Análise de Retenção Escolar em Cursos de Graduação na POLI/UPE Usando Mineração de Dados

Retention analysis in Undergraduate Courses at POLI/UPE using Data Mining

Katiane O. Alpes Silva[2]
orcid.org/0000-0003-4097-8217

Pedro J. Rodrigues Silva[2]
orcid.org/0000-0001-7500-360X

Alexandre M. A. Maciel[1]
orcid.org/0000-0003-4348-9291

[1]Escola Escola Politécnica de Pernambuco, Universidade de Pernambuco, Recife, Brasil.

E-mail:koas@ecomp.poli.br; pjrs@ecomp.poli.br; alexandre.maciel@upe.br

[2]Serviço Federal de Processamento de Dados (SERPRO).

DOI: 10.25286/repa.v7i2.2223

Como citar este artigo pela NBR 6023/2018: Katiane O. Alpes Silva; Pedro J. Rodrigues Silva; Alexandre M. A. Maciel. Análise de Retenção Escolar em Cursos de Graduação na POLI/UPE Usando Mineração de Dados. Revista de Engenharia e Pesquisa Aplicada, v.7, n. 2, p. 108-117, 2022

RESUMO

A retenção ou permanência prolongada é uma das situações que pode ocorrer ao longo da vida acadêmica do aluno, podendo ser vista como ineficiência do sistema escolar. A retenção, além de causar desperdício de dinheiro nas instituições de ensino superior públicas, também deixa de entregar para o mercado de trabalho mão de obra qualificada. Nas instituições de ensino superior brasileiras, a taxa média de permanência era 20% ao final do quinto ano da graduação de ingressantes no ano de 2010. Diante desse contexto, o trabalho proposto descreve o uso de mineração de dados para analisar a retenção em Cursos de Graduação na Escola Politécnica de Pernambuco (POLI) para melhorar a compreensão do cenário e possibilitar a definição e acompanhamento de ações na tentativa de diminuir a taxa de retenção e aumentar a quantidade de concluintes. A partir da realização de ações no combate à retenção, espera-se o uso mais eficiente dos recursos das instituições públicas de ensino de educação superior no Brasil e um aumento no quantitativo de profissionais aptos a ingressarem no mercado de trabalho.
PALAVRAS-CHAVE: Mineração de Dados Educacionais; Árvore de Decisão; Regressão Logística; Perceptron Multicamadas;

ABSTRACT

Retention is one of the situations that can occur throughout the student's academic life, which can be seen as an inefficiency of the brazilian educational system. Retention, in addition to causing waste of money in public higher education institutions, also fails to deliver qualified professionals to the labor market. In Brazilian higher education institutions, the average retention rate of students who entered in 2010 was 20% at the end of the fifth year of the undergraduate courses. In this context, the purpose of this article is to describe the adoption of data mining to analyze retention in undergraduate courses at Escola Politécnica de Pernambuco (POLI) to improve the understanding of the scenario and enable the definition and
monitoring of actions in an attempt to reduce the retention rate and increase the number of graduate students. Carrying out actions to avoid retention, it is expected a more efficient use of the resources of brazilian public institutions of higher education and an increase in the number of qualified professionals to enter the labor market.

KEY-WORDS: Educational Data Mining; Decision tree; Logistic Regression; Multilayer Perceptron;

1 INTRODUÇÃO

A Evasão, abandono e retenção são alguns dos aspectos relacionados à ineficiência do processo de aprendizagem e precisam ser analisados continuamente para melhoria das instituições de ensino. A evasão pode ser vista como a saída definitiva do curso sem intenção de voltar, enquanto o abandono é caracterizado pela descontinuidade da frequência, podendo culminar na evasão ou não **[1]**. Já a retenção pode ser configurada quando o aluno não conclui o curso no tempo máximo definido para integralização curricular **[2]**. Todos esses fenômenos precisam fazer parte de acompanhamento sistemático e periódico das instituições de ensino para que possam alcançar seus objetivos de garantir a aprendizagem e eficiência operacional.

Para financiar a Educação no Brasil, segundo dados do Instituto Nacional de Estudos e Pesquisas Educacional Anísio Teixeira (INEP) **[3]**, o Brasil investiu 5,1% do Produto Interno Bruto em 2017 em instituições educacionais públicas, estando acima da média de 4,1% gastos pelos países membros e parceiros da Organização para Cooperação e Desenvolvimento Econômico (OCDE). O investimento público anual no Brasil em 2017 foi US$4.661/por aluno em instituições educacionais públicas, contemplando desde os anos iniciais do ensino fundamental até o ensino superior. Ao se analisar a distribuição deste montante, quando se compara os gastos públicos por aluno nas etapas de ensino, verifica-se que no Brasil o valor empregado em 2016 por aluno na educação superior foi cerca de 4 vezes maior do que os direcionados aos ensinos fundamental e médio **[4]**.

Os recursos financeiros destinados à educação superior deveriam permitir a população o acesso às instituições de ensino superior e o resumo técnico do Censo da Educação Superior 2019 **[5]** informa que no ano de 2019 no Brasil 11,6% de instituições de ensino superior eram públicas e que foram responsáveis por disponibilizar 837.809 vagas. No entanto, este número de vagas somado à quantidade disponibilizada pelas instituições privadas de ensino superior se apresenta muito aquém do quantitativo da população entre 18 e 24 anos no Brasil que em 2019, na região Nordeste, era cerca de 6 vezes maior do que o número de vagas disponíveis para esta Região, configurando-se uma suboferta de vagas no ensino superior brasileiro. Esta quantidade insuficiente de vagas, talvez explique a grande concorrência registrada no Sistema de Seleção Unificada (SISU) **[6]**, a segunda edição de 2019 para ingresso na educação superior apresentou uma concorrência de mais de 100 alunos por vaga para o curso de Medicina, cerca de 44 para Direito e 30,14 para Administração.

Apesar do cenário de disponibilidade de vagas adverso e a concorrência acirrada para ingresso em um curso superior, os dados revelam que a grande maioria dos alunos que alcançam uma vaga na graduação no País desiste ou demora mais do que o tempo mínimo de integralização do curso para se formar. No resumo técnico do Censo da Educação Superior de 2019 é possível verificar que do total de ingressantes em 2010, ao final de 10 anos, apenas 40% concluíram o curso e 59% desistiram **[5]**. Em relação à permanência, ao acompanhar os ingressantes de 2010 a 2015, no final do quinto ano da graduação entre 18 a 20% dos ingressantes permaneceram cursando ou com matrícula trancada.

A partir de dados abertos disponibilizados pelo Portal do INEP foi possível traçar a trajetória de alunos ingressantes nos anos de 2010, 2011 e 2012 em cursos de graduação da Escola Politécnica de Pernambuco (POLI) e foi possível verificar que ao final do quinto ano de curso cerca de 66% dos alunos ingressantes em 2010 ainda permaneciam cursando ou com matrícula trancada. Já para os alunos ingressantes em 2011 e 2012, as taxas de retenção eram de aproximadamente 71% e 73% ao final do quinto ano, respectivamente.

Diante desse contexto, tanto a evasão quanto a permanência prolongada ou retenção em instituições públicas de ensino no Brasil precisam de acompanhamento para fins de definição de ações que possam reduzir estes tipos de ineficiência do sistema de ensino, evitando desperdício de dinheiro público e melhorando a formação escolar da população brasileira. Na tentativa de colaborar com o avanço no entendimento de aspectos relacionados à permanência prolongada ou retenção, este estudo tem o objetivo de analisar a retenção de alunos nos cursos de graduação da POLI/UPE utilizando Mineração de Dados para que a instituição possa compreender melhor o cenário da retenção e atuar proativamente em seu quadro de discentes, realizando ações preventivas que consigam melhorar os resultados acadêmicos e, assim, a eficiência do sistema educacional.

O artigo está estruturado de acordo com as seguintes seções: inicialmente a seção 2 elenca artigos relacionados ao tema. Em seguida, a seção 3 descreve a metodologia CRISP-DM (*CRoss-Industry Standard Process for Data Mining*) adotada neste trabalho. As seções seguintes seguem a

DOI: 10.25286/repa.v7i2.2223

estrutura prevista no CRISP-DM, sendo assim, na seção 4 e 5 são apresentados o entendimento do negócio e dos dados. Na seção 6 está descrita a preparação dos dados. Na sequência, a seção 7 detalha a modelagem dos dados e avaliação de desempenho. Por fim, são apresentadas as conclusões e sugestões de trabalho futuro.

2 TRABALHOS RELACIONADOS

Uma análise importante do problema do risco de retenção em cursos de graduação foi realizada utilizando a metodologia CRISP/DM com o intuito de extrair conhecimento usando mineração de dados **[7]**. Inicialmente foi feita uma distinção entre evasão e retenção, sabendo-se que ambas são responsáveis pela ineficiência do sistema de ensino. Sendo assim, a evasão foi configurada como o desligamento do estudante e a retenção foi definida como ocorrendo quando o estudante não concluiu seu curso no prazo normal especificado pela instituição de ensino. A solução proposta buscou identificar variáveis relevantes que pudessem refletir o risco de retenção no final do primeiro ano de graduação, possibilitando intervenção proativa na tentativa de diminuir a taxa de retenção dos alunos na instituição de ensino superior. O estudo foi realizado a partir da base de dados de 6 cursos dos anos de 1998 a 2008 da Universidade Federal de Pernambuco. Esses dados foram selecionados, pré-processados e transformados de acordo com as características da pesquisa. As técnicas de modelagem utilizadas foram regressão logística e indução de regras e tiveram sua qualidade avaliada por meio de indicadores, demonstrando que a solução de intervenção com consultoria para alunos com propensão a retenção precisa de apenas um ponto de ajuste para controlar o melhor ponto de operação do sistema.

Em **[8]** foi realizada avaliação do ensino médio dos institutos federais usando mineração de dados. Para tanto, usaram o conhecimento de especialistas do domínio, a metodologia CRISP/DM, as bases do ENEM e do censo escolar disponibilizadas pelo INEP. Após seleção, preparação e transformação dos dados empregaram árvore de decisão, regressão logística e indução de regras para extração do conhecimento. A qualidade dos modelos foi analisada, comprovando seu bom desempenho para prever e estimar a propensão ao sucesso de alunos dos institutos federais. Por fim, descobriram que fatores socioeconômicos influenciam fortemente o bom desempenho, assim como a formação do professor, a opção pela língua inglesa, a permanência do aluno na escola e suas perspectivas em relação ao ensino superior.

Foi realizada uma análise da evasão escolar no último ano do ensino fundamental nas escolas públicas estaduais e municipais de Pernambuco com base nos dados dos censos escolares de 2011 e 2012 **[9]**. O trabalho visou identificar fatores relevantes que levam estudantes a evadir e através deste conhecimento, permitir a criação de mecanismos para minimizar este grande problema da educação pública brasileira. Para atingir este intuito, foi utilizada a metodologia CRISP/DM para analisar as bases do censo escolar. Para identificar os fatores que poderiam acarretar a evasão foram aplicadas técnicas baseadas em Árvores de Decisão, Indução de Regras e Regressão Logística. Essas técnicas permitem alta legibilidade e validação por especialistas, geração de regras de associação do tipo "se-então" que são facilmente entendidas e permitem também quantificar quais variáveis explicativas são mais relevantes para definir a variável alvo, respectivamente. O estudo identificou uma forte relação entre evasão e as seguintes características: a idade dos alunos, o turno em que estudam, a estrutura e o tamanho da escola, o tempo de permanência em sala de aula, a existência de aula de religião ou educação física ou a dedicação dos professores.

A mineração de dados educacionais (MDE) e a *Learning Analytics* (LE) vem utilizando a inteligência artificial (IA) no campo das pesquisas educacionais. A LE busca compreender os sistemas em sua totalidade. Por outro lado, a MDE adota uma visão reducionista ao analisar componentes individuais, buscando novos padrões nos dados. Mesmo com estas diferenças, estas áreas continuam a escapar de uma definição mais precisa. A fim de identificar as diferenças e similaridades entre MDE e LE, foi conduzida uma modelagem para análise de tópicos de artigos relacionados à mineração de dados educacionais e *Learning Analytics* **[10]**. O estudo apontou que durante os anos de 2015 a 2019, os tópicos dos artigos nestas áreas foram mais focados na performance dos alunos nas plataformas de aprendizado e na modelagem do comportamento dos alunos. A tendência apontada pelo trabalho no campo da pesquisa educacional é uma convergência na utilização de aplicações avançadas de IA para extrair conhecimento de grandes bases de dados com o intuito de otimizar o ensino e a aprendizagem.

Um estudo de caso usando técnicas de mineração de dados para realizar previsões sobre evasão escolar nos cursos de graduação da Universidade de Brasília (UnB) foi realizado por **[11]**. Para condução deste estudo, usaram as bases de dados do Sistema Integrado da Graduação da UnB com dados referentes aos anos de 2006 até 2018. Para análise destes dados foi usada a metodologia CRISP-DM que fornece uma abordagem rápida, confiável e estruturada para processos
de mineração de dados. Após a preparação dos dados, um modelo de predição foi elaborado após a avaliação de alguns algoritmos de classificação: *Gradient Boost Machine*, *Generalized Linear Model*, *Support Vector Machines* (SVM) e *Random Forest*. O estudo concluiu que o modelo de predição utilizando *Gradient Boost Machine* apresentou a melhor acurácia e que de acordo com este modelo, fatores como quantidade de créditos do curso, tempo de duração do curso, a forma de entrada através de cota ou não e a soma de notas negativas são fatores relevantes que levam o estudante a evasão.

Foi realizada pesquisa na literatura de mineração de dados educacionais para compreender quais abordagens, modelos, conjuntos de dados, ferramentas, técnicas e medidas de desempenho estão sendo utilizadas para previsão de retenção de alunos na educação superior **[12]**. Para tanto, utilizaram o conceito de retenção para aqueles alunos que se mantêm matriculados até a conclusão de seu curso, onde a formatura é o objetivo final de sucesso. Por outro lado, utilizaram o termo evasão para os alunos que desistiram ou retiraram-se voluntariamente do curso. A pesquisa de literatura foi realizada na base de dados Scopus, utilizando-se alguns filtros de interesse, resultando em 19 artigos que foram revisados. Com base na análise desses artigos, observaram que não existe um padrão de fatores que afetam a retenção e cada instituição de ensino precisa buscar um modelo preditivo a partir de seu conjunto de dados. Também perceberam que os algoritmos de árvore de decisão e SVM foram os modelos mais usados, seguidos por análise de regressão, com destaque para a interpretabilidade humana fornecida pela árvore de decisão. Por fim, sugerem que vários modelos sejam testados e que os modelos de dados sejam expandidos a partir de dimensões originais combinadas.

Foi conduzida um estudo que investigou a aplicação de Mineração de Dados na tentativa de classificar o risco de evasão de estudantes em instituições de ensino superior, utilizando atributos socioeconômicos dos alunos informados no momento de ingresso na instituição **[13]**. Foi utilizada a metodologia CRISP-DM e os dados foram obtidos do Sistema de Informação e Gestão Acadêmica (SIGA) da Universidade Federal de Pernambuco (UFPE) dos anos de 2009 a 2011. Na modelagem de dados foram utilizados 13 atributos e os seguintes algoritmos para classificar o risco de evasão do aluno: Naive Bayes, Árvore de Classificação, *Random Forest*, Regressão Logística, SVM e KNN. Os algoritmos foram empregados em 5 cenários distintos, onde os 2 primeiros cenários simularam o uso de base com poucos registros disponíveis, variando apenas a forma de amostragem. Os cenários 3 e 4 usaram base de dados maiores, variando o tipo de amostragem. Por fim, o quinto cenário usou uma base com muitos registros, onde os dados de treino foram compostos pelos dados mais antigos e os dados de predição montado com os dados dos alunos mais recentes. Os resultados demonstraram que os algoritmos, de forma geral, apresentaram uma acurácia acima de 70% mostrando a viabilidade da abordagem, com destaque para o Naive Bayes, Regressão Logística e Árvore de Classificação. Um ponto importante destacado nas conclusões é que o desempenho dos modelos utilizando apenas dados de ingresso são menores do que modelos baseados em dados de desempenho acadêmico dos alunos.

3 METODOLOGIA

O CRISP-DM (*CRoss-Industry Standard Process for Data Mining*) é um processo padrão para mineração de dados que reúne boas práticas utilizadas para resolver problemas desta área de acordo com **[14]**.

Esta metodologia será utilizada no processo de mineração de dados, pois é considerada apropriada para este tipo de projeto apesar da evolução na Ciência de Dados e suas metodologias **[15]**. As etapas previstas para serem utilizadas no CRISP-DM estão elencadas abaixo.

- A primeira etapa consiste no entendimento do problema, ou seja, compreender os objetivos do projeto a partir da perspectiva de negócio.
- A segunda etapa foca no entendimento dos dados visando familiarizar-se com os dados, identificar problemas de qualidade de dados, descobrir os primeiros insights sobre os dados e detectar subconjuntos interessantes para formar hipóteses sobre informações ocultas.
- A etapa de Preparação dos dados cobre todas as atividades necessárias para construir o conjunto de dados final a partir dos dados

DOI: 10.25286/repa.v7i2.2223

brutos iniciais Serão realizadas tarefas para leitura da base de dados e de pré-processamento, tais como: seleção de atributos, transformação e limpeza dos dados para serem utilizados nas ferramentas de modelagem. A base de dados utilizada será o Censo da Educação Superior disponível no portal do INEP.

- A modelagem acontece na quarta etapa e nela são selecionadas e aplicadas várias técnicas de modelagem. Durante a análise dos modelos os parâmetros são calibrados para otimizar os modelos. Muitas vezes é necessário voltar à fase de preparação de dados para ajustes nos dados que são requeridos pelos modelos utilizados. Nesta etapa, conforme **[16]**, serão utilizadas árvores de decisão, regressão logística e indução de regras para descoberta de conhecimento a partir do conjunto de dados.
- A quinta etapa é a avaliação, antes de prosseguir para a implantação final do modelo, é importante avaliá-lo completamente e revisar as etapas executadas para criá-lo, para ter certeza de que o modelo atinja adequadamente os objetivos de negócios. Um ponto importante desta fase é determinar se há alguma questão comercial importante que não foi suficientemente considerada. Ao final desta fase deve ser tomada a decisão sobre o uso dos resultados da mineração de dados.
- A última etapa é a implantação e nela os novos conhecimentos descobertos são apresentados de uma forma que a organização possa utilizá-los. Dependendo dos requisitos, a fase de implantação pode ser tão simples como gerar um relatório ou tão complexa como implementar uma mineração de dados aplicável para toda a empresa.

4 ENTENDIMENTO DO NEGÓCIO

Os fenômenos de evasão e retenção nas Instituições de Ensino Superior no Brasil foram objeto de um estudo aprofundado conduzido pela Comissão Especial de Estudos sobre Evasão instituída em 1995 e sugerem que taxas de retenção maiores que 10% precisam de análise aprofundada pela instituição de ensino **[2]**. Os conceitos de evasão, retenção e conclusão utilizados neste trabalho seguem as definições encontradas no Resumo Técnico do Censo da Educação Superior 2019 para desistência, permanência e conclusão **[5]**, respectivamente:

- Permanência: corresponde aos alunos com situação de vínculo igual a "cursando" ou "matrícula trancada", ou seja, trata de alunos que possuem vínculos ativos com o curso e, portanto, deverão ser informados com qualquer situação de vínculo no ano subsequente (no mesmo curso e com a mesma data de ingresso).
- Desistência: corresponde aos alunos com situação de vínculo igual a "desvinculado do curso" ou "transferido para outro curso da mesma Instituição de Ensino Superior (IES)", ou seja, tais alunos encerraram seu vínculo com o curso e, portanto, não deverão ser informados no ano subsequente (no mesmo curso e com a mesma data de ingresso).
- Conclusão: corresponde aos alunos com situação de vínculo igual a "formado", ou seja, também encerraram seu vínculo com o curso e, portanto, não deverão ser informados no ano subsequente (no mesmo curso e com a mesma data de ingresso).

O estudo foi realizado com base nos dados da Escola Politécnica de Pernambuco (POLI) da Universidade de Pernambuco (UPE) disponibilizados no Portal INEP. A POLI oferece 7 cursos de Engenharia, além do curso Física de Materiais. Os cursos de engenharia oferecidos pela POLI são: Civil, Computação, Controle e Automação, Mecânica e Elétrica com as modalidades Eletrônica, Eletrotécnica e Telecomunicações. Nas bases selecionadas para realização deste trabalho, a denominação do curso Engenharia de Controle e Automação era Mecatrônica e os cursos Física de Materiais e Engenharia Mecânica não estavam presentes por terem sido criados em ano fora do período contemplado. Havia o curso de Mecânica Industrial que estava na situação "em extinção" e, por isso, também não foi incluído no escopo. Os cursos de Engenharia da POLI têm tempo previsto de integralização de 5 anos e duas entradas de aluno por ano, uma no primeiro semestre, comumente chamada de "Primeira entrada" e outra no segundo semestre, conhecida como "Segunda entrada".

5 ENTENDIMENTO DOS DADOS

As bases de dados utilizadas para realização deste estudo foram as do Censo da Educação Superior disponíveis no Portal de dados abertos do

INEP que tem objetivo de produzir e disseminar informações educacionais do Brasil **[17]**. O Censo da Educação Superior é realizado anualmente no início do ano seguinte ao de referência, registrando informações sobre todas as Instituições de Ensino Superior (IES), cursos, docentes, alunos e locais de oferta de cursos. Um dos objetivos das informações coletadas no censo é contribuir com o trabalho dos gestores de IES públicas ou privadas.

De acordo com os objetivos definidos neste estudo, as bases do censo da educação superior dos anos de 2010, 2011 e 2012 foram acessadas e foram selecionados alunos ingressantes nos cursos de graduação da POLI. Em seguida, as bases dos 6 anos posteriores aos anos de ingresso de 2010 e 2011 também foram acessadas para se verificar a situação do aluno ao longo de sua trajetória acadêmica, analisando dois cenários de acompanhamento da trajetória: o primeiro cenário de acompanhamento verificava a situação do aluno até o quinto ano do curso; já no segundo cenário, foi acrescentado mais um ano, acompanhando a situação do aluno até o sexto ano de curso. Por fim, os ingressantes de 2012 foram acompanhados por apenas mais 5 anos, pois a partir do ano de 2018 houve uma alteração nas bases do INEP onde não foi mais possível a identificação dos alunos entre os anos. É importante destacar que foram filtrados da amostra de dados os alunos transferidos, desvinculados ou falecidos por não fazerem parte da problemática educacional abordada neste trabalho. Sendo assim, após a seleção, acompanhamento da trajetória e filtragem, a amostra dos alunos ingressantes para cada cenário apresentou o quantitativo de alunos detalhado na tabela 1.

Tabela 1- Tamanho da Amostra por Ano de Ingressantes e Cenário de Acompanhamento
Fonte: Os autores.

ANO DE INGRESSO	ACOMPANHAMENTO	TOTAL DE ALUNOS DA AMOSTRA
2010	5 anos	229
2010	6 anos	216
2011	5 anos	336
2011	6 anos	268
2012	5 anos	349

As bases de dados do Censo da Educação Superior dos anos de 2010, 2011 e 2012 do INEP são formadas pelas seguintes entidades: Aluno, IES, Curso, Local de Oferta e Docente. A quantidade de atributos e tipo de atributo variam de acordo com o ano, mas existe um grupo comum de atributos ou com mesma semântica. Como o interesse deste estudo é analisar a retenção sob a ótica de características do aluno, o grão de análise será a entidade Aluno e atributos relevantes em outras entidades foram trazidos para o grão Aluno.

6 PREPARAÇÃO DOS DADOS

A preparação dos dados é uma etapa fundamental para que se alcancem resultados de qualidade na mineração de dados **[16]**. Sendo assim, as seguintes atividades de pré-processamento foram realizadas:

- Transformação de dados:
  - Construção da variável-alvo "Retido": Indica se o aluno está cursando ou com matrícula trancada após o prazo de acompanhamento definido. Destaca-se que alunos com data de ingresso na segunda entrada tiveram o tempo de acompanhamento acrescido em um semestre, pois o censo é anual e não possui em suas bases a data de conclusão do curso. Sendo assim, pode haver um viés na quantidade de formados para ingressantes da segunda entrada.
  - Alteração de valores do atributo dt_ingresso_curso: Primeira_Entrada e Segunda_Entrada.
  - Categorização dos Atributos: 0 = NAO, 1 = SIM e 2 = Não dispõe da informação.
- Limpeza de dados: atribuição do valor NINF (Não Informado) para campos *Null*.
- Redução de dados: exclusão de atributos: duplicados, irrelevantes, correlacionados, com valor único.

Após a realização das atividades de pré-processamento, a entidade Aluno, para cada base, passou a conter os atributos independentes detalhados na tabela 2 e mais a variável-alvo "Retido".

Atributo	2010	2011	2012
Nome do curso	X	X	X
Código da cor/raça do aluno	X	X	X
Informa o sexo do aluno	X	X	X

DOI: 10.25286/repa.v7i2.2223

Idade que o aluno completa no ano de referência do Censo	X	X	X
Data de ingresso do aluno no curso	X	X	X
Informa se o aluno participa de algum tipo de atividade extracurricular	X	X	X
Informa se o aluno ingressou no curso por meio de reserva de vagas	X	X	
Código do turno do curso ao qual o aluno está vinculado	X	X	X
Informa se o aluno ingressou no curso por vestibular		X	
Informa se o aluno recebe algum tipo de apoio social		X	
Informa se o aluno concluiu o Ensino Médio em escola pública			X

Tabela 2 – Atributos Independentes por Ano de Ingresso
Fonte: Os autores.

7 MODELAGEM DE DADOS E AVALIAÇÃO DE DESEMPENHO

A modelagem para extração de conhecimento a partir de dados foi realizada utilizando árvore de decisão. Em seguida, foram utilizados os algoritmos de classificação Regressão Logística, *Random Forest* e *Multi-Layer Perceptron* (MLP) para avaliar a possibilidade de realizar predições de alunos com propensão à retenção nas bases de dados.

- Extração de conhecimento - Árvore de Decisão
  A árvore de decisão foi usada para permitir explicitar o conhecimento de acordo com o domínio do problema e facilitar a compreensão dos especialistas no domínio **[13]**. Inicialmente, a partir dos dados multidimensionais rotulados com a variável-alvo "Retido", a árvore de decisão foi construída para cada um dos cenários de tempo de acompanhamento por ano e usando todo o conjunto de dados. Em seguida, as bases foram unificadas por tempo de acompanhamento e, para tanto, foi preciso realizar a etapa de preprocessamento novamente. Por fim, as regras mais relevantes foram extraídas baseadas em algumas medidas de interesse (Suporte, Confiança e *Lift*) que são comumente utilizadas para avaliar a qualidade de uma regra de associação **[18]**. Destacamos as regras e as medidas de interesse nas tabelas 3 e 4. As árvores de decisão e as medidas de interesse foram geradas usando a ferramenta *IBM SPSS Statistics*.

Ano/Acompanhamento	Regra	Suporte	Confiança	Lift
2010/5 anos	Alunos Tele, Eletrônica, Mecatrônica E Idade <= 19	15,7%	100%	1,41
2010/6 anos	Alunos Tele, Eletrônica, Mecatrônica E Idade <= 19	15,7%	100%	1,68
2011/5 anos	Alunos Eletrotécnica, Computação E Idade >= 20	10,4%	94,3%	1,11
2011/6 anos	Alunos Eletrônica, Eletrotécnica E Idade >= 20	13,4%	86,1%	1,3
2012/5 anos	Alunos Tele, Eletrônica, Mecatrônica, Computação E Segunda Entrada	23,5%	86,6%	1,02
Unificada/5 anos	Alunos Tele, Eletrônica, Mecatrônica, Computação, Eletrotécnica	49,5%	91,2%	1,12
Unificada/6 anos	Alunos Tele, Mecatrônica, Computação	21,3%	90,3%	1,42

Tabela 3 – Regras Destacadas -> Retido = SIM
Fonte: Os autores.

Ano/Acompanhamento	Regra	Suporte	Confiança	Lift
2010/5 anos	Alunos Civil, Eletrotécnica, Computação	73,4%	36,9%	1,26
2010/6 anos	Alunos Civil, Eletrotécnica, Computação	74,1%	5%	1,24
2011/5 anos	Alunos Civil E Idade >= 20	12,5%	40,5%	2,66
2011/6 anos	Alunos Civil E Segunda Entrada	24,3%	60%	1,79
2012/5 anos	Alunos Civil, Eletrônica E Segunda Entrada	10%	48,6%	3,26
Unificada/5 anos	Alunos Civil E Segunda Entrada	27%	34,4%	1,85
Unificada/6 anos	Alunos Civil	56%	50,9%	1,39

Tabela 4 – Regras Destacadas -> Retido = NÃO
Fonte: Os autores.

- Classificação

Para a utilização dos algoritmos de classificação, foi necessário retornar à etapa de preparação de dados e utilizar os dados numéricos. Também foi necessário transformar os atributos em variáveis *dummy* e aplicar a normalização entre 0 e 1. Foi utilizada a linguagem *Python* e algoritmos do pacote *scikit-learn*. Também foi utilizado o método *GridSearch* para identificar os melhores parâmetros e a validação cruzada na base de dados. Os algoritmos foram executados 30 vezes para se obter uma média dos indicadores de desempenho. A acurácia para cada classificador pode ser vista na tabela 5.

Ano/Acompanhamento	Regressão Logística	RandomForest	MLP
2010/5 anos	0,73	0,68	0,61
2010/6 anos	0,69	0,70	0,70
2011/5 anos	0,85	0,81	0,85
2011/6 anos	0,72	0,75	0,74
2012/5 anos	0,85	0,85	0,80

Tabela 5 – Desempenho dos Classificadores (Acurácia)
Fonte: Os autores.

8 CONCLUSÃO E TRABALHOS FUTUROS

A retenção em instituições de ensino superior brasileiras é um dos problemas educacionais que causam impactos sociais, econômicos e financeiros em toda a sociedade. Nesta perspectiva, é uma questão que precisa ser enfrentada continuamente pelos gestores educacionais na tentativa de encontrar soluções e garantir a eficiência do sistema. A fim de colaborar com o entendimento sobre este tema, a retenção foi abordada neste trabalho como a não conclusão do curso superior no tempo de integralização normal previsto e foi utilizado mineração de dados com o objetivo de identificar variáveis relacionadas ao perfil de alunos com propensão à retenção para dar subsídios à tomada de decisão em relação a ações que possam endereçar o problema.

A metodologia CRISP-DM (foi empregada e demonstrou-se bastante útil para realização do estudo. As etapas iniciais de entendimento do problema e dos dados foram realizadas a partir dos dados disponíveis no Portal INEP dos cursos de graduação da POLI da Universidade de Pernambuco. Em seguida, a preparação dos dados foi executada com foco na limpeza, redução e transformação de dados. Na etapa de modelagem foram utilizadas árvore de decisão para extração de conhecimento e Regressão Logística, *Random Forest* e MLP para classificação. Por fim, os resultados foram avaliados e foi possível perceber na base unificada que alunos ingressantes nos cursos de telecomunicações, computação e mecatrônica têm propensão à retenção. Por outro lado, alunos ingressantes no curso de engenharia civil têm menor probabilidade de se tornarem retidos. Em relação aos algoritmos de classificação, todos se mostraram interessantes apresentando acurácia acima de 80% em algumas configurações de base.

A principal contribuição do estudo foi a compreensão das "ilhas de sucesso" e das oportunidades de melhoria. Ademais, outros benefícios foram observados, tais como:

- Visão do cenário real da retenção dos cursos de graduação da Poli.
- Extração de conhecimento relacionado ao risco de retenção e de não retenção.
- Utilização de modelo de classificação para identificação de alunos com risco de retenção já no ingresso do curso.

As seguintes limitações foram identificadas no estudo:

- Censo da educação superior não possui um atributo que indique a data de conclusão do curso.
- Base do INEP com muitos atributos irrelevantes, com valores ausentes ou iguais.
- Impossibilidade de rastreamento dos alunos nas bases do INEP a partir de 2018.

Reconhecendo as restrições e limitações enfrentadas durante a realização do presente trabalho, compreende-se que novos trabalhos poderão evoluir a experiência e os conhecimentos extraídos adotando novas perspectivas de análise, a saber:

DOI: 10.25286/repa.v7i2.2223

- Utilização da base de dados SIGA com informações acadêmicas do desempenho dos alunos durante a realização do curso.
- Utilização da base de dados da Reitoria com informações socioeconômicas dos alunos para verificar se são atributos relevantes para explicação/previsão de retenção.
- Construção de um sistema informatizado que possa dar suporte à tomada de decisão sobre quais alunos precisam de acompanhamento para minimizar as taxas de retenção.
- Expansão da análise considerando as taxas de desistência de curso

REFERÊNCIAS

[1] INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA (INEP). **Informe estatístico do MEC revela melhoria do rendimento escolar**. Brasília, 2010. Disponível em: http://portal.inep.gov.br/artigo/-/asset_publisher/B4AQV9zFY7Bv/content/informeestatistico-do-mec-revela-melhoria-do-rendimento-escolar/21206#:~:text=A%20publica%C3%A7%C3%A3o%2C%20intitulada%20Informe%20Estat%C3%ADstico,e%2073%25%2C%20em%201996.&text=Esse%20percentual%20despencou%20para%2013,%2C9%25%2C%20em%201996. Acesso em: 27/06/2021.

[2] MINISTÉRIO DA EDUCAÇÃO. **Diplomação, Retenção e Evasão nos Cursos de Graduação em Instituições de Ensino Superior Públicas**. Brasília, 1996. Disponível em: https://www.andifes.org.br/wp-content/files_flutter/Diplomacao_Retencao_Evasao_Graduacao_em_IES_Publicas-1996.pdf. Acesso em: 27/06/2021.

[3] INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA (INEP). **Panorama da educação: Destaques do Education at a Glance 2020**. Brasília, 2020. Disponível em: https://download.inep.gov.br/publicacoes/institucionais/estatisticas_e_indicadores/panorama_da_educacao_destaques_do_education_at_glance_2020.pdf. Acesso em:20/06/2021.

[4] INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA (INEP). **Panorama da educação: Destaques do Education at a Glance 2019**. Brasília, 2019. Disponível em: https://download.inep.gov.br/publicacoes/institucionais/estatisticas_e_indicadores/panorama_da_educacao_destaques_do_education_at_a_glance_2019.pdf. Acesso em: 20/06/2021.

[5] INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA (INEP). **Resumo técnico do Censo da Educação Supeior 2019**. Brasília, 2021. Disponível em: https://download.inep.gov.br/publicacoes/institucionais/estatisticas_e_indicadores/resumo_tecnico_censo_da_educacao_superior_2019.pdf Acesso em: 27/09/2021.

[6] MINISTÉRIO DA EDUCAÇÃO. **Número de candidatos sobe 25,9% e o de inscrições cresce 24,3% em relação a 2018**. Brasília, 2019. Disponível em: http://portal.mec.gov.br/ultimas-noticias/410-sisu-535874847/77021-numero-de-candidatos-sobe-25-9-e-o-de-inscricoes-cresce-24-3-em-relacao-a-2018. Acesso em: 26/06/2021.

[7] SILVA, Hadautho R. B.; ADEODATO, Paulo J. L. **A data mining approach for preventing undergraduate students retention**. The 2012 International Joint Conference on Neural Networks (IJCNN), p. 1–8, 2012. DOI: https://doi.org/10.1109/IJCNN.2012.6252437

[8] FILHO, Rogério Luiz Cardoso Silva; ADEODATO, Paulo Jorge Leitão. **Solução de Mineração de Dados para Avaliação do Ensino Médio dos Institutos Federais a partir do Censo Escolar e do ENEM**, 2017. Disponível em: http://www.iadisportal.org/digital-library/solu%C3%A7%C3%A3o-de-minera%C3%A7%C3%A3o-de-dados-para-avalia%C3%A7%C3%A3odo-ensino-m%C3%A9dio-dos-institutos-federais-a-partir-do-censo-escolar-e-do-enem. Acessado em: 24/06/2021.

[9] BEZERRA, Camila et al. **Evasão Escolar: Aplicando Mineração de Dados para Identificar Variáveis Relevantes**, 2016. Disponível em: https://www.br-ie.org/pub/index.php/sbie/article/view/6795/4680. Acesso em: 23/06/2021.

[10] LEMAY, David J. et al. **Comparison of learning analytics and educational data mining: A topic modeling approach**. Computers and Education: Artificial Intelligence 2, 2021. DOI: https://doi.org/10.1016/j.caeai.2021.100016.

[11] RIBEIRO, Renato; CANEDO, E.D. **Using Data Mining Techniques to Perform School Dropout Prediction: A Case Study**. 17th

International Conference on Information Technology–New Generations, 2020. Advances in Intelligent Systems and Computing, vol 1134. DOI: https://doi.org/10.1007/978-3-030-43020o-7_28.

[12] SHUQFA, Zaid; HAROUS, Saad. **Data Mining Techniques Used in Predicting Student Retention in Higher Education: A Survey**. International Conference on Electrical and Computing Technologies and Applications (ICECTA), p. 1–4, 2019. DOI: https://doi.org/10.1109/ICECTA48151.2019.8959789.

[13] AMARAL, Marcelo Gomes. **Mineração de dados aplicada à classificação do risco de evasão de discentes ingressantes em instituições federais de ensino superior**. Dissertação (Ciência da Computação). Universidade Federal de Pernambuco, Recife, 2016.

[14] CHAPMAN, Peter et al. **CRISP-DM 1.0: Step-by-step data mining guide**, 2000. Disponível em: https://www.kde.cs.uni-kassel.de/wp-content/uploads/lehre/ws2012-13/kdd/files/CRISPWP-0800.pdf. Acessado em: 28/09/2021.

[15] Martínez-Plumed, Fernando et al. 2021. **CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories**. IEEE Transactions on Knowledge and Data Engineering 33, 2021, p. 3048–3061. DOI: https://doi.org/10.1109/TKDE.2019.2962680

[16] HAN, Jiawei; KAMBER, Micheline; PEI, Jian P. **Data Mining: Concepts and Techniques (3rd ed.)**. Morgan Kaufmann Publishers Inc., San Francisco, California, 2011.

[17] INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA (INEP). 2014. **Acesso à Informação Dados Abertos Microdados Censo da Educação Superior**. Disponível em: http://https://www.gov.br/inep/pt-br/acessoa-informacao/dados-abertos/microdados/censo-da-educacao-superior. Acessado em: 26/08/2021.

[18] GONÇALVES, Eduardo. **Regras de Associação e suas Medidas de Interesse Objetivas e Subjetivas**. INFOCOMP, 2005. Disponível em: https://www.researchgate.net/publication/301504294_Regras_de_Associacao_e_suas_Medidas_de_Interesse_Objetivas_e_Subjetivas. Acesso em: 01/09/2021.