisionada. Além disso, ela apresenta a aplicação dessas técnicas em conjuntos de dados escolares na literatura.

2.1. Seleção de Características em Dados Escolares

A seleção de características busca um subconjunto de características relevantes para aprimorar a predição de classes. Em função de um método de avaliação e um critério de parada, cada subconjunto de características é avaliado. A cada iteração de avaliação, as características redundantes e irrelevantes são removidas (Silva et al. 2022). Ademais, a seleção de características engloba três abordagens principais: filtro, *wrapper* e embutida.

A abordagem embutida é uma etapa do processamento de alguns algoritmos de classificação, por exemplo, a seleção de características para projetar árvores de decisão (Tan, Steinbach e V. Kumar 2005). Já a abordagem filtro usa um algoritmo sem vínculo com o classificador para medir o desempenho de cada subconjunto e escolher o melhor (Tan, Steinbach e V. Kumar 2005). Por fim, a abordagem *wrapper* faz a seleção de subconjuntos de características do conjunto de treinamento para treinar um algoritmo de classificação induzido, e medir o desempenho de cada subconjunto a fim de escolher o melhor (Freitas 2002). Na literatura, as três abordagens foram aplicadas para selecionar características em dados educacionais. O Quadro 1 lista as publicações e as técnicas que foram aplicadas para selecionar características em bases de dados escolares. Neste trabalho, foi aplicada a abordagem *wrapper* para selecionar características em dados escolares.

Quadro 1. Seleção de características na literatura

Publicações	**Técnicas de Seleção de Caraterísticas**
Ahmed, Al-Hamdani e Croock (2020)	Embutida
Ajibade, Ahmad e Zainal (2020)	Híbrida*
Abdullahi, Kenneth e Olalere (2021)	Filtro
Souza (2021)	Filtro
Hussain e Khan (2021)	Filtro
Nidhi, Majithia e Sharma (2021)	Filtro
K. Hengpraprohm, S. Hengpraprohm e Sudjitjoon (2022)	Filtro

* Aplicou as técnicas filtro e *wrapper*.

Fonte: Elaborado pelos autores

2.2. Seleção de Características em Dados Escolares

De acordo com Alamri et al. (2021), a aplicação do processo de mineração de dados com técnicas de aprendizagem de máquina no campo educacional é conhecida como mineração de dados educacionais. O termo mineração de dados refere-se à descoberta de conhecimento, tais como regras e padrões, em grandes volumes de dados (Baker, Isotani e Carvalho 2011). Já o termo aprendizado de máquina é referenciado como uma área da inteligência artificial. De acordo Baeza-Yates e Ribeiro-Neto (2011 como citado em Silva et al. 2022), o aprendizado de máquina busca desenvolver algoritmos que aprendem automaticamente a partir de conjuntos de dados. Na literatura, são apresentadas três abordagens de aprendizado de máquina: supervisionado, não supervisionado e semi-supervisionado.

Neste trabalho, foi aplicado o aprendizado de máquina supervisionado para classificar os desempenhos dos alunos. No aprendizado supervisionado, um conjunto de instâncias dividida em classes (categorias) é ofertado ao algoritmo. Essas instâncias são rotuladas com suas respectivas classes por humanos e formam um conjunto de dados. O conjunto de dados é usado para treinar e testar os modelos de aprendizagem de máquina. Por meio de um subconjunto de treinamento, o algoritmo modela uma função de classificação para predizer instâncias não rotuladas (Baeza-Yates e Ribeiro-Neto 2011).

A predição de desempenhos de alunos na literatura aplicou diferentes técnicas de aprendizagem supervisionada, tais como: ANN (*Artificial Neural Network*), DT (*Decision Tree*), DNN (*Deep neural network*), ECOC (*Error-Correcting Output Codes*), KNN (*K-Nearest Neighbors*), LR (*Linear Regression*), MLP (*Multilayer Perceptron*), NB (*Naïve Bayes*), RF (*Random Forest*) e SVM (*Support Vector Machines*). O Quadro 2 lista as publicações e as técnicas que foram aplicadas para classificar o desempenho dos alunos.

Quadro 2. Técnicas aplicadas para classificar o desempenho dos alunos na literatura

Publicacões	**Técnicas de Classificacão**
Vijayalakshmi e Venkatachalapathy (2019)	DT, DNN, KNN, NB, RF, SVM
Ajibade, Ahmad e Zainal (2020)	DT, KNN, NB
Ahmed, Al-Hamdani e Croock (2020)	DT
Abdullahi, Kenneth e Olalere (2021)	DT, ECOC, KNN
Souza (2021)	DT, NB, RF, SVM
Hussain e Khan (2021)	DT, KNN
Kour, R. Kumar e Gupta (2021)	ANN, LR
Nidhi, Majithia e Sharma (2021)	DT, LR, MLP, NB, RF
Chettaoui, Atia e Bouhlel (2022)	DT, KNN, NB, RF, SVM
K. Hengpraprohm, S. Hengpraprohm e Sudjitjoon (2022)	ANN, KNN, LR, RF

Fonte: Elaborado pelos autores

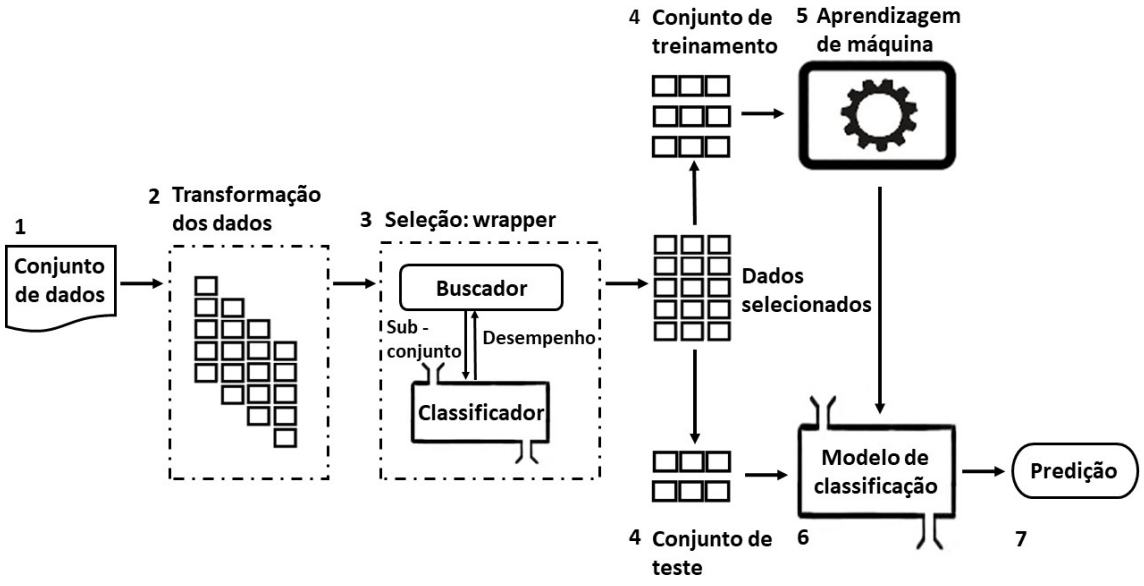
Neste Trabalho, foram aplicadas as técnicas de aprendizagem de máquina supervisionada DT, GTB (*Gradient Tree Boosting*) e SVM:

- A técnica DT cria regras de classificação em função do conjunto de treinamento. Essas regras representam graficamente uma árvore de decisão invertida. Cada tronco, da raiz à folha, é uma regra de classificação (Baeza-Yates e Ribeiro-Neto 2011);
- O GTB cria e combina modelos de árvores de decisão em função da confiança de cada um. Em um processo iterativo, os novos modelos são criados considerando os desempenhos dos modelos das iterações anteriores. Cada novo modelo visa acertar a predição de instâncias classificadas erroneamente, por meio da atribuição de pesos maiores a tais exemplos (Witten, Frank e Hall 2011);
- O SVM é uma técnica de classificação binária de vetores. Na fase de treinamento, o objetivo é maximizar a margem entre instâncias de duas classes distintas (vetores) e um hiperplano de decisão (H) no espaço n-dimensional. O valor de 'n' é a quantidade de características do conjunto de dados (Baeza-Yates e Ribeiro-Neto 2011).

3. Metodologia

Esta seção apresenta as técnicas utilizadas para predizer os desempenhos dos alunos. A Figura 1 apresenta a metodologia proposta, que foi desenvolvida utilizando a linguagem de programação *Python*. As seguintes bibliotecas em *Python* foram aplicadas: *sklearn* (Buitinck et al. 2013) e *xgboost* (Chen e Guestrin 2016).

Figura 1. Metodologia para criar o modelo de classificação



Fonte: Elaborado pelos autores

Em seguida são apresentados os objetos e métodos usados para projetar os modelos de classificação:

1. Conjunto de dados: o conjunto contém instâncias com o desempenho de cada aluno nas disciplinas de Português e Matemática;
2. Transformação dos dados: faz a normalização e a conversão dos dados para os algoritmos de aprendizagem de máquinas na etapa de mineração de dados.
3. Seleção de característica: faz a seleção de um conjunto de atributos que aprimora a distinção entre as instâncias de cada classe;
4. Divisão de treino e teste: reparte aleatoriamente o conjunto de vetores de instâncias em dois subconjuntos, treino e teste. O subconjunto de teste é aplicado para testar o modelo. O subconjunto de treino alimenta a validação cruzada estratificada com 10 pastas para cada modelo. Em cada pasta, subconjuntos de instâncias são selecionadas aleatoriamente por meio do subconjunto de treino. Desses, k-1 subconjuntos de instâncias são usados para treinar um modelo e, um subconjunto é aplicado para validar o mesmo;
5. Mineração de dados educacionais: faz a aprendizagem de máquina. Treina um modelo de classificação para predizer o desempenho de um aluno;
6. Modelo de classificação: é o produto final da aprendizagem de máquina em função do conjunto de treinamento. Os exemplos sem os desempenhos dos alunos (classe) do conjunto de teste são usados para testar os modelos;
7. Predição: é a classificação do desempenho do aluno feito pelo modelo, em função de uma instância do conjunto de testes. Essa predição é comparada com a classe verdadeira da instância para calcular o desempenho do modelo.

3.1. Conjunto de dados

Este trabalho aplicou um conjunto de dados público, denominado *Student Performance Data Set*, disponível no repositório *UCI Machine Learning* (Dua e Graff 2017). Esse conjunto de dados foi aplicado em Souza (2021) para prever o desempenho dos alunos. Ele contém informações de alunos do ensino médio de duas escolas portuguesas, distribuídas em 33 atributos: notas, faltas, e características demográficas e sociais. Além disso, ele é composto por dois subconjuntos de dados, relativos ao desempenho desses em Matemática e Língua Portuguesa. Assim como em Souza (2021), este trabalho junta os dois subconjuntos de dados das duas disciplinas. Logo, o conjunto contém mil e quarenta e quatro linhas (instâncias) e trinta e três colunas (atributos).

Os atributos do conjunto de dados, as descrições destes e seus valores correspondentes são apresentados a seguir, conforme proposto em Souza (2021):

1. *School*: escola do aluno (binário: ‘GP’ - Gabriel Pereira ou ‘MS’ - Mousinho da Silveira);
2. *Sex*: gênero do aluno (binário: ‘F’ - feminino ou ‘M’ - masculino);
3. *Age*: idade do aluno (numérico: de 15 a 22);
4. *Address*: tipo de endereço residencial do aluno (binário: ‘U’ - urbano ou ‘R’ - rural);
5. *Famsize*: tamanho da família (binário: ‘LE3’ - menor ou igual a 3 ou ‘GT3’ - maior que 3);
6. *Pstatus*: status de coabitação dos pais (binário: ‘T’ - morando junto ou ‘A’ - à parte);
7. *Medu*: escolaridade da mãe (numérico: 0 - nenhum, 1 - ensino fundamental - (4ª série), 2 - 5ª a 9ª série, 3 - ensino médio ou 4 - ensino superior);

8. *Fedu*: escolaridade do pai (numérico: 0 - nenhum , 1 - ensino fundamental - (4ª série), 2 - 5ª a 9ª série, 3 - ensino médio ou 4 - ensino superior);
9. *Mjob*: trabalho da mãe (nominal: 'professor', 'saude' (relacionado à saúde), 'serviços' (serviços civis - por exemplo, administrativo ou policial), 'em casa' ou 'outro');
10. *Fjob*: trabalho do pai (nominal: 'professor', 'saude' (relacionado à saúde), 'serviços' (serviços civis - por exemplo, administrativo ou policial), 'em casa' ou 'outro');
11. *Reason*: motivo para escolher esta escola (nominal: 'casa' (mora perto da escola), 'reputação' da escola, 'curso' de preferência ou 'outro');
12. *Guardian*: tutor do aluno (nominal: 'mãe', 'pai' ou 'outro');
13. *Traveltime*: intervalo gasto de casa para a escola (numérico: 1 - menor que 15 min., 2 - 15 a 30 min., 3 - 30 min. a 1 hora, ou 4 - maior que 1 hora);
14. *Studytime*: intervalo gasto de estudo semanal (numérico: 1 - menor que 2 horas, 2 - 2 a 5 horas, 3 - 5 a 10 horas ou 4 - maior que 10 horas);
15. *Failures*: número de reprovações anteriores nas aulas (numérico: n, se 1 <= n < 3, senão 4);
16. *Schoolup*: suporte educacional extra (binário: sim ou não);
17. *Famsup*: suporte educacional familiar (binário: sim ou não);
18. *Paid*: aulas extras pagas dentro da disciplina (matemática ou português) (binário: sim ou não);
19. *Activities*: atividades extracurriculares (binário: sim ou não);
20. *Nursery*: cursou creche (binário: sim ou não);
21. *Higher*: deseja cursar o ensino superior (binário: sim ou não);
22. *Internet*: acesso à internet em casa (binário: sim ou não);
23. *Romantic*: com um relacionamento romântico (binário: sim ou não);
24. *Famrel*: qualidade das relações familiares (numérico: de 1 - muito ruim a 5 - excelente);
25. *Freetime*: intervalo de livre depois da escola (numérico: de 1 - muito baixo a 5 - muito alto);
26. *Gooout*: saindo com os amigos (numérico : de 1 - muito baixo a 5 - muito alto);
27. *Dalc*: consumo de álcool no dia de trabalho (numérico: de 1 - muito baixo a 5 - muito alto);
28. *Walc*: consumo de álcool no fim de semana (numérico: de 1 - muito baixo a 5 - muito alto);
29. *Health*: estado de saúde atual (numérico: de 1 - muito ruim a 5 - muito bom);
30. *Absences*: número de faltas na escola (numérico: de 0 a 93);
31. *G1*: nota do primeiro período relacionada à disciplina de Matemática ou Português (numérico: de 0 a 20);
32. *G2*: nota do segundo período relacionada à disciplina de Matemática ou Português (numérico: de 0 a 20);
33. *G3*: nota final relacionada à disciplina de Matemática ou Português (numérico: de 0 a 20 - alvo de sáıda).

3.2. Transformação dos dados

A transformação de dados modifica ou converte os valores dos atributos (Silva et al. 2022). Essa etapa prepara os dados para as entradas dos algoritmos e pode melhorar a

aprendizagem de máquina. Assim como em Souza (2021), este trabalho transformou o atributo numérico G3 para níveis de classificação. Logo, cada faixa de notas recebeu uma classe: notas de 0 a 3 recebeu 'D'; 4 a 10 recebeu 'C'; 11 a 15 recebeu 'B'; e de 16 a 20 recebeu 'A'. Além disso, transformou os dados nominais para numéricos aplicando a classe *LabelEncoder* do pacote *sklearn.preprocessing*.

3.3. Algoritmo de seleção de características

A técnica *wrapper* aplicada para selecionar características foi o algoritmo da classe RFE (*Recursive Feature Elimination*) do pacote *sklearn.feature_selection*. A classe RFE aplica um estimador externo para atribuir pesos a recursos, com o objetivo de selecionar recursivamente um subconjunto de recursos do conjunto de dados cada vez menor, em função de um critério de parada. Com exceção dos parâmetros *estimator* e *n_features_to_select* do RFE, os valores padrões das demais entradas não foram alterados. O *estimator* recebe instâncias dos algoritmos de aprendizagem de máquina e o parâmetro *n_features_to_select* recebe um valor inteiro correspondente a quantidade de características a serem selecionadas (Guyon et al. 2002). Na prática, o algoritmo seleciona os atributos do conjunto de dados transformado.

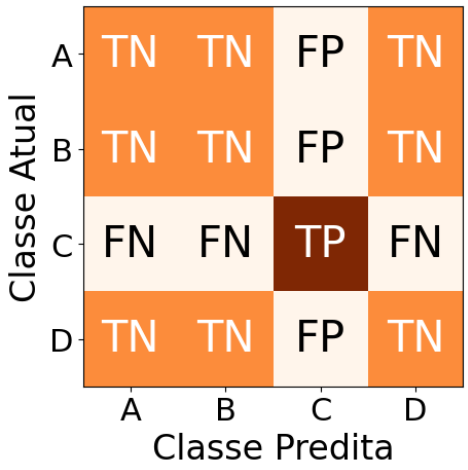
3.4. Algoritmos de aprendizado de máquina supervisionado

Os algoritmos aplicados para treinar os modelos de classificação e classificar amostras de teste foram: DTC (*Decision Tree Classifier*) do pacote *sklearn.tree*, uma algoritmo de DT; SVC (*Support Vector Classification*) do pacote *sklearn.svm*, um algoritmo de SVM; *XGB Classifier* (*XGB - Extreme Gradient Boosting*) do pacote *xgboost*, um algoritmo de GTB. Na fase de treinamento, os algoritmos recebem as amostras compiladas pelo algoritmo de seleção de características com as respectivas classes de desempenho acadêmico dos alunos. Já na fase de teste, as amostras compiladas pelo método de seleção de características sem as classes são submetidas aos algoritmos para predizer uma classe. A classe predita é comparada com a real, para medir o desempenho do modelo de classificação.

3.5. Avaliação de desempenho da tarefa de classificação

Os algoritmos aplicados para estimar os desempenhos dos modelos de classificação são do módulo *sklearn.metrics*. As macro-médias da precisão (PR), revocação (RE) e pontuação F1 foram calculadas em função dos acertos e erros da matriz de confusão. Os acertos são referenciados por verdadeiros positivos (TP) e verdadeiros negativos (TN). Já os erros são referenciados por falsos positivos (FP) e falsos negativos (FN). Um exemplo de matriz de confusão multi-classe para classe 'C' é exposta na Figura 2.

Figura 2. Matriz de confusão



Fonte: Elaborado pelos autores

As macro-médias estimam os desempenhos médios das métricas em função das classes, para cada modelo. A PR determina a probabilidade das classes predita pelos modelos está correta. Já a RE determina a probabilidade dos modelos predizer instâncias de cada classe corretamente. Por fim, a pontuação F1 combina e equilibra a importância entre precisão e revocação (Baeza-Yates e Ribeiro-Neto 2011). As macro-médias da PR, RE e pontuação F1 são definidas nas equações 1, 2 e 3, respectivamente. Ademais, os desempenhos medidos na validação cruzada são as médias das macro-médias das métricas usadas.

$$PR_{(Macro)} = \frac{1}{C}\sum_{i=1}^{C} \frac{TP}{(TP + FP)} \quad (1)$$

$$RE_{(Macro)} = \frac{1}{C}\sum_{i=1}^{C} \frac{TP}{(TP + FN)} \quad (2)$$

$$F1_{(Macro)} = \frac{1}{C}\sum_{i=1}^{C} \frac{2TP}{(2TP + FP + FN)} \quad (3)$$

A validação cruzada aplicada foi a estratificada. Nessa validação, o conjunto de dados é dividido em ‘k’ subconjuntos (pastas) e mantém a mesma distribuição de classes. Esses subconjuntos são aplicados como conjunto de treino ou teste, para validar a capacidade de generalização de cada modelo. O algoritmo *StratifiedKFold* do módulo *sklearn.model_selection* executou essa validação com 10 pastas.

4. Resultados e Análise

Nesta seção, são apresentados os resultados da avaliação experimental com o conjunto de dados públicos (Souza 2021) e após a seleção de características. Os resultados foram comparados em função da pontuação macro F1, visando verificar se a seleção de característica beneficiou ou prejudicou a capacidade de generalização dos modelos de classificação construídos. A Tabela 1 apresenta os resultados médios da validação cruzada com 10 pastas. O objetivo principal foi o de mostrar os resultados gerados no conjunto de dados públicos e os experimentos contendo o algoritmo RFE. As médias mais altas para cada métrica estão em negrito.

Tabela 1. Desempenho médio dos modelos de classificação - fase de treinamento

Modelo	Usou RFE	Macro PR	Macro RE	Macro F1
DTC	Não	0,7786 (±0,0793)	0,7513 (±0,0615)	0,7538 (±0,0528)
DTC	Sim	0,8160 (±0,0425)	0,7731 (±0,0513)	0,7838 (±0,0288)
SVC	Não	0,8108 (±0,0397)	0,7940 (±0,0635)	0,7958 (±0,0482)
SVC	Sim	**0,8484 (±0,0740)**	**0,8341 (±0,0671)**	**0,8350 (±0,0629)**
XGB	Não	0,8402 (±0,0634)	0,8110 (±0,0524)	0,8128 (±0,0418)
XGB	Sim	0,8384 (±0,0573)	0,8232 (±0,0490)	0,8223 (±0.0421)

Fonte: Elaborado pelos autores

A seleção de características com o algoritmo RFE foi benéfica para construir cada modelo de classificação, pois eles apresentaram médias maiores de desempenhos após o uso desse procedimento. O modelo que apresentou a maior média na pontuação macro F1 na fase de treinamento foi o modelo SVC com seleção de características. Além disso,os experimentos com o algoritmo RFE mostraram que cada método de aprendizagem de máquina testado, requer um conjunto de atributos específico, em um espaço menor de características, para manter um modelo de classificação com capacidade de generalização. A melhor seleção de características para cada algoritmo referenciado no parâmetro *estimator* da classe RFE e a quantidade de atributos a serem selecionados, são apresentados na Tabela 2. Ademais, uma seleção de características menor, em conjunto com um desempenho maior ou igual de uma outra seleção maior, foi o critério utilizado para definir o melhor conjunto de atributos para um modelo.

Tabela 2. Seleção de Características

Modelo	Qtd. de Atributos Selecionados	Atributos selecionados
DTC	3	*{'absences', 'g1', 'g2'}*
SVC	14	*{'school', 'age', 'address', 'famsize', 'studytime', 'failures', 'famsup', 'paid', 'nursery', 'higher', 'romantic', 'dalc', 'absences', 'g2'}*
XGB	26	*{'school', 'age', 'address', 'famsize', 'pstatus', 'fedu', 'mjob', 'fjob', 'reason', 'guardian', 'traveltime', 'studytime', 'failures', 'schoolsup', 'paid', 'activities', 'nursery', 'higher', 'internet', 'famrel', 'freetime', 'goout', 'dalc', 'absences', 'g1', 'g2'}*

Fonte: Elaborado pelos autores

Na fase de teste, os modelos DTC e SVC mantiveram o mesmo padrão de comportamento com e sem o algoritmo RFE, ou seja, a seleção de características beneficiou a construção dos modelos no conjunto de testes. Contudo, o modelo XGB com o algoritmo RFE não superou o de mesma técnica sem o RFE. Uma justificativa para esse resultado é a baixa representatividade de exemplos por classe. A coleta de mais exemplos por classe e o acréscimo desses no conjunto de dados poderia influenciar a seleção de características para o XGB, e reproduzir o desempenho obtido com o conjunto de treino, na fase de teste. A Tabela 3 apresenta os desempenhos dos modelos construídos a partir da seleção de características na fase de teste. O desempenho mais alto em cada métrica está em negrito.

Tabela 3. Desempenho de modelos de classificação - fase de teste

Modelo	Usou RFE	Macro PR	Macro RE	Macro F1
DTC	Não	0,7133	0,7152	0,7136
DTC	Sim	0,7888	0,7610	0.7681
SVC	Não	0,8547	0,7591	0,7951
SVC	Sim	**0,8906**	0,7999	0,8365
XGB	Não	0,8548	**0,8379**	**0,8454**
XGB	Sim	0,8472	0,8265	0,8359

Fonte: Elaborado pelos autores

5. Discussão

Em Souza (2021), a autora considerou que os atributos do conjunto de dados público referentes às notas (*g1* e *g2*) e a quantidade de faltas dos alunos (*absences*) foram mais influentes para predizer os desempenhos dos alunos, do que as características demográficas. Essa inferência foi baseada na análise do gráfico da árvore de decisão, que permitiu visualizar os atributos selecionados para construir o modelo, segundo o cálculo de entropia. Também neste trabalho, os atributos *g1*, *g2* e *absences* foram os mais relevantes no modelo de árvore de decisão DTC.

Contudo, os modelos de classificação são construídos de maneiras distintas, consoante à abordagem de aprendizagem. Assim, a construção de uma árvore de decisão, um modelo baseado em regras, aplica um certo subconjunto de atributos relevantes. Já a construção do modelo XGB, um comitê de classificadores, aplica um outro subconjunto de atributos relevantes. O mesmo ocorre na construção do SVM, um modelo baseado em funções. Assim sendo, os resultados dos experimentos confirmaram que o pressuposto de pesquisa é verdadeiro; para cada abordagem de modelo de classificação construído em função da seleção de características, existe um subconjunto de atributos mais influente, capaz de predizer os desempenhos dos alunos com mais exatidão. Esse subconjunto pode conter informações sobre notas, faltas, características demográficas ou sociais de alunos.

Portanto, os registros com informações sobre notas, faltas, características demográficas e sociais de alunos são relevantes para construir modelos de predição de desempenhos de alunos. Porém, a identificação das características mais relevantes para predizer os desempenhos dos alunos depende da abordagem de aprendizagem de cada modelo.

6. Considerações Finais

Este trabalho demonstrou que a identificação de características relevantes para predizer desempenhos de alunos em um conjunto de dados de classificação, depende da abordagem de aprendizagem de cada modelo. Essa constatação foi obtida após a verificação do pressuposto formulado, aplicando seleção de características RFE e os algoritmos de aprendizagem de máquina supervisionada: DTC, SVC e XGB.

A seleção de características com o algoritmo RFE demonstrou ser eficaz na identificação de características relevantes para construir modelos de classificação. Na fase de treinamento, as médias da pontuação macro F1 dos modelos aumentaram após o uso do algoritmo RFE. A melhor seleção de características para o algoritmo DTC foi com três características. O SVC alcançou os melhores desempenhos com quatorze características selecionadas. Por fim, o XGB alcançou o melhor desempenho com vinte e seis características selecionadas.

O modelo que alcançou a maior média na fase de treinamento foi o SVC com seleção de características, que obteve 0,8350 na pontuação macro F1. Na fase de teste, o modelo com a maior pontuação macro F1 foi XGB sem a seleção RFE, que alcançou 0,8454. Em razão da baixa representatividade de exemplos por classe na fase de teste, foi plausível considerar um empate entre o modelo XGB sem seleção RFE, com os modelos SVC e XGB que usaram a seleção. Os modelos XGB e SVC com seleção RFE alcançaram, respectivamente na pontuação macro F1, 0,8359 e 0,8365.

Esses resultados demonstraram que os registros contendo notas, faltas, características demográficas e sociais dos alunos de duas escolas de Portugal são

relevantes para predizer os desempenhos destes. Além disso, esses registros indicam novos estudos com registros escolares de alunos brasileiros, uma lacuna não investigada neste trabalho e pouco explorada no Brasil, devido à coleta de dados socioeconômicos de alunos ser complexa (Manhães e Cruz 2020).

Em trabalhos futuros será aplicada uma abordagem filtro com a medida estatística chi-quadrado, para medir a dependência entre características e classes. Também, serão aplicados algoritmos de aprendizagem de máquina supervisionada, variando os parâmetros de entrada, para aumentar os desempenhos de classificação destes. Por fim, será aplicada a metodologia proposta com dados da educação brasileira.

Referencias

Abdullahi, Wokili, Mary Ogbuka Kenneth e Morufu Olalere (2021). "Student performance prediction using a cascaded bi-level feature selection approach." Em: *Journal of Computer Science Research* 3.3. GS Search, pp. 16–28.

Ahmed, Saja Taha, Rafah Al-Hamdani e Muayad Sadik Croock (2020). "Developed third iterative dichotomizer based on feature decisive values for educational data mining". Em: *Indonesian Journal of Electrical Engineering and Computer Science* 18.1. GS Search, pp. 209–217.

Ajibade, Samuel-Soma M., Nor Bahiah Binti Ahmad e Anazida Zainal (2020). "Analysis of metaheuristics feature selection algorithm for classification". Em: *Hybrid Intelligent Systems - 20th International Conference on Hybrid Intelligent Systems (HIS 2020), Virtual Event, India, December 14-16, 2020*. Ed. by Ajith Abraham, Thomas Hanne, Oscar Castillo, Niketa Gandhi, Tatiane Nogueira Rios e Tzung- Pei Hong. Vol. 1375. Advances in Intelligent Systems and Computing. Springer, pp. 214–222. URL: https://link.springer.com/chapter/10. 1007/978-3-030-73050-5_21.

Alamri, Leena H., Ranim S. Almuslim, Mona S. Alotibi, Dana K. Alkadi, Irfan Ullah Khan e Nida Aslam (2021). "Predicting student academic performance using support vector machine and random forest". Em: *Proceedings of the 2020 3rd International Conference on Education Technology Management*. ICETM '20. GS Search. London, United Kingdom: Association for Computing Machinery, pp. 100–107. ISBN: 9781450388757. DOI: 10.1145/3446590.3446607.

Albreiki, Balqis, Nazar Zaki e Hany Alashwal (2021). "A systematic literature review of student' performance prediction using machine learning techniques". Em: *Education Sciences* 11.9. GS Search, pp. 1–27.

Baeza-Yates, Ricardo e Berthier Ribeiro-Neto (2011). M*odern information retrieval: the concepts and technology behind search*. 2ª ed. GS Search. Boston, MA: Addison-Wesley Publishing Company.

Baker, Ryan, Seiji Isotani e Adriana Carvalho (2011). "Mineração de dados educacionais: oportunidades para o Brasil". *Revista Brasileira de Informática na Educação* 19.02. GS Search, p. 00. ISSN: 1414-5685.

Buitinck, Lars, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt e Gaël Varoquaux (2013). "API design for machine learning software: experiences from the scikit-learn project". Em: *CoRR* abs/1309.0238. arXiv: 1309.0238. URL: http://arxiv.org/abs/ 1309.0238.

Chen, Tianqi e Carlos Guestrin (2016). "XGBoost: A Scalable Tree Boosting System". Em: *CoRR* abs/1603.02754. arXiv: 1603.02754. URL: http://arxiv.org/abs/1603.02754.

Chettaoui, Neila, Ayman Atia e Med Salim Bouhlel (2022). "Predicting students performance using eye-gaze features in an embodied Learning Environment". Em: *IEEE Global Engineering Education Conference, EDUCON 2022, Tunis, Tunisia, March 28-31, 2022*. Ed. por Ilhem Kallel, Habib M. Kammoun e Lobna Hsairi. IEEE, pp. 704–711. DOI: 10.1109/EDUCON52537.2022.9766783.

Costa, Rebeca Soler, Qing Tan, Frédérique Pivot, Xiaokun Zhang e Harris Wang (2021). "Personalized and adaptive learning: educational practice and technological impact". Em: *Texto Livre* 14.3. GS Search, e33445. ISSN: 1983-3652.

Dabhade, Pranav, Ravina Agarwal, K.P. Alameen, A.T. Fathima, R. Sridharan e G. Gopakumar (2021). "Educational data mining for predicting students' academic performance using machine learning algorithms". Em: *Materials Today: Proceedings* 47. GS Search, pp. 5260–5267. ISSN: 2214-7853. DOI: https://doi.org/10.1016/j.matpr.2021.05.646.

Dua, Dheeru e Casey Graff (2017). *UCI Machine learning repository*. URL: http://archive.ics.uci.edu/ml.

Freitas, Alex A (2002). *Data mining and knowledge discovery with evolutionary algorithms*. GS Search. Springer Science & Business Media.

Guyon, Isabelle, Jason Weston, Stephen Barnhill e Vladimir Vapnik (2002). "Gene selection for cancer classification using support vector machines". Em: *Machine learning* 46. GS Search, pp. 389–422. URL: https://link.springer.com/article/10.1023/A:1012487302797.

Hengpraprohm, Kairung, Supoj Hengpraprohm e Wannee Sudjitjoon (2022). "A study of factors affecting learning efficiency on higher education student performance evaluation dataset using feature selection techniques". Em: *Information Technology Journal* 18.2, pp. 34–43. URL: https://ph01.tci-thaijo.org/index.php/IT_Journal/article/view/251051.

Hussain, Shah e Muhammad Qasim Khan (2021). "Student-performulator: predicting students' academic performance at secondary and intermediate level using machine learning". Em: *Annals of data science* 10.3. GS Search, pp. 637–655. URL: https://link.springer.com/article/10.1007/s40745-021-00341-0.

Kabathova, Janka e Martin Drlik (2021). "Towards predicting student's dropout in university courses using different machine learning techniques". Em: *Applied Sciences* 11.7. GS Search. ISSN: 2076-3417. DOI: 10.3390/app11073130. URL: https://www.mdpi.com/2076-3417/11/7/3130.

Kampff, Adriana Justin Cerveira (mar. de 2010). "Mineração de Dados Educacionais para Geração de Alertas em Ambientes Virtuais de Aprendizagem como Apoio à Prática Docente". Em: *Informática na educação: teoria & prática* 12.2. DOI: 10.22456/1982-1654.12310. URL: https://seer.ufrgs.br/index.php/InfEducTeoriaPratica/article/view/12310.

Khanal, Shristi Shakya, PWC Prasad, Abeer Alsadoon e Angelika Maag (2020). "A systematic review: machine learning based recommendation systems for e-learning". Em: *Education and Information Technologies* 25.4. GS Search, pp. 2635–2664.

Kour, Sunpreet, Rakesh Kumar e Meenu Gupta (2021). "Analysis of student performance using Machine learning Algorithms". Em: *2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA)*. GS Search. IEEE, pp. 1395–1403.

Lemay, David John, Clare Baek e Tenzin Doleck (2021). "Comparison of learning analytics and educational data mining: a topic modeling approach". Em: *Computers and Education: Artificial Intelligence* 2. GS Search, p. 100016. ISSN: 2666-920X. DOI: https://doi.org/10.1016/j.caeai.2021.100016.

Manhães, Laci Mary Barbosa e SMS Cruz (2020). "Predição do desempenho acadêmico de alunos da graduação utilizando mineração de dados." Em: *Simpósio de Pesquisa Operacional e Logística da Marinha - Publicação Online* 3.1, pp. 2050– 2064. ISSN: 2175-6295. DOI: http://dx.doi.org/10.5151/spolm2019-148. URL: www.proceedings.blucher.com.br/article-details/34563.

Nidhi, Nidhi, Sachin Majithia e Neeraj Sharma (2021). "Predictive model for students' academic performance using classification and feature selection techniques". Em: *2021 2nd International Conference on Computational Methods in Science & Technology (ICCMST)*. GS Search. IEEE, pp. 106–111.

Romero, Cristóbal, Manuel-Ignacio López, Jose-María Luna e Sebastián Ventura (2013). "Predicting students' final performance from participation in on-line discussion forums". Em: *Computers Education* 68, pp. 458–472. ISSN: 0360-1315. DOI: https://doi.org/10.1016/j.compedu.2013.06.009. URL: https://www.sciencedirect.com/science/article/pii/S0360131513001607.

Santos, Fábio Paula (abr. de 2017). "Um modelo computacional de apoio à análise da opinião de alunos sobre práticas docentes por meio da mineração de dados educacionais". GS Search. Tese de doutorado. São Paulo, SP, BR: Universidade Presbiteriana Mackenzie.

Santos, Fabricia Damando (abr. de 2016). "Descoberta do desânimo de alunos em ambientes virtuais de ensino e aprendizagem: um modelo a partir da mineração de dados educacionais". GS Search. Tese de doutorado. Porto Alegre, RS, BR: Universidade Federal do Rio Grande do Sul - UFRGS.

Sekeroglu, Boran, Kamil Dimililer e Kubra Tuncal (2019). "Student performance prediction and classification using machine learning algorithms". Em: *Proceedings of the 2019 8th International Conference on Educational and Information Technology*. ICEIT 2019. GS Search. Cambridge, United Kingdom: Association for Computing Machinery, pp. 7–11. ISBN: 9781450362672. DOI: 10.1145/3318396.3318419.

Silva, Douglas H, Erick G Maziero, Muhammad Saadi, Renata L Rosa, Juan C Silva, Demostenes Z Rodriguez e Kostromitin K Igorevich (2022). "Big data analytics for critical information classification in online social networks using classifier chains". Em: *Peer-to-Peer Networking and Applications* 15.1. GS Search, pp. 626–641. URL: https://link.springer.com/article/10.1007/s12083-021-01269-1.

Souza, Vanessa Faria de (2021). “Mineração de dados educacionais com aprendizagem de máquina”. Em: *Revista Educar Mais* 5.4. GS Search, pp. 766–787.

Tan, Pang-Ning, Michael Steinbach e Vipin Kumar (2005). *Introduction to data mining, (First Edition)*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc. ISBN: 0321321367. URL: https://dl.acm.org/doi/10.5555/1095618.

Vijayalakshmi, V e K Venkatachalapathy (2019). “Comparison of predicting student’s performance using machine learning algorithms”. Em: *International Journal of Intelligent Systems and Applications* 11.12. GS Search, pp. 34–45.

Witten, Ian H., Eibe Frank e Mark A. Hall (2011). *Data mining: practical machine learning tools and techniques, 3rd Edition*. Morgan Kaufmann, Elsevier. ISBN: 9780123748560. URL: https://www.worldcat.org/oclc/262433473.

Yakubu, Mohammed Nasiru e Abubakar Mohammed Abubakar (abr. de 2022). “Applying machine learning approach to predict students’ performance in higher educational institutions”. Em: *Kybernetes* 51.2. GS Search, pp. 916–934. DOI: 10.1108/K-12-2020-0865.

Yufeia, Liu, Salmiza Salehb, Huang Jiahuic e Syed Mohamad Syed (2020). “Review of the application of artificial intelligence in education”. Em: *Integration (Amsterdam)* 12.8. GS Search, pp. 1–15.