2021 XVI Latin American Conference on Learning Technologies (LACLO) | 978-1-6654-2358-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/LACLO54177.2021.00020

Model for Prediction of Student Dropout in a Computer Science Course

Alexandre G. Costa
Universidade Federal de Pelotas
Centro de Desenvolvimento Tecnológico
Pelotas, Brazil
alexandre.costa@inf.ufpel.edu.br

Julio C. B. Mattos
Universidade Federal de Pelotas
Centro de Desenvolvimento Tecnológico
Pelotas, Brazil
julius@inf.ufpel.edu.br

Tiago Thompsen Primo
Universidade Federal de Pelotas
Centro de Engenharias
Pelotas, Brazil
tiago.primo@inf.ufpel.edu.br

Cristian Cechinel
Universidade Federal de Santa Catarina - UFSC
Araranguá-SC, Brasil
contato@cristiancechinel.pro.br

Roberto Muñoz
Universidad de Valparaiso, Chile
roberto.munoz@uv.cl

***Abstract*—This work presents a model that can predict the student's risk of dropout using data from the first three semesters attended by Computer Science Undergraduate students. Nowadays, Educational Management Systems store a large amount of data from the interaction of not only students and professors but also of students and the educational environment. Analyze and find patterns manually from a huge amount of data is hard, so Educational Data Mining (EDM) is widely used. This work uses the CRISP-DM methodology and data from Computer Science Undergraduate students from Federal University of Pelotas, Brazil. The results are shown for three algorithms: the Decision Tree algorithm presents a precision of 84.80%, a Recall of 85.80% and an AUC of 77.24%; the Random Forest algorithm presents a precision of 88.57%, a Recall of 90.14% and an AUC of 83.22%; the Logistic Regression algorithm presents a precision of 71.24%, a Recall of 94.28% and an AUC of 58.39%. The results indicate that it is possible to use a prediction model using only the data from the first three semesters of the course.**

***Index Terms*—educational data mining, learning analytics, prediction techniques.**

I. Introdução

O uso constante de Tecnologia da Informação e Comunicação (TICs) em diversas áreas vêm gerando um grande volume de dados. Tecnologias como redes sociais, AVAs, aplicativos embarcados, sensores e sistemas de informação em geral são alguns exemplos de recursos que vem aumentando o número de dados das mais diversas naturezas [1].

Atualmente, áreas como a educação produzem uma grande quantidade de dados relacionados a alunos e professores. Os dados são provenientes de sistemas de gestão de interações acadêmicas, projetos de pesquisa, ensino ou extensão originando volumes consideráveis de dados.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001. À Fundação de Amparo à pesquisa do Estado do RS (FAPERGS) e SEBRAE/RS 03/2021 - PROEdu. R. Munoz has been supported by ANID/CONICYT Fondecyt Regular 1211905.

A partir desse volume de dados está sendo proposta a analise da evasão escolar. A evasão é um problema que atinge não só as Instituições de Ensino Superior (IES) privadas, mas também as públicas. Segundo dados do INEP [2], em 2017, o índice de matrículas desvinculadas em todo o Brasil foi de 16,41%. Já para as IES públicas esse índice no mesmo período foi de 11,56%.

Ao comparar os índices da Universidade Federal de Pelotas com os do INEP [2] estes índices não mudam muito. Em 2017 o índice de matriculas desvinculadas na UFPEL foi de 11,53% que está abaixo do índice de 11,56% apresentado pelo INEP [2]. Observando a evasão curso a curso no ano de 2017 nota-se dados alarmantes. Existem cursos onde a taxa de evasão é de 29,07%. Esse fenômeno é conhecido na estatística como paradoxo de Simpson, uma tendência aparece em um determinado grupo de dados e desaparece quando estes grupo de dados são combinados. Dessa forma, olhando para os cursos individualmente vários deles apresentam uma taxa de evasão bem elevadas, mas quando essas taxas são combinadas a taxa de evasão não é significativamente relevante [3].

Embora com ajuda de ferramentas computacionais, analisar essa crescente quantidade de dados não é um trabalho humanamente viável. Para ajudar nesta questão a área de Descoberta de Conhecimento em Base de Dados (*Knowledge Discovery in Database* - KDD) é focada em extrair conhecimento em cima de grandes volumes de dados. O termo mais conhecido relacionado a essa área é a Mineração de Dados (MD) que é uma das etapas do processo de KDD.

Segundo [4] Mineração de Dados (MD) aplicada à educação é um campo interdisciplinar emergente mais conhecido como Mineração de Dados Educacionais (MDE). [5] define MDE como a área de investigação científica centrada no desenvolvimento de métodos para fazer descobertas dentro dos tipos de dados que vêm de ambientes educacionais e usando esses métodos para entender melhor as questões relacionadas aos alunos e a aprendizagem deles.

O objetivo deste trabalho é fornecer recursos para que

978-1-6654-2358-8/21/$31.00 ©2021 IEEE
DOI 10.1109/LACLO54177.2021.00020

os gestores acadêmicos possam criar políticas para fazer o enfrentamento a evasão escolar. Recursos esses que serão gerados a partir dos algoritmos resultantes deste trabalho.

Neste trabalho é explorada a utilização de MDE visando classificar e identificar perfis de alunos com tendência a evadir utilizando apenas dados acadêmicos dos três primeiros semestres de um curso presencial. Segundo [1] uma tarefa de classificação possui dois grupos: a) um grupo contem normalmente um atributo apenas que vai servir para fazer a predição de um valor (atributo-alvo) e b) Outro grupo corresponde aos atributos que vão servir para fazer a predição do valor (atributos de predição). Tarefas de classificação são largamente utilizadas para fazer a predição de alunos em risco de evasão escolar, como é discutido neste trabalho.

Como uma de suas motivações, este trabalho apresenta a modelagem dos dados produzidos pelos sistemas acadêmicos, para que estes se transformem em informações, e consequentemente conhecimento, sobre o perfil dos estudantes da universidade. Assim, este trabalho pretende auxiliar nas políticas de combate aos índices de evasão apresentados na Universidade Federal de Pelotas, principalmente nos cursos de exatas e engenharias. Além disso, isto é possibilitado pois existe uma grande quantidade de dados históricos de cursos de graduação presencias da UFPEL.

Este trabalho apresenta a investigação dos motivos que levam o alunos evadir com ajuda de técnicas de MDE através dos dados acadêmicos dos alunos do curso de Ciência da Computação da Universidade Federal de Pelotas que foi escolhido preliminarmente por possuir elevados índices de evasão. Para este propósito, foi analisado e coletado dados dos 3 primeiros semestres de alunos do curso de 2000 a 2018, para responder a seguinte questão de pesquisa: Quais os classificadores e técnicas podem ser utilizados nessa tarefa?

Este trabalho está organizado da seguinte forma. A seção 2 apresenta os trabalhos relacionados a predição de alunos em risco de evasão utilizando técnicas de MDE. A seção 3 descreve a metodologia CRISP-DM, utilizada para como guia para este trabalho. A seção 4 apresenta os experimentos e resultados alcançados neste trabalho, sendo dividida em duas grandes partes que apresentam a caracterização da evasão e análise de predição. Por fim, a seção 5 apresenta as conclusões e trabalhos futuros.

II. Trabalhos relacionados

O uso de técnicas de mineração de dados educacionais é consideravelmente recente e a metodologia utilizada ainda não é bem definida [6]. Além disso, questões como quais atributos devem ser selecionados ou qual algoritmo deve ser utilizado não são unânimes.

Manhães et. al. [7] provaram que é possível identificar alunos em risco de evasão através das primeiras notas semestrais dos alunos ingressantes. A base de dados utilizada no trabalho foi do sistema acadêmico da instituição e contou com alunos que cursaram Engenharia Civil na UFRJ de 1994 a 2005. O trabalho consistiu em 3 experimentos, onde foi modificado apenas a forma de treinamento dos algoritmos 10 fold *Cross-validation*, *train/test percentage split (data randomized)* e *supplied test set*, em que 2/3 para treinamento e o restante para teste. Cada experimento foi submetido a 10 algoritmos utilizados em trabalhos relacionados. Os experimentos alcançaram acurácia média entre 75% e 80%, além disso a predição incorreta de risco de evasão foi considerada como erro grave do classificador.

No trabalho de [8] é apresentado um estudo de fatores envolvidos no fenômeno de evasão escolar e descrevem a utilização de um sistema para MDE e Learning Analytics (LA) durante 18 meses em cursos de graduação na modalidade de Educação a Distância. Ao todo foram executados 4 experimentos onde foram acompanhados 603, 250, 925 e 713 alunos. Para cada estudo de caso que trata o trabalho foi utilizado a técnica *RNA Multilayer Perceptron*. O melhor resultado com relação a predição da evasão foi no experimento 4 onde a melhor média de acertos foi de 83,7%.

Santo, Siebra e Oliveira [9] propõem identificar precocemente estudantes de graduação EAD em risco de evasão através de técnicas de mineração de dados. Os autores coletaram dados do AVA e do Sistema de Controle Acadêmico (SCA). Na base de dados do AVA foram selecionadas as notas intermediárias da disciplina durante o semestre e no SCA foram utilizadas a situação da disciplina, média da disciplina, quantidade de reprovação no período ou semestre e média no período. Os dados foram submetidos a três algoritmos de classificação baseados em Arvore de Decisão (*SimpleCart*, J48 e *ADTree*), para construção do modelo preditivo. Os autores relataram que obtiveram um acurácia média de 80% na predição da evasão, utilizando a primeiras notas semestrais dos estudantes. Ainda destacam que no AVA chegaram a obter uma acurácia de 98,47% na predição do desempenho de uma disciplina.

Detoni, Cechinel e Araújo [10] apresentaram uma metodologia para classificar alunos usando apenas a contagem de interação de cursos EAD. Foram utilizados dados de disciplinas do primeiro e segundo semestre de dois cursos EAD que ocorreram em 2013. Os autores criaram dois modelos um que utilizou apenas o número absoluto de iterações dos alunos, e outro que foi adicionado atributos derivados do número de iterações. Foram aplicados quatro modelos de classificação aos dados coletados. No trabalho foi possível provar que a abordagem de atributos derivados do número de iterações dos alunos obteve resultados superiores a abordagem de trabalhos anteriores.

Queiroga, Cechinel e Araújo [11] apresentaram os resultados iniciais de um trabalho voltado para a predição precoce da evasão de alunos em um curso EAD utilizando MDE. Os autores coletaram os logs das iterações dos alunos de duas turmas de diferentes polos. Foram realizados 3 experimentos onde o primeiro e segundo experimentos usaram os dados de turmas de diferentes cidade e o terceiro reuniu os 2 conjuntos de dados. A cada experimento foi aplicado um conjunto de 9 algoritmos de aprendizagem de máquina usando a opção de *Cross-validation*. A conclusão do trabalho foi de que é viável a predição precoce do risco de evasão através da análise dos

logs das 4 primeiras semanas de uma turma.

Em [12] foi discutida a importância de atividades extracurriculares para prever o abandono escolar de estudantes de dois cursos de Bacharel em Ciências (Engenharia e Negócios). Foram coletados dados de 4840 alunos. Dois modelos foram treinados, uma incluindo todos os dados e outra removendo notas e créditos obrigatórios do valor das atividades. Ambos os modelos foram treinados e validados usando *Cross-validation*. A primeira árvore de decisão obteve a melhor precisão (93,94%). Para o segundo modelo obteve uma precisão de 79,29%. Os autores relataram que embora a previsão de abandono com dados acumulados mostre um melhor desempenho, o segundo modelo ajuda a resolver o problema de disponibilidade de dados dos alunos.

Kantorski et. al. [13] propõem prever a evasão de cursos de graduação presenciais em universidades públicas. Foram extraídos dados pessoais, acadêmicos, sociais e econômicos de alunos e construídos modelos de predição através de algoritmos de aprendizagem de máquina. Os autores destacam que a vantagem da proposta foi a otimização dos resultados pela combinação de vários modelos de mineração de dados para gerar uma única predição e isso permite um resultado mais abrangente. Nos testes alcançaram uma acurácia de 98% e mais de 70% de sucesso na predição de alunos que evadiram do curso.

Lanes e Alcântra [14] apresentaram um estudo que visa identificar estudantes que apresentam risco de evasão a partir do seu primeiro ano no curso de graduação. Os experimentos foram realizados com informações extraídas do sistema acadêmico da FURG. O conjunto de dados contou com 916 registros de 12 cursos de graduação de áreas distintas. Os dados foram discretizados e categorizados para gerar o *dataset* final. Foi utilizado a ferramenta Weka e aplicado o algoritmo J48 para processar o *dataset* e obter a arvore de decisão. Os resultados mostram que os potenciais alunos em risco de evadir podem ser identificados com acurácia de 90,7% usando o algoritmo J48.

A principal diferença deste trabalho é a aplicação de técnicas de visualização e MDE em cima de dados reais de alunos de um curso de graduação presencial. Outro ponto a ser levado em consideração é que utilizamos apenas um curso para não generalizar e evitar problemas como o paradoxo de Simpson. Também foi feito uma caracterização da evasão em cima de dados de 20 anos do curso de Ciência da Computação da UFPEL. Por fim, diferentemente dos outros trabalhos utilizamos dados apenas dos semestres iniciais. Em trabalho anterior [15] desdes autores também foram utilizados dados de alunos do curso de Ciência da Computação, porém neste trabalho o modelos foi consolidado. A diferença entre os trabalhos esta na abrangência dos dados, onde o trabalho anterior utilizou dados de alunos até 2020 e o corrente trabalho utiliza dados de alunos até 2018, pois verificou-se que não era possível um aluno ter notas até o terceiro semestre se ele entrou em 2020. Outra diferença foi no treinamento dos algoritmos, onde diferente do trabalho anterior que foi utilizado *split train test*, neste foi utilizado *cross-validation* como método de treino e teste. Este método evita que o modelos sofram com sobre-ajuste.

III. Metodologia

Este trabalho seguiu a metodologia CRISP-DM que tem como principal objetivo fornecer uma direção para conduzir o processo de KDD [1]. A seguir será apresentado uma breve descrição das seis fases da metodologia:

1) **Compreensão do negócio:** Esta é a fase onde se deve identificar o problema a ser resolvido. Esta fase compreende também uma descrição do *background*, dos objetivos e também dos critérios de sucesso [16].
2) **Compreensão dos dados:** É a fase responsável por fazer a análise exploratória de dados (AED). Esta fase tem que dizer como os dados foram adquiridos, qual o seu formato, qual foi a quantidade de dados, descrever cada atributo selecionado, fazer visualizações dos dados e além disso qualquer informação pertinente aos dados.
3) **Preparação dos dados:** Compreende as atividades de pré-processamento dos dados para a próxima fase. Normalmente se faz a seleção, limpeza, formatação dos dados, e ainda gera-se novos atributos derivados dos atributos existentes.
4) **Modelagem:** Corresponde a fase de aplicação dos algoritmos de mineração de dados selecionados sobre os dados preparados. É a etapa de mineração de dados do processo de KDD [1]. Nessa fase é criado um modelo para testar a sua qualidade e validade. É comum usar a taxa de erro como medida de qualidade do modelo em aprendizado supervisionado [16].
5) **Avaliação:** Consiste em avaliar o modelo gerado, examinando os passos seguidos e validando se realmente foram alcançados os objetivos elencados na fase de compreensão do negocio [16]. A partir da avaliação é possível propor revisões das fases anteriores e redefinir os próximos passos [1].
6) **Desenvolvimento:** É a fase onde se faz o planejamento e acompanhamento a serem realizadas com o modelo gerado pelas fases anteriores [1]. Esta fase não faz parte do escopo deste trabalho.

Neste trabalho foram utilizados os dados de alunos registrados no Cobalto, que é o sistema gestão da Universidade Federal de Pelotas. O Cobalto é um sistema integrado de gestão que faz a gestão acadêmica da universidade. Este sistema possui dados históricos de alunos, importados do sistema anterior da IES, denominado GOL.

IV. Resultados e Discussão

Esta seção apresenta os resultados conforme a metodologia proposta na seção III.

A. Caracterização da Evasão

Os dados extraídos do sistema correspondem aos alunos do curso de Ciência da Computação que ingressaram entre os anos 2000 e 2018. É importante destacar que os dados dos alunos correspondem aos três primeiros semestres do curso.

A Tabela I apresenta o número de alunos para a situação final ou atual. A situação “Cursando” representa todos os alunos que ainda estão dentro do período de integralização curricular e não sairam do curso. Já a situação “Retido” são todos os alunos que já passaram do período de integralização curricular e não tem saída. As situações “Formado” e “Evadido” são alunos que já tem uma saída registrada.

TABLE I
QUANTITATIVO DOS DADOS COLETADOS

Aluno				
Total	Cursando	Evadido	Formado	Retido
1514	286	786	330	112
100,00%	18,89%	51,91%	21,80%	7,40%

A partir da AED foram geradas algumas visualizações, onde são destacadas duas neste artigo. A figura 1 apresenta uma série histórica do curso de Ciência da Computação. O eixo x representa o ano e semestre de ingresso do aluno e o eixo y o número de alunos no período agrupados pela situação final do aluno. De 2000 até 2007 o curso tinha uma população de alunos formados maior do que a de alunos que evadiram. Depois de 2007 o número de alunos formados começa a ser inferior ao número de evadidos, além disso o número de alunos retidos fica cada vez mais significante com o passar do tempo.

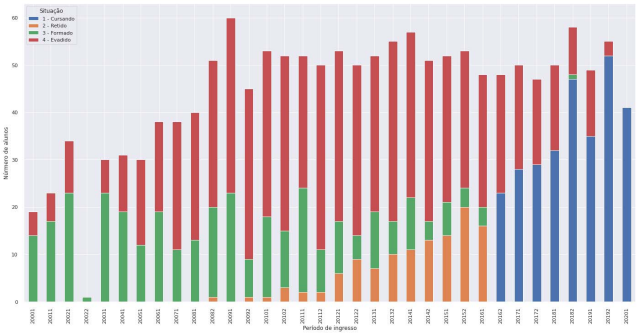

Fig. 1. Número de alunos e respectiva situação pelo período de ingresso.

A Figura 2 apresenta o número total de alunos no semestre de saída do aluno, agrupados em conjunto a respectiva situação. É importante notar que o número de alunos evadidos diminui ao passo que o semestre aumenta. Outro ponto que vale destacar é que o número de alunos formados cresce significativamente a partir do oitavo semestre e chega ao ponto mais alto no décimo semestre. Isso se deve ao período de integralização curricular que em alguns currículos foi de 8 semestres e outros de 9 semestres. Ainda é possível destacar que 19,33% evade até o terceiro semestre e 32,45% depois do terceiro semestre.

A Figura 3 apresenta a média geral dos alunos por semestre (três semestres iniciais).No terceiro semestre alunos que conseguiram concluir o curso tem uma média geral mais alta que alunos nas outras situações. Por outro lado, alunos que evadiram tiveram a média mais baixa nos 3 períodos iniciais do curso de Ciência da Computação.

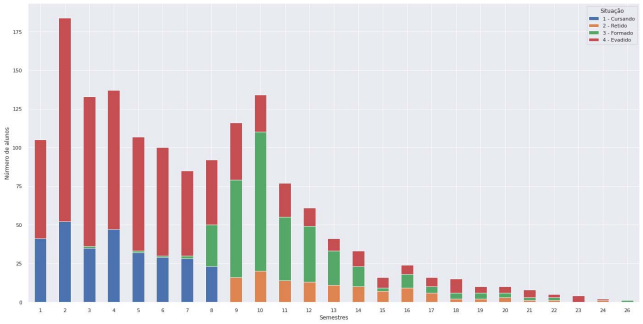

Fig. 2. Número de alunos e respectiva situação pelo semestre de saída do aluno.

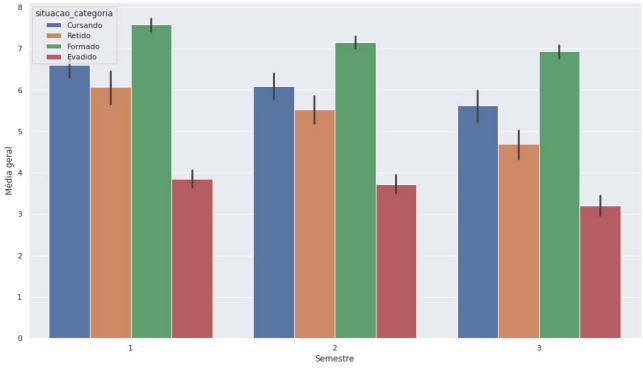

Fig. 3. Média geral dos alunos do Curso de Ciência da Computação nos três primeiros semestres.

Após a fase de compreensão dos dados iniciou-se a de preparação dos dados, nesta fase primeiro foi feita a separação dos atributos. A escolha dos atributos foi baseada nos atributos de trabalhos encontrados nos trabalhos relacionados. Na fase de limpeza dos dados foram removidos os valores nulos dos atributos. Em seguida foram removidas todas as linhas que possuíam disciplinas na situação dispensado, por não ser uma atividade realizada no curso pelo o aluno. Foi tomado o cuidado para que os atributos tivessem os seu valores de maneira homogenia.

Além disso foram gerados novos atributos derivados dos atributos coletados do sistema tais como: Idade, período que o aluno ingressou, semestre que o aluno cursou uma determinada disciplina, médias do primeiro, segundo e terceiro semestre, média dos três primeiros semestres, número de disciplinas cursadas no primeiro, segundo e terceiro semestre, e a média do número de disciplina cursadas no três primeiros semestres. Esses atributos foram gerados para tentar contextualizar melhor os dados retirados do sistema e obter um melhor desempenho dos algoritmos de classificação.

Após todo o processo de preparação dos dados o conjunto de alunos do estudo se limitou a 744 alunos, que era de 1514 alunos, e 19 atributos selecionados para o estudo incluindo o atributo alvo, conforme a Tabela II. Destes alunos, 66,5% (495) estavam na situação de evasão do curso e um total de

TABLE II
RELAÇÃO DE ATRIBUTOS SELECIONADOS.

Atributo	Descrição
med_1_sem	média do primeiro semestre
med_2_sem	média do segundo semestre
med_3_sem	média do terceiro semestre
evadiu	indica se o aluno evadiu ou não
genero_F	indica que o aluno é do gênero feminino
genero_M	indica que o aluno é do gênero masculino
flg_escola_publica	indica que o aluno vem de escola pública
flg_curso_superior	indica que o aluno possuía curso superior anterior
flg_beneficio	indique que o aluno possui benefício
naturalidade_ESTADO	indica que o aluno nasceu no Rio Grande do Sul
naturalidade_MESOPELOTAS	Indica que o aluno nasceu na meso-região de pelotas
naturalidade_MICROPELOTAS	indica que o aluno nasceu na micro-região
naturalidade_PAIS	indica que o aluno nasceu fora do Rio grande do sul, mas no Brasil
naturalidade_PELOTAS	indica que o aluno nasceu em Pelotas
idade_ate_17	idade de ingresso do aluno é menor que 17
idade_de_18_a_20	idade de ingresso do aluno está entre 18 e 20 anos
idade_de_21_a_23	idade de ingresso do aluno está entre 21 e 23 anos
idade_de_24_a_26	idade de ingresso do aluno está entre 24 e 26 anos
idade_de_27_a_56	idade de ingresso do aluno está entre 27 e 56 anos

33,5% (249) estavam na situação de conclusão do curso de 2000 a 2018.

B. Análise da Predição

Esta subseção corresponde a fase de Modelagem da metodologia proposta na seção III que é a fase onde os algoritmos de aprendizagem de máquina (AM) são aplicados ao conjunto de dados extraído do sistema acadêmico.

Para a implementação das rotinas mencionadas foi utilizado o *Cloud Google Colab*, *Python*, *Scikit-learn* e *pandas*. Para fazer o treinamento e teste do conjunto de dados foi utilizado o método de Validação Cruzada com K Conjuntos Estratificada (*Stratifield K-fold Cross-validation*) que se assemelha a Validação Cruzada com K Conjuntos, porém quando gera os subconjuntos mantém a mesma proporção do atributo classe [1]. Foi utilizado o método *cross_val_predict* da biblioteca *Scikit-learn* que por padrão faz estratificação dos conjuntos de dados e também o número de conjuntos foi configurado para 10.

Neste trabalho os dados foram processados pelos algoritmos de Floresta Aleatória, Árvore de decisão e Regressão Logística. Os algoritmos de Árvore de decisão e Floresta Aleatória foram escolhidos por gerarem modelos de fácil interpretação e Regressão Logística foi escolhido pelo seu desempenho e relevância encontrado em outros trabalhos. Na Tabela III são apresentados os resultados da execução dos três algoritmos utilizados neste trabalho. A partir da tabela é possível verificar também que todos os algoritmos tiveram algum grau de aprendizagem, visto que todos tiveram uma acurácia maior do que 66,5% que é a quantidade de alunos evadidos na base.

A figura 4 mostra a *Feature Importance* do modelo de Arvore de decisão. O atributo com maior relevância para prever se o aluno evadiu ou não foi a média do terceiro semestre, seguida pela média do segundo semestre e alunos que nasceram fora do Rio Grande do Sul. Outros atributos como gênero, se aluno teve curso superior anterior, e outros não tiveram muita relevância para a predição.

TABLE III
RESULTADO DA EXECUÇÃO DOS ALGORITMOS.

Algoritmo	Acurácia	Precisão	Revocação	AUC
Árvore de decisão	80.16%	84.80%	85.80%	77.24%
Floresta Aleatória	85.58%	88.57%	90.14%	83.22%
Regressão Logística	70.63%	71.24%	94.28%	58.39%

A figura 5 mostra a *Feature Importance* do modelo de Floresta Aleatória. O atributo com maior relevância para prever se o aluno evadiu ou não foi a média do terceiro semestre, seguida pela média do primeiro e segundo semestres, diferente do modelo de arvores de decisão a naturalidade do aluno não teve tanta relevância para a predição.

A figura 6 mostra a *Feature Importance* do modelo de Regressão Logística. O atributo com maior relevância para prever se o aluno evadiu ou não foi a média do terceiro e segundo semestres, seguidos pela alunos com idade entre 18 e 20 anos.

Os resultados do experimento mostraram que é possível fazer a predição de alunos em risco de evasão através das notas dos 3 primeiros semestres com uma acurácia de até 85,58%.

Para determinar qual modelo teve o melhor desempenho foi utilizada a métrica AUC, pois [17] mostram que a AUC é uma métrica superior a acurácia para esse tipo de problema. O algoritmo que obteve os melhores resultados foi o de Floresta Aleatória com 83,22% na área sob a curva ROC (AUC).

V. CONCLUSÕES

Este trabalho apresentou os resultados para a predição da evasão de alunos através de dados dos três primeiros semestres do curso de Ciência da Computação da UFPEL nos períodos de 2000 a 2018. Para fazer essa classificação foi utilizado o processo de KDD e algoritmos de aprendizagem de máquina.

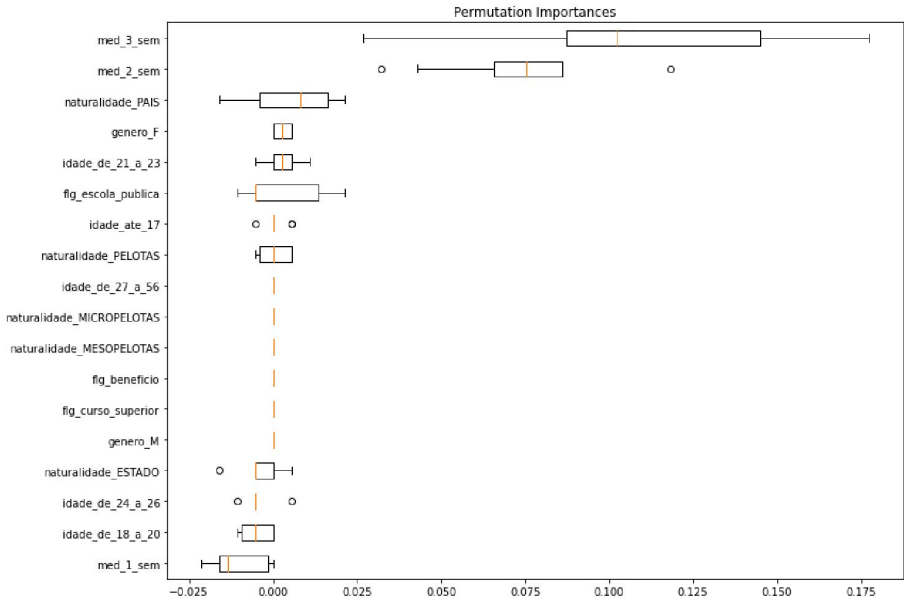

Fig. 4. *Feature Importance* do modelo de Árvore de Decisão.

Para a questão de pesquisa, foi apresentada uma acurácia de 85,58% e uma AUC de 83,22% para a predição de alunos em risco de evadir utilizando os dados pessoais e acadêmicos, considerando dados 744 alunos. O melhor resultado foi do algoritmo de Floresta Aleatória com uma revocação de 90,14% e um precisão de 88,57%.

Como trabalhos futuros pretende-se expandir a análise deste modelo de predição para outros cursos, primeiramente da área de exatas e engenharias e após demais cursos. Desta forma, será possível analisar a viabilidade ou não deste modelo em outros cursos e caso necessário adaptá-lo. Além disso, pretende-se desenvolver uma aplicação que forneça para os gestores de cursos a relação de alunos em risco de evadir para ajuda-los a planejar politicas de combate a evasão da Universidade Federal de Pelotas.

REFERENCES

[1] R. Goldschmidt, E. Bezerra, and E. Passos, "Data mining: conceitos, técnicas, algoritmos, orientações e aplicações," *Rio de Janeiro-RJ: Elsevier*, pp. 56–60, 2015.

[2] Inep, "Instituto nacional de estudos e pesquisas educacionais anísio teixeira. sinopses estatísticas da educação superior - graduação," Aug. 2018. [Online]. Available: http://portal.inep.gov.br/web/guest/sinopses-estatisticas-da-educacao-superior

[3] C. H. Wagner, "Simpson's paradox in real life," *The American Statistician*, vol. 36, no. 1, pp. 46–48, 1982.

[4] K. Koedinger, K. Cunningham, A. Skogsholm, and B. Leber, "An open repository and analysis tools for fine-grained , longitudinal learner data," *Proceedings of Int. Conference on Educational Data Mining*, pp. 157–166, 2008.

[5] R. Baker *et al.*, "Data mining for education," *International encyclopedia of education*, vol. 7, no. 3, pp. 112–118, 2010.

[6] G. W. Dekker, M. Pechenizkiy, and J. M. Vleeshouwers, "Predicting Students Drop Out: A Case Study," *the 2nd International Conference on Educational Data Mining*, pp. 41–50, 2009. [Online]. Available: http://www.win.tue.nl/ mpechen/research/edu.html.

[7] L. M. B. Manhães, S. M. S. da Cruz, R. J. M. Costa, J. Zavaleta, G. Zimbrão, S. M. S. da Cruz, R. J. M. Costa, J. Zavaleta, and G. Zimbrão, "Previsão de Estudantes com Risco de Evasão Utilizando Técnicas de Mineração de Dados," *Anais do XXII SBIE - XVII WIE*, pp. 150–159, 2011. [Online]. Available: http://www.br-ie.org/pub/index.php/sbie/article/view/1585/1350

[8] S. J. Rigo, W. Cambruzzi, J. L. Barbosa, and S. C. Cazella, "Aplicações de mineração de dados educacionais e learning analytics com foco na evasão escolar: oportunidades e desafios," *Revista Brasileira de Informática na Educação*, vol. 22, no. 1, pp. 132–146, 2014.

[9] R. N. dos Santos, C. d. A. Siebra, and E. S. Oliveira, "Uma Abordagem Temporal para Identificação Precoce de Estudantes de Graduação a Distância com Risco de Evasão em um AVA utilizando Árvores de Decisão," *Anais dos III Congresso Brasileiro de Informática na Educação (CBIE 2014)*, vol. 1, no. Cbie, p. 262, 2014.

[10] D. Detoni, C. Cechinel, and R. Araújo, "Modelagem e predição de reprovação de acadêmicos de cursos de educação a distância a partir da contagem de interações." *Revista Brasileira de Informática na Educação*, vol. 23, no. 3, 2015.

[11] E. Queiroga, C. Cechinel, and R. Araújo, "Um estudo do uso de contagem de interações semanais para predição precoce de evasão em educação a distância," in *Anais dos Workshops do Congresso Brasileiro de Informática na Educação*, vol. 4, no. 1, 2015, p. 1074.

[12] T. Hasbun, A. Araya, and J. Villalon, "Extracurricular activities as dropout prediction factors in higher education using decision trees," in *Advanced Learning Technologies, IEEE 16th International Conference on*, 2016, pp. 242–244.

[13] G. Kantorski, E. G. Flores, J. Schmitt, I. Hoffmann, and F. Barbosa, "Predição da evasão em cursos de graduação em instituições públicas," in *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*, vol. 27, no. 1, 2016, p. 906.

[14] M. Lanes and C. Alcântara, "Predição de alunos com risco de evasão: estudo de caso usando mineração de dados," in *Simpósio Brasileiro de Informática na Educação-SBIE*, vol. 29, no. 1, 2018, p. 1921.

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

[16] G. Melo and C. Viglioni, "Metodologia Para Previsão De Demanda Ferroviária," 2007.

[17] Jin Huang and C. X. Ling, "Using auc and accuracy in evaluating learning algorithms," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 3, pp. 299–310, 2005.

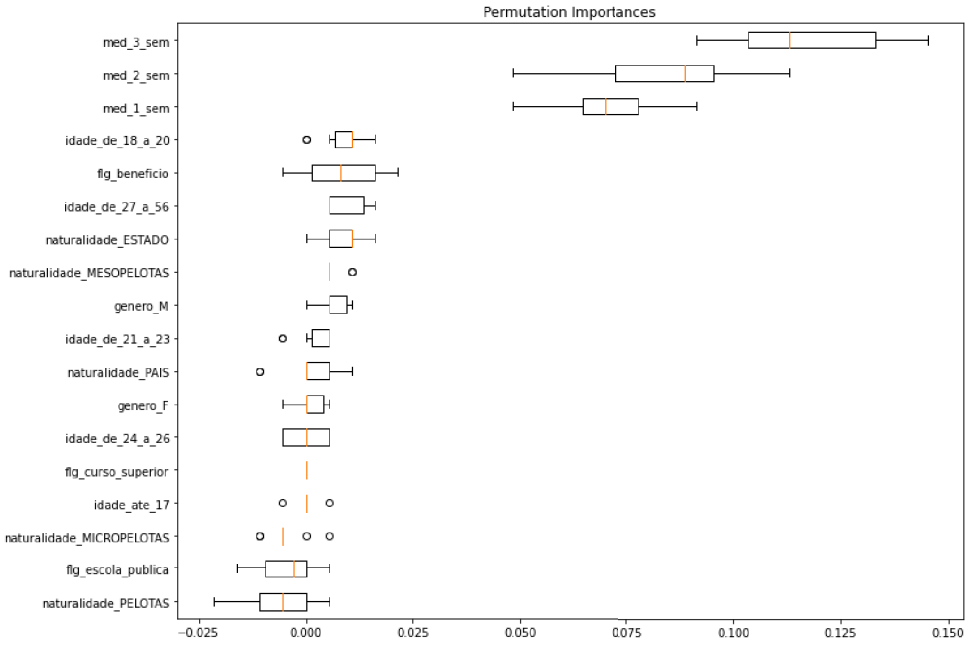

Fig. 5. *Feature Importance* do modelo de Floresta Aleatória.

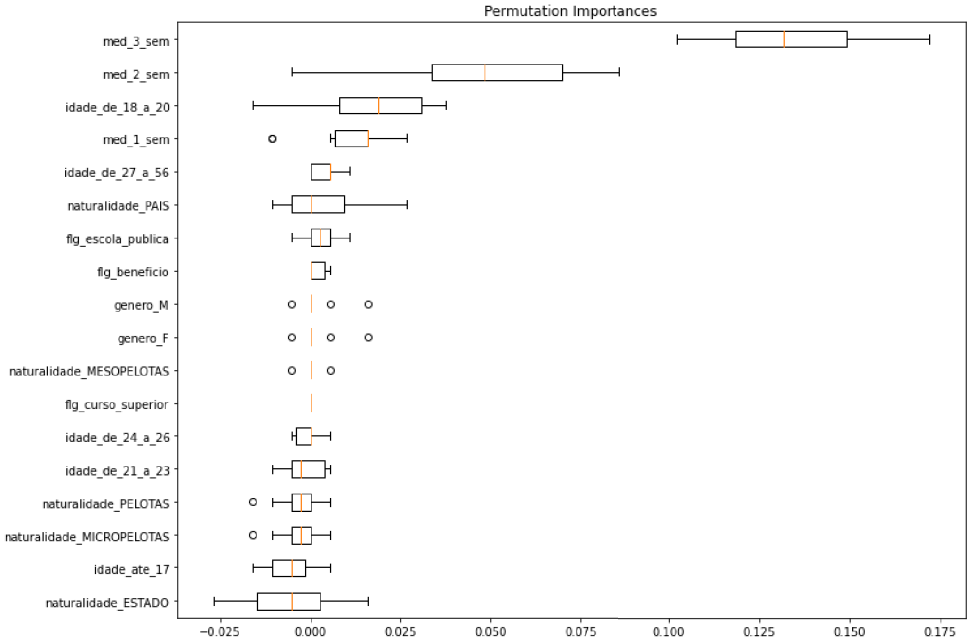

Fig. 6. *Feature Importance* do modelo de Regressão Logística.