Predição da evasão estudantil: uma análise comparativa de diferentes representações de treino na aprendizagem de modelos genéricos

Miriam Pizzatto Colpo[1,2], **Tiago Thompsen Primo**[1], **Marilton Sanchotene de Aguiar**[1]

[1]Programa de Pós-Graduação em Computação (PPGC)
Universidade Federal de Pelotas (UFPel)
Pelotas, RS, Brasil

[2]Instituto Federal de Educação, Ciência e Tecnologia Farroupilha (IFFar)
Santa Maria, RS, Brasil

{miriam.colpo, tiago.primo, marilton}@inf.ufpel.edu.br

Abstract. *In this work, different ways to represent the dropout behavior are evaluated in the development of generic models, aimed to predicting the dropout risk, in different semesters and courses, of face-to-face undergraduate students. From a careful pre-processing and the creation of distinct representations of training data, different machine learning models were built in order to evaluate which representation best contributes to the predictions performance. As a result, it was found that exemplifying the behavior of students in all semesters attended, in an accumulated and progressive way, benefited the learning of the predictive model, providing a accuracy of 80.1%.*

Resumo. *Neste trabalho são avaliadas diferentes formas de representar o comportamento de evasão no desenvolvimento de modelos genéricos, destinados a prever o risco de abandono, em diferentes semestres e cursos, de alunos de graduação da modalidade presencial. A partir de um cuidadoso pré-processamento e da criação de distintas representações de dados de treino, foram construídos diferentes modelos de aprendizado de máquina, a fim de avaliar qual representação melhor contribui para o desempenho das predições. Como resultado, verificou-se que exemplificar o comportamento dos alunos em todos os semestres cursados, de forma acumulada e progressiva, beneficiou a aprendizagem do modelo preditivo, provendo uma acurácia de 80.1%.*

1. Introdução

Em meio à pandemia de COVID-19, doença infecciosa causada pelo coronavírus da síndrome respiratória aguda grave 2 (do Inglês, *Severe Acute Respiratory Syndrome Coronavirus 2* – SARS-COV-2), o ensino, até então presencial, passou a ser conduzido de forma remota, com o intuito de promover o distanciamento social e diminuir a transmissão do agente patogênico [Santos and Zaboroski 2020]. Embora essencial ao contexto atual, o Ensino Remoto Emergencial (ERE) impôs desafios a estudantes, professores e instituições, que não estavam preparados, metodológica e estruturalmente, para essa forma alternativa de ensino [Santos and Zaboroski 2020]. Nesse contexto de dificuldades e mudanças, a evasão estudantil, definida pelo Ministério da Educação como "a

DOI: 10.5753/sbie.2021.218517

saída definitiva do estudante do curso de origem, sem concluí-lo" [Brasil 1996], passou a causar uma preocupação ainda maior. Dada a correlação entre o nível de escolarização e os ganhos salariais da população [Pontili et al. 2018], a evasão, além de causar danos sociais e acadêmicos, prejudica o desenvolvimento econômico do país, já fortemente prejudicado pela pandemia. Economicamente, a perda se torna ainda maior sob a perspectiva da rede pública de ensino, na qual a evasão significa recursos do Estado, não só financeiros, mas também de pessoal e infraestrutura, investidos sem o devido retorno [Silva Filho et al. 2007].

Com o objetivo de identificar o perfil de evasão em cursos de graduação e, assim, prever alunos em risco e fomentar a implementação de estratégias preventivas, diversas pesquisas fazem uso de técnicas de Mineração de Dados Educacionais (MDE) no desenvolvimento de modelos preditivos [Mduma et al. 2019][Colpo et al. 2020]. Naturalmente, esses estudos utilizam dados históricos de evasão para treinar algoritmos de Aprendizado de Máquina (AM) e, dessa forma, construir modelos baseados em experiências anteriores [Han et al. 2012]. Embora muitos desses modelos sejam projetados para prever o risco de abandono em um curso específico ou em um determinado estágio da trajetória acadêmica do aluno, a construção de modelos genéricos, destinados a predições periódicas e a estudantes de diferentes cursos, envolve decisões adicionais, incluindo a forma de representação dos dados de treino. Por isso, neste trabalho, além de um pré-processamento de dados voltado ao desenvolvimento de modelos genéricos, é realizada uma análise comparativa acerca de diferentes formas de representação dos registros de treino, de acordo com as classes dos alunos (evadido ou não evadido/formado) e variando entre as seguintes suposições: (*i*) a evasão tende a decorrer de um processo contínuo e não de uma causa isolada ou momentânea, podendo, então, ser caracterizada por diferentes registros dos estudantes, desde o início de suas trajetórias acadêmicas; e (*ii*) a evasão é um fenômeno dinâmico e, por isso, deve ser caracterizada apenas pelo comportamento que os estudantes apresentaram em seu último semestre.

A descrição deste estudo está organizada da seguinte forma: na Seção 2 são sintetizados trabalhos relacionados ao desenvolvimento de modelos preditivos de evasão, assim como as representações adotadas por eles. A metodologia seguida no desenvolvimento desta pesquisa, incluindo detalhes da extração e do pré-processamento dos dados, das representações consideradas, e das técnicas de AM utilizadas na construção dos modelos, são apresentadas na Seção 3. E, por fim, na Seção 4, são descritos e discutidos os resultados obtidos; além de apresentadas, na Seção 5, conclusões e considerações finais.

2. Trabalhos Relacionados

Como já mencionado, muitos trabalhos direcionam a predição de evasão a um estágio específico da trajetória acadêmica dos alunos, em geral, ao momento da primeira matrícula, sem considerar indícios do comportamento do estudante no curso atual, como em [Nagy and Molontay 2018] e [Baranyi et al. 2020]; ou após um número fixo de períodos cursados, considerando, neste caso, dados que caracterizem o desempenho inicial do aluno, como em [Costa et al. 2020] e [Yu et al. 2021]. Em comum, estes trabalhos costumam representar os dados de cada estudante a partir de um único registro, já que o momento de predição é fixo e, assim, sabe-se exatamente a estrutura de atributos a ser considerada. Se as predições serão executadas após os alunos completarem dois semestres, por exemplo, pode-se considerar como atributos as médias de notas do primeiro e do

segundo semestre. Além disso, caso os modelos sejam voltados a alunos de um curso específico ou a cursos com estruturas curriculares semelhantes, como em [Costa et al. 2020] e [Böttcher et al. 2021], respectivamente, é comum utilizar atributos absolutos, como o número de disciplinas cursadas ou aprovadas, na caracterização da situação dos alunos.

Já para modelos genéricos, destinados a prever a situação de alunos em diferentes estágios da trajetória acadêmica e de diferentes cursos, como em [Kang and Wang 2018] e [Ortigosa et al. 2019], é desejável a inclusão de atributos relativos, que permitam posicionar a situação do aluno em relação aos seus pares ou que tornem comparáveis os dados de todos os estudantes. Considerando estruturas curriculares discrepantes entre os alunos, a porcentagem de carga horária cumprida pode ser mais informativa que o número de disciplinas aprovadas, por exemplo. Além disso, representar cada aluno a partir de um único registro, contendo atributos específicos a cada semestre cursado pelo aluno, deixa de ser uma opção viável, já que as previsões devem contemplar alunos que estejam nos mais variados estágios de suas jornadas acadêmicas. Dessa forma, nestes trabalhos, os estudantes são, em geral, caracterizados por atributos que consideram apenas dados do último semestre (como a média semestral de notas) ou que agregam e sumarizam dados de seu histórico (como a média geral de notas).

Porém, para modelos genéricos, também faz-se necessário decidir como representar os dados de treinamento, a fim de melhor caracterizar/exemplificar o fenômeno da evasão, embora essa decisão não costume ser descrita claramente nas pesquisas. Em [Solis et al. 2018], são analisadas diferentes representações para a caracterização dos comportamentos de evasão e não evasão, incluindo perspectivas acerca da utilização ou não de múltiplos registros por aluno (um para cada semestre cursado). Neste trabalho, além de se considerar essas duas representações, são incluídas na análise as variações de representação por classe de aluno (múltiplos registros para alunos evadidos e único registro para alunos concluídos, e vice-versa). Adicionalmente, como cada forma de representação impacta no balanceamento dos exemplos (mesmo que se considere a mesma quantidade de alunos por classe, alunos evadidos geralmente cursam um número menor de semestres), realiza-se a sobreamostragem dos dados de treino, a fim de oportunizar uma avaliação mais justa. Por fim, ao contrário de [Solis et al. 2018], que utiliza atributos absolutos por semestre, neste estudo são considerados atributos relativos, a fim de contribuir para a generalização do aprendizado dos modelos preditivos.

3. Metodologia

Para o desenvolvimento deste trabalho, diversos procedimentos foram realizados, abrangendo desde a extração dos dados até o treinamento e a comparação dos modelos preditivos. Na Figura 1, é apresentada uma visão geral da metodologia seguida neste estudo, sendo suas etapas detalhadas a seguir, nas Subseções 3.1, 3.2 e 3.3.

3.1. Extração, pré-processamento e separação dos dados

Como indicado na Figura 1, na **primeira etapa** deste trabalho, foi realizada a extração de dados da base do Sistema Integrado de Gestão de Atividades Acadêmicas (SIGAA), utilizado no Instituto Federal de Educação, Ciência e Tecnologia Farroupilha (IFFar), desde 2014. Após o estudo/entendimento dos dados disponíveis, foram construídas consultas SQL (do Inglês, *Structured Query Language*) para viabilizar a coleta dos dados

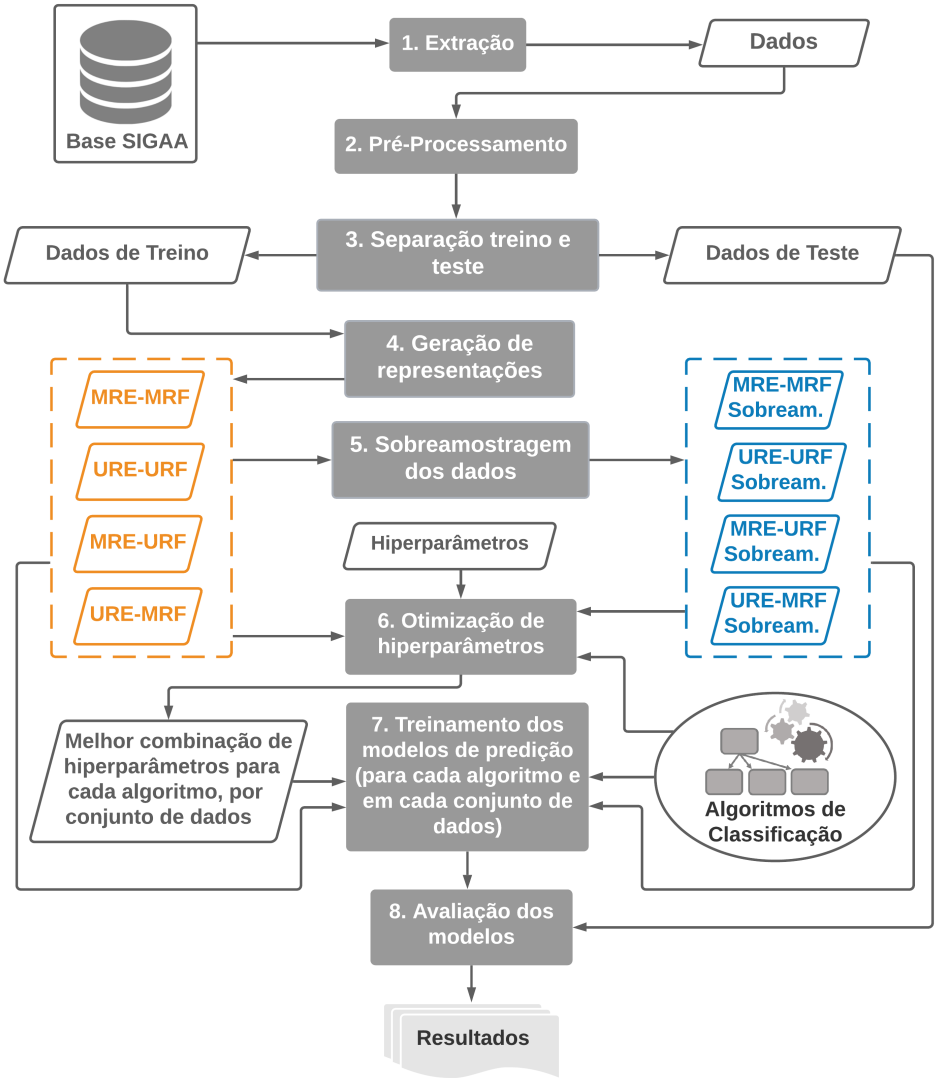

Figura 1. Visão geral da metodologia adotada.

de interesse. Assim, em maio de 2021, por meio de um *script* desenvolvido na linguagem *Python*, foram coletados dados de alunos que abandonaram ou concluíram seus cursos de graduação em 2019 ou 2020, incluindo informações acadêmicas, contextuais, econômicas, interacionais e sociais. Faz-se importante destacar que, dentre os alunos matriculados em algum semestre dos anos considerados, foram rotulados como evadidos os que, sem possuir registro de trancamento ou conclusão do curso, deixaram de se matricular no semestre subsequente.

Na Tabela 1, são apresentados os quantitativos de alunos que evadiram ou concluíram seus cursos, perante os totais de alunos matriculados em 2019 e 2020. Pode-se notar que, dentre os alunos que possuíam matrícula em 2019, 8.63% concluíram o curso e 23.86% evadiram. Já para os alunos matriculados em 2020, esses percentuais são de 7.74% e 25.73%, respectivamente. Além disso, é possível observar que os quantitativos de registros coletados são maiores que os de alunos formados ou evadidos. Isso porque cada aluno teve suas informações coletadas por semestre cursado, de forma progressiva e acumulada. Dessa forma, cada registro de um aluno provê informações relacionadas a um estágio (semestre) diferente de sua evolução no curso.

	2019		2020	
	Nº de Alunos	Nº de Registros	Nº de Alunos	Nº de Registros
Matriculados	5147	-	5153	-
Formados	444 (8.63%)	3483	373 (7.24%)	3173
Evadidos	1228 (23.86%)	3497	1326 (25.73%)	4160

Tabela 1. Quantitativos de alunos e registros, por ano.

Na **segunda etapa**, os dados extraídos foram pré-processados, com o auxílio das bibliotecas *Pandas* [McKinney 2010] e *Scikit-learn* [Pedregosa et al. 2011]. Nessa fase, atributos absolutos (como o número de disciplinas em que o aluno foi aprovado) foram transformados em atributos relativos (como a porcentagem de aprovação do aluno), a fim de padronizar e adequar suas informações a alunos dos diferentes cursos de graduação presenciais do IFFar, incluindo tecnólogos, licenciaturas e bacharelados. Ou seja, como os registros se relacionam a estruturas curriculares bastante distintas, o uso de informações relativas permite a equiparação dos dados e possibilita, assim, a construção de modelos mais generalizáveis. Além disso, para tratamento de dados faltantes, atributos categóricos tiveram seus valores vazios preenchidos com o termo "Não Informado"; e atributos numéricos do tipo econômico receberam o valor "-1.0" em seus dados faltantes, de modo a distinguir a inexistência de renda do aluno (0.0) da sua não declaração. Para os demais atributos numéricos, valores faltantes foram preenchidos por meio do método de imputação pela média dos vizinhos mais próximos (em inglês, *k-Nearest Neighbors Imputer*). E, por fim, como os algoritmos de AM do pacote *Scikit-learn* aceitam apenas atributos preditivos de tipo numérico, os atributos categóricos foram binarizados, por meio da técnica de codificação *one-hot*. Assim, além do atributo alvo (classe), que indica se o estudante evadiu ou não, cada registro de um aluno foi representado por 67 atributos, descritos na Tabela 2. Faz-se importante mencionar que todos esses atributos consideram dados do histórico do aluno, até o semestre relacionado ao registro. Ou seja, para um registro com período/semestre = 3, por exemplo, a média de notas é calculada a partir de todas as disciplinas cursadas pelo aluno, até o terceiro semestre.

Após serem pré-processados, na **terceira etapa** os dados foram separados em dois conjuntos distintos, a serem considerados, posteriormente, no treinamento e na avaliação dos modelos preditivos. Com o intuito de aproximar a avaliação à uma aplicação de predição futura, ao invés de dividir de forma estratificada toda a base de dados, decidiu-se separar os conjuntos de dados de acordo com o ano de conclusão ou abandono dos alunos. Ou seja, os registros de alunos evadidos ou formados em 2019 foram reservados para o treinamento dos modelos preditivos, de modo que a avaliação desses modelos seja realizada sobre os registros relacionados às evasões e conclusões ocorridas em 2020.

3.2. Geração de diferentes representações de treino e sobreamostragem

Após a definição dos conjuntos de treino e teste, foram geradas diferentes representações da base de treino, o que corresponde à **quarta etapa** da Figura 1. Essas representações se diferenciam por considerar ou não a inclusão de todos os registros (semestrais) dos alunos na base de treinamento, da seguinte forma:

- **Múltiplos Registros para Evadidos e Múltiplas Registros para Formados (MRE-MRF)**: representação em que todos os registros semestrais dos alunos (formados ou evadidos) são considerados na base de treino. Como os dados já foram extraídos dessa forma, essa representação corresponde à base de treino resultante

Tabela 2. Descrição dos atributos do conjunto de dados.

#	Descrição do Atributo	Aspecto
1	média de notas	Acadêmico
2	média de frequências	Acadêmico
3	% de carga horária integralizada	Acadêmico
4	% de aprovações	Acadêmico
5	% de exames	Acadêmico
6	média de participação em projetos por semestre	Acadêmico
7	% de dias em situação de trancamento	Acadêmico
8	% de dias em faltas justificadas ou estudos domiciliares	Acadêmico
9	período/semestre	Contextual
10-19	indicadores de pertencimento, para cada campus do IFFar	Contextual
20	indica se o aluno cursa um bacharelado	Contextual
21	indica se o aluno cursa uma licenciatura	Contextual
22	indica se o aluno cursa um tecnólogo	Contextual
23	indica se o curso do aluno é de Ciências Agrárias	Contextual
24	indica se o curso do aluno é de Ciências Biológicas	Contextual
25	indica se o curso do aluno é de Ciências Exatas e da Terra	Contextual
26	indica se o curso do aluno é de Ciências Humanas	Contextual
27	indica se o curso do aluno é de Ciências Sociais Aplicadas	Contextual
28	indica se o curso do aluno é de Engenharias	Contextual
29	indica se o curso do aluno é de Outras áreas	Contextual
30-36	indicadores de forma de ingresso	Contextual
37	renda total do núcleo familiar	Econômico
38	quantidade de membros do núcleo familiar	Econômico
39	pontuação no Cadastro Único	Econômico
40	renda per capta do núcleo familiar	Econômico
41	média do número de auxílios recebidos por semestre	Econômico
42	porcentagem de enquetes respondidas	Interacional
43	porcentagem de questionários respondidos	Interacional
44	porcentagem de fóruns respondidos	Interacional
45	porcentagem de tarefas respondidas	Interacional
46	média diária de ações em turmas virtuais	Interacional
47	idade ao ingressar no curso	Social
48	indicador de sexo	Social
49-53	indicadores de cor/raça	Social
54-57	indicadores de estado civil	Social
58	indica se cursou o ensino médio em escola pública	Social
59	tempo (em anos) entre a conclusão do ensino médio e o ingresso	Social
60	indica se reside na mesma cidade do campus	Social
61-67	indicadores de reserva de vaga	Social

do processo descrito na Subseção 3.1. Essa abordagem busca fornecer exemplos que caracterizem o comportamento dos estudantes, desde o início de suas trajetórias acadêmicas. Ou seja, mesmo que um aluno abandone seu curso no terceiro semestre, seus registros de primeiro e segundo semestre serão utilizados como exemplos de abandono, partindo do entendimento de que a evasão tende a decorrer de um processo e não de uma causa isolada ou momentânea;

- **Único Registro para Evadidos e Único Registro para Formados (URE-URF)**: representação que mantém na base de treino apenas um registro por aluno. Para alunos evadidos, o último registro é mantido, pressupondo que as informações do último semestre são mais relevantes para a caracterização da evasão, e que dados de semestres anteriores podem prover ruído. Já para alunos formados, mantém-se

um registro escolhido aleatoriamente. Isso porque a base de treino precisa conter exemplos de diferentes semestres para a generalização do aprendizado do modelo, o que não seria possível mantendo apenas dados do semestre de conclusão;
- **Múltiplos Registros para Evadidos e Único Registro para Formados (MRE-URF)**: combinação das representações anteriores, em que se mantém todos os registros semestrais dos alunos evadidos e apenas um registro, escolhido aleatoriamente, por aluno formado;
- **Único Registro para Evadidos e Múltiplos Registros para Formados (URE-MRF)**: combinação das duas primeiras representações, em que se mantém apenas os dados do último semestre para alunos evadidos, e todos os registros semestrais dos alunos formados.

Enquanto as duas primeiras representações correspondem a perspectivas avaliadas em [Solis et al. 2018], as outras duas variações foram acrescentadas neste estudo, a fim de contemplar todas as combinações possíveis de representação. Embora [Solis et al. 2018] também avalie a utilização de registros de alunos ativos (matriculados) na classe de alunos "não evadidos", essa abordagem não foi considerada neste estudo por duas razões: (*i*) considerar alunos ativos como exemplos negativos de evasão aumenta muito a probabilidade de ruído nos registros, uma vez que alunos matriculados possuem situação indefinida e podem estar prestes a evadir; e (*ii*) na avaliação de [Solis et al. 2018], essa abordagem já demonstrou os piores resultados.

Como cada representação envolve diferentes proporções de registros relacionados a alunos evadidos e formados, para uma avaliação mais justa, na **quinta etapa**, foi realizado o balanceamento completo dos dados de cada representação, dando origem a quatro novos conjuntos de treino, conforme ilustrado na Figura 1. Para isso, utilizou-se a técnica de sobreamostragem minoritária sintética (do inglês, *Synthetic Minority Over-sampling Technique* – SMOTE), disponível no pacote *Imbalanced-learn* [Lemaître et al. 2017].

3.3. Otimização, treinamento e avaliação dos modelos preditivos

Para a construção dos modelos preditivos, a partir dos conjuntos de treino de cada uma das representações, foram utilizados os algoritmos de classificação *Decision Tree* (DT) e *Random Forest* (RF), do pacote *Scikit-learn*. Enquanto o algoritmo de árvore de decisão foi escolhido por apresentar bons resultados e prover modelos de fácil interpretação [Han et al. 2012], o RF, que é um *ensemble* de árvores de decisão, foi escolhido por apresentar o melhor desempenho em [Solis et al. 2018].

Porém, novamente com o intuito de oportunizar uma comparação mais justa, antes de treinar cada um dos 16 modelos (2 algoritmos x 8 conjuntos de treino), foi realizada, na **sexta etapa**, a otimização automática de seus hiperparâmetros, por meio do método *GridSearchCV*, disponível no pacote *Scikit-learn*. Nessa fase, buscou-se encontrar a melhor combinação de hiperparâmetros relacionados ao limite de altura (*max_depth*) e à quantidade mínima de exemplos por folha (*min_samples_leaf*) para cada modelo. Tais hiperparâmetros foram considerados com o intuito de evitar o problema de *overfitting* [Han et al. 2012] e, assim, garantir maior capacidade de generalização aos modelos.

Após escolhidas, as melhores combinações de hiperparâmetros foram utilizadas, na **sétima etapa**, para o treinamento dos modelos preditivos, considerando seus respectivos algoritmos de classificação e conjuntos de dados de treinamento. Por fim, na **oitava**

etapa, os modelos construídos foram avaliados sobre o conjunto de teste, composto pelos dados extraídos de alunos evadidos ou formados em 2020. Os modelos foram avaliados a partir das métricas padrão de precisão e revocação por classe, além da acurácia, sendo os resultados apresentados a seguir, na Seção 4.

4. Resultados e Discussões

Na Figura 2 são apresentados, em forma de mapa de calor, os resultados dos modelos preditivos baseados em diferentes algoritmos de classificação e formas de representação dos dados de treino. Observe que, para facilitar a análise, foram apresentados apenas os modelos que obtiveram os melhores resultados, considerando ou não a aplicação da sobreamostragem SMOTE. A partir dos oito modelos selecionados, percebe-se que a maioria foi beneficiada pela sobreamostragem dos dados de treino; e, dentre os três que não foram, dois se baseiam na representação MRE-MRF, que, originalmente, já demonstra balanceamento entre os registros de alunos evadidos (E) e formados (F). Além disso, é possível notar que a maioria dos modelos (6/8) apresentou melhor desempenho ao utilizar ao menos um hiperparâmetro com configuração diferente da considerada como padrão nos algoritmos do *Sckit-learn*. Isso demonstra a importância de se considerar o balanceamento dos dados de treinamento e a otimização de hiperparâmetros, para comparações mais justas entre os modelos/representações.

Conjunto de Treino				**Modelo**			**Resultados**				
#	Rep.	SMOTE	Registros (E | F)	Alg.	Máx. altura	Mín. exemplos por folha	Precisão Evadidos	Precisão Formados	Revocação Evadidos	Revocação Formados	Acurácia
1	MRE-MRF	Não	3497 | 3483	DT	Padrão	2%	0.837	0.715	0.754	0.807	0.777
2	MRE-MRF	Não	3497 | 3483	RF	Padrão	Padrão	0.827	0.767	0.820	0.775	**0.801**
3	URE-MRF	Sim	3483 | 3483	DT	Padrão	0.5%	0.865	0.675	0.684	0.859	0.760
4	URE-MRF	Sim	3483 | 3483	RF	8	Padrão	**0.920**	0.661	0.638	**0.927**	0.763
5	URE-URF	Não	1228 | 0444	DT	3	0.5%	0.834	0.712	0.752	0.804	0.775
6	URE-URF	Sim	1228 | 1228	RF	Padrão	Padrão	0.858	0.724	0.757	0.836	0.791
7	MRE-URF	Sim	3497 | 3497	DT	15	Padrão	0.724	0.756	0.860	0.570	0.735
8	MRE-URF	Sim	3497 | 3497	RF	15	Padrão	0.733	**0.847**	**0.923**	0.560	0.766

Figura 2. Resultados dos modelos para diferentes representações e algoritmos.

Atentando para os resultados das métricas de precisão, revocação e acurácia, percebe-se que os modelos baseados no algoritmo RF, em geral, apresentaram melhor desempenho que os treinados a partir do classificador DT, o que já era esperado, considerando que *ensembles* são algoritmos mais robustos, que combinam diferentes modelos de aprendizado de máquina para aumentar a capacidade de generalização do aprendizado e, assim, melhorar os resultados de predição [Lee and Chung 2019]. Já em relação às representações usadas no treinamento, pode-se notar que os modelos treinados a partir das representações URE-MRF e MRE-URF apresentaram, para a classe positiva de evasão, as melhores precisões e piores revocações; e as melhores revocações e piores precisões; respectivamente. Esses resultados não se mostram atrativos, ao se considerar que, no problema de previsão de evasão, a ponderação entre a revocação e a precisão da classe positiva é muito importante. Embora seja desejável recuperar o máximo possível de alunos em risco, uma alta revocação deixa de ser benéfica se, dentre os alunos preditos como evasores, existir uma parcela considerável de falsos positivos (baixa precisão), pois isso dispersaria a atenção dos alunos que realmente precisam de acompanhamento.

Dessa forma, o modelo RF, treinado a partir do conjunto de dados com representação MRE-MRF (segundo modelo da Figura 2), demonstrou o melhor comportamento, tendo sido capaz de predizer como de evasores 82.0% dos registros de teste que realmente se relacionavam a alunos evadidos (revocação da classe positiva); e de acertar a predição de 82.7% dos registros classificados como de alunos evadidos (precisão da classe positiva); o que resultou em uma taxa geral de acertos (acurácia) de 80.1%. Esses resultados confirmam, a partir de uma avaliação mais justa, o concluído por [Solis et al. 2018]: incluir registros de todos os semestres cursados pelos alunos (evadidos e formados) é a melhor forma de exemplificar os comportamentos de evasão e conclusão, para a construção de modelos preditivos. Por fim, embora apresente o segundo melhor desempenho geral, o modelo RF baseado na representação URE-URF (sexto modelo da Figura 2) também pode ser considerado uma opção viável, caso se trabalhe com uma quantidade de dados muito elevada e o tempo de treinamento seja um problema.

Embora baseie suas predições em um conjunto de modelos interpretáveis (DTs), um modelo RF é classificado, devido sua complexidade, como "caixa preta", sendo de difícil interpretação. Porém, para verificar indícios da importância das variáveis e de suas possíveis relações com a predição, foi gerada uma matriz de correlação, considerando o atributo alvo/classe (indicador de evasão) e os dez principais atributos do modelo RF, baseado na representação MRE-MRF. Essa matriz é apresenta na Figura 3, tendo os atributos preditivos ordenados por importância, de forma decrescente. Note que os valores de correlação, representados por cores na matriz, vão de -1 a 1, -1 significando uma correlação total e inversa, 0 a inexistência de correlação, e 1 uma correlação total e direta [Costa et al. 2020]. Dessa forma, pode-se observar que a média de notas, a porcentagem de aprovações, a média de frequências, e a porcentagem de carga horária integralizada são os quatro atributos mais importantes do modelo, todos possuindo uma forte correlação inversa com a evasão. Ou seja, quando menor o desempenho e a assiduidade do aluno, maior sua tendência de evasão. Da mesma forma, quanto menor a carga horária integralizada (quanto menos o aluno tiver avançado no curso), maior o risco de abandono.

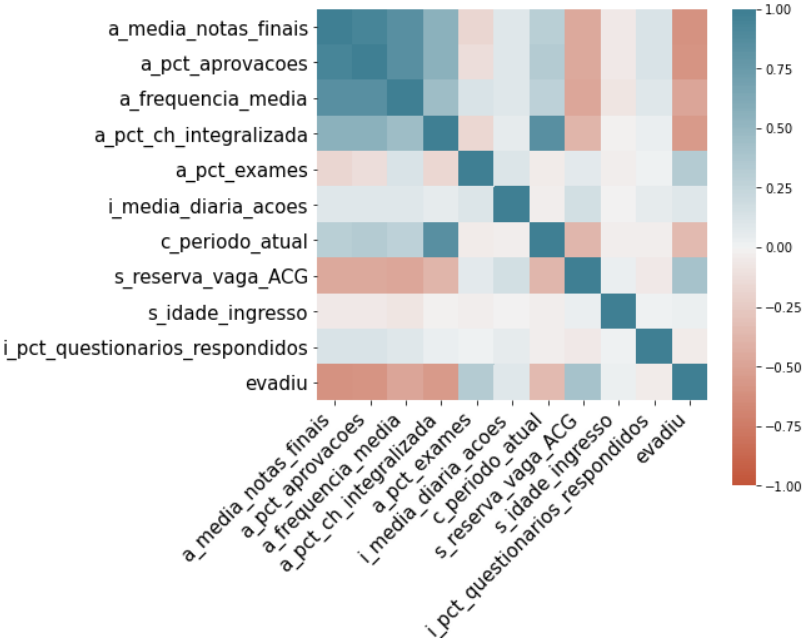

Figura 3. Matriz de correlação com os principais atributos do melhor modelo.

5. Conclusão e Considerações Finais

Neste trabalho foram avaliadas diferentes formas de representar/exemplificar o comportamento de alunos evadidos e formados no desenvolvimento de modelos genéricos, destinados a prever a evasão, em diferentes semestres e cursos, de estudantes de graduação, na modalidade presencial. Para isso, após extraídos, os dados passaram por um pré-processamento cuidadoso, no qual atributos absolutos foram transformados em atributos relativos, a fim de tornar as informações de alunos de diferentes cursos comparáveis entre si e, assim, propiciar uma maior capacidade de generalização aos modelos. Após, diferentes representações de dados de treino foram geradas, a fim de investigar se registros dos comportamentos dos alunos em semestres anteriores à evasão ou conclusão contribuem ou não para a caracterização dos padrões de abandono e conclusão. Além disso, como cada representação envolvia diferentes proporções de registros entre alunos evadidos e formados, foi realizada a sobreamostragem das bases de treino de cada representação, para uma avaliação mais justa.

Como resultado, a representação que considera registros de todos os semestres cursados pelos alunos, caracterizando, de forma acumulada e progressiva, seus comportamentos de evasão ou conclusão, demonstrou maior contribuição para o treinamento/aprendizagem dos modelos preditivos. Considerando o modelo de melhor resultado, gerado a partir do algoritmo de classificação *Random Forest*, a taxa de acertos das predições superou 80%, sendo esse um bom resultado, dada a complexidade do problema de evasão e de sua predição por meio de modelos genéricos. Além disso, a análise de correlações dos principais atributos do modelo indicaram que quanto maior/melhor a frequência e o desempenho de um aluno, e quanto mais ele tiver avançado no curso (maior a carga horária integralizada), menor sua tendência de evasão. Embora não permitam o entendimento de quais padrões/regras, especificamente, o modelo aprendeu, esses indícios vão ao encontro da realidade educacional, uma vez que é sabido que os semestres iniciais costumam acumular as maiores taxas de evasão e que o baixo desempenho acadêmico, possivelmente decorrente de dificuldades de aprendizagem, contribui para o abandono estudantil.

Em trabalhos futuros, outros algoritmos de aprendizado de máquina, assim como métodos para seleção automática de atributos, podem ser aplicados e avaliados, com o intuito de melhorar os resultados das predições de evasão. Adicionalmente, como modelos caixa-preta costumam apresentar desempenhos superiores, pode-se buscar um melhor entendimento dos padrões de evasão e das decisões de predição, por meio da adoção de técnicas destinadas a aumentar a transparência e a explicabilidade de modelos complexos. Por fim, pode ser desenvolvido um sistema que, baseado no modelo preditivo, analise o risco de abandono para alunos ativos/matriculados, a fim de auxiliar os gestores educacionais na implementação de medidas preventivas.

Agradecimentos: O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001, e do Programa Institucional de Incentivo à Qualificação Profissional em Programas Especiais (PIIQPPE) do IFFar. Os autores agradecem à CAPES, ao PPGC/UFPel e ao IFFar.

Referências

Baranyi, M., Nagy, M., and Molontay, R. (2020). Interpretable deep learning for university dropout prediction. In *Proceedings of the 21st Annual Conference on Information Technology Education*, SIGITE '20, page 13–19, New York, NY, USA. Association for Computing Machinery.

Brasil (1996). Diplomação, retenção e evasão nos cursos de graduação em instituições de ensino superior públicas. Technical report, Ministério da Educação, Comissão Especial de Estudos sobre a Evasão nas Universidades Públicas Brasileiras: ANDIFES; ABRUEM; SESu/MEC, Brasília, DF.

Böttcher, A., Thurner, V., Häfner, T., and Hertle, J. (2021). A data science-based approach for identifying counseling needs in first-year students. In *2021 IEEE Global Engineering Education Conference (EDUCON)*, pages 420–429.

Colpo, M. P., Primo, T. T., Pernas, A. M., and Cechinel, C. (2020). Mineração de Dados Educacionais na Previsão de Evasão: uma RSL sob a Perspectiva do Congresso Brasileiro de Informática na Educação. In *Anais do XXXI Simpósio Brasileiro de Informática na Educação*, pages 1102–1111, Porto Alegre, RS, Brasil. SBC.

Costa, A. G., Queiroga, E., Primo, T. T., Mattos, J. C. B., and Cechinel, C. (2020). Prediction analysis of student dropout in a computer science course using educational data mining. In *2020 XV Conferencia Latinoamericana de Tecnologias de Aprendizaje (LACLO)*, pages 1–6.

Han, J., Kamber, M., and Pei, J. (2012). *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers, Waltham, MA, 3rd edition.

Kang, K. and Wang, S. (2018). Analyze and predict student dropout from online programs. In *Proceedings of the 2nd International Conference on Compute and Data Analysis*, ICCDA 2018, page 6–12, New York, NY, USA. Association for Computing Machinery.

Lee, S. and Chung, J. (2019). The machine learning-based dropout early warning system for improving the performance of dropout prediction. *Applied Sciences (Switzerland)*, 9(15).

Lemaître, G., Nogueira, F., and Aridas, C. K. (2017). Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, 18(17):1–5.

McKinney, W. (2010). Data Structures for Statistical Computing in Python. In Stéfan van der Walt and Jarrod Millman, editors, *Proceedings of the 9th Python in Science Conference*, pages 56 – 61.

Mduma, N., Kalegele, K., and Machuve, D. (2019). A survey of machine learning approaches and techniques for student dropout prediction. *Data Science Journal*, 18:14:1–10.

Nagy, M. and Molontay, R. (2018). Predicting dropout in higher education based on secondary school performance. In *2018 IEEE 22nd International Conference on Intelligent Engineering Systems (INES)*, pages 389–394.

Ortigosa, A., Carro, R. M., Bravo-Agapito, J., Lizcano, D., Alcolea, J. J., and Blanco, O. (2019). From lab to production: Lessons learnt and real-life challenges of an

early student-dropout prevention system. *IEEE Transactions on Learning Technologies*, 12(2):264–277.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Pontili, R., Staduto, J., and Henrique, J. (2018). Abandono e atraso escolar e sua relação com indicadores socioeconômicos: uma análise para a região sul do brasil. *Gestão & Regionalidade*, 34(101):4–22.

Santos, J. R. and Zaboroski, E. (2020). Ensino remoto e pandemia de COVID-19: Desafios e oportunidades de alunos e professores. *Interacções*, 16(55):41–57.

Silva Filho, R. L. L., Motejunas, P. R., Hipolito, O., and Lobo, M. B. C. M. (2007). A evasão no ensino superior brasileiro. *Cadernos de Pesquisa*, 37(132):641–659.

Solis, M., Moreira, T., Gonzalez, R., Fernandez, T., and Hernandez, M. (2018). Perspectives to predict dropout in university students with machine learning. In *2018 IEEE International Work Conference on Bioinspired Intelligence (IWOBI)*, pages 1–6.

Yu, R., Lee, H., and Kizilcec, R. F. (2021). Should college dropout prediction models include protected attributes? In *Proceedings of the Eighth ACM Conference on Learning @ Scale*, L@S '21, page 91–100, New York, NY, USA. Association for Computing Machinery.