Mineração de Dados Educacionais para identificação de Perfil de Retenção em um curso de Ciência da Computação

Joana Pacheco Rolim[1], Rômulo César Silva[1]

[1]Universidade Estadual do Oeste do Paraná (UNIOESTE) – Foz do Iguaçu, PR – Brazil

joana.rolim@unioeste.br, romulocesarsilva@gmail.com

Abstract. *The phenomenon of retention in undergraduate courses is a cause for concern and negative impacts for both the university and society. This work proposes the use of Educational Data Mining and Machine Learning techniques to identify the retention profile of undergraduate students in Computer Science at Unioeste, Foz do Iguaçu campus, applying the KNN and SVM algorithms. The results showed that there is a relationship between data such as gender, age and retention, making it useful for the institution to adopt strategies aimed at avoiding low student performance.*

Resumo. *O fenômeno da retenção no ensino superior é motivo de preocupações e impactos negativos tanto para a universidade e sociedade. Este trabalho propõe a utilização de técnicas de Mineração de Dados Educacionais e Aprendizagem de Máquina para identificar o perfil de retenção de alunos de graduação em Ciência da Computação da Unioeste, campus de Foz do Iguaçu, aplicando os algoritmos KNN e SVM. Os resultados demonstraram que existe relação entre os dados como sexo, idade e a retenção, sendo útil para a instituição adotar estratégias que visem evitar o baixo rendimento estudantil.*

1. Introdução

Uma educação de qualidade é fundamental para o desenvolvimento do ser humano, e formação de cidadãos conscientes e atuantes socialmente. Por outro lado, ingressar em uma universidade pode ser desafiador e isto pode levar muitos estudantes a interromper temporária ou definitivamente o curso de graduação matriculado. Apesar da retenção ser um problema que afeta grande parte da esfera acadêmica, ainda é pouco difundido ou pesquisado, no qual não se identificam trabalhos sobre os determinantes acadêmicos da retenção nas instituições de Ensino Superior público no Brasil, principalmente se comparado ao problema da evasão. Isso se deve ao fato de a evasão ser mais facilmente percebida pelos gestores, em virtude da ausência de estudantes, gerando vagas ociosas. Já na retenção, o estudante permanece na instituição de ensino, e sua vaga continua ocupada, porém o papel da universidade em formar um cidadão para a sociedade capacitado e qualificado não está sendo cumprido de maneira eficaz [Manhães 2015].

A retenção no Ensino Superior acarreta perdas tanto para o estudante, visto que ao demorar ao se formar ele está deixando de receber os benefícios proporcionados pela graduação, quanto para a sociedade, uma vez que recursos repassados são perdidos, resultando na falta de profissionais disponíveis no mercado de trabalho. Segundo dados do Censo da Educação Superior feito pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), os cursos das áreas de exatas, que incluem Matemática, Computação e Engenharias, apresentaram taxas de evasão acumulada entre 60% e 90% para o período de 2009 a 2014. Esse cenário alarmante traz consigo

preocupações, dado que os cursos das áreas de Computação e Engenharia estão diretamente relacionados com parte significativa da geração de inovações tecnológicas e com o aumento de produtividade [Saccaro, França & Jacinto 2019]. Destaca-se também que, entre os anos de 2006 e 2016, o número de matrículas na Educação Superior aumentou 62,8%, com uma média anual de 5,0% de crescimento. Entretanto, a quantidade de alunos concluintes, em 2014, por exemplo, era de 44,7% nas instituições públicas, indicando altas taxas de evasão e de retenção.

O fenômeno da retenção na graduação é motivo de preocupação tanto nas instituições privadas quanto nas públicas. Seus impactos negativos incluem a redução de profissionais que chegam ao mercado de trabalho e o desperdício de recursos humanos e econômicos, dado ao número expressivo de acadêmicos que evadem do curso, comprometendo o retorno à sociedade. De modo geral, os cursos da área de exatas apresentam altos índices de retenção, como é o caso do curso de Bacharelado em Ciência da Computação. As instituições de ensino, na busca por fornecer ajuda e assistência aos estudantes de graduação, realizam geralmente atividades de tutoria e monitoria, no entanto, o acompanhamento do desempenho é feito de modo subjetivo, uma vez que envolve e depende da experiência acadêmica e envolvimento dos docentes. Segundo Faúndez, Muñoz & Cornejo (2012), a instituição deve investir em uma educação permanente de seus docentes, estimulando habilidades pedagógicas e investigativas, na identificação de riscos associados a estudantes. É importante salientar, que em geral, os docentes desenvolvem e participam de inúmeras atividades fora do ambiente de sala de aula, como laboratórios de pesquisa, o que resulta em dificuldades de reconhecer as necessidades de cada estudante de maneira individual. Assim, é necessário investigar a fundo e propor medidas que viabilizem o acompanhamento dos acadêmicos, identificando seus desafios, objeções e possíveis riscos.

A Mineração de Dados (do inglês *Data Mining* - DM) é parte fundamental nesse processo. O uso de técnicas de Mineração de Dados dentro do contexto educacional é chamado de Mineração de Dados Educacionais (do inglês *Educational Data Mining* - EDM), definida como uma área que explora a Estatística, Aprendizagem de Máquina e algoritmos de Mineração de Dados aplicados a diferentes tipos de dados educacionais [Romero & Ventura 2010]. A quantidade e solidez dos dados obtidos é essencial para garantir a efetividade da análise. A fim de assegurar a qualidade e padronização para utilização, a base de dados passa por um pré-processamento minucioso e sistemático. EDM, ao explorar as informações coletadas, permite compreender os alunos de forma eficaz e adequada, acarretando ganhos e melhorias educacionais.

O objetivo deste trabalho é identificar o perfil de retenção dos estudantes que necessitam de apoio no curso de Bacharelado em Ciência da Computação – Unioeste, campus de Foz do Iguaçu, tendo como base o desempenho acadêmico do aluno em algumas disciplinas anuais e semestrais eleitas com os maiores índices de reprovação, a partir da 2º série da graduação, entre os anos de 2007 a 2019, fazendo uso de técnicas de Mineração e Aprendizagem de Máquina. Através de um conjunto de dados real, foi avaliada a relação entre o resultado (aprovação ou reprovação) e média final do aluno na disciplina, além de dados como sexo e idade que cursou a disciplina analisada. O estudo justifica-se pela importância que possui a identificação do perfil de retenção do estudante em um curso de Computação, entendendo que o sucesso acadêmico deve ir adiante do acesso ao Ensino Superior. A intenção que o move vai além da constatação acerca da relação entre os dados analisados e resultado do estudante nas disciplinas, diferenciando-

se pela busca de elementos e padrões para a criação de medidas e alertas na instituição de ensino, visando minimizar possíveis riscos de atraso.

São apresentados a seguir alguns trabalhos relacionados, a metodologia de análise de dados empregada neste estudo, os algoritmos e ferramentas utilizados, discussões sobre os resultados obtidos, e por último, as considerações finais.

2. Trabalhos Relacionados

No Brasil, a retenção é definida como a permanência longa e demorada do aluno na universidade, ou seja, com uma duração maior do que a planejada pelo currículo da instituição, ocorrendo, segundo Lamers et al. (2017), por motivo de suspensão, cancelamento ou trancamento de matrícula ou repetência, resultando em um maior tempo para o estudante finalizar o curso. [Lima Júnior et al. 2019]. De acordo com Campello e Lins (2008) e Mainier et al. (2006), o atraso no término da graduação pode levar o aluno a evasão. Assim, essas duas definições estão bastantes associadas. Em vista disso, os fatores determinantes da retenção estão estritamente ligados aos fatores que levam à evasão.

Um estudo que buscou encontrar relação entre a nota de ingresso de estudantes e o seu desempenho nas disciplinas de Cálculo Diferencial e Integral I, Física Aplicada à Computação I, Cálculo Vetorial e Geometria Analítica do primeiro período do curso de Ciência da Computação da Universidade Federal da Paraíba (UFPB), fazendo uso de ferramentas e algoritmos de Aprendizado de Máquina, obteve precisão acima de 70% e seus autores concluíram que predizer o desempenho dos alunos é viável e necessário como base para a criação e o planejamento de políticas e estratégias que visem diminuir o número de reprovações no curso de Ciência da Computação [de Brito, Júnior, Queiroga & do Rêgo 2014].

Estudos brasileiros evidenciam que diversos fatores internos e externos à instituição determinam a retenção. Dentre fatores externos, ressaltam-se as dificuldades dos estudantes em conciliar as aulas e o trabalho [Vanz et al. 2016]. Já em relação aos fatores internos, a falta de apoio financeiro e a falta ou restrição de programas acadêmicos, como auxílios, bolsas e monitorias, englobam o cenário. Esses fatores também estão associados aos cursos, como as dificuldades nas disciplinas, metodologias dos professores e falta de prática do curso [Silva 2017]. Andriola e Araújo (2018) destacam a importância da seleção de indicadores educacionais pelos gestores, como tentativa de obter um diagnóstico da instituição, estabelecendo estrategicamente um plano de ações e projetos. Donoso-Díaz, Iturrieta e Traverso (2018), nessa mesma linha, salientam a importância de medidas que visem minimizar as taxas de retenção, entre elas está o desenvolvimento de Sistemas de Alerta Precoce, que buscam detectar alunos que possuem risco de atraso ou abandono. Dessa maneira, é possível adotar medidas que possam influir na decisão dos alunos de não atrasar ou de abandonar o curso.

Nesse contexto apresentado, que a presente pesquisa busca contribuir com as possíveis políticas e estratégias de redução da retenção na graduação, ao identificar o perfil de retenção dos alunos em um curso de Computação, e fatores determinantes, embasando a criação de sistemas e ações que visem detectar a retenção, reduzindo a sua ocorrência, e, consequentemente, a evasão durante a graduação.

3. Metodologia

Para este estudo, foram coletados os seguintes dados: sexo, idade, data de nascimento, nacionalidade, procedência, forma de ingresso, se ocupa ou não vaga de cotista, notas nas disciplinas, média, resultado final, bem como o status que consiste em formado, cancelado por abandono, cursando ou transferido. Tais dados são referentes aos anos 2007 a 2019 do curso de Bacharelado em Ciência da Computação da Unioeste, campus de Foz do Iguaçu, tendo sido fornecidos pela Secretaria Acadêmica da Universidade. O curso é estruturado em 4 (quatro) séries ou anos, sendo algumas disciplinas anuais e outras semestrais. Optou-se por fazer a identificação de perfil de retenção para disciplinas a partir da 2ª série, em função de não se ter disponível as informações de variáveis socioeconômicas e desempenho dos alunos durante o ensino médio.

3.1. Coleta e Pré-processamento dos dados

Como passo inicial, foi realizado um tratamento e filtragem na base de dados, com o objetivo de se garantir a coerência e consistência, bem como manter uma padronização nos dados [Hand 2007]. Para tal fim, foram usadas as ferramentas de edição disponibilizadas pelo Microsoft Excel (versão 2016), resultando em um *dataset* com 231 alunos. Devido ao agravamento da pandemia de COVID-19, a adoção de ensino remoto, e o consequente atraso na conclusão do ano letivo, foram excluídos da análise alunos que ingressaram no ano de 2020. Foi construído um *ranking* por série (1ª, 2ª, 3ª e 4ª) das disciplinas da grade do curso de acordo com os índices de reprovação. Para esse cálculo, foi levado em conta somente acadêmicos cuja situação era cursando ou formado entre os anos de 2007 a 2019, isto é, excluindo desistentes ou transferidos para outra instituição, bem como alunos reprovados por não cumprirem o requisito mínimo da Universidade de presença as aulas, ou seja, 75% da carga horária total da disciplina. Para este estudo, decidiu-se pesquisar, inicialmente, o perfil de retenção para as duas disciplinas com maior taxa de retenção por série do curso, que conforme o *ranking* obtido são:

- 2ª série: Introdução à Arquitetura de Computadores (34,45%) e Cálculo Numérico (30,61%).
- 3ª série: Redes de Computadores (35,16%) e Engenharia de Software (17,95%).
- 4ª série: Inteligência Artificial (27,38%) e Engenharia de Software II (9,28%).

Também foram gerados os seguintes atributos derivados para cada disciplina: PssFsc_Sexo (sexo do estudante), Idade_Prdletivo (idade do estudante no respectivo ano letivo), Acd_MdFinal (média final do aluno na disciplina) e Acd_Resultado (resultado na disciplina).

3.2. Algoritmos e Ferramentas Utilizados

Para a fase de aplicação, foram pesquisados e definidos dois algoritmos de aprendizagem supervisionada amplamente usados: *Support Vector Machine* (SVM), que trata os dados de acordo com o centro esférico mais próximo e gera padrões [Aggarwal 2018], e *K-Nearest Neighbor* (KNN), baseado em distâncias, no qual dados similares se encontram no mesmo espaço [Igual & Seguí 2017]. Optou-se por esses algoritmos por sua adequação à tarefa de classificação em um *dataset* com dados completamente rotulados.

A linguagem de programação *Python* foi utilizada, juntamente com a biblioteca *Pandas* para a manipulação de dados, os pacotes *Numpy* e *Scipy*, que oferecem uma gama de funções matemáticas para manipular as estruturas, a biblioteca *Matplotlib*, que oferece suporte gráfico para representar e visualizar os resultados, e também a biblioteca *scikit-learn*, que possui os algoritmos implementados e suporte para classificação, geração de acurácias e melhores parâmetros como *GridSearchCV*. O ambiente escolhido foi o *Google Colaboratory*, um serviço gratuito de nuvem, composto por uma série de células que podem ser do tipo texto ou código executável. Todos os dados tabulados foram exportados em formato *CSV*, e transformados em numéricos para aplicação nos algoritmos. Foi gerado um dicionário descritivo dos dados em cada *notebook* ou arquivo do ambiente, separados e divididos por disciplina, com base no *ranking* de retenção. As variáveis sexo, idade do estudante no ano letivo analisado, média final na disciplina foram consideradas como sendo as entradas, e o resultado, indicado por aprovado ou reprovado, como sendo o *target*.

Para cada análise, separou-se 30% dos conjuntos de exemplos em função do resultado do aluno na disciplina, em um conjunto de treinamento e testes de maneira aleatória, segundo o esquema *Train Test Split*. Para o valor de K do algoritmo KNN, e suas métricas foram adotadas duas táticas: a) raiz do tamanho do conjunto de teste, e se o resultado for par, basta subtrair 1, usando a métrica euclidiana padrão; b) testes com valores aleatórios entre 1 e 50, bem como usando *GridSearch,* que fornece o erro gerado e as melhores métricas e valores, ajustando e otimizando os parâmetros. Para os parâmetros do SVM, as táticas adotadas foram: a) *Kernel* padrão definido como *"rbf"* e valor de C como 2.0; b) ajuste das métricas, *kernel* e valor de *gamma* com base em testes usando *GridSearch*. Para ambos os algoritmos foram realizadas duas variações de parâmetros, ou seja, métricas iniciais que consistiam em valores padrões para o algoritmo KNN (*n_neighbors=7, metric='euclidean'*) e SVM (C=2.0, *kernel='rbf'*), bem como as métricas melhoradas com base em testes de erro e variações através do *GridSearchCV*, automatizando o processo de ajustes do algoritmo, destacando que, para as métricas do SVM, os melhores valores de C e do *kernel*, se mostraram iguais para todas as disciplinas analisadas (C=0.1, *kernel='poly'*), ocorrendo variações somente no parâmetro *gamma*. Como método de avaliação e desempenho dos algoritmos, foi usada a pontuação de classificação de precisão (*accuracy_score*), que retorna a fração de amostras classificadas corretamente, sendo o melhor desempenho indicado por 1, também, foi gerado um relatório de classificação com as principais métricas de avaliação (*classification_report*). Além disso, utilizou-se a validação cruzada (*cross validation*) para avaliar o classificador gerado. Nessa abordagem, o conjunto de treino e teste foi dividido, e esse processo repetido, totalizando 10 modelos, gerando a média de todos os resultados, ou seja, é calculada a pontuação 10 vezes consecutivas com diferentes divisões cada vez, retornando a porcentagem média dessas pontuações. Essa estratégia objetiva garantir maior solidez, estabilidade e precisão ao modelo.

4. Resultados e Discussões

As Tabelas 1 e 2 apresentam os resultados de acurácia obtidos para os dois algoritmos selecionados, realizando-se ajustes nas métricas para cada algoritmo, visando gerar uma pontuação otimizada e melhorada. Verificou-se que as acurácias dos algoritmos foram maiores que 92% para os parâmetros e métricas iniciais, e maiores que 94% com métricas e parâmetros melhorados. Assim, os dois algoritmos se mostraram eficazes.

Tabela 1. Acurácias do algoritmo KNN

Disciplina	Score Inicial	Score Melhorado	Cross Validation Inicial (%)	Cross Validation Melhorado (%)
IAC	1	1	100	100
CN	1	1	99,49999	100
REDES	1	1	99,49999	99,49999
ES1	1	1	98,75	98,75
REQUISITO IES	0,96226	0,98113	98,88888	99,44444
IA	1	1	98,23529	98,23529
ES2	1	1	99,23076	100

Tabela 2. Acurácias do algoritmo SVM

Disciplina	Score Inicial	Score Melhorado	Cross Validation Inicial (%)	Cross Validation Melhorado (%)
IAC	1	1	98,09523	100
CN	0,95	1	98,00001	100
REDES	0,98275	1	96,94736	99,49999
ES1	0,97959	1	94,41176	100
REQUISITO IES	0,98113	1	95,52287	100
IA	0,98	1	96,47058	99,41176
ES2	0,92307	1	97,62820	100

Além disso, em relação à faixa etária, é notável que os maiores índices de reprovações ocorrem na faixa dos 19 aos 23 anos. Dentre esses resultados, a relação entre o aumento da retenção e a idade era esperada, uma vez que a faixa etária em questão é a mais expressiva na graduação. O número de indivíduos do sexo feminino matriculadas no curso, é consideravelmente menor, acredita-se que isso ocorra, conforme estudos de Wilson (2003), em decorrência das características estereotipadas do cientista da computação: sexo masculino, antissocial, obsessivo e fascinado com a máquina. A Tabela 3 sumariza os resultados obtidos em relação a variável sexo, considerando apenas uma reprovação por aluno e a amostra total de matriculados nas disciplinas. Apesar, do número de estudantes do sexo feminino ser expressivamente menor, ainda assim é registrada uma porcentagem alta de reprovações, principalmente nas séries iniciais do curso. Contudo, é importante salientar que ao longo das séries do curso, houve uma redução no número de acadêmicas matriculadas. Tal situação pode ser explicada, em parte, pelo estudo realizado por Lima (2013), na qual professores afirmaram que percebem que, para estudantes do sexo feminino, a aprendizagem em certas disciplinas é mais demorada, principalmente as que envolvem lógica e cálculos.

Tabela 3. Relação entre o sexo e quantidade de reprovações

	Masculino		Feminino	
Disciplina	Porcentagem de retenções (%)	Amostra Total	Porcentagem de retenções (%)	Amostra Total
IAC	34	124	40	25
CN	33	124	20	25
REDES	35	109	32	25
ES1	17	123	21	19
REQUISITO IES	13	131	7	27
IA	29	106	23	17
ES2	8	102	6	15

Além disso, quando analisado exclusivamente os grupos de requisitos de disciplinas, nota-se que apenas uma pequena parcela de alunos retidos na disciplina considerada pré-requisito, apresentou dificuldades e reprovou na disciplina do ano seguinte, o que é justificável, uma vez que uma disciplina é o desdobramento da anterior. É importante destacar, que os alunos retidos em uma disciplina analisada, conforme *ranking* pré-definido, acabaram retidos na outra disciplina eleita com o maior índice de reprovação da mesma série, e, quando comparado os grupos das duas disciplinas selecionadas por série, é notável que alunos retidos na série anterior, apresentaram reprovações, em sua grande maioria, em ambas as disciplinas eleitas da série seguinte, evidenciando um ciclo vicioso e sucessivo de retenções, como por exemplo, 5 alunos que reprovaram em Introdução à Arquitetura de Computadores e Cálculo Numérico (2ª série), foram retidos em Redes de Computadores (3ª série), e 8 alunos que reprovaram em Redes de Computadores e Engenharia de Software 1 (3ª série), foram retidos em Inteligência Artificial (4ª série).

Em relação a variável média das notas dos alunos nas disciplinas, e, realizando uma média final, temos:

- 2ª série: Introdução à Arquitetura de Computadores (23 pontos) e Cálculo Numérico (33 pontos).
- 3ª série: Redes de Computadores (33 pontos), Engenharia de Software 1 (40 pontos) e pré-requisito Introdução a Engenharia de Software (37 pontos).
- 4ª série: Inteligência Artificial (37 pontos) e Engenharia de Software 2 (40 pontos).

Fazendo uma análise individual do perfil de retenção de cada disciplina, observa-se que em Introdução à Arquitetura de Computadores, a moda das médias, ou seja, o valor que mais se repete é 0, com a maior porcentagem de reprovações entre o sexo feminino e na faixa etária dos 19 anos. Além disso, é importante destacar, que alunos repentes na disciplina, tiveram dificuldades e reprovações, em sua maioria, nas disciplinas da 1ª série, sendo Computação 1, Física Geral e Eletricidade Básica, Probabilidade e Estatística e, Lógica e Matemática Discreta. Em Cálculo Numérico, o perfil do aluno retido é do sexo masculino, entre 18 e 19 anos, e a moda das médias é 33 pontos, além disso, os alunos tiveram reprovações, especialmente nas disciplinas de Física Geral e Eletricidade Básica, Geometria Analítica e Álgebra Linear e Cálculo Diferencial e Integral, da 1ª série.

Entre as disciplinas da 3ª série, como Redes de Computadores, por uma diferença mínima, a quantidade de meninos retidos foi maior, principalmente na faixa dos 22 anos, e a frequência dentre as médias foi de 26 pontos, destacando também, que os alunos apresentaram reprovações nas disciplinas de Sistemas Digitais, Conceitos de Linguagens de Programação e Cálculo Numérico, da 2ª série do curso. Já em Engenharia de Software 1, o sexo feminino registrou maior porcentagem de reprovação, entre 19 e 20 anos, com uma frequência de média em 54 pontos, e marcado por reprovações, em sua maioria, nas disciplinas de Introdução à Arquitetura de Computadores, Conceitos de Linguagens de Programação e Algoritmos e Estrutura de Dados, da 2ª série. Em relação a 4ª e última série da graduação, temos Inteligência Artificial, com um maior registro de retenções no sexo masculino, na faixa dos 23 anos de idade e com a moda das médias em 31 pontos, com reprovações, nas disciplinas da 3ª série, de Redes de Computadores em sua maioria, Sistemas Operacionais e Engenharia de Software 1. E, por fim, Engenharia de Software 2, registrou os maiores índices de reprovação no sexo masculino, na faixa dos 21 anos, com a moda das médias em 28 pontos, e reprovações principalmente nas disciplinas de Redes de Computadores, Projeto e Análise de Algoritmos e Sistemas Operacionais, da 3ª série do curso.

A precisão na classificação manteve-se sempre alta, indicando assim, que a tarefa de prever a retenção com base nos dados obtidos é efetiva e revelando que dados como como sexo, idade, média em função do resultado, indicado por aprovado e reprovado, serem de grande influência no processo de predição, o que permite identificar perfis específicos de alunos com dificuldades e direcionar ações pedagógicas a fim de se reverter quadros de retenções.

5. Considerações Finais e Trabalhos Futuros

Neste artigo, foi apresentado os resultados da aplicação de técnicas de Aprendizado de Máquina na predição de reprovação de estudantes em disciplinas da 2ª à 4ª série no curso de Ciência da Computação da Unioeste, campus de Foz do Iguaçu, com base no desempenho acadêmico em disciplinas de anos anteriores, e, considerando pré-requisitos. Através da aplicação dos algoritmos KNN e SVM, obteve-se uma precisão superior a 92%, utilizando como atributos o sexo, idade em que o estudante cursou a disciplina, média final do acadêmico e resultado indicado por aprovado e reprovado, além de se verificar fatores determinantes associados a retenção, representados pela faixa etária dos 19 aos 23 anos, sexo e notas finais.

Os resultados apresentados mostram a viabilidade de se prever o fenômeno da retenção. Evidentemente pode-se cogitar de outras variáveis, distintas das analisadas neste trabalho, serem capazes de influenciar o desempenho do aluno, tais como a motivação pessoal, porém mais difíceis de serem medidas precisamente. A possibilidade de prever o perfil de retenção do estudante é bastante útil para o ambiente universitário, corpo docente e discente, que podem ajustar os meios pedagógicos, através de estratégias e planejamentos, visando diminuir o número de reprovações, reduzindo assim consequentemente a evasão dos alunos. É fundamental implementar políticas que levem em conta a dimensão desse problema para a sociedade, buscando um acompanhamento presente aos estudantes durante a graduação, identificando e investigando os motivos de retenção, para melhor auxiliá-los, visando o sucesso desses, e eficiência da instituição de Ensino. Como trabalhos futuros, pode-se citar a expansão para outras disciplinas com taxas menores de retenção ou outros cursos de graduação, e a inclusão de variáveis

socioeconômicas e desempenho dos alunos durante o ensino médio visando prever a retenção em disciplinas a partir da 1ª série.

Referências

Andriola, W. B. & Araújo, A. C. (2018). Uso de indicadores para diagnóstico situacional de Instituições de Ensino Superior. *Ensaio: Avaliação e Políticas Públicas em Educação*, Rio de Janeiro, v. 26, n. 100, p. 645-663. https://doi.org/10.1590/s0104-40362018002601062

Aggarwal, C. C. (2018). Neural Networks and Deep Learning (1st ed.). *Springer Nature*. https://doi.org/10.1007/978-3-319-94463-0

Campello, A. V. C. & Lins, L. N. (2008) Metodologia de análise e tratamento da evasão e retenção em cursos de graduação de instituições federais de ensino superior. *In*: ENCONTRO NACIONAL DE ENGENHARIA DE PRODUÇÃO, 28., 2008, Rio de Janeiro. *Anais* [...] Rio de Janeiro: ABEPRO, p. 1-13. Disponível em:http://www.abepro.org.br/biblioteca/enegep2008_TN_STO_078_545_11614.pdf

de Brito, D., Júnior, I., Queiroga, E., & do Rêgo, T. (2014). Predição de desempenho de alunos do primeiro período baseado nas notas de ingresso utilizando métodos de aprendizagem de máquina. *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação - SBIE), 25*(1), 882. doi:http://dx.doi.org/10.5753/cbie.sbie.2014.882

de França, R. S. & do Amaral, H. J. C. (2013). Mineração de Dados na Identificação de Grupos de Estudantes com Dificuldade de Aprendizagem no Ensino da Programação. RENOTE, v. 11, n. 1.

Donoso-Díaz, S., Iturrieta, T. N. & Traverso, G. D. (2018). Sistemas de Alerta Temprana para estudiantes en riesgo de abandono de la Educación Superior. *Ensaio: Avaliação e Políticas Públicas em Educação*, Rio de Janeiro, v. 26, n. 100, p. 944-967.https://doi.org/10.1590/s0104-40362018002601494

Faúndez, F., Muñoz, K. & Cornejo, F. (2012). *Percepción sobre el modelo educativo basado en competencias y su contribuición a la retención de estudiantes de la universidad de Talca*. In: CONFERÊNCIA LATINOAMERICANA SOBRE EL ABANDONO EN LA EDUCACIÓN SUPERIOR, 2., 2012, Porto Alegre. Anais... Porto Alegre: PUC-RS, 2012. p. 738-744.

Hand, D.J. (2007). Principles of Data Mining. *Drug-Safety* 30, 621–622. https://doi.org/10.2165/00002018-200730070-00010.

Igual, L., & Seguí, S. (2017). Introduction to Data Science A Python Approach to Concepts, Techniques and Applications (1st ed.).. *Springer Nature*. https://doi.org/10.1007/978-3-319-50017-1

Lamers, J. M. S. et al. (2017). Retenção e evasão no ensino superior público: estudo de caso em um curso noturno de odontologia. *Educação em Revista*, Belo Horizonte, v. 33, e154730. https://doi.org/10.1590/0102-4698154730

Lima Junior, P. et al. (2019). Taxas longitudinais de retenção e evasão: uma metodologia para estudo da trajetória dos estudantes na Educação Superior. *Ensaio: Avaliação e*

Políticas Públicas em Educação, Rio de Janeiro, v. 27, n. 102, p. 157-178. https://doi.org/10.1590/s0104-40362018002701431

Lima, Michelle Pinto. (2013). As mulheres na Ciência da Computação. Revista Estudos Feministas [online], v. 21, n. 3, pp. 793-816. Epub 28 Jan 2014. ISSN 1806-9584. https://doi.org/10.1590/S0104-026X2013000300003

Mainier, F. B. et al. (2006) A contribuição da disciplina de introdução à engenharia química no diagnóstico da evasão. *Ensaio: Avaliação e Políticas Públicas em Educação*, Rio de Janeiro, v. 14, n. 51, p. 261-277. https://doi.org/10.1590/S0104-40362006000200008

Manhães, L. M. B. (2015). Predição do Desempenho Acadêmico de Graduandos Utilizando Mineração de Dados Educacionais (Publication No. XVII) [Doctoral dissertation, Universidade Federal do Rio de Janeiro]. https://www.cos.ufrj.br/uploadfile/1426690008.pdf

Romero, C., & Ventura, S. (2010). Educational Data Mining: A Review of the State of the Art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C, 40*(6), 601-618.

Saccaro, A., França, M. T. A., & Jacinto, P. d. A. (2019). Fatores Associados à Evasão no Ensino Superior Brasileiro: um estudo de análise de sobrevivência para os cursos das áreas de Ciência, Matemática e Computação e de Engenharia, Produção e Construção em instituições públicas e privadas., 49(2), 337-373. https://doi.org/10.1590/0101-41614925amp

Silva, G. S. (2017) *Retenção e evasão no ensino superior no contexto da expansão: o caso do curso de engenharia de alimentos da UFPB*. Dissertação (Mestrado em Políticas Públicas, Gestão e Avaliação da Educação Superior) - Centro de Educação, Universidade Federal da Paraíba, João Pessoa.

Vanz, S. A. S. et al. (2016). Evasão e retenção no curso de Biblioteconomia da UFRGS. *Avaliação: Revista de Avaliação da Educação Superior (Campinas)*, Sorocaba, v. 21, n. 2, p. 541-568. https://doi.org/10.1590/S1414-40772016000200012

Wilson, Fiona. (2003). "Can compute, won't compute: women's participation in the culture of computing." *New Technology, Work and Employment*, v. 18, n. 2, p. 127-142.