A Framework for prediction of dropout in distance learning through XAI techniques in Virtual Learning Environment

Herbert da Silva Costa[1,2], Anderson Cordeiro Cardoso[2], Cristiane Mendes Netto[2], David Correa Martins-Jr[3], Sérgio Nery Simões[1]

[1]Mestrado em Computação Aplicada – Instituto Federal do Espírito Santo (IFES)
Campus Serra – 29166-630 – Serra – ES – Brasil

[2]Núcleo de Ensino a Distância – Universidade Vale do Rio Doce (Univale)
Rua Israel Pinheiro, 2000 – 35020-220 – Gov. Valadares – MG – Brasil

[3]Centro de Matemática, Computação e Cognição – Universidade Federal do ABC
Av. dos Estados, 5001 – Bangú – 09210-580 – Santo André - SP – Brasil

{herbert,anderson,cristiane}.costa@univale.br

david.martins@ufabc.edu.br,sergio@ifes.edu.br

Abstract. *A challenge in the distance learning modality is to avoid student's dropout which, according to the ABED, can vary between 21% and 50%. To this end, several data mining methods are applied, using student's interaction data in the Virtual Learning Environment. However, a relevant problem is to select the best features (variables/attributes) for early prediction student's dropout. In this paper, we propose a framework that uses explainable AI methods (XAI-SHAP) to find out attributes with greater predictive power on VLE integrated with third-party CMS. After selection, the proposed model achieved results of recall 0.96 and precision 0.95, compatible with the state of the art, but using a smaller set of attributes and a database with a smaller number of instances.*

Resumo. *Um desafio na modalidade EaD é combater a evasão, que segundo a ABED, varia entre 21 e 50%. Para este fim, diversos métodos de mineração de dados foram aplicados, utilizando dados de interações dos alunos no AVA. Contudo, um problema relevante é selecionar as melhores características (variáveis/atributos) para predição da evasão. Neste artigo, propomos um arcabouço que utiliza métodos de explicabilidade (XAI-SHAP) para selecionar atributos com maior poder preditivo em VLE que utilizam CMS terceirizados. Após a seleção, o modelo proposto alcançou resultados de recall 0,96 e precision 0,95, compatíveis com o estado da arte, porém utilizando um conjunto menor de atributos e uma base de dados com menor número de instâncias.*

1. Introdução

A modalidade de educação a distância (EaD) assumiu um papel importante na capacitação e na formação continuada das pessoas [Ramos et al. 2020]. O Censo EaD realizado em 2019 pelo Instituto Nacional de Estudos e Pesquisas Educacionais (INEP), demonstra o avanço da EaD através do número de alunos registrados em cursos desta modalidade em todo país [INEP 2020]. Segundo este estudo, somente em 2019, o número de ingressantes nos cursos de graduação na EaD foi de 1.592.184 alunos, o que representa um aumento

de 15,93% em relação a 2018 que foi de 1.373.321 alunos. A EaD tem como pontos positivos a possibilidade de ofertar um grande número de cursos para um número ilimitado de estudantes, proporcionando liberdade para o aluno estudar onde e quando quiser e na velocidade mais adequada ao seu perfil. Apesar disso, o número de alunos que concluem estes cursos (aqueles que não evadem) ainda é considerado baixo em todo mundo [Jin 2021]. De acordo com [ABED 2020], isto também ocorre no Brasil, pois tanto nas Instituições de Ensino Superior (IES) públicas como nas privadas, a taxa de evasão para graduação fica entre 21 e 50% na maior parte dos cursos, enquanto que no ensino presencial esta mesma taxa fica em torno de 22% e permanece inalterada ao longo dos anos. É neste contexto que surgiu o interesse em aplicar técnicas de Mineração de Dados no ambiente educacional (MDE), que segundo [Rabelo et al. 2017], trata-se de uma área tradicional que envolve a análise e extração de conhecimento de um grande volume de dados disponibilizado pelo Ambiente Virtual de Aprendizagem (AVA). Uma das abordagens mais utilizadas para registrar os dados de interação do aluno no AVA é chamada de *clickstream*, isto é, o histórico de todos os cliques e visualizações realizadas pelo aluno é registrado em *logs*. É através da mineração realizada nestes *logs*, disponíveis nos bancos de dados de AVAs como o Moodle, que grande parte dos modelos preditivos são treinados.

As plataformas AVAs gerenciam todas as atividades e interações entre professores e alunos e entre os próprios alunos. Em inglês, estes ambientes são também conhecidos por *Learning Management System* (LMS) ou *Virtual Learning Environment* (VLE), unindo funcionalidades de oferta de conteúdo – *Content Management Systems* (CMS) – com a avaliação das atividades dos alunos. No Brasil, AVAs mais utilizados como o Moodle praticamente unem todas estas siglas em uma única plataforma, mas em muitos casos, as IESs terceirizam a oferta de conteúdo, integrando seus AVAs com um ambiente de CMS terceirizado através do protocolo *Learning Tools Interoperability* (LTI).

Para o caso de CMS terceirizado, algumas variáveis propostas pelos autores [Rabelo et al. 2017, Ramos et al. 2018, Liu et al. 2020], não estão disponíveis pelo fato de registros de acessos a conteúdos, exercícios, materiais didáticos, vídeos e outros recursos metodológicos estarem em um ambiente que não possibilita mapeamento de tais interações. Como consequência disso, apenas quatro variáveis (características) (`med_geral_ac_sema_aluno`, `med_ac_manha_sema_aluno`, `med_ac_tarde_sema_aluno` e `med_ac_noite_sema_aluno`) são recuperadas do AVA, reduzindo o poder preditivo dos modelos utilizados neste trabalho, que são considerados o estado da arte (classifcadores *ensemble*) na área de classificação em MDE [Alamri et al. 2019, Kostopoulos et al. 2021, Panagiotakopoulos et al. 2021, Adnan et al. 2021]. Desta forma, nossa hipótese de trabalho é que realizando a integração de novos dados em conjunto com a engenharia de atributos – utilizando a abordagem de *eXplainable Artificial Intelligence* (XAI), por meio da combinação de recursos do pacote *SHapley Additive exPlanations* (SHAP) e um método *wrapper* chamado BorutaSHAP – pode-se aumentar o poder preditivo dos classificadores.

2. Materiais e Métodos

O percurso metodológico escolhido para este trabalho foi uma adaptação do processo de KDD (*Knowledge Discovery in Databases*) para o contexto educacional conforme sugere [Ramos et al. 2020], sendo apresentado na Figura 1. Este processo começa com a escolha da base de estudos (banco de dados do *Moodle*), seguida da realização de um pré-processamento visando realizar a limpeza e o correto mapeamento das variáveis

que registram as informações fortemente relacionadas com o alto risco de evasão [Ramos et al. 2018]. Em seguida, são aplicadas técnicas de mineração de dados para encontrar padrões e, por fim, realizar a interpretação dos resultados.

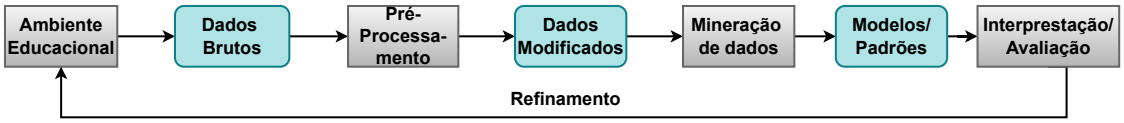


Figura 1. Processo adaptado do KDD

A partir desta estrutura de etapas proposta na Figura 1, foi criado o arcabouço ilustrado na Figura 2, que permite o emprego de métodos de aprendizado de máquina em um processo iterativo com o objetivo de encontrar o melhor conjunto de características (variáveis/atributos) que permita aumentar o poder preditivo dos classificadores analisados. Este arcabouço sugere que uma vez iniciado o experimento, a partir da **Fase 1**, sejam realizadas nas **Fases 2 e 3** um total de 30 (trinta) iterações de treinamento e teste para cada um dos seis modelos de classificação adotados neste trabalho. Em cada uma das iterações, que começa na **Fase 2**, é realizado um *hold out* dos dados na proporção de 70% para treino e 30% para teste, sendo estes dados estratificados de acordo com a classe de evasão e normalizados quando necessário. Os dados são balanceados utilizando a técnica SMOTE[1] [Chawla et al. 2002]. Na **Fase 3**, os dados de treinamento são submetidos à validação cruzada de Monte Carlo (método *RepeatedStratifiedKFold* com *kfold* = 5 e *n_repeats* = 3) e os estimadores resultantes desta validação são aplicados ao conjunto de teste. Ao final da **Fase 3**, são realizadas análises das métricas de desempenho adotadas (veja subseção 2.2.2) juntamente com o teste de significância da diferença de seus resultados pelo teste de Wilcoxon [Demir and Sahin 2022]. Na **Fase 4**, com os modelos de melhor desempenho escolhidos, a proposta é aplicar métodos de seleção de características utilizando a abordagem de XAI através da combinação de dois recursos do pacote SHAP, sendo o primeiro para análise de importância e impacto que cada característica possui na predição (gráfico de sumário) e o segundo para verificar a necessidade de retirar variáveis com alto grau de similaridade entre elas (dendrograma). Ao final desta fase aplica-se um método *wrapper* de seleção de atributos chamado BorutaSHAP [Gramegna and Giudici 2022] para confirmar as conclusões obtidas através da análise dos gráficos de SHAP, podendo ser iniciada uma nova iteração do processo para adição/remoção de variáveis, caso o resultado seja insatisfatório, ou confirmada a escolha do modelo que conseguiu o desempenho mais satisfatório baseado na estratégia estabelecida, juntamente com o conjunto de variáveis encontrado até esta fase e que melhor explica o fenômeno da evasão.

O desequilíbrio causado por dados desbalanceados, como é o caso do problema apresentado neste trabalho, de alunos frequentes ou **não evadidos** e os alunos **evadidos**, é especialmente um problema se houver interesse no bom desempenho da classe minoritária. Alcançar uma boa acurácia geral não necessariamente implica que o modelo tenha um bom desempenho. De fato, para uma avaliação adequada do desempenho do

[1] *Synthetic minority oversampling technique*, a qual cria exemplos sintéticos para cada uma das classes minoritárias até que o número de suas instâncias fique equivalente à majoritária

modelo, é necessária a análise de várias métricas de avaliação combinadas, como *precision*, *recall* e *f1-score*. Especialmente a pontuação *f1-score* é mais adequada para os casos em que a recuperação bem sucedida da classe minoritária seja de vital importância [Mahani and Ali 2019]. A métrica *f1-score*, que respresenta a média hormônica entre *recall* e *precision*, é muito importante para a estratégia das IESs, que em sua maioria pretendem identificar com eficácia os alunos que realmente estejam sujeitos à evasão (verdadeiros positivos). O arcabouço proposto orienta na escolha de modelos preditivos que tenham o melhor desempenho da métrica *f1-score* baseado na aplicação de métodos que auxiliam na identificação de variáveis que permitem o melhor entendimento do problema de evasão, podendo ser adotado por IESs que utilizam o AVA com oferta de conteúdo próprio, ou mesmo aquelas que utilizam o AVA integrado com plataformas de CMS terceirizadas.

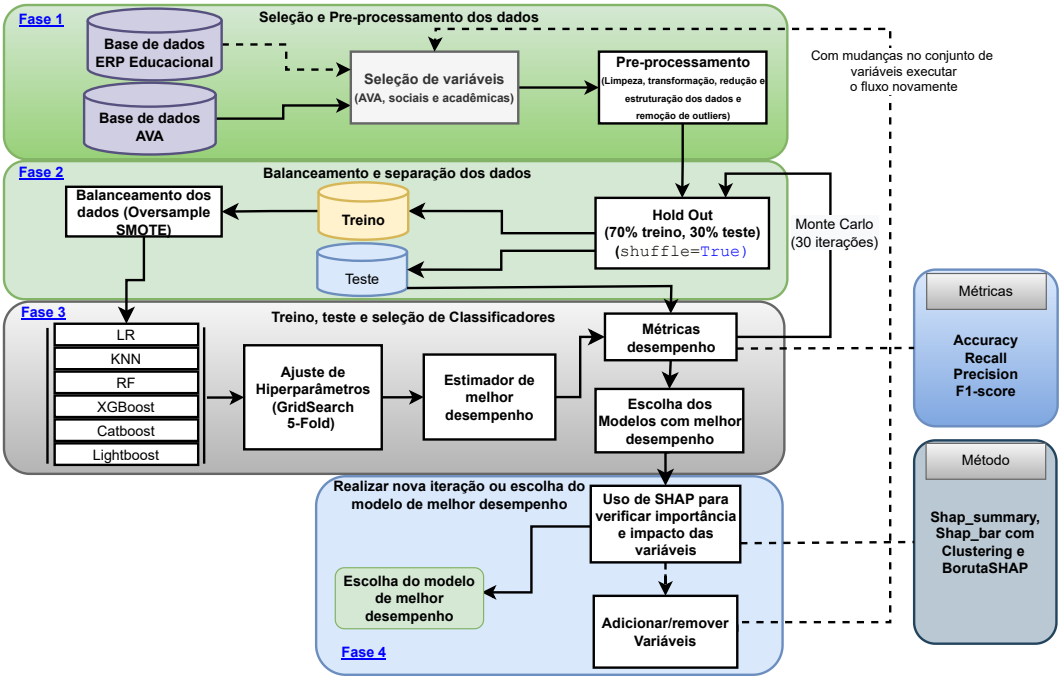


Figura 2. Arcabouço proposto para predição de Evasão

A etapa de **mineração dos dados** do processo KDD, além da etapa de **interpretação e avaliação**, permite a utilização das abordagens XAI para auxiliar os profissionais de ciência de dados a entender alguns padrões encontrados pelos modelos de mineração no tratamento das características e na importância de cada uma para predizer o comportamento futuro [Nalepa et al. 2021]. O pacote SHAP oferece um conjunto de métodos para exercer análises de XAI para diversos tipos de modelos [Linardatos et al. 2021], sendo uma abordagem que atua na interpretação de importâncias local e global de cada variável preditora de um modelo. Funciona de forma semelhante à teoria dos jogos, na qual as características são consideradas jogadores que recebem atributos denominados *Shapley values* como uma medida unificada de importância. Também contribui para que as interpretações sejam efetuadas e explicadas de forma gráfica. Desse modo, as caixas pretas dos diversos modelos são abertas e exploradas [Lundberg and Lee 2017]. Além disso, esta abordagem já se mostrou bastante eficiente no suporte a seleção de características (seleção *Embedded*), permi-

tindo um melhor desempenho dos modelos conforme descrevem [Marcílio and Eler 2020, Gramegna and Giudici 2022], e garantindo maior consistência na comparação destas variáveis para os diferentes modelos testados. Ou seja, a importância dos atributos selecionados tende a não variar quando o modelo é alterado [Lundberg et al. 2019].

2.1. Descrição dos dados

A abordagem de *clickstream* foi utilizada neste trabalho para a captura dos dados de interações do aluno com o AVA. Foram utilizados dados de uma IES privada, disponíveis entre os anos 2018 e 2021. Nesta IES os cursos de graduação em EaD, listados na Tabela 1(a) com suas respectivas taxas de evasão, são noturnos e semipresenciais, sendo que a maior parte das aulas acontecem de forma *online* através do AVA e o aluno frequenta a IES por no máximo duas vezes na semana para as aulas práticas. Estes cursos são organizados por módulos, sendo dois módulos por semestre e cada um destes módulos possui duração de 10 (dez) semanas. A análise realizada nestes dados foram feitas de acordo com os módulos ou semanas e não por períodos letivos, como é comum verificar em outros trabalhos.

Tabela 1. Dados da IES

(a) Cursos EAD e evasão

	Quantidade de Alunos		
Curso	**Evadidos**	**Total Curso**	**% Evadidos**
ADMINISTRACAO	14	30	46,67%
DESIGN GRAFICO	15	44	34,09%
ENGENHARIA CIVIL	18	49	36,73%
RECURSOS HUMANOS	11	42	26,19%
SISTEMAS DE INFORMAÇÃO	42	106	39,62%
PEDAGOGIA	2	9	22,22%

(b) Número de Instâncias do *dataset*

	Base de dados	
Classe	**Instâncias**	**%**
Não Evadido	2152	77,75%
Evadido	616	22,25%
Total	2768	100,00%

As instâncias recuperadas no banco de dados da IES, foram agrupadas por curso, período (módulo), disciplina e aluno, sendo possível verificar o risco de evasão por qualquer um destes grupos. A Tabela 1(b) demonstra a quantidade destas instâncias que corresponde a um total de 2.768, divididas em **não evadido** e **evadido** com valores absolutos (e percentuais) de 2.152 (77,75%) e 616 (22,25%) respectivamente. As variáveis utilizadas em cada um dos experimentos realizados (ver Seção 3) estão listadas na Tabela 2.

Estas variáveis estão segmentadas em três conjuntos: (i) o conjunto denominado "Mapeamento inicial" sugerido por [Ramos et al. 2016], contendo os dados disponibilizados para extração nos *logs* do AVA (*Moodle*) desta IES, que utiliza um CMS terceirizado; (ii) O segundo conjunto ("Novas variáveis mapeadas do AVA"), foi criado através da execução de tarefas de mineração de dados nos *logs* do AVA, que resultou na descoberta do registro de acesso do aluno ao módulo de CMS terceirizado via LTI e acesso à plataforma de diferentes lugares (acesso por diferentes IPs). Foi acrescentada também uma variável sobre o tempo gasto pelo aluno na utilização do ambiente, que segundo [Alamri et al. 2019], é muito importante para o resultado dos modelos; (iii) o terceiro conjunto foi extraído do Sistema de Gestão Acadêmica (ERP Educacional) da IES, baseado em sugestões encontradas na literatura para dados acadêmicos e sociais do aluno.

2.2. Metodologia e inicialização dos experimentos

O arcabouço proposto, ilustrado na Figura 2, emprega métodos de aprendizado de máquina em um processo iterativo com o objetivo de encontrar o melhor conjunto de

Tabela 2. Variáveis utilizadas nos experimentos

Variável	Descrição
Mapeamento inicial (Tabela/Ação Moodle: mdl_logstore_standard_log/loggedin)	
med_geral_ac_sema_aluno	Média semanal da quantidade de acessos do aluno ao ambiente no período (módulo)
med_ac_manha_sema_aluno	Média semanal de acessos do aluno ao ambiente por turno (Manhã), por período (módulo)
med_ac_tarde_sema_aluno	Média semanal de acessos do aluno ao ambiente por turno (Tarde), por período (módulo)
med_ac_noite_sema_aluno	Média semanal de acessos do aluno ao ambiente por turno (Noite), por período (módulo)
Novas variáveis mapeadas do AVA (Tabela/Ação Moodle: mdl_logstore_standard_log/loggedin e mod_lti)	
med_difip_ac_sema_aluno	Qtde de diferentes locais (IP´s) a partir dos quais a(o) aluna(o) acessou o ambiente por semana
tmp_medutil_semanahr	Tempo médio semanal de utilização da plataforma pelo aluno (em horas)
med_ac_lti_sema_disc	Média semanal de todos os acessos à LTI, por semana
med_ac_lti_aluno_sema_disc	Média semanal de acessos do aluno à LTI, por semana
Variáveis acadêmicas e sociais (ERP Educacional)	
idade	Idade do aluno
periodo_modulo	Período/Módulo que o aluno está cursando
qtde_rep_curso	Qtde de reprovações no curso em andamento
qtde_cursos_concluidos_instituicao	Qtde de Cursos que o aluno concluiu na IES
qtde_evasao_instituicao	Qtde de cursos em que o aluno evadiu na IES
sexo	Gênero do aluno
qtde_rep_prim_modulo	Qtde de reprovações do aluno no primeiro módulo do curso
qtde_reprov_disc	Qtde de reprovações na disciplina cursada
qtde_trancamento_curso	Qtde de trancamentos do aluno no curso em andamento
qtde_reprov_ult_modulo	Qtde de reprovações do aluno no último módulo
qtde_aprov_prim_modulo	Qtde de aprovações do aluno no primeiro módulo

variáveis que permita aumentar o poder preditivo dos classificadores analisados. Neste trabalho, as iterações foram realizadas por meio de dois experimentos, sendo que o primeiro utilizou apenas as variáveis que puderam ser mapeadas do AVA, conforme sugerido por alguns trabalhos recentes e que estão descritas no primeiro conjunto de variáveis da Tabela 2 (“Mapeamento inicial”). No segundo experimento, foram avaliadas alternativas de busca de novas variáveis (AVA/acadêmicas e sociais) ou remoção de outras já existentes nos conjuntos com o objetivo de melhorar o desempenho dos modelos. Foi definida uma configuração inicial para que todos os experimentos fossem realizados utilizando as mesmas regras estabelecidas.

2.2.1. Configuração inicial para os experimentos

O problema apresentado é de classificação binária, para o qual ficou definido que a classe 0 (zero) representa as instâncias com comportamento de **não evasão** e a classe 1 (um) as de **evasão**. Os dados de todos os cursos listados na Tabela 1(a) foram utilizados nos experimentos, sendo aplicados os passos e os métodos definidos pelo arcabouço proposto (Figura 2).

2.2.2. Métricas utilizadas

Para o problema de classificação binária aplicada à evasão, as métricas *recall* e *precision* são as que estrategicamente precisam ter resultados mais expressivos, principalmente na cobertura das predições de alunos com alto risco de evasão. As métricas selecionadas para medir o desempenho dos modelos foram **(i)** ***Accuracy***, que indica o percentual de acerto na relação das classificações corretas dentre todas as classificações, ou seja, o quão bom o modelo é para a previsão correta de dois grupos, como é o caso de alunos não evadidos e alunos evadidos; **(ii)** ***Recall*** que permite analisar se os valores de verdadeiro positivo, neste caso a classe 1 (um), foram corretamente classificados; **(iii)** ***Precision***

demonstra se os valores de verdadeiro positivo foram corretamente classificados dentre todas as classificações positivas.Valores baixos para esta métrica indicam um número mais alto de falsos positivos; e por fim **(iv) *f1-score*** que é a média harmônica entre os valores de *recall* e *precision*, que acabam sendo inversamente proporcionais. Esta métrica informa o quão preciso é o classificador (quantas instâncias ele classifica corretamente) e também o quão robusto ele é (cobre um número significativo de instâncias com classes positivas classificadas corretamente).

3. Resultados Experimentais

Para avaliar o impacto das variáveis no desempenho das predições dos classificadores, realizamos dois experimentos:

- **Experimento I**: SEM INTEGRAÇÃO DE DADOS, ou seja, utilizando apenas as quatro variáveis do ambiente AVA do mapeamento inicial mencionadas na Tabela 2;
- **Experimento II**: COM INTEGRAÇÃO DE DADOS, realizando a união das variáveis do experimento I com as variáveis extraídas do próprio AVA relacionadas à LTI e também com as variáveis acadêmicas e sociais do ERP educacional; ou seja, acrescentando as variáveis que estão nos quadros "*Novas variáveis mapeadas do AVA*" e "*Variáveis acadêmicas e sociais (ERP Educacional)*" da Tabela 2.

Para ambos os experimentos, treinamos e testamos seis modelos de classificação (Regressão Logística-RegLog - LR, *KNN, RandomForest* - RF, *XGBoost* - XGB, *CatBoost - CAT* e *LightBoost* - LGB), com ajuste de hiperparâmetros, sendo dois tradicionais (utilizados como *baseline)* e quatro *ensembles*. Para cada configuração de hiperparâmetros, aplicamos validação cruzada Monte Carlo com 30 repetições para avaliar a robustez dos resultados. Na Tabela 3 é possível verificar o resultado dos testes realizados, na qual podemos observar que, ao final do experimento II, os modelos *ensembles* se sobressaíram em relação aos demais, com *f1-score* chegando a 0,96. Este resultado ficou bem alinhado aos trabalhos de **(i) [Alamri et al. 2019]** que utilizando a abordagem de *clickstream*, alcançaram um *f1-score* de 0,93 com apenas duas variáveis (*Time spent* e *Number of accesses*) utilizando um *dataset* que possui registros de aproximadamente 2.200.000 interações de mais de 110.000 estudantes que utilizavam o AVA; **(ii) [Liu et al. 2020]**, utilizando correlação híbrida para seleção de características e classificadores treinados com base no comportamento dos estudantes ao longo do tempo, com registros de interações de mais de 112.000 alunos matriculados em 39 cursos e um conjunto de 16 variáveis, alcançaram um valor de *f1-score* de 0,90; **(iii) [Adnan et al. 2021]** com engenharia de atributos, considerando os diferentes estágios dos cursos e adicionando dados sociais e de desempenho, utilizando um *dataset* aberto disponibilizado pela *Open University, UK* com dados de 7 cursos e 32.593 alunos matriculados, obtiveram *f1-score* de 0,91 com *RandomForest*, enquanto que **(iv) [Panagiotakopoulos et al. 2021]**, com dados de 936 alunos, tendo 11 variáveis com dados sociais, 7 com dados de interação extraídos do AVA, e 2 com notas das avaliações realizadas no próprio AVA, alcançaram 0,96 de *f1-score* utilizando *LightBoost*.

A Tabela 3 mostra que o Experimento I (sem integração de dados) mostrou que com poucas características (apenas quatro), os modelos apresentaram um desempenho pior que o estado da arte. Já o Experimento II (com integração de dados), ou seja, após a adição de novas características (do próprio AVA+Academicas/sociais) e a seleção da-

Tabela 3. Comparativo de desempenho com novas variáveis mapeadas do AVA mais as acadêmicas e sociais (ERP Educacional) - Experimentos I e II

Autores	Modelo	Recall	Precision	f1-score	accuracy
[Alamri et al. 2019]	XGBoost	0,93	0,94	**0,93**	0,93
	RandomForest	0,93	0,94	**0,93**	0,93
[Liu et al. 2020]	XGBoost	-	-	**0,90**	-
	RegLog	-	-	0,85	-
[Adnan et al. 2021]	RandomForest	0,91	0,92	**0,91**	0,91
	KNN	0,90	0,90	0,90	0,90
[Panagiotakopoulos et al. 2021]	LightBoost	0,95	0,98	**0,96**	0,96
	RandomForest	0,94	0,97	0,95	0,94
	RegLog	0,95	0,96	0,95	0,94
Este Trabalho	**Modelo**	**Recall**	**Precision**	**f1-score**	**accuracy**
Experimento I – sem Integração de Dados (com mapeamento inicial Tabela 2 com hiperparâmetros ajustados)	KNN	0,83	0,89	**0,85**	0,91
	RandomForest	0,85	0,84	0,84	0,89
	XGBoost	0,85	0,84	0,84	0,89
	LightBoost	0,84	0,84	0,84	0,89
	CatBoost	0,84	0,84	0,84	0,89
	RegLog	0,77	0,70	0,70	0,75
Experimento II – com Integração de Dados (com variáveis selecionadas via SHAP conf. Figura 5 com hiperparâmetros ajustados)	CatBoost	0,96	0,95	**0,96**	0,98
	LightBoost	0,96	0,95	**0,96**	0,98
	XGBoost	0,96	0,95	0,95	0,97
	RandomForest	0,95	0,95	0,95	0,97
	KNN	0,85	0,86	0,85	0,92
	RegLog	0,75	0,65	0,66	0,75

quelas mais importantes pelo SHAP, mostrou que os modelos *ensemble* alcançaram desempenho geral compatível nas métricas (e.g. *f1-score*) em relação a alguns trabalhos do estado da arte. Portanto, os resultados destes experimentos indicam que a integração de dados (Experimento II) combinada com o auxílio de métodos de explicabilidade (e.g. SHAP) possibilitou selecionar características ainda mais relevantes para a construção de classificadores com desempenho bastante competitivo, mesmo utilizando uma quantidade relativamente pequena de características e instâncias frente aos trabalhos do estado da arte.

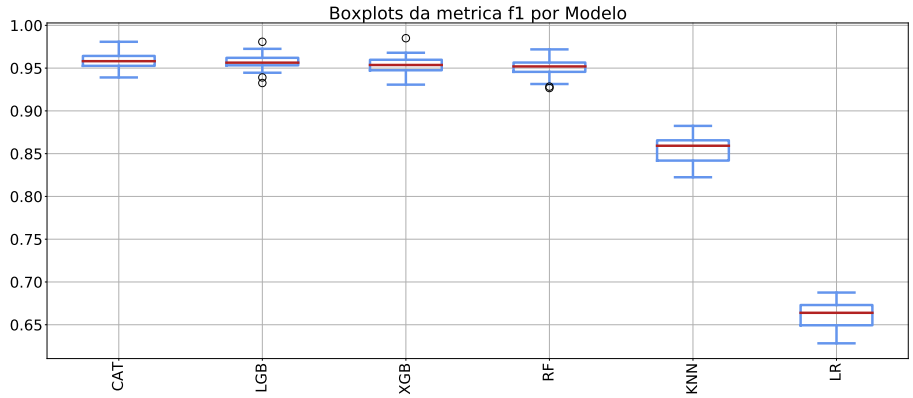
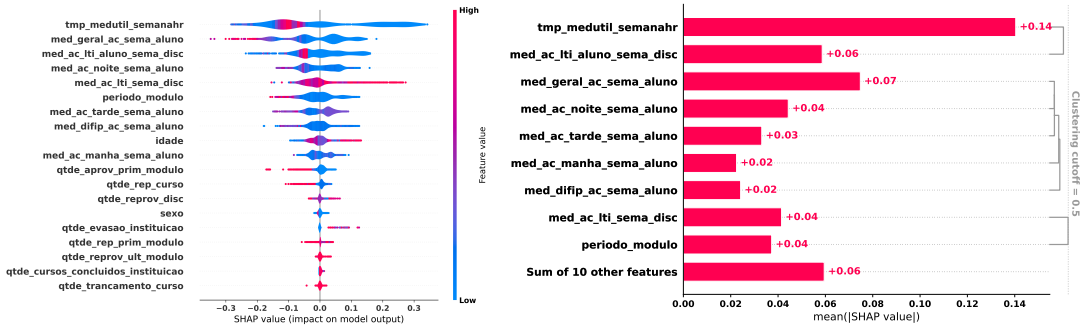


Figura 3. *Box Plots* dos testes do experimento II

Os resultados dos testes do Experimento II, também podem ser vistos na Figura 3, que contém os (*boxplots*) da métrica *f1-score* ordenados por valores decrescentes da mediana. Tais valores foram obtidos através da validação cruzada Monte-Carlo para cada modelo avaliado. Conforme podemos observar, os quatro modelos *ensembles* (*CatBoost*, *LightBoost*, *XGBoost* e *RandomForest*)[2] obtiveram resultados muito próximos, o que configura praticamente um empate técnico.

[2]Com os seguinte hiperparâmetros: *CatBoost*: {'learning_rate': 0.15, 'l2_leaf_reg': 1, 'iterations': 300,

Para realizar a análise de importância global das variáveis através do método SHAP, selecionamos o modelo *LightBoost* por ter um processo de treinamento mais eficiente em tempo de processamento, conforme constatado por [Daoud 2019]. A Figura 4a apresenta um sumário que combina a importância de cada variável com os impactos para cada uma delas na predição do modelo. Cada ponto no gráfico de resumo é um valor de *Shapley* para uma variável e uma instância. A posição no eixo y é determinada pela variável e no eixo x pelo valor de seu impacto na predição do modelo. A cor representa o valor da variável do mais baixo (azul) para o mais alto (vermelho) e os pontos sobrepostos são dispostos na direção do eixo y, fornecendo uma noção da distribuição dos valores de impacto por variável. As variáveis são ordenadas de acordo com sua importância de cima para baixo no gráfico. Como o problema deste experimento é de classificação binária, o impacto pode ser positivo, ou seja, tendendo a valores positivos em x, e neste caso aproximando-se da classe 1 (evadidos) ou negativo, cujo impacto tende a se aproximar da classe 0 (Não evadidos).



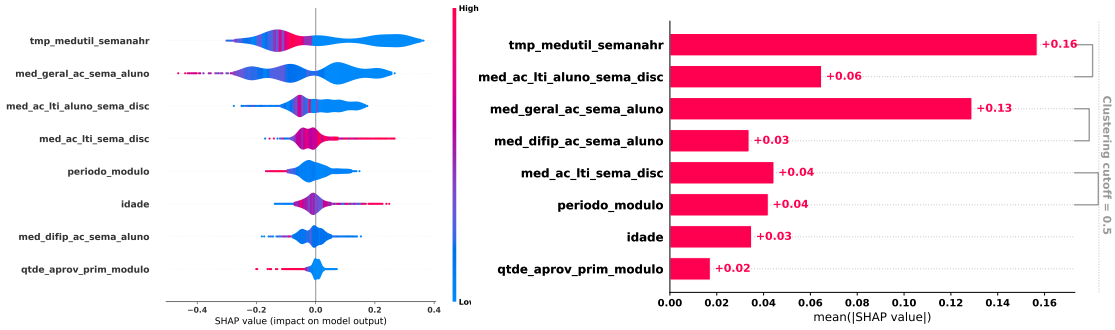
(a) Sumário de Importâncias **(b) Dendrograma de importâncias**

Figura 4. Importância global de todas as variáveis (*LightBoost*) - Experimento II

A Figura 4b apresenta uma informação relevante para a análise de explicabilidade: um agrupamento de variáveis por nível de similaridade entre eles, formando um dendrograma à direita das barras de cada variável. Esse dendrograma permite verificar o grau de dependência entre variáveis, dentre as quais as variáveis que forem redundantes com outras se tornam candidatas a serem removidas. Neste gráfico com dendrograma do modelo escolhido, é possível perceber um forte agrupamento entre as variáveis de acesso por turno (`med_ac_manha_sema_aluno`, `med_ac_tarde_sema_aluno` e `med_ac_noite_sema_aluno`) e a variável `med_geral_ac_sema_aluno`, demonstrando uma possível redundância entre elas, sendo importante realizar testes com sua remoção para avaliar o impacto na predição. A Figura 4a também demonstra que existem outras variáveis candidatas à remoção (as de menor impacto). A partir da variável `qtde_rep_curso` foram retiradas as oito características (contadas de cima para baixo) do gráfico de sumário, para verificação do impacto no desempenho dos classificadores. Com isso, constatamos que mesmo retirando essas variáveis, o desempenho do modelo permaneceu praticamente o mesmo, confirmando que elas estavam sendo redundantes neste caso. Antes de remover definitivamente essas variáveis, utilizamos também uma abordagem *Wrapper* de seleção de atributos, aplicando o método BorutaShap [Gramegna and Giudici 2022] que corroborou o resultado SHAP

'depth': 7}, *LightBoost*: {'num_leaves': 300, 'n_estimators': 200, 'max_depth': 50, 'learning_rate': 0.1}, *XGBoost*: {'n_estimators': 400, 'min_child_weight': 1, 'max_depth': 10, 'learning_rate': 0.16} e *RandomForest*: {'n_estimators': 300, 'min_samples_split': 4, 'min_samples_leaf': 1, 'bootstrap': False}

obtido anteriormente. A Figura 5 demonstra como ficou a distribuição das variáveis por ordem de importância e impacto (Figura 5a) e por agrupamento de similaridade (Figura 5b), após a remoção.



(a) Sumário de Importâncias **(b) Dendrograma de importâncias**

Figura 5. Importância global (*LightBoost*) - Ao finalizar o experimento II

4. Conclusão

Neste artigo, apresentamos um arcabouço para selecionar os modelos com maior poder preditivo de evasão em cursos de Ensino a Distância (EaD) que utilizam ambientes LMS (ou VLE) juntamente com CMS terceirizados. Para isso, o arcabouço desenvolvido recebe como entrada um conjunto de dados extraídos do próprio AVA e também do ERP Educacional utilizado pela IES. Em seguida, aplica-se um *pipeline* que contém métodos de XAI-SHAP visando selecionar características com maior poder preditivo que sejam capazes de explicar o fenômeno da evasão em Educação à Distância.

O arcabouço realiza o pré-processamento dos dados, testa automaticamente diversos tipos de classificadores com diversas combinações de hiperparâmetros e, com auxílio de técnicas de explicabilidade, encontra os modelos de maior desempenho. Para cada configuração, realizamos 30 experimentos, e alcançamos o resultado médio de aproximadamente 0,96 na métrica *f1-score* para os modelos de classificação *LightBoost*, *CatBoost*, *XGBoost* e *Random Forest*, indicando que a seleção das características relevantes contribuiu para o bom desempenho dos algoritmos. É importante destacar que o resultado alcançado, foi compatível com a maioria dos trabalhos recentes citados na Tabela 3, mesmo contando com um número bastante reduzido de características e de registros de interações, indicando que as características selecionadas realmente foram as mais relevantes para o desempenho do modelo.

Portanto, concluímos que através do uso dos métodos de XAI, aplicados através de recursos do pacote SHAP, é possível alcançar eficiência e consistência na comparação de importância e impacto que cada característica possui em relação ao problema proposto e não a um modelo preditivo específico, revelando um conjunto enxuto de características juntamente com o modelo mais adequado a este conjunto.

4.1. Considerações finais e trabalhos futuros

Apesar dos bons resultados alcançados através do uso do arcabouço proposto, em termos de desempenho e de escolha de características, é importante realizar experimentos com base de dados que possua mais exemplos, a fim de diminuir possíveis limitações impostas

às conclusões deste estudo em virtude do número reduzido de amostras. Além disso, a participação do usuário final como especialista da área de EaD pode ser muito importante para garantir a evolução do modelo preditivo, sendo possível através de sua interpretação e avaliação dos resultados do modelo, evitar que decisões sejam tomadas de forma equivocada. Para que isso seja possível é importante o uso do critério *post hoc* de XAI, que permite a aplicação de métodos de explicabilidade após o treinamento do modelo, ou seja, no processo de predição. Neste tipo de critério, a análise de importância das características pode ser aplicada a um número reduzido de instâncias ou mesmo a apenas uma instância a partir do resultado do modelo.

Como trabalho futuro, sugerimos o desenvolvimento de uma aplicação computacional com uma interface simples que permita a visualização de resultados descritivos e preditivos. Estes resultados podem ser disponibilizados ao usuário final com recursos de explicabilidade através de gráficos do SHAP, possibilitando que o usuário contribua para a melhoria do modelo preditivo, fornecendo *feedbacks* fundamentados em sua experiência sobre o problema e criando condições para possíveis otimizações que possam ser efetuadas nos resultados da fase de **interpretação e avaliação** proposta pelo KDD.

Referências

ABED (2020). Censo ead 2019/2020 associação brasileira de educação a distância - abed.

Adnan, M., Habib, A., Ashraf, J., Mussadiq, S., Raza, A. A., Abid, M., Bashir, M., and Khan, S. U. (2021). Predicting at-risk students at different percentages of course length for early intervention using machine learning models. *IEEE Access*, 9:7519–7539. IEEE Access.

Alamri, A., Alshehri, M., Cristea, A. I., Pereira, F. D., Oliveira, E., Shi, L., and Stewart, C. (2019). Predicting MOOCs dropout using only two easily obtainable features from the first week's activities. In *Lecture Notes in Computer Science*, volume 11528, pages 163–173. arXiv.org.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357.

Daoud, E. A. (2019). Comparison between XGBoost, LightGBM and CatBoost using a home credit dataset. *International Journal of Computer and Information Engineering*, 13(1):6–10.

Demir, S. and Sahin, E. K. (2022). Comparison of tree-based machine learning algorithms for predicting liquefaction potential using canonical correlation forest, rotation forest, and random forest based on CPT data. *Soil Dynamics and Earthquake Engineering*, 154:107130.

Gramegna, A. and Giudici, P. (2022). Shapley feature selection. *FinTech*, 1(1):72–80. Number: 1 Publisher: Multidisciplinary Digital Publishing Institute.

INEP (2020). Apresentação da coletiva de imprensa censo da educação superior 2019.

Jin, C. (2021). Dropout prediction model in MOOC based on clickstream data and student sample weight. *Soft Computing*, 25(14):8971–8988.

Kostopoulos, G., Panagiotakopoulos, T., Kotsiantis, S., Pierrakeas, C., and Kameas, A. (2021). Interpretable models for early prediction of certification in MOOCs: A case study on a MOOC for smart city professionals. *IEEE Access*, 9:165881–165891. IEEE Access.

Linardatos, P., Papastefanopoulos, V., and Kotsiantis, S. (2021). Explainable AI: A review of machine learning interpretability methods. *Entropy*, 23(1):18. Number: 1 Publisher: Multidisciplinary Digital Publishing Institute.

Liu, K., Tatinati, S., and Khong, A. W. H. (2020). A weighted feature extraction technique based on temporal accumulation of learner behavior features for early prediction of dropouts. In *2020 IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE)*, pages 295–302.

Lundberg, S. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. *arXiv:1705.07874 [cs, stat]*. arXiv:1705.07874 [cs, stat].

Lundberg, S. M., Erion, G. G., and Lee, S.-I. (2019). Consistent individualized feature attribution for tree ensembles. arXiv:1802.03888.

Mahani, A. and Ali, A. R. B. (2019). *Classification Problem in Imbalanced Datasets*. IntechOpen. Publication Title: Recent Trends in Computational Intelligence.

Marcílio, W. E. and Eler, D. M. (2020). From explanations to feature selection: assessing SHAP values as feature selection mechanism. In *2020 33rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, pages 340–347.

Nalepa, G. J., Bobek, S., Kutt, K., and Atzmueller, M. (2021). Semantic data mining in ubiquitous sensing: A survey. *Sensors*, 21(13):4322. Number: 13 Publisher: Multidisciplinary Digital Publishing Institute.

Panagiotakopoulos, T., Kotsiantis, S., Kostopoulos, G., Iatrellis, O., and Kameas, A. (2021). Early dropout prediction in MOOCs through supervised learning and hyperparameter optimization. *Electronics*, 10(14):1701. Number: 14 Publisher: Multidisciplinary Digital Publishing Institute.

Rabelo, H., Burlamaqui, A., Valentim, R., Rabelo, D. S. d. S., and Medeiros, S. (2017). Utilização de técnicas de mineração de dados educacionais para predição de desempenho de alunos de EaD em ambientes virtuais de aprendizagem. 28(1):1527. Simpósio Brasileiro de Informática na Educação - SBIE.

Ramos, J. L. C., Santos, L. F. L., Silva, J. C. S., and Rodrigues, R. L. (2020). Identificação de perfis de interação de estudantes de educação a distância por meio de técnicas de agrupamentos. In *Anais do Simpósio Brasileiro de Informática na Educação*, pages 932–941. SBC.

Ramos, J. L. C., Silva, J., Prado, L., Gomes, A., and Rodrigues, R. (2018). Um estudo comparativo de classificadores na previsão da evasão de alunos em EAD. 29(1):1463. Simpósio Brasileiro de Informática na Educação - SBIE.

Ramos, J. L. C., Silva, J., Rodrigues, R., Gomes, A. S., and Souza, F. d. F. d. (2016). Mapeamento de dados de um LMS para medida de construtos da distância transacional. 27(1):1056. Simpósio Brasileiro de Informática na Educação - SBIE.