Uso de Ontologias no Suporte a Aplicação de Machine Learning: um Caso no Domínio de Evasão Escolar

Eduardo Moura da Silva[1], Filipe Wall Mutz[2] and Fabiano Borges Ruy[1]

[1]*Programa de Pós-graduação em Computação Aplicada (PPComp) - Instituto Federal do Espírito Santo (IFES), Av. dos Sabiás, 330 - Morada de Laranjeiras, Serra - ES, Brasil*

[2]*Departamento de Informática, Universidade Federal do Espírito Santo (UFES), Av. Fernando Ferrari, 514 - Goiabeiras, Vitória - ES, Brasil*

Abstract

With technological advances in the area of Artificial Intelligence, studies and applications combining machine learning and ontologies are increasingly present. In this convergence, ontologies enable a better understanding of the domain and the data, which is essential for providing semantic integration of different data sources and for favoring a proper application of ML techniques. This combination allows creating solutions that generate ML models able for dealing with different data sources. This paper presents a case in the field of school dropout, demonstrating how this combination allows the development of technological resources for predicting school dropout that can be applied to different educational institutions.

Keywords

Ontologies, Machine Learning, School Dropout, Semantic Integration

1. Introdução

Segundo Gaioso [1], a evasão escolar é um fenômeno social, definido como a interrupção no ciclo de estudos. Para Araújo et al. [2], a evasão é caracterizada pelo abandono do curso, rompendo com o vínculo estabelecido, não renovando o compromisso ou sua manifestação de continuar no estabelecimento de ensino. A Comissão Especial de Estudos sobre a Evasão nas Universidades Públicas Brasileiras [3], designada pelo MEC, define evasão como a saída definitiva do aluno de seu curso de origem, sem concluí-lo. Percebe-se diferentes amplitudes nos conceitos de evasão, mas de forma geral, tem-se como fator comum a interrupção no ciclo de estudos.

Este trabalho adota um conceito baseado nas definições de [1] e [3], considerando a evasão escolar como um fenômeno social, definido como a interrupção no ciclo de estudos que dê origem à saída definitiva do curso, da instituição ou do sistema escolar.

A evasão escolar é uma questão que preocupa não apenas os gestores escolares, mas também o Estado e a sociedade em geral, pois é um fenômeno que gera impacto para todos. Esse é um problema que preocupa as instituições de ensino públicas e privadas, pois as saídas de alunos provocam graves consequências sociais, acadêmicas e econômicas [4]. Para a instituição, a

Proceedings of the 15th Seminar on Ontology Research in Brazil (ONTOBRAS) and 6th Doctoral and Masters Consortium on Ontologies (WTDO), November 22-25, 2022

eduardo.silva@ifes.edu.br (E. M. d. Silva); filipe.mutz@ufes.edu.br (F. W. Mutz); fabianoruy@ifes.edu.br (F. B. Ruy)

evasão acarreta ociosidade do espaço físico, de professores, de funcionários e de equipamentos, o que, nas instituições públicas, se reflete em desperdícios dos investimentos do governo e, nas particulares, perdas financeiras em relação às mensalidades [5]. Para os estudantes, por sua vez, a evasão pode representar o atraso ou cancelamento de um sonho, perda de oportunidades de trabalho, de crescimento pessoal e de melhoria de renda, entre muitas outras consequências.

Visto que a evasão escolar é um relevante problema que causa grandes impactos na sociedade, há a necessidade de compreender esse fenômeno e tomar ações para mitigá-lo. Nesse sentido, dados acadêmicos e sociais dos alunos cumprem um importante papel, permitindo compreender o contexto e as situações que podem levar a evasão. Tais dados podem ser trabalhados de diversas maneiras: com a utilização de ferramentas de *business intelligence* para geração de relatórios e *insights*, com uso de técnicas de mineração de dados, e com a aplicação de técnicas de Inteligência Artificial (IA), como as de Aprendizado de Máquina, que podem antecipar situações a serem tratadas.

Aprendizado de Máquina (*Machine Learning - ML*) envolve um conjunto de técnicas que empregam um princípio de inferência denominado indução, no qual é possível obter conclusões genéricas a partir de um conjunto particular de exemplos [6]. Algoritmos de ML se propõem a resolver tarefas pela identificação de padrões em dados de um determinado domínio e diversas técnicas são utilizadas para tentar selecionar soluções que generalizem para dados ainda não conhecidos.

Beltran et al. [7] ressaltam que é importante conhecer bem os dados disponíveis para aplicar corretamente as técnicas de ML. Modelos conceituais, como ontologias, são ferramentas que auxiliam a compreensão do domínio estudado. Ontologias podem representar de forma consistente o conhecimento de um domínio, por meio de seus conceitos e relações, indicando a interpretação desejada ao domínio, independentemente das aplicações específicas [8]. Elas são artefatos de representação da informação muito úteis para integração de dados e para garantir interoperabilidade semântica das informações que representam [9]. Assim, ontologias estão sendo cada vez mais usadas para fornecer conhecimento em análises baseadas em similaridade e modelos de ML. Os métodos empregados para combinar ontologias e ML ainda são novos e estão sendo ativamente desenvolvidos [10].

Tal combinação é favorável especialmente em aplicações em que se deseja processar múltiplas fontes de dados. No contexto de evasão escolar, técnicas de integração semântica e ML podem se complementar. Enquanto uma abordagem de integração semântica produz uma representação consistente de dados escolares independentemente de quais ou quantas fontes, técnicas de ML podem favorecer a identificação de padrões de comportamentos úteis à gestão escolar. Como é um fenômeno presente em diferentes contextos escolares, é fundamental que soluções tecnológicas construídas com objetivo de minimizar esse fenômeno possam ser aplicadas a diferentes entidades educacionais. Um dos grandes desafios para a criação de soluções com essa característica está relacionado ao fato de que os dados advêm de fontes distintas e estão armazenados segundo formatos, esquemas e, muitas vezes, semânticas diferentes.

Este trabalho apresenta um *case* no domínio de evasão escolar que demonstra como o uso de ontologias pode apoiar a aplicação de técnicas de ML para realizar a predição da evasão escolar, e como tal combinação favorece a aplicação a diferentes instituições de ensino. Como contribuições, vale citar: uma Ontologia de Evasão Escolar que permite um melhor entendimento do domínio e a geração de repositórios de dados padronizados; uma abordagem que propõe

os passos para a criação de recursos tecnológicos combinando ontologias e ML; e um *case* apresentando resultados de predições de Evasão Escolar a partir de duas bases de dados públicas.

O artigo está organizado da seguinte forma: a Seção 2 aborda alguns conceitos importantes relacionados às tecnologias base deste artigo: ontologias e *machine learning*; a Seção 3 apresenta o desenvolvimento de um case de integração e aplicação de ML para predição de Evasão Escolar usando uma ontologia como suporte; a Seção 4 discute os trabalhos correlatos; e a Seção 5, as considerações finais.

2. Ontologias e ML

De acordo com Studer et al. [11], uma ontologia é uma especificação formal e explícita de uma conceituação compartilhada. É uma teoria lógica utilizada para capturar os modelos pretendidos de uma conceituação e excluir os não pretendidos [12]. Ou seja, uma teoria utilizada para especificar e explicitar uma conceituação.

O uso de ontologias pode facilitar a integração de dados de várias maneiras, incluindo representação de metadados, verificação automática de dados, conceituação global, suporte para consultas semânticas de alto nível e se estende além das abordagens tradicionais de uso de elementos de dados comuns e modelos de dados comuns [13].

As ontologias têm se mostrado como importante ferramenta para realização de integração de dados [14] [8], entretanto o processo de criação de ontologias não é trivial. Além do conhecimento do domínio, é importante aplicar técnicas de Engenharia de Ontologias e recursos tais como ontologias de fundamentação e uma linguagem de modelagem adequada. Visando maior expressividade na representação do domínio e facilidades na criação e evolução da ontologia, bem como na criação de repositórios de dados padronizados.

Este trabalho adota UFO e OntoUML. UFO [15] é uma ontologia de fundamentação que define distinções úteis para compreender e representar um domínio. São providas distinções básicas tais como sortais rígidos (*kind, subkind*) e antirrígidos (*role, phase*), mediadores em relações materiais (*relator*), além de não sortais para generalizações (*category, rolemixin*). A linguagem OntoUML [16] captura essas distinções em uma extensão da UML, e tem sido usada na construção de modelos conceituais em diversos domínios [17].

Com o estabelecimento de uma ontologia no domínio e escopo desejado, esta pode ser utilizada como referência semântica para a integração de dados, promovendo a interoperabilidade de dados a partir de uma base comum para interpretação e redução de inconsistências conceituais [14]. Os dados, independentemente de suas origens, podem ser mapeados para um repositório comum que provê uniformidade para aplicações diversas [8].

Mais especificamente, um repositório baseado em uma ontologia provê uma estrutura adequada, semanticamente enriquecida, para a aplicação de algoritmos de ML. Além de permitir lidar de maneira homogênea com múltiplas fontes de dados, a ontologia também fornece um melhor entendimento do domínio e das características inerentes aos dados, facilitando, na aplicação de ML, a transformação de campos específicos de entrada para otimizar o processamento e a seleção dos algoritmos a serem aplicados.

As características dos dados e do problema abordado levam à aplicação de algoritmos de classificação que estão contidos no aprendizado de máquina supervisionado. No Aprendizado

Supervisionado, para cada amostra apresentada ao algoritmo de aprendizado é necessário definir a saída que o modelo deve produzir para uma dada entrada [18]. Quando as saídas são discretas, esse problema é chamado de classificação e para valores contínuos, é chamado de regressão. Em classificação, cada exemplo é composto por uma entrada (e.g., uma imagem, um áudio ou vetor de valores (atributos)), e por uma classe de saída associada. O objetivo do algoritmo é construir um modelo capaz de determinar corretamente a classe de exemplos diferentes daqueles usados durante o treinamento.

Um problema de classificação pode ser definido formalmente da seguinte maneira: dado um conjunto de exemplos de treinamento composto por pares (x_i, c_j), no qual x_i representa um vetor de atributos de entrada, e c_j sua classe associada, deve-se encontrar uma função que mapeie cada x_i para sua classe associada c_j, tal que $i = 1, 2, ..., n$, em que n é o número de exemplos de treinamento, e $j = 1, 2, ..., m$, em que m é o número de classes do problema [19].

Com algoritmos de classificação é possível realizar algumas predições no contexto da evasão escolar. Por exemplo, Fernando Filho et al. [20] apresentam um estudo de caso no qual realizou-se predições de evasão escolar com o uso de técnicas de classificação. Contudo a aplicação dessas técnicas, assim como no exemplo citado, são muitas das vezes direcionadas para uma base de dados específica.

De acordo com Kulmanov et al. [10], com o rápido crescimento de métodos para construir modelos preditivos, em particular métodos de ML, ontologias podem agora desempenhar um papel no fornecimento sistemático de conhecimento de domínio para habilitar ou melhorar os modelos preditivos.

O uso de uma abordagem de integração de dados baseada em ontologia permite não apenas padronizar as definições de variáveis de dados por meio de um vocabulário comum e controlado, mas também torna as relações semânticas entre variáveis de diferentes fontes explícitas e claras para todos os usuários [13].

3. Uso de ontologias no suporte à integração de dados e predição da evasão escolar

A partir da ontologia, para realizar o ciclo completo desde a captura dos dados de fontes distintas até a predição da evasão escolar, foram necessárias atividades tais como a identificação das fontes de dados, o acesso e captura dos dados, a transformação dos dados para o modelo comum, e experimentação com técnicas de ML. Tais atividades foram organizadas em uma abordagem, conforme apresenta a Figura 1.

Em um trabalho anterior [21], foi utilizada uma versão da abordagem que contemplava apenas análise de dados, mas não o uso de técnicas de ML. Agora a abordagem está mais completa, com melhorias e extensões voltadas a construir e aplicar modelos de ML para realizar predições.

O suporte provido pela ontologia para abstrair a origem dos dados e possibilitar a realização de predições ocorre dentro da abordagem. É a partir da ontologia que os dados podem ser padronizados em repositórios usando o mesmo *schema*, para posterior geração e aplicação de modelos de ML. A Figura 1 apresenta etapas da abordagem, descritas independentemente de domínio, e aplicadas no case de evasão escolar.

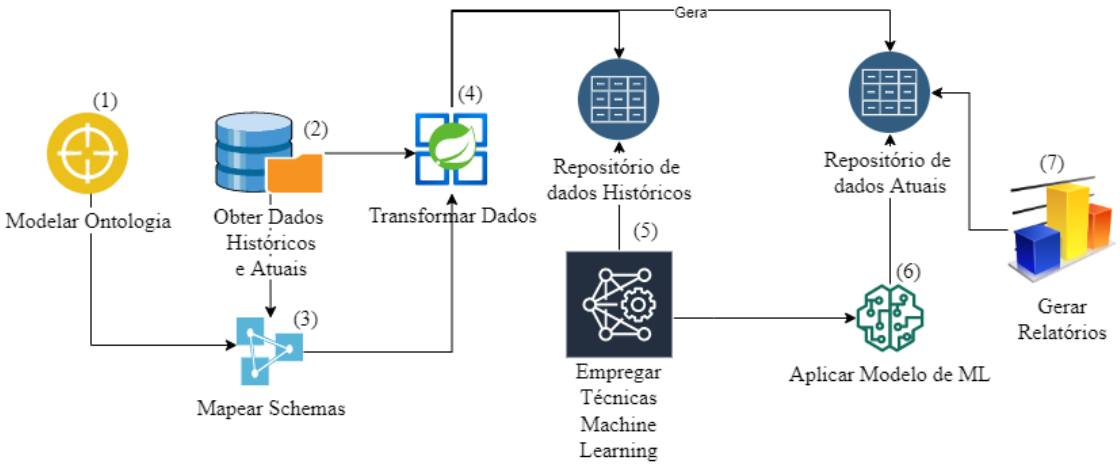

Figura 1: Abordagem para aplicação de ML em dados de diferentes fontes.

1. Modelar Ontologia - cria uma Ontologia de Referência, construída a partir da conceituação do domínio.
2. Obter Dados - consiste em acessar / capturar os dados disponíveis do domínio a partir de suas fontes. Para aplicação de técnicas de ML é importante o acesso a dados históricos, pois é o que permite a identificação dos padrões por parte dos algoritmos. Portanto, nesta etapa deve-se ter acesso a dados históricos e dados atuais para os quais se deseja realizar a predição.
3. Mapear *Schemas* - realiza o mapeamento semântico entre os *schemas* das Fontes de Dados (2) e a Ontologia de Referência (1), indicando qual é a relação dos tipos dos dados com os correspondentes conceitos e propriedades da ontologia.
4. Transformar Dados - processa os dados, conforme o mapeamento, para um formato baseado na ontologia. Para aplicação das técnicas de ML, são criados dois repositórios, um para os dados históricos e outro para os dados atuais, ambos populados com instâncias da ontologia criadas a partir de diferentes recortes dos dados da base.
5. Empregar Técnicas de ML - baseado no repositório de dados históricos, faz a aplicação das técnicas de aprendizado de máquina para geração do modelo de ML.
6. Aplicar Modelo de ML - realiza a predição para os dados atuais baseada no modelo gerado pela aplicação das técnicas de ML.
7. Gerar Relatórios - gera relatórios com dados das predições realizadas pelo modelo de ML.

Com a utilização de uma ontologia como interlíngua, os mesmos tipos de informações podem ser extraídos de forma transparente a partir de diferentes origens de dados. Por exemplo, aplicando a abordagem para uma Instituição A, será gerado um repositório padronizado pela ontologia para essa instituição; de posse desse repositório podem ser aplicadas técnicas de ML para realizar predições acerca de seus dados. Em seguida, sem a necessidade de adequações na aplicação, pode-se aplicar a abordagem para uma Instituição B, gerando seu próprio repositório. Como o formato dos repositórios é padronizado pela ontologia, as mesmas técnicas de ML

poderão ser aplicadas para geração dos modelos de ML e para a realização de predições para a Instituição B ou quaisquer outras no mesmo contexto.

3.1. Ontologia de Evasão Escolar

Aplicando-se a abordagem, a primeira atividade é a criação da Ontologia de Referência, neste caso, no domínio de Evasão Escolar. Seu propósito é representar as principais características do domínio, necessárias para possibilitar a identificação de padrões e um melhor entendimento dos fatores que influenciam na decisão das pessoas evadirem. Ela é utilizada para abstrair a origem e formato dos dados e para gerar repositórios padronizados que permitam a aplicação de técnicas de ML para múltiplas fontes, gerando modelos capazes de identificar alunos com maiores potenciais de evasão. Assim, a ontologia pode ser utilizada para projetar uma solução de IA que possa ser utilizada em diferentes instituições, com variados tipos e níveis de escolaridade.

A ontologia foi construída com base no método *SABiO (Systematic Approach for Building Ontologies)* [22]. Para a fase de captura e formalização foi utilizado UFO-A [23] e os conceitos foram definidos a partir de referências como: Lei 9.394/1996 - Lei de Diretrizes e Bases da Educação; Decreto 9.235/2017 - que dispõe sobre o exercício das funções de regulação, supervisão e avaliação das instituições de educação superior; Constituição Federal Brasileira de 1988; glossário do Censo da Educação Superior, realizado pelo INEP [24]; apresentação de resultados do Censo da Educação Superior [25]; tabela de classificação de áreas de conhecimento da Capes [26]; e publicações, tais como [27] e [28]. Como linguagem de modelagem foi utilizada OntoUML [23].

A ontologia, apresentada na Figura 2, é uma evolução da versão publicada em [21], a partir da qual foram adicionados conceitos para representar variados níveis de ensino (além do ensino superior), e outros conceitos no módulo acadêmico para ser possível representar notas e frequências dos alunos; além de informações sobre as probabilidades de evasão. Os conceitos que foram acrescentados estão representados em cores mais claras dentro de cada módulo, conforme pode ser visualizado na Figura 2.

No módulo Organização (em amarelo), o principal conceito é **Instituição Educacional**, que representa os tipos de organizações que estão inseridas no domínio, que são organizações que oferecem atividades educacionais. Elas possuem uma **Categoria Administrativa** e podem ser dos tipos **Universidade, Faculdade, Instituto** ou, voltadas à educação básica, **Escola** ou **Colégio**. Estas instituições englobam um ou mais **Níveis Escolares**, ou seja, podem ser da educação superior e/ou da educação básica (infantil, fundamental e médio). Para ilustrar instâncias desses conceitos: a *Escola Estadual de Ensino Fundamental Manuel Lopes* é uma instância de Escola, possui Categoria Administrativa *pública estadual*, e é uma Instituição Nível *fundamental*.

No módulo Socioeconômico (em azul), o principal conceito é de **Pessoa**, que apresenta algumas características que podem ser úteis para a identificação de padrões de evasão escolar: **Cor**, **Estado Civil**, **Idade**, **Gênero** e **Renda**. Por exemplo, *Maria da Silva* é uma instância de Pessoa, que possui Estado Civil *solteira*, Cor *parda*, Renda per capta de *R$ 1200*, Idade de *13* anos e Gênero *feminino*.

No módulo Acadêmico (em verde), os principais conceitos são **Turma**, **Aluno** e **Matrícula**. A turma pode ser **Turma da Educação Básica** ou **Turma da Educação Superior**, é formada

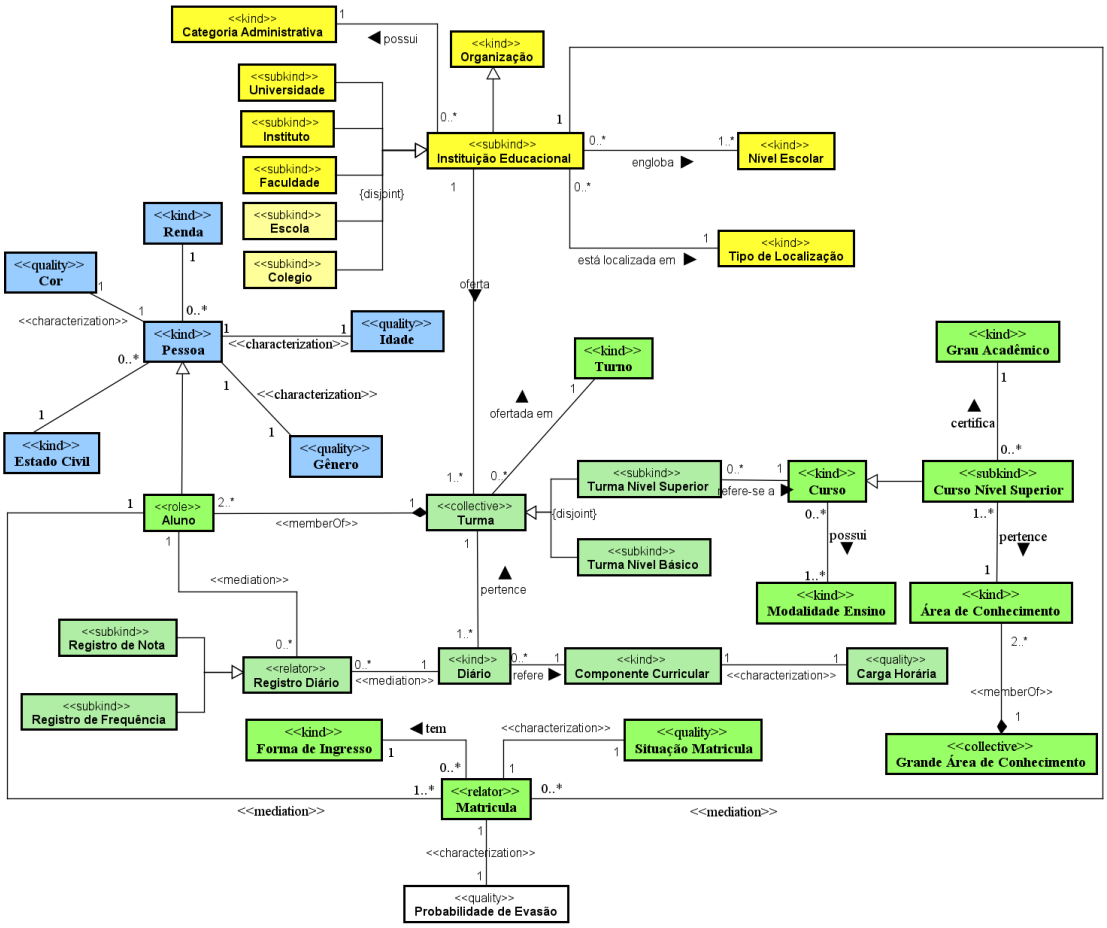

Figura 2: Ontologia de Evasão Escolar.

por um conjunto de **Alunos** e possui **Diários** para os seus **Componentes Curriculares**. Os diários são os locais onde se fazem **Registros de Notas** e **Registros de Frequências** dos alunos da turma. Como instâncias desses conceitos tem-se por exemplo, Turma da educação básica *6º ano A*, que tem como um de seus Alunos *Maria da Silva* e possui o Diário *23297* do Componente Curricular *Matemática*, neste diário há um Registro de Nota *9,5* para *Maria da Silva* em uma atividade avaliativa.

Aluno é o papel (*role*) que uma Pessoa assume ao fazer Matrícula em uma Instituição Educacional. A **Matrícula** é responsável por estabelecer uma relação (*relator*) de vínculo entre o **Aluno** e a **Instituição Educacional**, e possui uma **Situação de Matrícula** (Desvinculado do curso, Formado, Matriculado), por meio da qual se saberá, por exemplo, se um aluno evadiu ou concluiu.

Por fim, o conceito **Probabilidade de Evasão** foi adicionado à ontologia como uma forma de representar os resultados das predições realizadas pelo modelo de ML, complementando o modelo e permitindo um suporte na aplicação de ML e outras técnicas.

Sexo	Renda	Cor / Raça	Idade	...	Código da Matrícula	Data de Matrícula	Situação da Matrícula	...	...	Turno
F	2,5<RFP<=3,5	Preta	17	x	66777851	01/02/2016	CONCLUÍDA	y	z	Integral
F	0<RFP<=0,5	Preta	17	x	66777845	01/02/2016	CONCLUÍDA	y	z	Integral
M	1<RFP<=1,5	Não Declarada	23	x	85115376	01/02/2018	EM_CURSO	y	z	Integral
M	NÃO DECLARADA	Não Declarada	19	x	66440888	01/02/2016	EM_CURSO	y	z	Integral
M	0,5<RFP<=1	Não Declarada	19	x	66776361	01/02/2016	DESLIGADA	y	z	Integral
M	0,5<RFP<=1	Não Declarada	18	x	66776337	01/02/2016	DESLIGADA	y	z	Integral

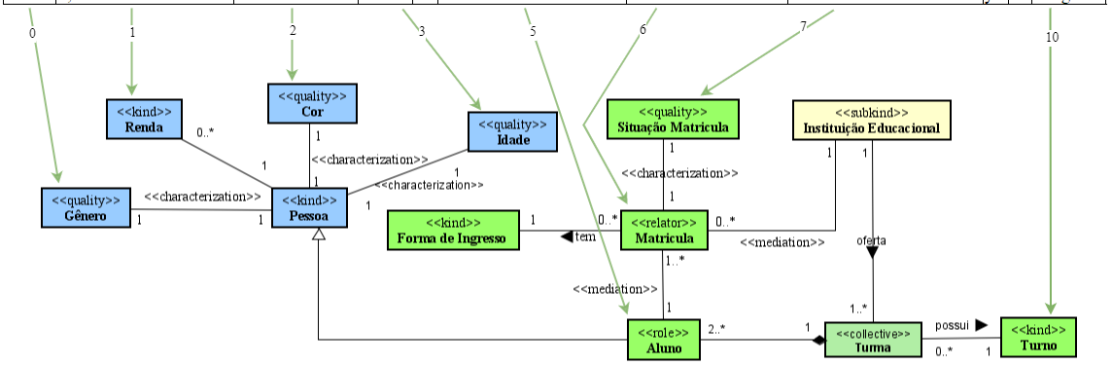

Figura 3: Exemplo de como é realizado o mapeamento dos dados para ontologia.

As definições dos conceitos da Ontologia de Evasão Escolar e suas respectivas referências estão disponíveis nesta página[1].

3.2. Mapeamento dos dados para a ontologia

Como fontes de dados para a aplicação foram utilizados microdados do ano base 2019, da Plataforma Nilo Peçanha (PNP)[2] e do Censo da Educação Superior do INEP[3]. Os dados da PNP do ano base de 2020[4] foram utilizados como referência a dados atuais, base para o qual são realizadas as predições e gerados os resultados.

Foi definido um padrão de mapeamento dos dados para os conceitos da ontologia, em que, inicialmente, a estrutura dos dados de cada fonte foi estudada para melhor compreensão. Em seguida, em cada base de dados (PNP e Censo), para cada conceito, foi identificado o dado com a semântica correspondente (os termos foram utilizados como apoio para encontrar as correspondências, mas elas são definidas pelo significado do dado/conceito). A Figura 3 ilustra o mapeamento. Neste processo de mapeamento, é gerado manualmente um arquivo para cada fonte de dados no formato CSV (*comma-separated values*), que foi selecionado por ser mais adequado aos algoritmos a serem aplicados. Assim, como resultado do mapeamento, para cada origem de dados é gerado um arquivo, no qual a primeira coluna é o índice do dado na sua origem e a segunda coluna é o respectivo conceito ou propriedade da ontologia.

Na Figura 3 estão representados alguns dados da PNP (na tabela) mapeados para conceitos da Ontologia de Evasão Escolar. As setas representam os mapeamentos entre a fonte de dados e a

[1]Conceitos da Ontologia de Evasão Escolar: https://github.com/ontologia/conceitos-evasao-escolar/wiki/Conceitos-da-Ontologia-Evas%C3%A3o-Escolar---v2

[2]Microdados PNP ano base 2019: http://dadosabertos.mec.gov.br/pnp/item/118-2019-microdados-matriculas

[3]Censo Educação Superior ano base 2019: https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/censo-da-educacao-superior

[4]Microdados PNP ano base 2020: http://dadosabertos.mec.gov.br/pnp/item/134-2020-microdados-matriculas

ontologia, com os números representando os índices das colunas de dados em sua fonte. Por exemplo, o campo *Código de Matrícula* é mapeado para o conceito **Aluno**, pois é o identificador do aluno no domínio. Se um aluno possuir mais de uma matrícula ao longo do tempo, haverá duas instâncias do *role* **Aluno**. Já o campo *Data de Matrícula* é mapeado para o *relator* **Matrícula**.

De posse dos dados e dos arquivos de mapeamento é possível gerar os repositórios de instâncias da ontologia. Para isso foi desenvolvida uma aplicação que executa o passo 4 da abordagem, Transformar Dados, gerando os repositórios padronizados pela ontologia. Importante ressaltar que para cada origem de dados será gerado um repositório com o *schema* padronizado pela ontologia. As duas bases utilizadas para o experimento possuem grandes volumes de dados, reunindo informações de diversas instituições de ensino do Brasil, por isso foi importante a definição de uma arquitetura de alto desempenho para a criação dos repositórios.

3.3. Treinamento e Avaliação de Modelos de ML

Uma vez que os dados das múltiplas fontes estão uniformizados e integrados com base na ontologia, eles são utilizados na etapa 5 para treinar e avaliar algoritmos de ML. Os algoritmos recebem como entrada os dados socioeconômicos e de contexto acadêmico de um estudante e geram como saída uma classificação indicando se a previsão é de o aluno evadir ou não. Tais predições foram obtidas a partir de dados históricos das instituições. Os experimentos foram realizados de forma independente usando as bases de dados da PNP e do Censo. Para cada base, os dados de um ano foram utilizados para treinamento e avaliação dos modelos, e dados do ano seguinte para demonstração de seu uso. O código-fonte desta etapa foi desenvolvido usando a linguagem de programação *Python* e a biblioteca *scikit-learn* [29].

A validação cruzada aninhada [30] foi utilizada para busca de hiperparâmetros e avaliação dos modelos. São exemplos de hiperparâmetros o número de k vizinhos mais próximos, utilizados pelo algoritmo KNN, e o número m de árvores de decisão presentes no algoritmo *Boosting*. Apesar de ter um maior custo computacional, essa validação leva a uma estimativa mais correta da performance dos modelos em um ambiente de produção. Nesta técnica, é realizado um primeiro nível de *K-Fold cross validation* em que a cada iteração um *fold* é utilizado como conjunto de teste e os demais como treinamento. Para cada conjunto de teste, é realizado um segundo nível de *K-Fold cross validation* sobre os dados separados para treinamento. A cada iteração do segundo nível, um *fold* é utilizado como conjunto de validação e os demais como treinamento. O objetivo do segundo nível é selecionar os hiperparâmetros que maximizem a performance média nos *folds* de validação [30]. Ao final do *loop* interno (o segundo nível), o conjunto completo de treinamento e os hiperparâmetros são utilizados para treinar os modelos que, em seguida, são avaliados usando o conjunto de teste. As métricas reportadas são as médias dentre os *folds* de teste. É importante enfatizar que os dados de teste permanecem intocados pelo *loop* interno.

Um conjunto de preprocessamentos foi aplicado sobre os atributos oriundos dos conceitos do domínio: dados faltantes foram preenchidos utilizando a moda dos valores; para substituição de dados categóricos foram realizados experimentos com *OneHotEncoder* e com *OrdinalEncoder* (a primeira estratégia aumenta a dimensão dos dados e não apresentou melhoria no desempenho, portanto foram substituídos por números inteiros); foi realizada uma operação de normalização para mapear os valores para o intervalo $[0, 1]$; por fim, foram selecionados atributos com maior

	Accuracy	Precision	Recall	F1-Score
Decision Trees	0.610000	0.500000	0.340000	0.180000
KNN	0.730000	0.540000	0.230000	0.180000
Random Forest	0.650000	0.520000	0.310000	0.160000
Extra Trees	0.610000	0.510000	0.340000	0.170000
Ada Boost	0.600000	0.540000	0.480000	0.270000
Bagging	0.610000	0.530000	0.440000	0.250000
Gradient Boosting	0.680000	0.560000	0.280000	0.140000

(a) PNP

	Accuracy	Precision	Recall	F1-Score
Decision Trees	0.770000	0.540000	0.340000	0.220000
KNN	0.890000	0.630000	0.100000	0.160000
Random Forest	0.830000	0.560000	0.250000	0.220000
Extra Trees	0.790000	0.540000	0.300000	0.210000
Ada Boost	0.610000	0.530000	0.540000	0.210000
Bagging	0.610000	0.530000	0.540000	0.210000
Gradient Boosting	0.900000	0.710000	0.050000	0.090000

(b) Censo

Figura 4: Tabelas de métricas

score em um Teste F usando as anotações [31]. Os parâmetros de preprocessamento foram obtidos no *loop* interno da validação cruzada aninhada.

Os algoritmos de ML avaliados foram KNN, Decision Trees, Bagging, Random Forests, Extra Trees, AdaBoost e Gradient Boosting [31]. A maioria desses modelos são *ensembles* de *Decision Trees* que são conhecidos por exibirem bom desempenho com dados tabulares [32, 33]. *Ensembles* combinam modelos de forma a obter resultados melhores do que aqueles que seriam alcançados individualmente [33]. Uma diversidade de projeções (seleções) sobre os atributos, seleções de amostras de treinamento e técnicas de fusão das predições dos modelos podem ser utilizadas para alcançar este objetivo.

Os algoritmos de ML foram avaliados utilizando matrizes de confusão e as métricas derivadas *accuracy*, *precision*, *recall* e *f1-score* [31], com os melhores resultados estão indicados em verde na Figura 4. Ela apresenta os valores das métricas para cada algoritmo, sendo que a Tabela 4a se refere a uma instituição educacional da base da PNP e a Tabela 4b se refere a uma instituição da base do Censo.

Importante ressaltar que com a mesma aplicação desenvolvida para empregar as técnicas de ML, foi possível gerar os modelos para as duas fontes de dados, pois as técnicas são aplicadas sobre os repositórios que estão padronizados pela ontologia.

Após o treinamento e a avaliação preliminar dos modelos, um deles foi selecionado e utilizado para predição da evasão usando dados atuais de uma instituição da base PNP. O algoritmo *Bagging* foi selecionado porque é um dos que obteve um melhor desempenho considerando a métrica *f1-score*, que é indicada para bases desbalanceadas. Essa métrica é a média harmônica entre a precisão e a revocação, métricas que valorizam os acertos de alunos que de fato evadiram.

O modelo selecionado é utilizado para calcular a probabilidade da evasão escolar para cada aluno da base. Esses dados são inseridos no repositório por meio do conceito **Probabilidade de Evasão** que foi adicionado na ontologia com essa finalidade. Assim, após a aplicação do modelo de ML, o repositório é consultado para gerar um relatório que permite visualizar quem são os alunos que o modelo está apontando com risco de evasão e quais as probabilidades disso acontecer segundo o modelo. A Figura 5 apresenta uma visão deste relatório.

As colunas do relatório consistem em conceitos da Ontologia de Evasão Escolar, e os dados são instâncias desses conceitos. Por exemplo, na primeira linha do relatório tem-se o Aluno de identificação *87799278* de Gênero *M* (masculino) e Idade de *64* anos, possui Renda familiar per

Aluno	Genero	Renda	Idade	Curso	Turno	Predicao	ProbabilidadeEvasao
87799278	M	1<RFP<=1,5	64	ESPECIALIZAÇÃO - RECURSOS NATURAIS	NOTURNO	1	0.74
65667204	M	RFP>3,5	66	SISTEMAS DE INFORMAÇÃO	INTEGRAL	1	0.73
73851170	M	RFP>3,5	57	TÉCNICO EM INFORMÁTICA	NOTURNO	1	0.68
88144998	M	RFP>3,5	23	ENGENHARIA DE MINAS	INTEGRAL	1	0.61
86832725	M	0,5<RFP<=1	34	TÉCNICO EM SEGURANÇA DO TRABALHO	VESPERTINO	1	0.61
87032197	M	1<RFP<=1,5	36	ADMINISTRAÇÃO	NOTURNO	1	0.61
87799510	M	0,5<RFP<=1	27	ESPECIALIZAÇÃO - RECURSOS NATURAIS	NOTURNO	1	0.61
93706312	M	2,5<RFP<=3,5	16	TÉCNICO EM MEIO AMBIENTE	VESPERTINO	0	0.48
58467776	F	1,5<RFP<=2,5	27	ENGENHARIA DE CONTROLE E AUTOMAÇÃO	NOTURNO	0	0.48
91749986	M	0,5<RFP<=1	21	ENGENHARIA DE MINAS	INTEGRAL	0	0.48

Figura 5: Amostragem do relatório de resultados das predições de evasão.

capita *1<RFP<=1,5* (1 a 1,5 salários mínimos), é aluno do Curso *ESPECIALIZAÇÃO - RECURSOS NATURAIS* que é do Turno *Noturno*, Predição igual a *1*, segundo o modelo de ML, a tendência é de que esse aluno evada, com Probabilidade de Evasão de *74%*.

4. Trabalhos Correlatos

O trabalho proposto por Carchedi et al. [9], apresenta a *Ontology For Learning Analytics (Onto4LA)*, criada com o objetivo de facilitar a integração de dados educacionais e garantir interoperabilidade semântica. Embora haja correlação no que diz respeito ao uso de ontologias para permitir análise de dados de diferentes fontes escolares (a Onto4LA lida com conceitos como: *Interaction, Profile, Message, Log etc.*), essa difere de nossa proposta por ter sido projetada para análise de dados da educação a distância, sendo modelada em uma perspectiva de eventos, enquanto a Ontologia de Evasão Escolar está mais focada em aspectos estruturais da educação em geral. Além disso, não foi feito uso de *machine learning*.

Outro trabalho [34] está relacionado no que diz respeito a fazer uso de ontologias a aplicação de técnicas de ML. Propõe uma ontologia para representar o perfil de um aluno e com base nessa ontologia criar um sistema de apoio à decisão por meio de tarefas de predição multi objetivas. A abordagem baseia-se na eficiência da ontologia para prover interoperabilidade semântica e dos benefícios das técnicas de ML para construir um sistema inteligente para objetivos multifuncionais de suporte à decisão. O que se percebe de diferencial em relação a tal trabalho, está no fato de que aqui são utilizados diferentes algoritmos de ML, enquanto eles utilizam apenas árvore de decisão. Como eles integram o modelo de ML à ontologia por meio de *Semantics Web Rule Language (SWRL)*, pode haver desafios para realizar o mapeamento de outros modelos baseados em outros algoritmos, pois como o modelo já teria sido treinado antes deste passo, isso não implica em melhor desempenho das predições realizadas. Portanto, uma abordagem com um módulo ou novos conceitos na ontologia para armazenar os resultados das predições se mostra mais interessante e facilita a aplicação de outros algoritmos.

Foi identificado ainda o trabalho [35], que propõe um sistema de recomendação baseado em ontologia e aprimorado com técnicas de ML para orientar estudantes do ensino médio a escolherem cursos e universidades. Faz uso de métodos baseados em semântica e técnicas de

ML para aprimorar etapas do processo de recomendação. O trabalho utiliza ML e ontologias, e a ontologia fornece suporte ao sistema, desempenhando um papel importante, pois é usada para modelar o conhecimento do domínio. Porém a ontologia não apresenta uma relação significativa com a aplicação das técnicas de ML. No trabalho aqui proposto, a interoperabilidade provida pela ontologia é usada como base para a aplicação das técnicas de ML. Isso permite algumas vantagens como: (i) a utilização de diversas fontes de dados, (ii) a consistência do processo de aplicação das técnicas ML, mesmo se as fontes de dados forem alteradas; e (iii) o suporte do conhecimento do domínio provido pela ontologia e mapeamentos realizados para tomada de decisão referente aos atributos de entrada e algoritmos de transformação.

5. Considerações finais

Sabe-se que os impactos da evasão escolar são grandes, para a sociedade, para as instituições de ensino e, principalmente, para os alunos. Segundo Bezerra [36], as consequências da evasão escolar na vida dos educandos têm sido visíveis na sociedade contemporânea, trazendo altos índices de violência, criminalidade e envolvimento com drogas, comprometendo seus sonhos e seus projetos futuros. Com isso, é muito importante a criação de recursos tecnológicos capazes de contribuir para compreender as causas e reduzir a evasão.

Acredita-se que a análise de dados sobre a evasão escolar é algo fundamental para contribuir para um melhor entendimento dos fatores que levam à evasão, bem como para definição de políticas e estratégias que visem minimizar esse fenômeno. Aplicando soluções como esta, é possível que gestores escolares tenham acesso a relatórios com informações de alunos que estão mais propensos a evadir, o que pode contribuir para a definição de estratégias pedagógicas mais direcionadas a esses alunos. Com os repositórios padronizados pela ontologia é possível a realização de consultas que permitam a identificação de padrões de evasão, por exemplo, identificar que os alunos de determinada faixa de renda são os que mais evadem.

Além de uma nova versão da Ontologia de Evasão Escolar, este trabalho apresentou uma abordagem aplicada a um estudo de caso que combina o uso de ontologias e ML, que envolve o mapeamento e transformação eficiente dos dados para uma base consistente e padronizada, de onde podem ser realizadas consultas homogêneas, bem como permite a aplicação de técnicas de ML para diferentes instituições de ensino. Isso colabora para que a solução proposta possa ser aperfeiçoada e utilizada para mitigar o risco de evasão escolar. A possibilidade de aplicação de ML para diferentes fontes se mostrou promissora, uma vez que as soluções que fazem uso dessas técnicas são geralmente aplicadas para uma base de dados específica.

Como trabalhos futuros pretende-se ampliar a abrangência da Ontologia de Evasão Escolar, para que seja possível representar outros conceitos e deixar os repositórios mais completos. Por exemplo, adicionar outros dados socioeconômicos como a escolaridade dos pais, e mais características dos alunos (e.g., bolsistas, PcD, histórico familiar). Além disso, otimizar o desempenho dos algoritmos a fim de melhorar a predição dos modelos; e disponibilizar uma *Application Programming Interface (API)* para que seja possível verificar a predição de evasão de um aluno em tempo real.

Referências

[1] N. P. d. L. Gaioso, O fenômeno da evasão escolar na educação superior no brasil, Brasília, DF: Universidade Católica de Brasília (2005) 20.

[2] E. d. J. M. d. Araújo, et al., Evasão no proeja: estudo das causas no instituto federal de educação, ciência e tecnologia do maranhão/ifma (2012).

[3] M. Brasil, Comissão especial de estudos sobre a evasão nas universidades públicas brasileiras, http://www.dominiopublico.gov.br/download/texto/me001613.pdf 15 (1997) 2007.

[4] C. A. D. Santos Baggi, D. A. Lopes, Evasão e avaliação institucional no ensino superior: uma discussão bibliográfica, Avaliação: Revista da Avaliação da Educação Superior (Campinas) 16 (2011) 355–374.

[5] G. Tontini, S. A. Walter, Pode-se identificar a propensão e reduzir a evasão de alunos?: ações estratégicas e resultados táticos para instituições de ensino superior, Avaliação: Revista da Avaliação da Educação Superior (Campinas) 19 (2014) 89–110.

[6] A. C. Lorena, A. C. de Carvalho, Uma introdução às support vector machines, Revista de Informática Teórica e Aplicada 14 (2007) 43–67.

[7] C. A. R. Beltran, J. C. Xavier-Júnior, C. A. Barreto, C. O. Neto, Plataforma de aprendizado de maquina para detecção e monitoramento de alunos com risco de evasão, in: Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE), volume 30, 2019, p. 1591.

[8] G. C. Britto, F. B. Ruy, C. L. Azevedo, Um ambiente para integração de dados abertos relativos à despesa pública (2020).

[9] L. C. Carchedi, J. Souza, E. Barrére, F. Mendonça, Onto4la: uma ontologia para integração de dados educacionais, in: Anais dos Workshops do Congresso Brasileiro de Informática na Educação, volume 7, 2018, p. 439.

[10] M. Kulmanov, F. Z. Smaili, X. Gao, R. Hoehndorf, Semantic similarity and machine learning with ontologies, Briefings in bioinformatics 22 (2021) bbaa199.

[11] R. Studer, V. R. Benjamins, D. Fensel, Knowledge engineering: principles and methods, Data & knowledge engineering 25 (1998) 161–197.

[12] N. Guarino, D. Oberle, S. Staab, What is an ontology?, in: Handbook on ontologies, Springer, 2009, pp. 1–17.

[13] H. Zhang, Y. Guo, Q. Li, T. J. George, E. Shenkman, F. Modave, J. Bian, An ontology-guided semantic data integration framework to support integrative data analysis of cancer survival, BMC medical informatics and decision making 18 (2018) 129–147.

[14] P. M. C. Campos, Designing a Network of Reference Ontologies for the Integration of Water Quality Data, Ph.D. thesis, M. Sc. Thesis, Federal University of Espírito Santo, 2019. Available: https …, 2019.

[15] G. Guizzardi, A. Botti Benevides, C. M. Fonseca, D. Porello, J. P. A. Almeida, T. Prince Sales, Ufo: Unified foundational ontology, Applied ontology (????) 1–44.

[16] G. Guizzardi, C. M. Fonseca, A. B. Benevides, J. P. A. Almeida, D. Porello, T. P. Sales, Endurant types in ontology-driven conceptual modeling: Towards ontouml 2.0, in: International conference on conceptual modeling, Springer, 2018, pp. 136–150.

[17] R. Pergl, T. P. Sales, Z. Rybola, Towards ontouml for software engineering: from domain ontology to implementation model, in: International Conference on Model and Data

Engineering, Springer, 2013, pp. 249–263.
[18] T. B. Ludermir, Inteligência artificial e aprendizado de máquina: estado atual e tendências, Estudos Avançados 35 (2021) 85–94.
[19] R. Cerri, A. CARVALHO, Aprendizado de máquina: breve introdução e aplicações, Cadernos de Ciência & Tecnologia 34 (2017) 297–313.
[20] H. Fernando Filho, D. Siqueira, B. Leal, Predição de evasão utilizando técnicas de classificação: Um estudo de caso do instituto federal do ceará, in: Anais da VIII Escola Regional de Computação do Ceará, Maranhão e Piauí, SBC, 2020, pp. 141–148.
[21] E. M. da Silva, F. B. Ruy, F. W. Mutz, Abordagem para análise de múltiplas fontes de dados de evasão escolar, Anais do Computer on the Beach 13 (2022) 149–156.
[22] R. de Almeida Falbo, Sabio: Systematic approach for building ontologies., in: ONTO. COM/ODISE@ FOIS, 2014.
[23] G. Guizzardi, Ontological foundations for structural conceptual models (2005).
[24] I. N. de Estudos e Pesquisas Educacionais Anísio Teixeira, Glossário do censo da educação superior 2013, Sinopse Estatística (2014).
[25] I. N. de Estudos e Pesquisas Educacionais Anísio Teixeira, Apresentação dos resultados do censo da educação superior 2017, Sinopse Estatística (2018).
[26] D. A. D. P. D. COORDENAÇÃO, D. PESSOAL-CAPES, Tabela de áreas de conhecimento/avaliação, Brasília, DF (2014).
[27] A. M. Catani, J. d. OLIVEIRA, A educação superior, Organização do ensino no Brasil: níveis e modalidades na Constituição Federal e na LDB 2 (2002) 73–84.
[28] C. de Almeida Lima, M. A. Vieira, F. M. da Costa, J. F. D. Rocha, O. V. Dias, Correlação entre perfil sociodemográfico e acadêmico e formas de ingresso na graduação em enfermagem, Revista de Enfermagem UFPE on line 9 (2015) 7986–7994.
[29] B. Scikit-Learn, Biblioteca scikit-learn, 2020. URL: https://scikit-learn.org/.
[30] T. W. Rauber, A. L. da Silva Loca, F. de Assis Boldt, A. L. Rodrigues, F. M. Varejão, An experimental methodology to evaluate machine learning methods for fault diagnosis based on vibration signals, Expert Systems with Applications 167 (2021) 114022.
[31] K. P. Murphy, Machine learning: a probabilistic perspective, MIT press, 2012.
[32] H. E. Kiziloz, Classifier ensemble methods in feature selection, Neurocomputing 419 (2021) 97–107.
[33] Z.-H. Zhou, Ensemble methods: foundations and algorithms, CRC press, 2012.
[34] T. Hamim, F. Benabbou, N. Sael, An ontology-based decision support system for multi-objective prediction tasks, International Journal of Advanced Computer Science and Applications 12 (2021).
[35] C. Obeid, I. Lahoud, H. El Khoury, P.-A. Champin, Ontology-based recommender system in higher education, in: Companion Proceedings of the The Web Conference 2018, 2018, pp. 1031–1034.
[36] V. L. M. Bezerra, Narrativas das histórias de vida da evasão escolar dos estudantes da educação de jovens e adultos–eja (2019).