Revista de Sistemas de Informação da FSMA
n. 29 (2022) pp. 12-24

SISTEMAS DE INFORMAÇÃO
http://www.fsma.edu.br/si/sistemas.html

TRILHA PRINCIPAL

Aplicando Mineração de Dados Para Predição da Evasão Escolar no Ensino Superior em uma Instituição Federal de Ensino

Julio César Belenke dos Santos, *Instituto Federal de Santa Catarina, campus Caçador*,
Ademir Goulart, *Universidade Federal de Santa Catarina*,
Samuel da Silva Feitosa, *Universidade Federal da Fronteira Sul, campus Chapecó* e
Cristiano Mesquita Garcia, *Instituto Federal de Santa Catarina, campus Caçador*

***Resumo*—Mineração de Dados é um processo que objetiva encontrar conhecimento útil e descobrir padrões anteriormente desconhecidos a partir de dados. A Evasão Escolar ainda é um dos desafios a serem atacados em ambientes de Educação Superior. Este trabalho apresenta uma ferramenta que faz uso de um modelo inteligente capaz de predizer potenciais evasores em uma instituição de Ensino Superior através de algoritmos de *Machine Learning*. Para realizar as predições, foram utilizadas as técnicas de Árvore de Decisão e Rede Neural. A Árvore de Decisão apresentou o melhor resultado, alcançado 84% de acerto para a classe de evasores e acurácia de 87%, enquanto a Rede Neural alcançou 82% de acurácia com 78% de precisão para a classe de evasores. Com os dados obtidos da instituição, as características mais relevantes no auxílio à predição foram a média do número de faltas e a idade.**

***Palavras-chave*—Evasão escolar, Mineração de Dados, Classificação, Ensino Superior.**

Applying Data Mining for School Dropout Prediction in Higher Education at a Federal Institute for Education

***Abstract*—Data Mining is a process that seeks to extract useful knowledge and uncover patterns from data. School dropout is still one of the challenges to be tackled in the Higher Education environment. This paper presents a tool that uses a model that aims to predict a potential dropout of undergraduate students of a higher education institution using Machine Learning algorithms. In order to perform the predictions, we used the Decision Tree and Neural Network techniques, where the former achieved the best performance, with 84% precision and 87% accuracy in detecting dropout, while the second achieved 82% accuracy with 78% precision. Besides, given the data obtained from the institution, the most important features that help prediction school dropout are the average number of classes skipped in previous semester and the student's age.**

***Index Terms*—School Dropout, Data Mining, Classification, Higher Education.**

Autor correspondente: Último Autor, *cristiano.garcia@ifsc.edu.br*

I. Introdução

TÉCNICAS de *Machine Learning* têm sido empregadas em diversas áreas, como medicina [1][2], indústria de alimentos [3], mercado financeiro [4][5] e educação [6]. Diversos trabalhos ao redor do mundo utilizaram técnicas de *Machine Learning* com objetivo de reduzir e predizer a evasão escolar. Há diversos trabalhos encontrados na literatura que atacam este problema, na Turquia [7], na Holanda [8] e nos Estados Unidos [9]. No Brasil, alguns trabalhos também foram desenvolvidos. Em [10], os autores conseguiram identificar 87% dos evasores utilizando redes neurais e relatam ter reduzido a evasão em 11% em cursos de graduação EaD com a aplicação do sistema. Em [11], os autores obtiveram acerto de 90% na identificação dos estudantes evasores utilizando *Rotation Forest*, levando em consideração coleta de dados de 10 anos (entre 2005 e 2015) de estudantes de Sistemas de Informação na EACH-USP. No trabalho [12], os autores alcançaram entre 75% e 80% de precisão na previsão da situação final do estudante no curso utilizando notas semestrais dos mesmos estudantes no primeiro semestre. Em [13], os autores fizeram um estudo bibliométrico visando identificar trabalhos relacionando Mineração de Dados e evasão escolar, tendo seus resultados publicados em 2015. Os trabalhos [11] e [12] utilizaram Weka [14] como ferramenta de experimentação.

Técnicas de *Machine Learning* são utilizadas dentro de processos de *Knowledge Discovery in Databases* (KDD). KDD é o processo de extração de dados que visa a obtenção de informação que não é óbvia, anteriormente desconhecida e com potencial de ser útil [15].

Neste trabalho, busca-se analisar e prever a evasão escolar em um ambiente de Ensino Superior, utilizando o processo de KDD. Este estudo de caso foi realizado com os dados de estudantes de um campus de um Instituto Federal (IF) no Brasil, que contém dois cursos de graduação: Sistemas de Informação e Engenharia de Produção.

Neste artigo é apresentado o desenvolvimento de um modelo computacional que utiliza *Machine Learning* para auxiliar na identificação de estudantes com potencial de

evasão. Mais especificamente, são expostos os seguintes tópicos:

- Um Mapeamento Sistemático sobre o tema;
- A coleta e a seleção os dados a serem utilizados;
- A etapa de pré-processamento dos dados da base de dados do IF para eliminar inconsistências;
- O desenvolvimento e avaliação de modelos de *Machine Learning*, por meio dos dados coletados, para realizar a classificação de estudantes como potenciais evasores;
- A identificação das variáveis que mais contribuem na identificação da evasão;
- A implementação de um protótipo que realiza a predição da evasão a partir de dados dos estudantes.

A sequência deste texto está organizada da seguinte forma: a Seção II apresenta os conceitos básicos relativos a este projeto, incluindo a caracterização da evasão escolar, do processo de *Knowledge Discovery in Databases*, e do funcionamento dos processos e algoritmos de *Machine Learning*. Na Seção III é apresentado um Mapeamento Sistemático da literatura, o qual embasou esta pesquisa. Na Seção IV é apresentada a metodologia seguida por este projeto. A Seção V apresenta os principais resultados obtidos a partir dos modelos desenvolvidos e o protótipo implementado. Por fim, a Seção VI apresenta as considerações finais e os trabalhos futuros.

II. Fundamentação Teórica

A. Evasão Escolar

A evasão escolar é um problema recorrente e um desafio a ser solucionado. Para as instituições públicas pode-se dizer que a evasão escolar é um desperdício de recursos, enquanto que para as privadas, representa a diminuição no seu orçamento ou importante perda de receitas sem as mensalidades dos alunos evasores [16]. Podem-se observar alguns motivos de desistência do estudante, tais como: (a) a falta de orientação vocacional; (b) imaturidade do estudante; (c) reprovações sucessivas; (d) dificuldades financeiras; (e) falta de perspectiva de trabalho; (f) ausência de laços afetivos na universidade; (g) ingresso na faculdade por imposição familiar; e (h) casamentos não-planejados e nascimento de filhos [17]. É importante ressaltar que o problema da evasão escolar possui características locais: cada instituição pode ser suas particularidades, mesmo que estejam na mesma região [18].

Segundo [19], o termo "evasão" no contexto de uma Instituição de Ensino Superior, pode significar a saída definitiva de um estudante do seu curso sem que o mesmo tenha conseguido concluir, ou seja, evasão escolar. Porém, por existirem várias formas pelas quais o estudante pode evadir-se de um curso, os autores em [19] separaram em três categorias: evasão de curso, evasão da instituição e evasão do sistema, detalhadas a seguir:

- **Evasão do curso:** o estudante pode evadir do curso por diversas situações: abandono (neste caso refere-se ao aluno não se matricular), desistência (oficial), transferência ou re-opção (mudança de curso), exclusão por norma institucional;
- **Evasão da instituição:** o aluno se desliga da instituição;
- **Evasão do sistema:** o aluno abandona de forma definitiva ou temporária o sistema educacional.

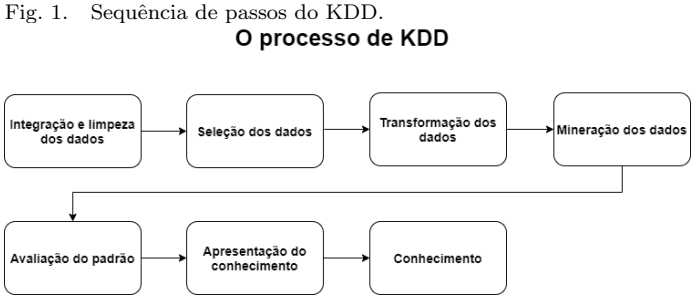
Outras correntes de pensamento preferem o termo "abandono escolar" à "evasão escolar", pois evasão escolar teria uma conotação unilateral, responsabilizando apenas o estudante pela evasão, tornando este ato um "ato solitário" [20][21]. Este trabalho utiliza o termo "evasão escolar" pois os objetivos giram em torno de predizer a evasão escolar com *Machine Learning*, não de discutir o fenômeno em si.

Uma das ferramentas que vêm sendo utilizadas para resolver uma série de problemas em diversas áreas de conhecimento é o processo de KDD. O processo fornece passos que podem auxiliar no encontro de padrões em dados, podendo ser útil para predizer a desistência de estudantes, baseado nas informações dos mesmos.

B. Knowledge Discovery in Databases

SEGUNDO [22], o processo de KDD é uma sequência de passos que tem por objetivo a descoberta de conhecimento. Importante ressaltar que o KDD é um processo iterativo, significando que a execução do processo em si não ocorre de forma linear do começo ao fim, sendo necessário, por muitas vezes, voltar a estágios anteriores para depois seguir adiante novamente [23]. A Figura 1 exibe os passos definidos na literatura.

Fig. 1. Sequência de passos do KDD.



Fonte: Elaborada pelos autores.

É importante entender o que acontece em cada passo do KDD, conforme mostrado na Figura 1. Os passos são:

1) **Limpeza de dados:** são removidos os dados que estão inconsistentes na base de dados ou fora do padrão;
2) **Integração de dados:** integração onde várias fontes de dados podem ser combinadas, de forma a enriquecer os dados originais;
3) **Seleção dos dados:** os dados relevantes para a solução do problema são extraídos desta base de dados.
4) **Transformação nos dados:** transformação ou consolidação dos dados em formas que possam ser utilizadas pelas técnicas de Mineração de Dados;
5) **Mineração de Dados:** processo onde são buscados padrões de dados úteis. Podem ser citados como exem-

plo de técnicas de *Machine Learning* para classificação: algoritmo de classificação bayesiano, de vizinho mais próximo (i.e. KNN) e árvore de decisão [24], entre outros;

6) **Avaliação dos padrões:** identifica os padrões interessantes entre os apresentados pela etapa anterior de mineração de dados;
7) **Apresentação do conhecimento:** utiliza técnicas de representação e visualização para apresentar o conhecimento adquirido aos usuários.

Embora não esteja explícito na Figura 1, o KDD é um processo iterativo e, como tal, em qualquer etapa é possível retornar à etapas anteriores para refinamento dos processamentos.

C. Machine Learning

MACHINE *Learning* (ML) é a programação de computadores que utiliza algum critério de desempenho por meio dos dados, podendo ser dados históricos, de experiência passada, ou de outros tipos. O modelo resultante pode ser tanto preditivo – para fazer predições do futuro –, como pode ser descritivo, para obter conhecimento de dados, ou para ambas as situações [25].

A etapa de ML dentro do processo de KDD está ligada aos aspectos de treinamento do modelo e, logo após, à avaliação dos padrões encontrados, verificando se há um nível de performance e confiança aceitáveis. Caso contrário, os passos anteriores a ele são repetidos.

Machine Learning tem como principal objetivo fazer com que os computadores aprendam padrões automaticamente por meio de modelos estatísticos, e até a fazer decisões inteligentes baseadas em dados [22].

Neste trabalho, são utilizados algoritmos de classificação para prever se o estudante tem a possibilidade (ou não) de evasão de seu curso. Esta tarefa (classificação) é a mais apropriada para este caso, pois é desejável que o atributo-alvo seja se o aluno tem potencial de evadir ou não. Como o objetivo principal deste trabalho é a detecção de tendência de evasão escolar a nível individual, a informação do atributo-alvo guiará o aprendizado supervisionado na detecção de padrões de evasão, podendo assim identificar novos estudantes com comportamentos similares. Já os atributos analisados podem ser vários dentro do contexto escolar de uma Instituição de Ensino Superior, como por exemplo, dados socioeconômicos, escolares ou pessoais. Abaixo, são explicados alguns conceitos para prosseguimento na leitura do trabalho.

1) Classificação: A classificação consiste em separar os dados de modo que estejam associados a uma categoria exclusiva conhecida como classe. Em sua configuração mais básica, cada conjunto de informações deve pertencer exclusivamente e ser associado a apenas uma classe [24]. Como exemplo de algoritmos, podem ser citados os de Classificação Bayesiano, Árvore de Decisão, Redes de Crença Bayesiana, Classificação por Retropropagação, Máquinas de Vetores de Suporte, Redes Neurais e outros.

De acordo com o estudo de Alban [26], o método mais utilizado e presente em 22 dos 28 artigos filtrados com técnicas de Mineração de Dados é a técnica de Árvore de Decisão, que foi usada para classificação. Além das Árvores de Decisão, uma técnica bastante utilizada é a Rede Neural. Tendo em vista que ambos os modelos mencionados foram utilizados neste trabalho, cabe aqui defini-los para melhor compreensão do leitor.

2) Rede Neural: O termo Rede Neural remete a um sistema artificial de neurônios que são inspirados nos neurônios biológicos. Ela pode abranger desde um simples nó, dado que todo nó corresponde a um neurônio, bem como a um conjunto de nós conectados em uma grande rede. Conforme vai aprendendo com os dados disponíveis, a Rede Neural vai atualizando o peso entre os nós. Este peso é multiplicado para cada respectiva entrada e representa a importância atribuída a cada informação [27].

3) Backpropagation: O algoritmo de retro-propagação, também conhecido como diferenciação de modo reverso, possui por objetivo geral encontrar valores para pesos, de forma que forneçam uma estimativa ótima de uma função que modela os dados de treinamento. Ou seja, encontrar um conjunto de pesos para minimizar a saída da função de perda. Dentro de uma Rede Neural, há um modelo no qual cada um dos neurônios ou nós está ligado a um nó anterior, contando com uma camada de entrada na qual são recebidos os dados. A camada oculta possui este nome por não ser nem de entrada nem de saída. Uma Rede Neural pode ter várias camadas ocultas, com diferentes números de neurônios. A camada de saída é responsável por emitir um valor de saída, com base na ponderação dos valores recebidos da(s) camada(s) escondida(s) [27]. Embora seja comum a utilização de múltiplas camadas escondidas, é dito que uma Rede Neural é capaz de aproximar a qualquer função não-linear (com nível arbitrário de acurácia) com apenas uma camada escondida [28].

4) Árvore de Decisão: De acordo com [29], Árvore de Decisão é "uma estrutura em forma de árvore na qual cada nó interno corresponde a um teste de atributo; cada ramo representa um resultado do teste e os nós folhas representam classes ou distribuições de classes". Ainda, "seu nó mais elevado é chamado nó raiz, e cada caminho entre a raiz e um nó folha representa uma regra de classificação" [29]. Estas estruturas são usadas comumente para classificação, mas podem ser utilizadas também em outras tarefas, como regressão.

III. Estado da Arte da Área Pesquisada

O processo de pesquisa e seleção dos trabalhos relacionados foi realizado com base em um Mapeamento Sistemático sobre as pesquisas com propostas para auxiliar na identificação de estudantes com potencial para evasão escolar utilizando a Mineração de Dados. Como resultado, foram identificados e selecionados os principais trabalhos de pesquisa no tema deste artigo, considerando ainda os métodos utilizados para resolver o problema em questão.

A. Mapeamento Sistemático da literatura

Durante a etapa de Mapeamento Sistemático da literatura, é preciso realizar as tarefas de definição de questões

de pesquisa e *string* de busca, realização da pesquisa de estudos primários relevantes, triagem dos documentos, *keywording* dos resumos, e extração dos dados. A ferramenta Parsifal[1] foi utilizada para automatizar o processo, definir a *string* de busca, buscar e armazenar os artigos, além de realizar as classificações pertinentes com os critérios selecionados.

A questão de pesquisa utilizada foi "Como a Mineração de Dados tem sido aplicada para prever potencial evasão escolar?", a partir da qual foram extraídos termos e palavras relacionados que auxiliaram na confecção da *string* de busca para realizar as consultas nas bases de dados selecionadas. Podem ser visualizadas estas palavras com seus sinônimos na Tabela I.

TABELA I
TABELA COM PALAVRAS-CHAVE E SINÔNIMOS

Palavra-chave	Sinônimos
Data Mining	KDD, Machine Learning, ML, Knowledge Discovery Database
Predict	avoid, foresee, predicting, prediction
School dropout	college dropout, drop out student, dropping student, scholar evasion, school abandonment, school leavers, school retention, university dropout, university evasion

Fonte: Elaborada pelos autores.

Na Tabela II, são listadas as bases de dados em que foram pesquisados os artigos, juntamente com o número de artigos que foram retornados com esta busca. Estas bases foram utilizadas devido ao seu destaque em termos de conteúdo na área de computação e aprendizado de máquina. A mesma *string* de busca foi utilizada para as três bases de dados, sendo ela:

```
(‘‘predict’’ OR ‘‘avoid’’ OR ‘‘foresee’’ OR
‘‘predicting’’ OR ‘‘prediction’’)
AND
(‘‘data mining’’ OR ‘‘KDD’’ OR
‘‘knowledge discovery database’’ OR
‘‘machine learning’’ OR ‘‘ML’’)
AND
(‘‘school dropout’’ OR ‘‘college dropout’’ OR
‘‘drop out student’’ OR ‘‘dropping student’’ OR
‘‘scholar evasion’’ OR ‘‘school abandonment’’ OR
‘‘school leavers’’ OR ‘‘school retention’’ OR
‘‘university dropout’’ OR ‘‘university evasion’’).
```

TABELA II
TABELA COM AS BASES DE ARTIGOS

Base de Artigos	Qtde. de trab. encontrados
ACM Digital Library	13
IEEE Digital Library	7
Scopus	31

Fonte: Elaborada pelos autores.

Após a definição das bases de dados e da *string* de busca, foram definidos critérios de inclusão e exclusão para filtrar os artigos estritamente relacionados a esta pesquisa.

1) Critérios de inclusão: Para iniciar a seleção dos artigos resultantes da pesquisa nas bases de dados, foram definidos alguns critérios de inclusão. O trabalho deveria conter:

- Nova tecnologia para prever evasão escolar;
- Processo, método ou técnica para previsão automática de evasão escolar;
- Sistema para prever evasão escolar usando Mineração de Dados.

2) Critérios de exclusão: De modo a selecionar os artigos relacionados aos propósitos deste projeto, foram excluídos trabalhos que se enquadram nos critérios de exclusão apresentados na Tabela III:

TABELA III
TABELA COM OS CRITÉRIOS DE EXCLUSÃO

Critério de exclusão	Trab. recusados
O estudo não é um estudo primário	3
O estudo não faz parte da área de pesquisa	15
O estudo não foi aplicado para o ensino superior	6
O estudo é duplicado	7
O autor usou uma base de dados de terceiros, ou de um repositório online. Não possui um *dataset* próprio	1

Fonte: Elaborada pelos autores.

A pesquisa nas bases de dados resultou em 51 artigos, considerando as três bases de dados utilizadas. Após a aplicação dos critérios de inclusão e exclusão restaram 19 artigos, os quais são apresentados na Tabela IV.

Todos os trabalhos se referem a maneiras de auxiliar na predição da evasão escolar por meio da Mineração de Dados ou da aplicação de técnicas de *Machine Learning*, as quais subsidiaram as etapas seguintes deste estudo.

B. Análise dos Trabalhos Selecionados

A última etapa do Mapeamento Sistemático da literatura foi a extração de dados dos trabalhos selecionados. Os algoritmos mais utilizados em grande parte dos trabalhos são, em ordem de frequência: Árvore de Decisão (13 artigos), Naïve Bayes e derivados (7 artigos), Regressão Logística (6 artigos), Floresta Aleatória (6 artigos), Rede Neural (5 artigos), K-ésimo vizinho mais próximo (3 artigos) e Máquinas de Vetores de Suporte (3 artigos). Vale ressaltar que alguns trabalhos utilizaram mais de um método, o que explica a soma das ocorrências superarem o número de trabalhos selecionados.

Os dados utilizados neste trabalho foram fornecidos pelo setor responsável pela Tecnologia da Informação da Instituição. Com base na análise dos artigos estudados, as técnicas identificadas foram estudadas de forma mais aprofundada, de modo a tirar vantagem do Mapeamento da área de pesquisa, especialmente do assunto sobre Redes Neurais e Árvores de Decisão. Estes algoritmos foram selecionados devido à sua frequência de aparição nos trabalhos elencados no Mapeamento Sistemático (Árvore de

[1]https://parsif.al/

TABELA IV
Tabela com os artigos selecionados

ID	Título do artigo	Autores/Autor(a)
A1	Data Mining Applied in School Dropout Prediction [30]	A. Viloria, J. G. Gualiny, W. N. Núñez, H. H. Palma e L. N. Núñez
A2	Early Prediction of University Dropouts – A Random Forest Approach [31]	A. Behr, M. Giese, H. Teguim, K. Theune
A3	Predictive Models for Imbalanced Data: A School Dropout Perspective [32]	T. Barros, P. A. Souza Neto, I. Silva, L. A. Guedes
A4	A Multinomial and Predictive Analysis of Factors Associated with University Dropout [33]	T. Fernández-Martin, M. Solis, M. T. Hernández-Jiménez, T. E. Moreira-Mora
A5	Ensemble Regression Models Applied to Dropout in Higher Education [34]	P. M. da Silva, M. N. C. A. Lima, W. L. Soares, I. R. R. Silva, R. A. de A. Fagundes, F. F. de Souza
A6	Supervised Learning in the Context of Educational Data Mining to Avoid University Students Dropout [35]	K. J. de O. Santos, A. G. Menezes, A. B. de Carvalho, C. A. E. Montesco
A7	Predictive Modelling of Student Dropout using Ensemble Classifier Method in Higher Education [36]	N. Hutagaol, S. Suharjito
A8	Integration of Data Technology for Analyzing University Dropout [37]	A. Viloria, J. G. Padilla, C. Vargas-Mercado, H. H. Palma, N. O. Llinas, M. A. David
A9	University Dropout: A Prediction Model for an Engineering Program in Bogotá, Colombia [38]	A. Acero, J. C. Achury, J. M. Piñero
A10	Predicting Early Dropout Students is a Matter of Checking Completed Quizzes: The Case of an Online Statistics Module [39]	J. Figueroa-Cañas, T. Sancho-Vinuesa
A11	Educational Data Mining: an Application of Regressors in Predicting School Dropout [40]	R. L. S. do Nascimento, R. B. das Neves Junior, M. A. de A. Neto, R. A de A. Fagundes
A12	Using Academic Analytics to Predict Dropout Risk in Engineering Courses [41]	J. Lima, P. Alves, M. J. Pereira, S. Almeida
A13	Prediction of University Dropout through Technological Factors: A Case Study in Ecuador [42]	M. S. A. Taipe, D. M. Sánchez
A14	Data-driven System to Predict Academic Grades and Dropout [43]	S. Rovira, E. Puertas, L. Igual
A15	Automatic Feature Selection for Desertion and Graduation Prediction: A Chilean Case [44]	B. Peralta, T. Poblete, L. Caro
A16	Prediction of University Desertion through Hybridization of Classification Algorithms [45]	C. F. Rocha, Y. F. Zelaya, D. M. Sánchez, A. F. Pérez
A17	Dropout Prediction at University of Genoa: A Privacy Preserving Data Driven Approach [46]	L. Oneto, A. Siri, G. Luria, D. Anguita
A18	Optimization of Weight Backpropagation with Particle Swarm Optimization for Student Dropout Prediction [47]	E. Y. Sari, A. S. Kusrini
A19	Applying Data Mining Techniques to Predict Student Dropout: A Case Study [48]	B Perez, C. Castellanos, D. Correal

Fonte: Elaborada pelos autores.

Decisão) e pela sua conhecida e poderosa capacidade de aproximação de funções (Redes Neurais).

IV. Metodologia

NESTA seção, é apresentada a metodologia utilizada neste trabalho, desde a coleta até a análise dos resultados obtidos com os algoritmos. Ademais, detalhes do protótipo são explicitados ao fim.

A. Coleta e Tratamento dos Dados

Foram obtidos dados institucionais anonimizados dos estudantes dos cursos de graduação do IF, para aplicação das etapas de KDD, como forma de auxiliar na predição da evasão escolar.

A seleção das colunas (informações) foi feita a partir do estudo dos artigos elencados no Mapeamento Sistemático, onde foram identificadas as informações mais utilizadas para a predição da evasão escolar. As informações selecionadas foram: nível de estudo e ocupação dos pais, sexo, idade, performance acadêmica no primeiro ano, frequência das aulas, data de matrícula, ano de inscrição, mora com (pais, sozinho), componentes da família (mãe, pai ou ambos), tipo de casa (própria, alugada, outra), trabalha atualmente (sim, não), renda familiar, entre outras. Porém, parte destas informações não estavam disponíveis no banco de dados consultado.

O Sistema Gerenciador de Banco de Dados (SGBD) utilizado pelo IF é o PostgreSQL e, após estudar estes diagramas, foram elencadas as informações mais importantes passíveis de serem extraídas. A partir disto, foi desenvolvida um *script* SQL (*Structured Query Language*) para obtenção dos dados de acordo com os diagramas recebidos do setor responsável.

Em um contexto geral, para a manipulação, tratamento, leitura e visualização dos dados e para a aplicação de algoritmos de *Machine Learning*, foi utilizada a linguagem de programação Python em conjunto com as bibliotecas Pandas[2] e Scikit-Learn[3].

Os algoritmos de *Machine Learning* selecionados, i.e. Árvore de Decisão e Rede Neural Artificial, possuem res-

[2]https://pandas.pydata.org/
[3]https://scikit-learn.org/stable/

trições, sendo necessário realizar uma transformação nos dados de categórico para numérico.

B. Codificação dos Dados

Ao trabalhar com dados e técnicas de *Machine Learning*, o responsável deve conhecer as limitações destas e algumas técnicas de tratamento de dados. Pode-se citar alguns exemplos de técnicas de codificação de variáveis categóricas como: *One Hot Encoding*, *Ordinal Encoding*, *Binary Encoding*, entre outras [49]. A técnica de *One Hot Encoding* foi utilizada para conversão da coluna de "estado_civil". Ao utilizar esta técnica ao invés de *Ordinal Encoding*, é possível caracterizar cada estado sem a necessidade de imputar uma ordem entre os valores, o que pode enviesar o aprendizado.

C. Configuração do Treinamento

Overfitting é um dos problemas mais comuns e recorrentes na etapa de treinamento de algoritmos de *Machine Learning*. Uma das formas de testar a capacidade de generalização de um modelo é através da separação dos dados em conjuntos de treinamento e teste. Assim, é possível avaliar seu desempenho por meio das métricas apropriadas [50].

Inicialmente, é realizada a separação dos dados em conjuntos de treinamento e teste, onde o modelo computacional, após treinado, precisa classificar corretamente a parcela de dados que foi definida para teste. Por exemplo, se houver a definição de 70/30, 70% do *dataset* será utilizado para o treinamento do modelo e 30% será definido como teste para verificar as taxas de acerto daquele modelo recém-treinado. Os dados são misturados antes da divisão, de forma que a sequência dos dados não interfira no processo de treinamento. Neste trabalho, foi utilizada a divisão 70/30 para validação do modelo.

D. Feature Engineering

Feature Engineering é um processo que usa do conhecimento do domínio para encontrar ou extrair variáveis úteis dos dados. Dentro deste universo, existe um conceito muito usado chamado *Feature Transformation* que possui como objetivo construir novas variáveis a partir de variáveis já existentes [51].

As novas colunas criadas a partir de *Feature Engineering* foram:

(a) "idade": gerada a partir da coluna de data de nascimento;
(b) "TAS" (taxa de aprovação no semestre): calculada pelo número de disciplinas em que o aluno foi aprovado, dividido pelo total de disciplinas que o aluno cursou no semestre;
(c) as colunas "penúltimo_TAS" e "último_TAS" têm seu valor recebido do "TAS" no penúltimo e último semestres do aluno respectivamente;
(d) a "média da TAS" que se refere à média geral do "TAS" no qual o cálculo foi descrito anteriormente, i.e., a taxa de aprovação ao longo do curso;
(e) "diferenca_anos": contabiliza a diferença entre o ingresso no Ensino Superior e o término do ensino médio em anos;
(f) foi feito um agrupamento (da perspectiva de banco de dados) pelo id, ano e período e calculado a média a partir do número de faltas, utilizando a média do penúltimo e último semestre, gerando as colunas "penúltimo_faltas" e "último_faltas".

E. Validação do Modelo

Os modelos foram avaliados por meio de métricas de desempenho do algoritmo de *Machine Learning* para classificação, sendo as mais comuns: revocação, acurácia, precisão e *F1 Score*. Dependendo do contexto e aplicação, essas métricas podem assumir outros nomes [24].

$$Acurácia = \frac{VP + VN}{T} \tag{1}$$

A Acurácia se refere à proporção de instâncias que foram classificadas corretamente no modelo em relação ao número total de instâncias. Na fórmula de acurácia (Equação 1), VP se refere aos "Verdadeiros-Positivos", VN, aos "Verdadeiros-Negativos" e T se refere ao total de instâncias que foram preditas pelo modelo [24]. Importante notar que: (a) $T = VP + VN + FP + FN$, onde FN significa "Falsos-Negativos" e FP significa "Falsos-Positivos"; (b) "Falsos-Negativos" e "Falsos-Positivos" são considerados no cálculo da Acurácia. Logo, quanto maior a proporção de erros (ou seja, quanto maior FP e FN), menor o valor da Acurácia. Da mesma forma, quanto maior o valor da Acurácia, melhor, pois significa mais acertos por parte do modelo.

$$Precisão = \frac{VP}{VP + FP} \tag{2}$$

A Precisão se propõe a responder a seguinte pergunta: "entre aqueles que foram classificados como de uma determinada classe, quantos realmente são?". Na Equação 2, VP representa "Verdadeiros-Positivos", FP é definido como "Falsos-Positivos' [24]. A Precisão é calculada individualmente em relação à cada classe. Pela Equação, pode-se perceber que, quando maior a existência de "Falsos-Positivos", menor a Precisão. Para a Precisão, obviamente, quanto maior o valor obtido, melhor.

$$Revocação = \frac{VP}{VP + FN} \tag{3}$$

A Revocação se refere à frequência com que o classificador encontra os exemplos de uma determinada classe. Na Equação 3, FN significa "Falsos-Negativos" [24]. A Revocação também é calculada em relação à cada classe. Através da Equação, pode-se perceber que, quanto menor a presença de "Falsos-Negativos", maior a Revocação. Logo, para Revocação, quanto maior o valor, melhor.

Como as métricas avaliam aspectos diferentes da classificação, a análise dos resultados precisa levá-las em consideração de forma conjunta.

F. Desenvolvimento do Protótipo

Como já mencionado anteriomente, todo o tratamento de dados e o treinamento do modelo (utilizando Rede Neural e Árvore de Decisão) foram feitos utilizando Pandas e Scikit-Learn, na linguagem Python. Para o desenvolvimento do protótipo de predição da evasão escolar, foi utilizada a biblioteca Streamlit[4].

A biblioteca Streamlit é uma ferramenta voltada para o desenvolvimento rápido de interfaces *Web* para modelos de *Machine Learning*, também chamados de *data apps*. Sendo assim, é apenas necessário programar a interface levando em consideração as entradas para o modelo computacional treinado e programar a forma de se exibir a saída do modelo. A biblioteca Streamlit ainda fornece a possibilidade de oferta de *data apps* hospedados na nuvem, serviço este oferecido por meio do site da ferramenta.

V. Resultados

NESTA seção, são apresentados os Modelos de *Machine Learning* implementados para o problema proposto neste trabalho, além de sumarizar os principais resultados obtidos.

A. Modelo

Após a definição das colunas a serem utilizadas pelo modelo e depois da etapa de *Feature engineering* – que adiciona novas colunas –, o próximo passo executado foi o de agrupar os dados para cada linha representar apenas um aluno, dando subsídios para o treinamento do modelo. Por fim, foi realizada a etapa de validação do modelo, visando avaliar as métricas de desempenho obtidas.

Após o agrupamento e a limpeza dos dados (remoção de linhas com dados faltantes) e identificação de colunas com dados faltantes acima de 10%, restaram 12 colunas com informações sobre alunos e as disciplinas cursadas por eles. Como os dados são anonimizados, cada aluno é representado por um identificador numérico, o que permitiu o agrupamento de informações mantendo a integridade dos dados. Ao final, o *dataset* englobou 380 registros, sendo que cada linha representa um único aluno. Apesar da quantidade ainda pequena de dados, especialmente para técnicas como Rede Neural, a intenção é expandí-la de acordo com a admissão de novos alunos. Na Tabela V são apresentadas as colunas com informações contidas no *dataset* final. Os tipos int64 se referem a variáveis do tipo inteiro, os tipos float64 se referem ao conjunto dos números reais e o tipo object se referem a objetos de classe como *String* e outros.

As colunas "estado_civil_descricao" e "tipo_raca_descricao" precisaram ser convertidas para numéricas por meio da técnica *One Hot Enconding* devido ao fato dos algoritmos de Árvore de Decisão e Rede Neural só aceitarem dados com tipos numéricos na implementação do Scikit-Learn. No caso da Rede Neural, para o atributo-alvo (classe) com o caso de classificação binária, também foi necessária a conversão de categórica para numérica, definindo 0 como indicativo da "não-evasão" e 1 para o caso do aluno ter se evadido ("evasão"). Apesar da utilização de "tipo_raca_descricao" ter o potencial de reprodução de preconceitos, este campo foi utilizado apenas para teste. Esta informação não é utilizada no modelo final, sendo excluída pelos autores por apresentar queda no desempenho do modelo de Árvore de Decisão.

[4]https://www.streamlit.io/

TABELA V
Informações após tratamento dos dados.

Nome da coluna	Nº Registros	Tipo
id_pessoa	380	int64
TAS	380	float64
último_TAS	380	float64
penúltimo_TAS	380	float64
média_TAS	380	float64
penúltimo_faltas	380	float64
último_faltas	380	float64
estado_civil_descricao	380	object
diferença_anos	380	float64
evadiu	380	int64
idade	380	float64
tipo_raca_descricao	380	object

Fonte: Elaborada pelos autores.

Para os casos de dados faltantes, os mesmos foram substituídos por valores estatísticos (mais especificamente pela média) nas colunas "penúltimo_TAS", "penúltimo_faltas" e "idade". Embora existam vários outros métodos mais complexos de imputação de dados faltantes (e.g. [52]), a frequência era muito baixa, o que justifica o uso de um método mais simples. Além disso, pode ser observado na análise de importância dos atributos que aqueles que sofreram imputação não estão entre os mais importantes para o modelo, com exceção de idade.

B. Evolução

Inicialmente, para a etapa de treinamento e testes dos modelos, foi adotada a estratégia de separação dos dados 70/30, onde as colunas utilizadas foram:

(a) "TAS", calculada pela taxa de aprovação do estudante durante todo seu tempo no curso, representado pela seguinte fórmula: $TAS = \frac{disciplinas\_aprovadas}{disciplinas\_cursadas}$;
(b) "penúltimo_TAS", que tem seu valor calculado semelhante ao "TAS", porém relativo apenas ao penúltimo semestre ativo do aluno;
(c) "último_TAS", que tem seu valor calculado semelhante ao "TAS", porém relativo apenas ao último semestre ativo do aluno;
(d) "média_TAS", que se refere à taxa de aprovação média por semestre.

Apesar das técnicas utilizadas neste trabalho serem capazes de realizar seleção de características, este conjunto de informações acima mencionado foi utilizado por descreverem o histórico do estudante, . O atributo-alvo denominado "E=evadiu" ou "NE=não evadiu", se refere à

classe. Os parâmetros utilizados nas técnicas de *Machine Learning* são mencionados a seguir.

1) Rede Neural: Os parâmetros estabelecidos para a Rede Neural são os parâmetros padrão do Scikit-Learn, com exceção do *random_state*, que recebe um valor fixo para que os resultados se mantenham reprodutíveis. Outro parâmetro que foi definido foi o número de iterações (ou épocas) para o algoritmo de *Machine Learning* convergir o modelo (*max_iter*). O valor de *random_state* apenas precisa ser definido e reutilizado, enquanto o número 1000 para *max_iter* foi selecionado após testes por suportar a convergência da rede. Desta forma, a instância do modelo foi criada como a seguir.

$$MLPClassifier(max\_iter = 1000, random\_state = 12)$$

Após a criação do objeto do classificador, foi realizado o treinamento do modelo utilizando os dados que foram previamente separados para esta etapa. Para avaliar o desempenho do modelo inicial, que utilizou-se das quatro colunas mencionadas acima, foi realizada a comparação entre os valores preditos pelo modelo para com os valores esperados. A seguir, a Tabela VI apresenta todas as execuções (com adições e alterações em colunas do *dataset*) utilizando Rede Neural, visando melhorar o desempenho das métricas.

TABELA VI
Tabela com a evolução das métricas da Rede Neural

Nº	Coluna alteradas ou adicionadas	Prec.	Revoc.	Acur.	Classe
1	Primeira execução	0.71	0.76	0.69	"NE"
		0.67	0.61		"E"
2	Adição da "idade"	0.77	0.81	0.76	"NE"
		0.75	0.71		"E"
3	Adição de "estado_civil" e "raça"	0.80	0.76	0.76	"NE"
		0.72	0.76		"E"
4	Adição de "diferença_anos"	0.85	0.81	0.82	"NE"
		0.78	0.82		"E"
5	Adição de "penúltimo_faltas" e "último_faltas"	0.77	0.95	0.82	"NE"
		0.92	0.65		"E"
6	Remoção da coluna "raça"	0.82	0.86	0.82	"NE"
		0.81	0.76		"E"

Fonte: Elaborada pelos autores.

Na última modificação, pela remoção da coluna "raça", houve uma variação de 11 pontos percentuais para baixo na precisão, enquanto a revocação aumentou na mesma proporção, embora a acurácia do modelo tenha se mantido. Isso significa que o modelo passou a errar pouco ao predizer estudantes com potencial de evasão (causando alta precisão), mas ao mesmo tempo, assinalou muitos estudantes com tal potencial como estudantes sem chance de evadir (baixa revocação). Considerando apenas a acurácia, pode-se argumentar em favor da utilização desta coluna por fatores estritamente matemáticos. No entanto, a utilização de dados sensíveis em modelos de aprendizado de máquina vem sendo debatida com maior frequência nos últimos anos, como em [53]. O uso de informações passíveis de reprodução de preconceitos ou desigualdades deve ser realizado com cautela ou mesmo evitado. Para este modelo em específico, mesmo que fosse o mais adequado para o problema, a informação de "raça" não seria utilizada.

2) Arvore de Decisão: Os parâmetros utilizados para a Árvore de Decisão são os parâmetros que vêm por padrão na biblioteca do Scikit-Learn, com exceção do parâmetro *random_state*, definido para reprodutibilidade.

A próxima etapa após a configuração da Árvore de Decisão, acontece o treinamento com o conjunto de dados de treino definidos na separação do *dataset*.

Na Tabela VII, são listadas todas execuções junto das modificações no conjunto de dados e os valores das métricas de desempenho obtidas ao longo das alterações. Estão também a precisão e revocação para cada classe e a acurácia referente ao modelo.

TABELA VII
Tabela com a evolução das métricas na Árvore de Decisão

Nº	Coluna alteradas ou adicionadas	Prec.	Revoc.	Acur.	Classe
1	Primeira execução	0.69	0.68	0.66	"NE"
		0.62	0.63		"E"
2	Adição da "idade"	0.65	0.71	0.63	"NE"
		0.60	0.53		"E"
3	Adição de "diferença_anos"	0.74	0.81	0.74	"NE"
		0.73	0.65		"E"
4	Adição de "penúltimo_faltas" e "último_faltas"	0.82	0.86	0.82	"NE"
		0.81	0.76		"E"
5	Adição de "estado_civil" e "raça"	0.82	0.86	0.82	"NE"
		0.81	0.76		"E"
6	Remoção da "raça"	0.83	0.90	0.84	"NE"
		0.87	0.76		"E"

Fonte: Elaborada pelos autores

Na Tabela VII, pode-se verificar a primeira execução com as mesmas cinco primeiras colunas mencionadas anteriormente, onde estão várias métricas que são geradas pela biblioteca do *Scikit-Learn*. Nas execuções seguintes da Árvore de Decisão, foram adicionadas as colunas de "idade", "diferença_anos", "penúltimo_faltas" e "último_faltas". Como citado anteriormente na seção da Rede Neural, houve uma progressão a cada adição de colunas nas métricas. Com exceção do caso do "estado_civil" e "raça" juntos que não resultaram em uma melhora na performance do modelo.

A partir da última modificação de informações com a Árvore de Decisão (remoção da coluna com os dados de raça dos estudantes), houve uma grande melhora, que resultou no melhor modelo entre todas as execuções. Esta afirmação pode ser verificada pelas métricas de precisão, revocação e acurácia maiores do que as dos outros modelos. Logo, este foi o modelo escolhido para aplicação neste trabalho. A coluna raça foi utilizada apenas para fins de testes e é importante para o responsável estar atento a dados socialmente sensíveis durante o processo de KDD, para evitar que o modelo geral reproduza preconceitos.

Além disso, deve ser ressaltado que a remoção desta coluna melhorou a performance da Árvore de Decisão, sugerindo que o raça pouco influi na decisão da evasão escolar.

Considerando as informações iniciais e a evolução dos modelos com base nas colunas adicionadas e removidas, o conjunto de colunas final foi: (a) idade; (b) diferença_anos; (c) estado_civil; (d) média_TAS; (e) último_faltas; (f) penúltimo_faltas; (g) último_TAS; (h) penúltimo_TAS; e (i) TAS, totalizando 9 colunas.

C. Protótipo

Na Figura 2, é mostrada a tela principal do aplicativo desenvolvido no *Streamlit*. Esta interface interage com o modelo, que busca prever a possibilidade de evasão de um estudante de cursos superiores do Instituto Federal utilizando técnicas de *Machine Learning*. A interação com o aplicativo acontece por meio do menu lateral, abaixo do texto "Parâmetros de Entrada".

À esquerda da Figura 2 está a parte da interface *Web* em que os parâmetros de entrada para o modelo são definidos. Estes campos possuem uma interatividade maior, como barra para ajuste de valores. É necessário ajustar os campos para que o modelo possa realizar a predição. A cada ajuste de valores, o modelo processa as entradas e fornece uma classificação. O protótipo está disponível como *data app* em https://bit.ly/predicao-evasao.

Como processo geral para implantação e execução do modelo de Árvore de Decisão via protótipo, têm-se:

1) Após definição da técnica melhor qualificada para o problema, o modelo foi treinado com 100% dos dados, serializado e salvo em arquivo, utilizando o módulo Pickle[5] para Python;
2) Ao executar o protótipo, foi programado para o arquivo referente ao modelo ser desserializado e carregado;
3) Com o modelo carregado, qualquer alteração de parâmetro realizado via interface do Streamlit gera um novo conjunto de informações a ser aplicado ao modelo, que então exibirá a saída (predição) na tela principal da interface.

Com o objetivo de interpretar o funcionamento da Árvore de Decisão, a Figura 3 foi gerada através da biblioteca *Shap*[6]. As duas informações com maior contribuição para a decisão do modelo são a idade do aluno e o número de faltas do último semestre, e as que possuem menos são alguns tipos de estado civil que possuem pouca ou zero contribuição. As importâncias das informações são calculadas internamente pela Árvore de Decisão, a partir da redução da impureza dos nós ao longo do modelo gerado. A utilização da biblioteca *Shap* apenas com o modelo de Árvore de Decisão se deve ao fato de modelos gerados por esta técnica serem explicáveis por padrão. Modelos de Rede Neural são conhecidos por serem "caixa-preta", significando que são necessárias outras técnicas para obter uma explicação das decisões do modelo.

[5]https://docs.python.org/3/library/pickle.html
[6]https://shap.readthedocs.io/en/latest/

VI. Conclusões

O presente trabalho teve como objetivo apresentar uma ferramenta para predição de potenciais casos de evasão em um campus de um Instituto Federal, utilizando técnicas de aprendizagem de máquina (Redes Neurais e Árvore de Decisão). A criação e aplicação de um modelo inteligente para a predição da evasão em um contexto institucional específico é importante, pois este é um problema onde as características locais possuem forte influência no resultado. Além disso, por meio dos explicadores de modelos de aprendizagem de máquina, foi possível identificar as variáveis mais relevantes para o referido problema, o que pode ser um indicativo para outros campi instalados em contextos similares, apesar da característica local do problema. Adicionalmente, também foi desenvolvido um protótipo funcional, para que os setores responsáveis pelo acompanhamento da evasão possam utilizar e tomar ações antes que a evasão do estudante se efetive.

A Tabela VIII sumariza os resultados obtidos pelo modelos desenvolvidos, onde, a partir do algoritmo Árvore de Decisão constatou-se melhor performance, obtendo acurácia de 84% e precisão 87% relativo à classe de evasores em comparação com o algoritmo de Redes Neurais que obteve 82% de acurácia e 78% de precisão para a mesma classe de evasores. Já na métrica de Revocação, o algoritmo que utiliza Redes Neurais apresentou melhor resultado, com 82% em comparação com os 78% obtidos pelo modelo que utiliza Árvore de Decisão.

TABELA VIII
Desempenho dos melhores modelos da Rede Neural e Árvore de Decisão

Modelo	Nº	Precisão	Revocação	Acurácia
Rede Neural	4	0.78	0.82	0.82
Árv. de Decisão	6	0.87	0.78	0.84

Fonte: Elaborada pelos autores.

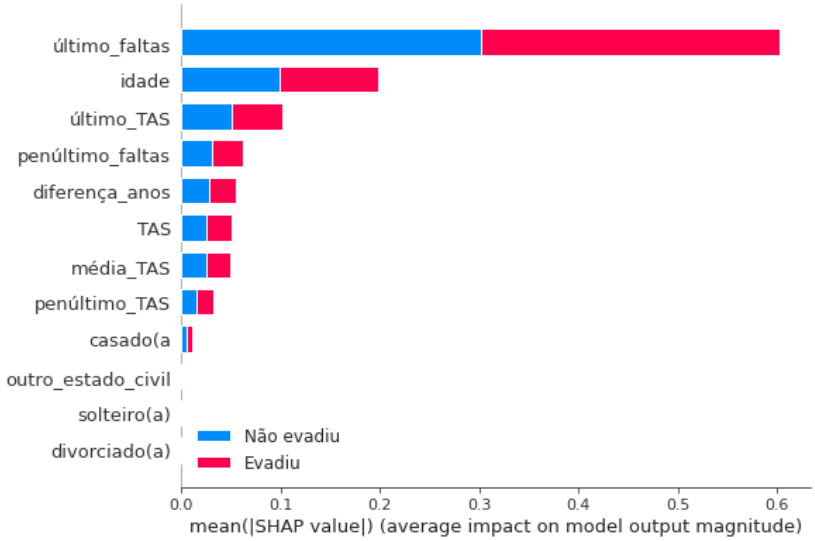
Como pôde-se constatar, os resultados obtidos foram levemente melhores para o algoritmo de Árvore de Decisão em comparação com o de Rede Neural em termos de acurácia. Não há diferença estatística significativa entre os modelos, sendo uns dos pontos a pesar na escolha da Árvore de Decisão para seu uso protótipo, além da maior precisão, sua explicabilidade, natural desta técnica. Uma limitação identificada neste trabalho foi o baixo volume de dados históricos existente no campus, uma vez que ambos os cursos superiores foram instalados recentemente, fazendo com que o modelo não conseguisse se aproximar dos 100% nas métricas testadas. Entretanto, a hipótese de pesquisa e os objetivos deste trabalho foram cumpridos, subsidiando a criação de um modelo para prever a potencial evasão dos estudantes por meio dos dados institucionais, que pode (futuramente) apresentar melhores resultados, ao incorporar novos dados históricos.

Este trabalho abre margem para diversos trabalhos futuros. Primeiro, é possível reaproveitar toda a fase de mineração de dados para toda a instituição, uma vez que todos os campi utilizam o mesmo sistema de informação. Segundo, é possível reutilizar as variáveis identificadas e as

Fig. 2. Conteúdo do protótipo desenvolvido

Fonte: Elaborada pelos autores.

Fig. 3. Importância das colunas da Árvore de Decisão



Fonte: Elaborada pelos autores.

colunas criadas a partir de engenharia de características para predição de evasão tanto para diferentes campi, como também para outras instituições. Terceiro, com a introdução de um conjunto maior de dados históricos, há a possibilidade de se aplicar diversos outros algoritmos de aprendizagem de máquina, de modo a buscar uma aproximação da solução ótima. Por fim, a instituição já demonstrou interesse em integrar o sistema em suas rotinas, e o aprimoramento do protótipo desenvolvido também indica um ramo de continuidade dessa pesquisa.

Para este trabalho, foram utilizados os dados de estudantes do próprio campus, porém anonimizados. O setor responsável justificou aderência à Lei Geral de Proteção de Dados (LGPD Lei 13709/2018 [54]) para a não disponibilização de informações que identificassem os estudantes. A disponibilização do modelo como *data app* é possível atualmente, porém as informações de entrada precisam ser calculadas manualmente pelo usuário.

A disponibilização do modelo como um sistema satélite para a própria instituição é tratada como trabalho futuro, pois depende de acesso aos dados completos dos estudantes do campus em "tempo real", seja via *Web Services* ou *views* de Banco de Dados. Este acesso sendo permitido, seria possível, via sistema, realizar os tratamentos necessários e aplicar ao modelo os dados de todos os estudantes ativos de Ensino Superior do campus para identificar aqueles com tendência à evasão escolar, chamando a atenção do setor responsável pela permanência e êxito dos estudantes. Além disso, haveria ainda a possibilidade de atualização do modelo de *Machine Learning* de forma automática, a cada transição de semestre, o que resultaria em novas fontes de trabalho, especialmente relacionados à temas como *concept drift* [55], que envolve mudança de padrões nos dados ao longo do tempo.

Referências

[1] S. Tsumoto and S. Hirano, "Risk Mining in Medicine: Application of Data Mining to Medical Risk Management," *Fundamenta Informaticae*, vol. 98, no. 1, pp. 107–121, 2010.

[2] D Rajesh, "Application of Spatial Data Mining for Agriculture," *International Journal of Computer Applications*, vol. 15, no. 2, pp. 7–9, 2011.

[3] A. Ropodi, E. Panagou, and G.-J. Nychas, "Data Mining derived from Food Analyses using Non-Invasive/Non-Destructive Analytical Techniques; Determination of Food Authenticity, Quality & Safety in Tandem with Computer Science Disciplines," *Trends in Food Science & Technology*, vol. 50, pp. 11–25, 2016.

[4] D. Enke and S. Thawornwong, "The Use of Data Mining and Neural Networks for Forecasting Stock Market Returns," *Expert Systems with Applications*, vol. 29, no. 4, pp. 927–940, 2005.

[5] C. Garcia, A. Esmin, D. Leite, and I. Škrjanc, "Evolvable Fuzzy Systems from Data Streams with Missing Values: With Application to Temporal Pattern Recognition and Cryptocurrency Prediction," *Pattern Recognition Letters*, vol. 128, pp. 278–282, 2019.

[6] C. Romero and S. Ventura, "Data Mining in Education," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 3, no. 1, pp. 12–27, 2013.

[7] E. Yukselturk, S. Ozekes, and Y. K. Turel, "Predicting Dropout Student: An Application of Data Mining Methods in an Online Education Program," *European Journal of Open, Distance and e-learning*, vol. 17, no. 1, pp. 118–133, 2014.

[8] G. W. Dekker, M. Pechenizkiy, and J. M. Vleeshouwers, "Predicting Students Drop Out: A Case Study," *International Working Group on Educational Data Mining*, 2009.

[9] D. Raju and R. Schumacker, "Exploring Student Characteristics of Retention that Lead to Graduation in Higher Education using Data Mining Models," *Journal of College Student Retention: Research, Theory & Practice*, vol. 16, no. 4, pp. 563–591, 2015.

[10] 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. 01, p. 132, 2014.

[11] L. A. Digiampietri, F. Nakano, and M. de Souza Lauretto, "Mineração de Dados para Identificação de Alunos com Alto Risco de Evasão: Um Estudo de Caso," *Revista de Graduação USP*, vol. 1, no. 1, pp. 17–23, 2016.

[12] L. M. B. Manhães, 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," in *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-sbie)*, vol. 1, 2012.

[13] R. D. Machado, E. O. B. Nara, J. N. C. Schreiber, and G. A. Schwingel, "Estudo Bibliométrico em Mineração de Dados e Evasão Escolar," in *Congresso Nacional de Excelência em Gestao*, vol. 11, 2015, pp. 1–21.

[14] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA Data Mining Software: An Update," *ACM SIGKDD explorations newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[15] R. Goldschmidt, E. Passos, and E. Bezerra, *Data Mining*. Elsevier Brasil, 2015.

[16] R. L. L. Silva Filho, P. R. Motejunas, O. Hipólito, and M. B. d. C. M. Lobo, "A Evasão no Ensino Superior Brasileiro," *Cadernos de Pesquisa*, vol. 37, pp. 641–659, 2007.

[17] N. P. d. L. Gaioso, "O Fenômeno da Evasão Escolar na Educação Superior no Brasil," *Brasília, DF: Universidade Católica de Brasília*, p. 20, 2005.

[18] M. Lobo, "Panorama da Evasão no Ensino Superior Brasileiro: Aspectos Gerais das Causas e Soluções," *Abmes Cadernos*, vol. 25, pp. 9–58, 2012.

[19] B. M. da Educação. Secretaria da Educação SESU, "Diplomação, Retenção e Evasão nos Cursos de Graduação em Instituições de Ensino Superior Públicas," *Avaliação: Revista de Rede de Avaliação Institucional da Educação Superior. Campinas*, vol. 1, no. 2, pp. 55–65, 1996.

[20] A. A. Steimbach, "Juventude, Escola e Trabalho: Razões da Permanência e do Abandono no Curso Técnico em Agropecuária Integrado," 2012.

[21] L. B. Pelissari, "O Fetiche da Tecnologia e o Abandono Escolar na Visão de Jovens que procuram a Educação Profissional Técnica de Nível Médio," 2012.

[22] J. Han, J. Pei, and M. Kamber, *Data mining: concepts and techniques*. Elsevier, 2011.

[23] U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, *et al.*, "Knowledge Discovery and Data Mining: Towards a Unifying Framework," in *KDD*, vol. 96, 1996, pp. 82–88.

[24] T. S. Korting, "C4.5 Algorithm and Multivariate Decision Trees," *Image Processing Division, National Institute for Space Research–INPE São José dos Campos–SP, Brazil*, p. 22, 2006.

[25] E. Alpaydin, *Introduction to Machine Learning*. MIT press, 2020.

[26] M. Alban and D. Mauricio, "Predicting University Dropout through Data Mining: A Systematic Literature," *Indian Journal of Science and Technology*, vol. 12, no. 4, pp. 1–12, 2019.

[27] K. Gurney, *An Introduction to Neural Networks*. CRC press, 2018.

[28] K. Hornik, M. Stinchcombe, and H. White, "Multilayer Feedforward Networks are Universal Approximators," *Neural Networks*, vol. 2, no. 5, pp. 359–366, 1989.

[29] D. G. Ferrari and L. N. d. C. Silva, *Introdução a Mineração de Dados*. Saraiva Educação SA, 2017.

[30] A. Viloria, J. G. Guliany, W. N. Núñez, H. H. Palma, and L. N. Núñez, "Data Mining Applied in School Dropout Prediction," in *Journal of Physics: Conference Series*, IOP Publishing, vol. 1432, 2020, p. 012 092.

[31] A. Behr, M. Giese, K. Theune, *et al.*, "Early Prediction of University Dropouts – A Random Forest Approach," *Jahrbücher für Nationalökonomie und Statistik*, vol. 240, no. 6, pp. 743–789, 2020.

[32] T. M. Barros, P. A. Souza Neto, I. Silva, and L. A. Guedes, "Predictive Models for Imbalanced Data: A School Dropout Perspective," *Education Sciences*, vol. 9, no. 4, p. 275, 2019.

[33] T. Fernández-Martín, M. Solís-Salazar, M. T. Hernández-Jiménez, and T. E. Moreira-Mora, "A Multinomial and Predictive Analysis of Factors Associated with University Dropout," *Revista Electrónica Educare*, vol. 23, no. 1, pp. 73–97, 2019.

[34] P. M. da Silva, M. N. Lima, W. L. Soares, I. R. Silva, A. d. A. Roberta, and F. F. de Souza, "Ensemble Regression Models Applied to Dropout in Higher Education," in *2019 8th Brazilian Conference on Intelligent Systems (BRACIS)*, IEEE, 2019, pp. 120–125.

[35] J. d. O. Kelly, A. G. Menezes, A. B. de Carvalho, and C. A. Montesco, "Supervised Learning in the Context of Educational Data Mining to Avoid University Students Dropout," in *2019 IEEE 19th International Conference on Advanced Learning Technologies (ICALT)*, IEEE, vol. 2161, 2019, pp. 207–208.

[36] N. Hutagaol and S. Suharjito, "Predictive Modelling of Student Dropout Using Ensemble Classifier Method in Higher Education," *Advances in Science, Technology and Engineering Systems Journal*, vol. 4, no. 4, pp. 206–211, 2019.

[37] A. Viloria, J. G. Padilla, C. Vargas-Mercado, H. Hernández-Palma, N. O. Llinas, and M. A. David, "Integration of Data Technology for Analyzing University Dropout," *Procedia Computer Science*, vol. 155, pp. 569–574, 2019.

[38] A. Acero, J Achury, and J Morales, "University Dropout: a Prediction Model for an Engineering Program in Bogotá, Colombia," in *8th Research in Engineering Education Symposium: Making Connections, REES*, 2019, pp. 483–490.

[39] J. Figueroa-Cañas and T. Sancho-Vinuesa, "Predicting Early Dropout Student Is a Matter of Checking Completed Quizzes: the Case of an Online Statistics Module," in *LASI-SPAIN*, 2019, pp. 100–111.

[40] R. L. S. do Nascimento, R. B. das Neves Junior, M. A. de Almeida Neto, and R. A. de Araújo Fagundes, "Educational Data Mining: an Application of Regressors in Predicting School Dropout," in *International Conference on Machine Learning and Data Mining in Pattern Recognition*, Springer, 2018, pp. 246–257.

[41] J. Lima, P. Alves, M. Pereira, and S. Almeida, "Using Academic Analytics to Predict Dropout Risk in Engineering Courses," in *17th European Conference on e-Learning ECEL 2018*, Academic Conferences and Publishing International Limited, 2018, pp. 316–321.

[42] M. S. Alban and D. Mauricio, "Prediction of University Dropout Through Technological Factors: a Case Study in Ecuador," *Revista Espacios*, vol. 39, no. 52, 2018.

[43] S. Rovira, E. Puertas, and L. Igual, "Data-driven System to Predict Academic Grades and Dropout," *PLoS one*, vol. 12, no. 2, e0171207, 2017.

[44] B. Peralta, T Poblete, and L Caro, "Automatic Feature Selection for Desertion and Graduation Prediction: a Chilean Case," in *2016 35th International Conference of the Chilean Computer Science Society (SCCC)*, IEEE, 2016, pp. 1–8.

[45] C. F. Rocha, Y. F. Zelaya, D. M. Sánchez, and F. A. F. Pérez, "Prediction of University Desertion through Hybridization of Classification Algorithms," in *SIMBig*, 2017, pp. 215–222.

[46] L. Oneto, A. Siri, G. Luria, and D. Anguita, "Dropout Prediction at University of Genoa: a Privacy Preserving Data Driven Approach," in *ESANN*, 2017.

[47] E. Y. Sari, A. Sunyoto, *et al.*, "Optimization of Weight Backpropagation with Particle Swarm Optimization for Student Dropout Prediction," in *2019*

4th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE), IEEE, 2019, pp. 423–428.

[48] B. Perez, C. Castellanos, and D. Correal, "Applying Data Mining Techniques to Predict Student Dropout: a Case Study," in *2018 IEEE 1st colombian conference on applications in computational intelligence (colcaci)*, IEEE, 2018, pp. 1–6.

[49] K. Potdar, T. S. Pardawala, and C. D. Pai, "A Comparative Study of Categorical Variable Encoding Techniques for Neural Network Classifiers," *International journal of computer applications*, vol. 175, no. 4, pp. 7–9, 2017.

[50] P. Domingos, "A Few Useful Things to Know About Machine Learning," *Communications of the ACM*, vol. 55, no. 10, pp. 78–87, 2012.

[51] G. Dong and H. Liu, *Feature Engineering for Machine Learning and Data Analytics*. CRC Press, 2018.

[52] C. Garcia, D. Leite, and I. Škrjanc, "Incremental Missing-data Imputation for Evolving Fuzzy Granular Prediction," *IEEE Transactions on Fuzzy Systems*, vol. 28, no. 10, pp. 2348–2362, 2019.

[53] C. O'neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Broadway Books, 2016.

[54] Brasil, "Lei Geral de Proteção de Dados Pessoais (LGPD)," *Disponível em <http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm>*, 2018.

[55] J. P. Barddal, H. M. Gomes, and F. Enembreck, "SFNClassifier: A Scale-free Social Network Method to Handle Concept Drift," in *Proceedings of the 29th Annual ACM Symposium on Applied Computing*, 2014, pp. 786–791.