Uma Abordagem de Descoberta de Conhecimento para Desvendar Causas da Evasão Acadêmica: Um Estudo de Caso na UFERSA

Leonardo Torres Marques, UECE, leonardo.torresmarques@gmail.com
Bruno Torres Marques, UFC, brunotores@alu.ufc.br
Carlos Alexandre Morais Silva, UFC, carlosalexandresilva100@gmail.com

Resumo: O abandono da universidade é um dos problemas mais intrigantes e cruciais da educação. Esse problema permeia os vários níveis e modalidades de educação e gera perdas para todos os envolvidos no processo educacional. Portanto, é essencial o desenvolvimento de métodos eficientes para prever o risco de abandono de estudantes, permitindo que as instituições adotem ações proativas para minimizar a situação. Assim, objetiva-se com este trabalho, apresentar uma abordagem de descoberta de conhecimento em banco de dados desenvolvida para a previsão de grupos de alunos em risco de abandono nos cursos presenciais de ensino superior. A abordagem foi validada, utilizando dados de ex-alunos de um curso superior (Ciência da Computação) da UFERSA campus UFERSA, utilizando modelos de classificação.

Palavras-Chave: abordagem de descoberta de conhecimento, evasão escolar, classificação, previsão.

Knowledge Discovery Approach to Uncover Causes of Academic Dropout: A Case Study at UFERSA

Abstract. Dropping out of the university is one of the most intriguing and crucial problems in education. This problem permeates the various levels and modalities of education and generates losses for everyone involved in the educational process. Therefore, it is essential to develop efficient methods to predict the risk of student dropout, allowing institutions to take proactive actions to minimize the situation. Thus, the objective of this work is to present a knowledge discovery approach in a database developed for the prediction of groups of students at risk of dropping out in on-campus higher education courses. The approach was validated, using data from alumni of a higher education course (Computer Science) at the UFERSA campus UFERSA, using classification models.

Keywords: knowledge discovery approach, school dropout, classification, forecast.

1. Introdução

A evasão escolar é um caso social complexo, definido como o encerramento do ciclo de estudos (GAIOSO, 2005). Esse problema aflige as instituições de ensino em geral, sejam públicas ou privadas, uma vez que a saída dos alunos ocasiona externalidades negativas nos âmbitos sociais, acadêmicos, econômicos e ambientais, repercutindo de maneira negativa nas políticas públicas e privadas de investimento e desenvolvimento educacional.

Compreender a evasão escolar no ensino superior pressupõe explorar e entender os processos de mudanças pelos quais passam os estudantes durante seu período de formação universitária. É fundamental depreender que a educação superior pode provocar mudanças nos estudantes em diversos níveis como no pessoal, cognitivo, profissional, afetivo e social (SCALI et al., 2009).

O estudo das causas da evasão escolar e a tomada de medidas preventivas estão fortemente ligadas ao contexto de cada instituição de ensino. A identificação dos fatores que influenciam a evasão escolar e a atribuição de uma ordem de importância para estes fatores é

um trabalho complexo, que está diretamente ligado à análise do conjunto de alunos (MANHÃES et al., 2011).

Uma solução propícia para o estudo das causas da evasão escolar, é o uso da descoberta de conhecimento, por meio de técnicas de Mineração de Dados (MD), denominado de Mineração de Dados Educacionais (MDE) (ROMERO; VENTURA, 2007). Essa área de pesquisa é responsável pelo desenvolvimento de métodos para entender melhor os alunos e o contexto em que eles aprendem (ROMERO; VENTURA, 2010).

Contudo, o processo de descoberta de conhecimento é complexo e composto por diversas etapas que não são triviais e passíveis de serem utilizadas uniformemente em qualquer aplicação. Além disso, para facilitar sua aplicação e melhorar a acurácia dos resultados, tanto o processo quanto suas etapas, devem ser adaptados de acordo com os diferentes domínios de aplicação (Kovanovic et al., 2015).

Com isso, o objetivo deste trabalho é propor uma abordagem de descoberta de conhecimento, que visa identificar precocemente alunos com tendência de evadir no ensino superior. Nesta abordagem, são definidas etapas que englobam desde a formulação dos tipos de fatores para exploração, até atividades de modelos preditivos para desvendar as possíveis causas que levam alunos de nível superior a evadir. A abordagem foi validada com dados reais de alunos de um curso de Ciência da Computação (CC) da UFERSA do campus UFERSA.

O restante deste trabalho está estruturado da seguinte forma: Na Seção 2, discutem-se os trabalhos relacionados concernentes ao problema da evasão escolar utilizando MD. Na Seção 3, discorre-se a respeito dos procedimentos e métodos. Na Seção 4, é descrita a abordagem de descoberta de conhecimento. Na Seção 5, é apresentada a validação da abordagem de descoberta de conhecimento. Por fim, na Seção 6, são apresentadas as considerações finais e discussão sobre trabalhos futuros.

2. Trabalhos Relacionados

A previsão do desempenho acadêmico é um objeto de estudo, explorado por diversos pesquisadores. Pesquisas mais antigas empregam métodos estatísticos ou outros procedimentos para compreender o problema (Johnston, 1998). O emprego de técnicas de MD sobre dados educacionais é relativamente recente conforme destaca Baker e Yacef (2009). Contudo, a maioria dos trabalhos correlatos estão restritos a identificar resultados em pequenos contextos relativos a apenas uma disciplina de um determinado curso.

Neste trabalho, empregou-se como base para propor etapas e atividades da abordagem de descoberta de conhecimento o Mapeamento Sistemático da Literatura (MSL) apresentado em Marques et al., (2019). Buscou-se no trabalho elencar ferramentas, técnicas e fatores indutores utilizados para prever as causas do problema (evasão escolar) nos últimos dez (10) anos. A partir dos resultados, evidenciou-se que das ferramentas investigadas nos artigos selecionados, destacam-se a *Weka*, *Python* e *Software R*; como as que tiveram melhores avaliações de acordo com três (3) critérios definidos pelos autores (licença, documentação e são *open sources*).

Quanto as técnicas, destaca-se que as encontradas foram diversificadas, no entanto a que mais apareceu foi a técnica de classificação. Ainda vale ressaltar que os algoritmos de classificação que apareceram nos trabalhos foram: *Naive Bayes* (NB), *Support Vector Machine* (SVM), *Multi-Layer Perceptron* (MLP), *K-Nearest Neigh-Bours* (KNN), *Jrip*, *OneR*, *J48*, *PART* e *AdaBoost*.

No que diz respeito aos fatores indutores para evasão escolar para desvendar possíveis causas da evasão escolar, foram utilizados dados de características individuais dos alunos, e, em seguida, dados inerentes às instituições. Para o levantamento de dados, destacam-se três (3) formas: utilização de dados acadêmicos, por questionários e entrevistas.

3. Procedimentos e Métodos

A abordagem de descoberta de conhecimento proposta neste trabalho, é embasada no MSL apresentado na Seção 2, de maneira geral as etapas propostas para a abordagem, bem como as ferramentas, técnicas e fatores indutores para a evasão escolar utilizados para a validação da abordagem foram acordados no trabalho de Marques et al., (2019).

Deve-se ressaltar que a abordagem é composta de cinco (5) etapas, onde as duas (2) primeiras foram definidas pelos autores deste trabalho com base nos resultados do MSL e as três (3) últimas etapas são adaptadas do processo *Knowledge Discovery in Databases* (KDD). O KDD, também conhecido como processo de descoberta de conhecimento, é utilizado quando se objetiva identificar conhecimento útil em bases de dados. Para Fayyad et al. (1996), o KDD consiste em um processo não trivial, responsável por identificar padrões em dados de forma a serem válidos, potencialmente úteis e compreensíveis.

Para validar a abordagem, foi utilizada uma base de dados de ex-alunos de um curso de CC da UFERSA campus UFERSA. A definição das variáveis utilizadas para a coleta dos dados, a forma de coleta e todo o procedimento seguinte de processamento de dados, transformação de dados e construção dos modelos foram seguidos de acordo com a abordagem proposta neste trabalho.

4. Abordagem de Descoberta de Conhecimento

Visando facilitar a descoberta das possíveis causas da evasão escolar no ensino superior, nesta Seção se apresenta uma abordagem de descoberta de conhecimento como mostrado na Figura 1. Nesta abordagem, uma vez definidos os fatores que serão estudados e, posteriormente, adquiridos os dados, estes dados passam por um tratamento programado através de um encadeamento de atividades nas etapas de processamento e transformação, que tenciona aumentar a qualidade do Banco de Dados (BD).

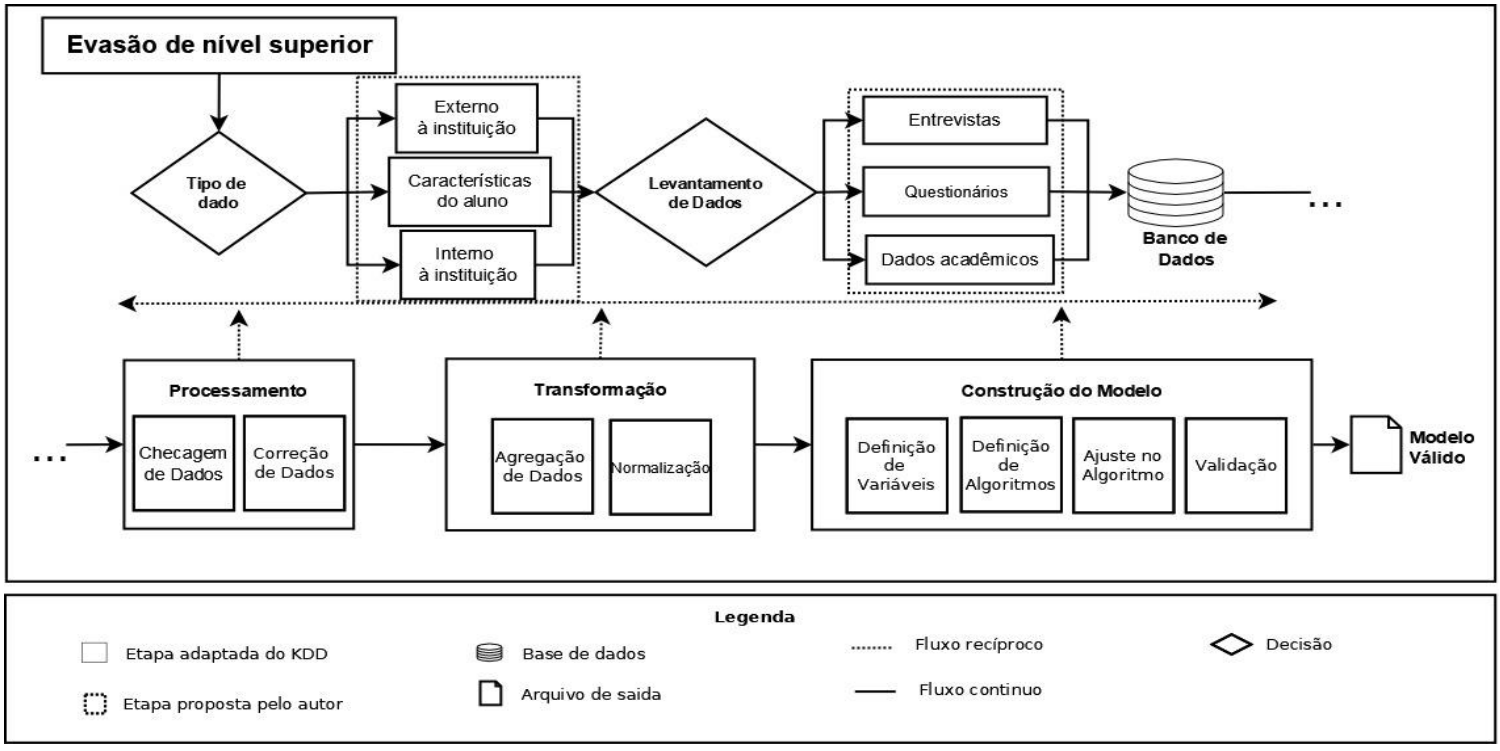


Figura 1 - Abordagem de descoberta de conhecimento para identificar causas da evasão.

Logo após, um modelo de predição é construído por meio da execução de algoritmo de MD. O especialista avalia o modelo desenvolvido, se os resultados se mostrarem relevantes, o mesmo pode implementar o modelo e usá-lo em previsões futuras. Senão, retorna-se às etapas anteriores com o intuito de aumentar a qualidade dos resultados. O conjunto de etapas incorporadas nesta abordagem são expostas adiante.

A abordagem tem início a partir de uma decisão de qual tipo de fatores indutores devem ser utilizados para desvendar as causas da evasão escolar, como visto na Seção 2, as causas do problema se enquadram a três (3) grandes categorias: fatores inerentes às características individuais do aluno, fatores internos às instituições e fatores externos às instituições. Vale salientar, que não necessariamente é obrigatório decidir por apenas uma categoria, o especialista pode decidir por uma ou mais categorias para utilizar nas buscas das causas da evasão escolar em determinada instituição.

Uma vez que se estabelece que tipos de fatores utilizar, inicia-se a segunda etapa da abordagem, que é decidir como será o levantamento dos dados e de onde extraí-los. Na Seção 2 foi possível observar que, nos trabalhos selecionados no MSL, empregou-se pelo menos uma das seguintes formas para levantamento de dados: dados acadêmicos, questionários ou entrevistas. Assim sendo, nesta etapa se decide por qual dessas formas é possível utilizar e a mais conveniente, assim como na etapa anterior, nesta etapa também é possível decidir por utilizar uma (1) ou mais das opções disponíveis.

Nos casos em que se decide pela utilização de questionários e/ou entrevistas, recomenda-se o emprego de questões objetivas de múltiplas escolhas fechadas, essa estratégia evita ou minimiza trabalhos posteriores com processamento e transformação de dados. Uma vez que, a característica fundamental destas questões é evitar o caráter subjetivo das respostas, o oposto disto leva as dificuldades de tratamento de dados.

Nas próximas Subseções, são apresentadas a terceira, quarta e quinta etapa da abordagem em mais detalhes. Para cada uma dessas etapas foram definidas atividades ou subetapas, sendo essas definidas com base também no MSL. O objetivo de definir subetapas, foi facilitar a resolução de um problema maior em subproblemas menores.

4.1 Atividades de Processamento

Sendo assim, definida a base de dados que será utilizada, principia-se a etapa de processamento, que pode aumentar a qualidade do BD, por meio da checagem e da correção dos dados. A checagem dos dados é uma atividade mutável que tem como finalidade revelar os atributos com valores nulos, raros e impossíveis. Recursos visuais como gráficos, implementações de código em uma linguagem de programação e consultas em *Structured Query Language* (SQL) a partir de um BD, dentre outras ferramentas, fornecem aos especialistas subsídios de detectar, de maneira eficiente, falhas em repositórios de dados de grandes volumes.

Recomenda-se como passo inicial nesta atividade (checagem dos dados), averiguar a existência ou não de valores nulos, uma vez que é o procedimento mais trivial pela facilidade de reconhecer quando acontece esse tipo de problema. Em seguida, indica-se verificar a presença ou não de valores raros, esse procedimento pode ser realizado por meio da utilização da tabela de frequência absoluta, em que o propósito é verificar dados que se repete em menor frequência. E, por fim, realiza-se a certificação da existência de valores impossíveis. Neste ponto é necessário identificar o conjunto de variáveis dependentes e independentes, com as variáveis dependentes, verifica-se a correlação destas com as independentes. De maneira geral, para realização de cada procedimento desta atividade o especialista deve ter total conhecimento do domínio de aplicação dos dados e como está disposta a base de dados a ser tratada.

Na correção dos dados, o especialista toma suas decisões sobre os atributos, de acordo com a situação detectada na atividade anterior. Ele pode resolver desconsiderar o atributo

suspeito, corrigir manualmente atributos, formatar casas decimais, definir uma constante global (média ou mediana), coletar informações necessárias para modelar ou estimar ruído (por exemplo, regressão ou inferência), dentre outros possíveis procedimentos (GARCÍA et al., 2017).

4.2 Atividades de Transformação

Com os dados processados, inicia-se a etapa de transformação cujo objetivo é aumentar a performance de cada algoritmo. Uma agregação é realizada com o intuito de reduzir o conjunto de dados e aumentar o desempenho na construção de modelos de predição. A redução é definida em concordância com a necessidade do especialista: por exemplo, se for considerado a elaboração de um modelo de previsão para definir causas da evasão escolar, pode-se transformar os dados de renda em faixas salariais, o que indubitavelmente reduziria substancialmente o conjunto de dados e aumentaria a performance na execução dos algoritmos, uma vez que diversos registros seriam unidos em poucas faixas.

Algoritmos de MD, em geral, constroem modelos preditivos de formas diferentes e podem ter um melhor comportamento e exatidão nos resultados, se executados em um conjunto de entrada apropriado: por exemplo, um método pode construir um modelo para predição mais acurado (ou mais rápido), se um atributo salário for representado na forma nominal “mil reais”, em vez da forma numérica mil (1000) (GARCÍA et al., 2017). Diante do exibido, o especialista pode estabelecer diferentes maneiras para simbolizar alguns dos atributos processados sem afetar a integridade deles e usá-los como entrada para os algoritmos na investida de conseguir melhores resultados.

Antes de iniciar a construção dos modelos, propõe-se normalizar todos os atributos em uma escala de dados para simplificar e acelerar a execução dos algoritmos. A normalização lida com a mudança e padronização da dimensão de escalas dos dados, assim sendo, atributos podem ser normalizados com casas decimais deliberadas após a vírgula ou mecanismos que definem uma única escala para todos os atributos (por exemplo, de 0 a 1) (GARCÍA et al., 2017) . Em ambos os casos, a normalização pode aumentar o desempenho dos modelos, posto que permite ao algoritmo executar de forma mais simples omitindo cálculos extensos.

4.3 Atividades da Construção do Modelo

A partir dos dados transformados, pode-se executar os algoritmos automáticos (ou semiautomáticos) e congruentes para predição da possibilidade de um aluno evadir. Os padrões de evasão escolar são diversos tornando difícil definir um único método para uma predição. Por isso, profusos algoritmos devem ser examinados no desenvolvimento de modelos de predição. Desta forma, dispõe-se quatro (4) atividades essenciais para auferir resultados plausíveis: análise dos atributos de entrada, definição do algoritmo, ajuste do algoritmo e validação.

Na análise de atributos se indica detectar uma considerável coleção de entrada para os algoritmos. Atributos em demasia, insignificantes na entrada do modelo, afligem os seus resultados de acerto e performance. Então, pode-se detectar nessa análise que um determinado algoritmo poderia executar em poucos segundos e conseguir melhores resultados de predição se fosse combinado em sua entrada somente seis (6) atributos ao invés de doze (12), por exemplo. Efetivamente, precisa-se analisar e delinear um relevante conjunto de entrada para cada algoritmo, de tal modo a atingir uma verticalidade entre desempenho e acurácia nos resultados.

A definição do algoritmo consiste em estabelecer um algoritmo adequado ou o mais acurado possível para um conjunto de dados previamente definido. É importante salientar que não existe um algoritmo ideal para todos os problemas, sendo assim, é possível que nessa

atividade seja necessário uma série de testes para decidir qual se adequará melhor. Como passo inicial para decidir por qual algoritmo utilizar, indica-se analisar na base de dados a dimensão e a quantidade de dados, visto que o desempenho dos algoritmos está condicionado as características das bases de dados. Por exemplo, um determinado algoritmo pode se comportar melhor com uma base que tem uma quantidade demasiada de variáveis (SVM), enquanto isso, outro algoritmo pode funcionar melhor com uma quantidade de variáveis reduzida (MLP).

Quanto aos testes dos algoritmos, orienta-se dividir o conjunto global de dados em dois (2) conjuntos, o primeiro conjunto se trata dos dados de treino, em que o propósito é treinar o modelo preditor. Enquanto isso, o segundo conjunto é destinado aos testes propriamente dito, em que se averigua a precisão de acerto dos modelos. Ambos os conjuntos são definidos de forma aleatória, empregando-se uma função responsável por tal procedimento (SILVA; SPATTI; FLAUZINO, 2010).

Definida a coleção de dados, pode-se ajustar os parâmetros de cada algoritmo e analisar os resultados. O ajuste dos algoritmos depende de uma série de fatores, como o tamanho da amostra de dados, projeções dos sinais e a quantidade de variáveis. Compete ao especialista executar uma série de testes com o propósito de detectar um plano adequado de configuração nas medidas para cada um dos algoritmos. Uma possibilidade de teste é examinar a acurácia do modelo com todos os atributos iniciais e, logo em seguida, fazer o mesmo processo anterior com os melhores atributos, esses atributos podem ser definidos, por exemplo, com a Lei de Entropia (GARCÍA et al., 2017).

Posteriormente à execução dos algoritmos, um modelo de predição é concebido e deve ser validado por um especialista, o modelo de predição mencionado anteriormente é o resultado de todo o processo utilizado para desvendar as causas da evasão escolar, ou seja, desde à formulação de questões até a validação por meio das técnicas estatísticas. Os resultados da predição de evasão escolar podem ser avaliados por muitas métricas estatísticas (*recall*, *f1-score*, precisão e etc.) que especificam uma representação numérica para o erro. Conquanto, ainda não existe uma única abordagem global para avaliar modelos de predição, em virtude que cada métrica municiona uma diferente visão do erro para o especialista.

5. Validação da Abordagem de Conhecimento

Com o intuito de averiguar o quão precisa a abordagem de descoberta de conhecimento proposta na Seção 4 é, foi realizado um estudo de caso com dados de ex-alunos (concludentes ou evadidos) do curso de CC da UFERSA campus UFERSA. Salienta-se que foram contempladas todas as etapas da abordagem, desde a definição dos fatores e variáveis a partir de questões exploradas em um questionário até a validação dos modelos de MD empregados.

Como definido na abordagem apresentada na Figura 1, a primeira etapa da abordagem consiste em definir quais fatores indutores para evasão escolar devem ser abordados na pesquisa, sendo assim, foram definidas variáveis que abrangessem fatores das três (3) categorias possíveis. No Quadro 1 são apresentadas as perguntas direcionadas aos ex-alunos e o fator à qual pertence.

Na segunda etapa, definiu-se a forma de coleta ou levantamento dos dados, decidiu-se pelo questionário, uma vez que foi o meio mais fácil de levantamento para o contexto do estudo de caso. O questionário foi aplicado de forma *on-line* no período que compreendeu 01/04/2019 à 30/04/2019, enfatiza-se que a instituição disponibilizou os *e-mails* dos ex-alunos para que fosse feito o contato. Sendo assim, destaca-se que foram obtidas respostas de cem (100) ex-alunos, entre as quais cinquenta e duas (52) foram de evadidos e quarenta e oito (48) de concludentes.

Quadro 1 - Fatores e variáveis exploradas e respectivas descrições.

Características individuais do aluno	
#1	Qual sua situação (evadido ou concludente)?
#2	Em que ano ingressou no curso?
#3	Em que ano concluiu ou evadiu do curso?
#4	Tinha experiência prévia na área?
#5	Estudou em escola pública ou privada?
#6	Que tipo de escola estudou (militar, federal e etc)?
#7	Qual cidade morava antes de ingressar no curso?
#8	Ao ingressar no curso mudou para a cidade do curso?
#9	Qual a renda mensal familiar do ex-aluno ao ingressar no curso?
#10	Qual a raça do ex-aluno (branco, negro, pardo etc)
#11	Como foi o aproveitamento do ex-aluno no 1º semestre?
#12	Qual estado civil dos pais dos ex-alunos?
#13	Qual grau de instrução da mãe do ex-aluno?
#14	Qual grau de instrução do pai do ex-aluno?
Fatores Internos à Instituição	
#15	Morou na vila acadêmica durante a graduação?
#16	Você ganhou bolsa durante a graduação?
#17	Você teve acompanhamento institucional?
#18	Opinião em relação a grade curricular do curso?
#19	Opinião em relação a cadeia de pré-requisitos das disciplinas?
#20	Qual sua opinião em relação a infraestrutura da instituição para realizar o curso?
#21	Qual sua opinião quanto à qualidade dos professores do curso?
Fatores Externos à Instituição	
#22	Você acha que as carreiras relacionadas ao curso são socialmente reconhecidas?
#23	Você acredita que os profissionais no curso são remunerados adequadamente?

Uma vez que os dados foram levantados, as atividades da terceira etapa da abordagem foram iniciadas (checagem dos dados e correção dos dados), a etapa de processamento. Vale destacar que foi utilizada a linguagem *Python* e suas bibliotecas (*pandas*, *matplotlib* e *sklearn*) para a realização das atividades, pois, como visto na Seção 2 foi uma das ferramentas resultantes do MSL e que de acordo com os critérios de avaliação, o autor deste trabalho juntamente com outros especialistas, julgaram a mais adequada.

Na atividade de checagem dos dados, tratou-se de buscar quais variáveis apresentavam valores nulos, raros ou impossíveis. Constatou-se, que oito (8) variáveis apresentavam pelo menos um dos problemas supracitados, as variáveis que apresentaram problemas foram: #1, #2, #3, #4, #5, #6, #10 e #12. Essas variáveis tiveram que ser corrigidas na atividade de correção dos dados, dentre outras manipulações, um exemplo, apresenta-se na Figura 2 com a variável "#1", salientando que na Figura 2(a) são apresentados os dados antes da correção e na Figura 2(b) o resultado após a correção de dados.

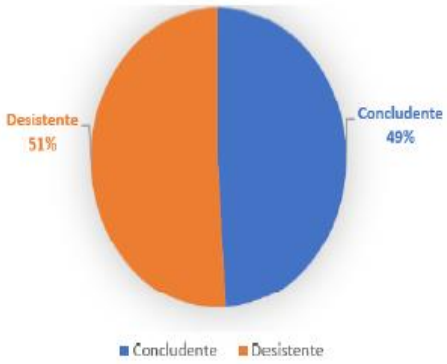
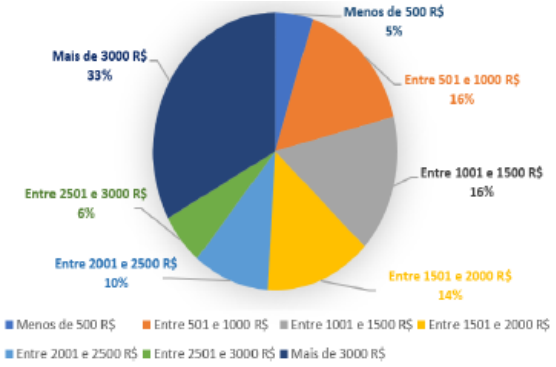


Figura 2 (a) – Antes da correção. **Figura 2 (b)** – Depois da correção.

Figura 2 – Correção de dados da variável "#1".

Dado que as variáveis em que se observou problemas foram corrigidas, iniciou-se as atividades (agregação de dados e normalização) da quarta etapa do processo, a etapa de transformação. Na atividade de agregação de dados, percebeu-se que era conveniente agregar dados em oito (8) variáveis (#2, #3, #6, #7, #9, #11, #13 e #14); pois, essas variáveis tinham dados dispersos, o que diminui substancialmente a eficiência dos algoritmos. Na Figura 3, apresenta-se um exemplo com a variável "#9", salientando que na Figura 3(a) são apresentados os dados antes da agregação e na Figura 3(b) o resultado após a agregação de dados.



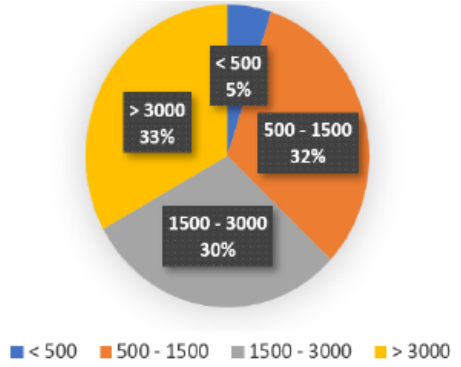
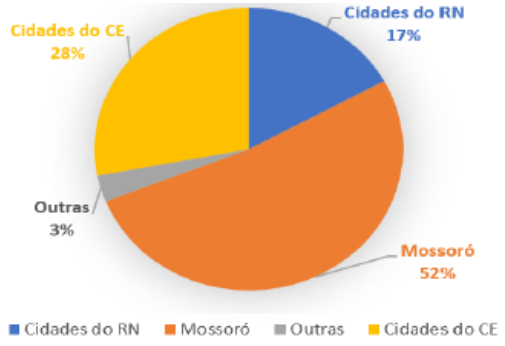


Figura 3 (a) – Antes da agregação. **Figura 3 (b)** – Depois da agregação.

Figura 3 – Agregação de dados da variável "#9".

Após o tratamento adequado dos dados na atividade de agregação, a última atividade da etapa de transformação, consistiu em normalizar todos as variáveis em uma escala de dados para simplificar e acelerar a execução dos algoritmos. Dessa forma, na Figura 4 é possível observar um exemplo das manipulações que foram realizadas nesta atividade. Mostra-se na Figura 4(a) a variável "#7" antes da normalização e na Figura 4(b) o resultado após à normalização de dados.



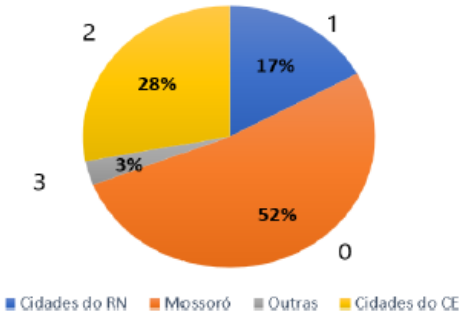


Figura 4 (a) – Antes da normalização. **Figura 4 (b)** – Depois da normalização.
Figura 4 – Normalização de dados da variável "#7".

A partir das manipulações realizadas até então, a etapa seguinte consistiu na construção do modelo preditivo. Neste caso, a primeira atividade foi definir uma coleção de variáveis de entrada para os algoritmos. Assim, utilizou-se a Lei da Entropia para definir as melhores variáveis, esta lei significa que toda a energia em um sistema eremítico se move do estado ordenado para o estado desordenado (RIFKIN, 1980). Dessa forma, o estado mais ordenado, em que a concentração é maior, é visto o estado da entropia mínima.

De forma resumida, representa o ganho de informação de uma variável para encontrar o valor real da variável objetivo, isso significa que os sistemas de maior grau de ganho de informação são mais constituintes nos dois (2) grupos de concludentes e evadidos. Neste trabalho, considerou-se que as melhores variáveis são aquelas que obtiveram ganho de informação igual ou superior a zero vírgula zero dois (0,02), salienta-se que as variáveis (#4 e #23) obtiveram valores inferiores ao mínimo desejado, sendo portanto desconsideradas para o estudo, pois, ambas contribuem minimamente para o modelo e podendo até diminuir a acurácia do modelo.

O passo seguinte foi definir os algoritmos. Como visto na Seção 2, a técnica mais utilizada no estudo das causas de evasão escolar, foi a classificação. Dessa forma, neste trabalho, os algoritmos de classificação utilizados, foram: *J48*, KNN, SVM e *Adaboost*; tais algoritmos foram escolhidos com base nos resultados do MSL. Em seguida, tratou-se de ajustar as variáveis resultantes da atividade "definição de variáveis" em cada um dos algoritmos, salienta-se que o conjunto global de dados foram divididos em dois grupos (treino e teste), ressaltando que foram em 30% e 70% respectivamente. No Quadro 2 é possível verificar o desempenho de cada algoritmo, deve-se destacar que o algoritmo SVM foi o que alcançou melhor resultado, obtendo noventa e oito por cento (98%) na média.

Quadro 2 – Acurácia dos Algoritmos.

Algoritmo	**Concludentes**			**Evadidos**			**Média**
	precisão	**recall**	**F1-score**	**precisão**	**recall**	**F1-score**	
J48	0.96	0.90	0.93	0.91	0.96	0.93	0.93
KNN	0.92	0.92	0.92	0.92	0.92	0.92	0.92
SVM	0.98	0.98	0.98	0.98	0.98	0.98	0.98
Adaboost	0.97	0.97	0.97	0.97	0.97	0.97	0.97

Dessa forma, evidencia-se que o algoritmo SVM foi o que obteve melhor resultado na classificação de alunos que evadem e concluem o curso de UFERSA. No geral, a média de acerto considerando todos os algoritmos ficou em torno de noventa e cinco por cento (95%), com esses resultados é possível destacar o bom desempenho da abordagem.

6. Considerações Finais

Este trabalho consiste em uma pesquisa que objetiva propor uma abordagem de descoberta de conhecimento inovadora, para identificar, proativamente, de forma contínua e precisa, os alunos considerados no grupo de risco de abandono, em salas de aula do ensino superior. Para validação da abordagem, utilizou-se um BD com informações coletadas de ex-alunos do curso de CC da UFERSA campus UFERSA.

A partir do modelo gerado, um sistema para alertar coordenadores de curso, docentes e os demais departamentos da instituição sobre os alunos que estão potencialmente em risco de desistir pode ser implementado. Como exemplo de ação possível, propõe-se que, aqueles alunos encontrados em risco, sejam designados para um tutor, a fim de lhes fornecer apoio acadêmico e orientação para motivar e tentar evitar o insucesso escolar.

Finalmente, como o próximo passo desta pesquisa, pretende-se realizar mais experimentos usando dados de diferentes cursos, para testar se os mesmos resultados de acurácia da abordagem são obtidos com diferentes bases de dados. Como trabalhos futuros, destaca-se o seguinte: adicionar novas variáveis que possibilitem prever o fracasso do aluno o mais rápido possível e propor estratégias para ajudar os alunos identificados dentro do grupo de risco.

Por fim, este estudo trata da previsão de ações resultantes de resoluções e decisões de seres humanos. Assim, reconhece as limitações da metodologia e as possíveis falhas, dados que as previsões fora do determinismo completo sobre à evasão escolar, podem ser resultado de um processo estocástico. Portanto, considerando os resultados, é notório que a abordagem é inovadora para prever alunos no grupo de risco de abandono escolar no ensino superior, contribuindo com uma lacuna identificada nas produções da comunidade científica.

Referências Bibliográficas

BAKER, R. S. and YACEF, K. (2009). The state of educational data mining in 2009: A review and future visions. JEDM— **Journal of Educational Data Mining**, 1(1):3–17.

FAYYAD, U., PIATETSKY-SHAPIRO, G., and SMYTH, P. (1996). From data mining to knowledge discovery in databases. **AI magazine**, 17(3):37–37.

GAIOSO, N. D. L. O fenômeno da evasão escolar na educação superior no brasil. Brasília, DF: **Universidade Católica de Brasília**, 2005.

SCALI, D. F. et al. Evasão nos cursos superiores de tecnologia: a percepção dos estudantes sobre seus determinantes. [sn], 2009.

MANHÃES, L. M. B. et al. 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)**. [S.l.: s.n.], 2011. v. 1, n. 1.

MARQUES, L. T., QUEIROZ, P. G. G., DE CASTRO, A. F., MARQUES, B. T., and SILVA, J. C. P. (2019). Mineração de dados auxiliando na descoberta das causas da evasão escolar: Um mapeamento sistemático da literatura. **RENOTE-Revista Novas Tecnologias na Educação**, 17(3):194–203.

ROMERO, C.; VENTURA, S. Educational data mining: A survey from 1995 to 2005. **Expert systems with applications**, Elsevier, v. 33, n. 1, p. 135–146, 2007.

ROMERO, C.; VENTURA, S. Educational data mining: a review of the state of the art. **IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)**, Ieee, v. 40, n. 6, p. 601–618, 2010.

KOVANOVIC, V. et al. Penetrating the black box of time-on-task estimation. In: ACM. **Proceedings of the fifth international conference on learning analytics and knowledge**. [S.l.], 2015. p. 184–193.

RIFKIN, J. Entropy: a new world view.[social and political implications of the second law of thermodynamics]. **Viking Press, New York**, NY, 1980.

SILVA, I. N. D.; SPATTI, D. H.; FLAUZINO, R. A. Redes neurais artificiais para engenharia e ciências aplicadas curso prático. **São Paulo: Artliber**, 2010.