Revista ARACÊ
ISSN: 2358-2472

USO DE ALGORITMOS DE MACHINE LEARNING PARA GESTÃO UNIVERSITÁRIA: REVISÃO SISTEMÁTICA

https://doi.org/10.56238/arev6n4-071

Data de submissão: 05/11/2024 **Data de publicação:** 05/12/2024

Aline Pacheco Primão
Universidade Federal de Santa Catarina, Brasil, aline.pacheco.pr@gmail.com

Alexandre Marino Costa
Universidade Federal de Santa Catarina, marinocad@gmail.com

Diego Rossa
Brasil
diegorossa67@gmail.com

Leonardo Flach
Universidade Federal de Santa Catarina, Brasil, leonardo.flach@gmail.com

RESUMO
Esta pesquisa tem por objetivo realizar uma revisão sistemática sobre a aplicação de Machine Learning (ML) na gestão universitária. O estudo permite analisar pesquisas anteriores sobre o uso de algoritmos de ML mais utilizados e destacados como os melhores, de modo a uma discussão e agenda. As buscas foram feitas nas bases científicas internacionais Scopus e Web of Science, a partir das palavras-chave Machine Learning, University, Higher Education. Foram selecionados para a amostra 32 artigos para a análise. Os resultados apontam que a maioria das pesquisas utilizam mais de um algoritmo de ML para realizar as previsões e que o Support Vector Machine (SVM) é o algoritmo destacado na maior parte das pesquisas como o de melhor desempenho. Outra conclusão identificada é que boa parte dos artigos avaliam o risco de evasão escolar, desempenho acadêmico do aluno, prever o resultado dos alunos e analisar a evasão a fim de evitar desistência, abandono ou atrito de cursos pelo aluno.

Palavras-chave: Gestão Universitária; Machine Learning; Algoritmos.

Revista ARACÊ
ISSN: 2358-2472

1 INTRODUÇÃO

Entre 2009 e 2019 e o Brasil registrou um aumento de 43,7% nas matrículas no ensino superior, e quando falamos em instituições federais este crescimento foi de 59,1% (INEP, 2019). Com isso, a gestão destas organizações é de grande relevância para alcançar resultados positivos nas tomadas de decisão, na busca do gasto público correto e na excelência universitária.

A Gestão Universitária (GU) pode ser entendida como sendo o ato de administrar a universidade. Segundo Schlickmann (2013) a gestão universitária está em construção, pois ela ainda está associada com a administração e outros campos como a educação, com isso, existe uma difusão de conhecimento e corpos teóricos (Schlickmann, 2013). A Gestão Universitária é considerada complexa, pois estas instituições possuem processos burocráticos predominantes e passam por frequentes mudanças por influências políticas e de grupos de interesse (Gomes et al.,2013).

Para colaborar com a GU, surge a Mineração de Dados Educacionais (MDE), que apesar de ser um campo emergente, vem ganhando atenção nos últimos anos, por poder ser utilizada para gerar informações que ajudem na tomada de decisão do processo educacional (Sultana et al., 2017). No entanto, as instituições de ensino possuem um volume muito grande de dados acadêmicos, e para produzir as buscas e análises destes dados é necessária uma investigação baseada na extração do conhecimento, para isso, os algoritmos de Machine Learning (ML), aprendizado de máquina, são opções praticáveis (Silva et al., 2020).

Os estudos com Machine Learning utilizam-se de modelos estatísticos, e têm sido usados para previsão de risco de algum evento ocorrer (Silva et al., 2020). Visto que, a partir da revisão bibliográfica, observa-se que um algoritmo de ML não conseguem obter uma boa performance em todos os tipos de aplicações, e que muitas vezes, é necessário congregar mais de um algoritmo para o ganho de desempenho, faz-se necessário analisar os algoritmos mais utilizados e verificar os que melhores se encaixam na pesquisa. Delen (2010) e Adekitan e Salau (2019) afirmam que um conjunto de algoritmos apresentam um desempenho melhor do que modelos individuais (Delen, 2010); (Adekitan & Salau, 2019).

A partir da introdução anterior, esta pesquisa bibliométrica justifica-se: 1º) Avaliar a produção das pesquisas em Machine Learning e o que elas trazem de ganho para a Gestão Universitária; 2º) Avaliar as áreas que mais são utilizadas as técnicas de MDE; 3º) Analisar os algoritmos de ML para as pesquisas relacionadas e os que trazem melhores resultados para previsão dentro da GU.

O artigo está apresentado da seguinte forma: 1) Introdução, onde foi fornecido uma breve apresentação do tema pesquisado e justificativa desta pesquisa; 2) Métodos, onde é apresentado o objetivo da pesquisa, como a revisão foi realizada, bases de buscas e critérios de seleção e exclusão

Revista ARACÊ
ISSN: 2358-2472

dos artigos; 3) Resultados, onde é exposta as características, síntese e gráfico dos resultados dos artigos; e 4) Discussão, onde é evidenciado os resultados relevantes, limitações, interpretações gerais e implicações futuras.

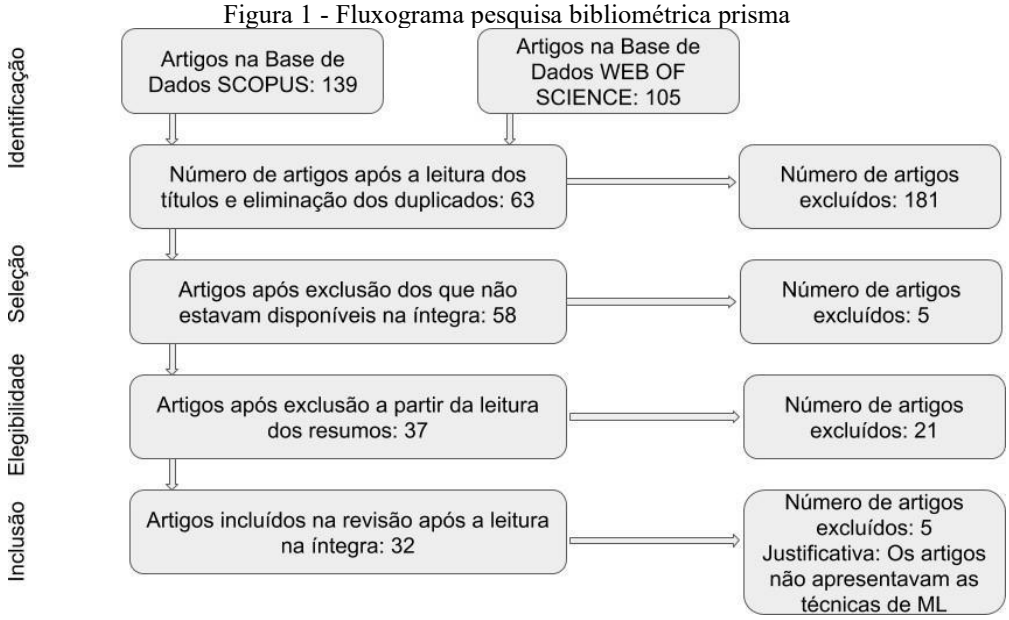
2 MÉTODO DE PESQUISA

Com o objetivo de analisar pesquisas anteriores sobre o tema proposto, ou temas relacionados e fazer uma discussão da literatura foi realizada a pesquisa bibliométrica a partir das seguintes palavras-chave: Machine Learning AND University AND Higher Education nas bases de dados Scopus e Web of Science a partir do ano de 2010. A busca foi realizada em dezembro de 2020, em todas as línguas.

Foram encontrados 244 artigos, retirados os repetidos e os que não se enquadraram na pesquisa, após ler os títulos restaram 63 artigos. Logo após, foram retirados os que não estavam disponíveis na íntegra, permanecendo 58 artigos. Na sequência com a leitura dos resumos, foram excluídos os que não eram apresentados como foco na gestão universitária, ficando 37 artigos. E por fim, após a leitura na íntegra, foram excluídos 5 por não apresentarem os algoritmos de Machine Learning, restando assim um total de 32 artigos.

Esta pesquisa utilizou a recomendação Prisma de Moher et al. (2009) e apresentou o fluxograma da Figura 1.

Figura 1 - Fluxograma pesquisa bibliométrica prisma



Fonte: Elaboração própria.

Na seção 3 serão abordados os resultados da pesquisa bibliométrica, trazendo as características dos artigos selecionados, as sínteses e gráficos.

Revista ARACÊ
ISSN: 2358-2472

3 DISCUSSÃO E ANÁLISE DE RESULTADOS

Com a pesquisa observou-se que os anos de 2019 e 2020 foram os mais evidentes, fornecendo 26 dos 32 artigos. Também se verificou que 29 destes eram na língua inglesa, 2 em espanhol e 1 em português.

O Continente asiático é o mais significativo nas publicações com 21 representações, sendo que a China é retratada em 5 artigos e Taiwan em 4, seguidos da Europa com 12 e América com 9 representações. Observa-se que um artigo pode ser representado por mais de um país ou continente. Apesar de apenas uma das pesquisas estar na língua portuguesa, o Brasil está representado em 3 artigos. Destaca-se Costa et al. (2017), em que na data da busca nas bases de dados, possuía 113 citações. A Figura 2 representa a disposição dos artigos por continente.

Figura 2 - Divisão dos artigos por continente

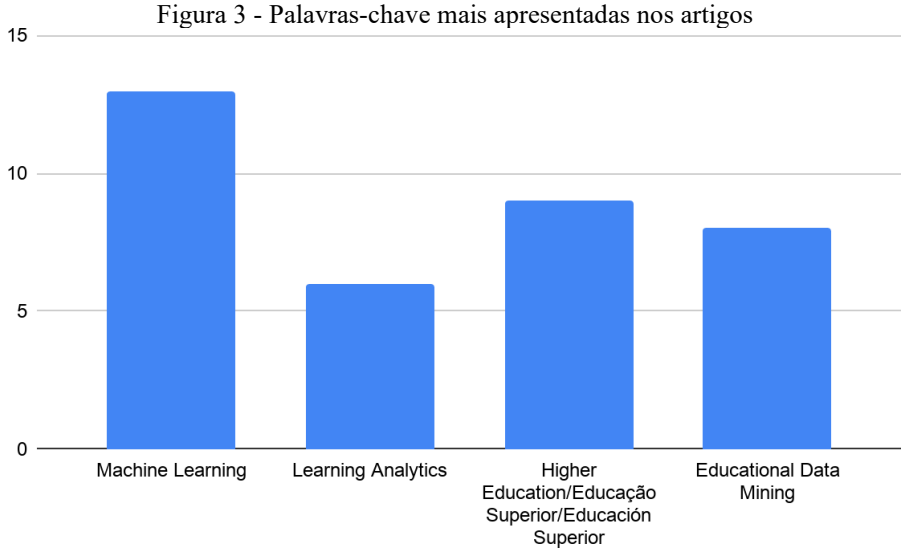
Continente	Quantidade
Ásia	21
Europa	12
América	9
África	4
Oceania	1

Fonte: Elaboração própria (2021).

As palavras-chave mais utilizadas foram Machine Learning com 13 citações, Higher Education/Educação Superior/Educación Superior com 9 citações, Educational Data Mining com 8 citações e Learning Analytics com 6 citações. A Figura 3 representa o número de vezes que a palavra-chave foi citada em um artigo.

Revista ARACÊ
ISSN: 2358-2472

Figura 3 - Palavras-chave mais apresentadas nos artigos



Fonte: Elaboração própria (2021).

No tocante aos objetivos dos artigos, 6 deles tem como objetivo principal prever resultados acadêmicos dos alunos, 6 analisar a evasão para evitar desistência, abandono ou atrito, 5 prever alunos em risco de reprovação, 2 prever notas dos alunos, 2 analisar a procrastinação de alunos, 2 prever o desempenho acadêmico, 1 revisar textos em pesquisas de opiniões de alunos, 1 prever número de alunos com baixo engajamento nos cursos, 1 propor sistema para prever caminho para alunos no ano preparatório da faculdade, 1 explorar os comportamentos de aprendizado gerados pelos alunos, 1 propor um modelo preditivo com taxas de precisão aprimoradas a partir da análise de dados de alunos, 1 estudar que tipos de modelos de previsão tem melhor desempenho e 1 promover o uso intuitivo de técnicas de análise de evasão. A Tabela 1 mostra os artigos por objetivos similares.

Revista ARACÊ
ISSN: 2358-2472

Tabela 1 - Artigos por objetivos similares

OBJETIVOS	QUANTIDADE
Analisar a procrastinação de alunos	2
Prever notas dos alunos	2
Prever alunos em risco de reprovação	5
Prever resultados acadêmicos dos alunos	6
Revisar textos em pesquisas de opiniões de alunos	1
Avaliar dados técnicos de alunos	1
Conhecer número efetivo de alunos numa plataforma	1
Analisar a evasão para evitar desistência, abandono ou atrito	6
Prever número de alunos com baixo engajamento nos cursos	1
Propor sistema para prever caminho para alunos no ano preparatório da faculdade	1
Explorar os comportamentos de aprendizado gerados por alunos	1
Propor um modelo preditivo com taxas de precisão aprimoradas a partir da análise de dados de alunos	1
Prever o desempenho acadêmico	2
Estudar que tipos de modelo de previsão tem melhor desempenho	1
Promover o uso intuitivo de técnicas de análise de evasão	1

Fonte: Elaboração própria.

Com as pesquisas relacionadas, verificamos que uma das maiores preocupações da Gestão Universitária é a evasão escolar, principalmente quando tratado de instituições públicas, pois estas precisam garantir resultados relevantes e afirmar a diplomação de estudantes para o mercado de trabalho (Andifes, 1996). Do mesmo modo, conseguir prever o desempenho escolar, permitindo que as instituições desenvolvam ações antecipadas para ajudar alunos a melhorarem suas notas e, consequentemente, formar profissionais mais capacitados e diminuir a evasão escolar é muito importante para as instituições de ensino superior.

Para Delen (2010), as altas taxas de evasão escolar afetam o planejamento de matrículas e trazem uma sobrecarga de trabalho para recrutar novas alunos, já para os alunos, desistir antes de obter um diploma representa que o potencial humano não foi explorado e também causando um baixo retorno dos investimentos da instituição (Delen, 2010). Para Lau (2003 apud Delen, 2010) uma das prováveis

Revista ARACÊ
ISSN: 2358-2472

causas do abandono escolar pode ser a dificuldade de adaptação à universidade, afetando a classificação, a reputação e o financeiro da instituição (Lau, 2003 apud Delen, 2010).

Quanto à temporalidade dos dados utilizados para as pesquisas percebe-se uma grande distinção. Delen (2010), por exemplo, utilizou dados de 5 anos com os calouros de uma universidade para fazer o gerenciamento de retenção de alunos, já Sultana et al. (2017) analisou os alunos do primeiro ano de faculdade de engenharia elétrica e ciência da Computação para prever o desempenho. Ezz e Elshenawy (2019) analisaram o comportamento dos alunos no curso preparatório para a faculdade, Hai-Tao et al. (2020) desenvolveu um modelo para prever o desempenho dos alunos antes mesmo do início do curso, Deo et al. (2020) utilizou dados de 6 anos em múltiplos cursos para investigar e propor um modelo de previsão de desempenho, Adekitan e Salau (2019) utilizaram dados para analisar o desempenho nos primeiros 3 anos da graduação para conseguir prever o resultado final dos alunos, já Costa et al. (2017) utilizou 2 tipos de dados de alunos que fizeram cursos de programação introdutória de 10 e 16 semanas para prever a falha precoce dos alunos.

3.1 ALGORITMOS DE MACHINE LEARNING

Dos algoritmos mais utilizados nas pesquisas, 18 artigos utilizaram Support Vector Machine (SVM), 16 Random Forest (RF), 16 Decision Tree (DT), 15 Neural Network (NN), 10 Logistic Regression (LR), 6 K-Nearest Neighbor (KNN) e 5 Multi-Layer Perceptron (MLP). A Tabela 2 apresenta a quantidade de vezes que os algoritmos de Machine Learning foram utilizados nas pesquisas, onde a maioria das pesquisas utilizou mais de um algoritmo.

Tabela 2 - Algoritmos de machine learning nos artigos

ALGORÍTMO	QUANTIDADE
SVM - Support Vector Machine	18
RF - Random Forest	16
DT - Decision Tree	16
LR - Logistic Regression	10
ANN - Artificial Neural Network ou NN - Neural Network ou DNN - Deep Neural Network ou PNN - Probabilistic Neural Network	15
KNN - K-Nearest Neighbor	6
MLP - Multilayer Perceptron	5

Fonte: Elaboração própria.

Revista ARACÊ
ISSN: 2358-2472

Delen (2010) utilizou Artificial Neural Network, Decision Tree, Support Vector Machine e Logistic Regression para realizar a previsão, sendo que SVM foi o algoritmo que teve o melhor desempenho. O autor afirma que dados balanceados trouxeram melhores desempenhos do que os não balanceados, independentes do algoritmo utilizado. Outra observação de Delen (2010) é que os algoritmos de Decision Tree fornecem uma visão mais transparente de onde e como fazem comparando com Support Vector Machine (Delen, 2010).

Adekitan e Salau (2019) aplicaram seis algoritmos (Probabilistic Neural Network, Random Forest, Decision Tree, Naive Bayes, Tree Ensemble e Regressão Logística) de forma independentes para prever a Média Cumulativa de Notas Final dos alunos com dados dos três primeiros anos de graduação, a abordagem que apresentou o melhor desempenho foi Regressão Logística (Adekitan & Salau, 2019). Por fim, os autores combinaram todos os algoritmos em um modelo para obter os benefícios de cada um juntos. Os autores utilizaram as notas do ensino médio, nível de participação das aulas, assiduidade, notas intermediárias, relatórios de laboratórios, notas de tarefas de casa, pontuação de seminários, conclusão de tarefas e notas gerais para prever a evasão escolar no ensino superior (Adekitan & Salau, 2019).

Silva et al. (2020) empregou Regressão Logística, K-Nearest Neighbors, Naive Bayes, Support Vector Machines, Decision Tree Based Methods (C trees, Baggins, Random Forest, Boosting) e Penalized Methods (Ridge, Lasso, Elastic Net) para avaliar as melhores variáveis para a performance de modelos.

Costa et al. (2017) selecionou os algoritmos Naive Bayes, Decision Tree, Artificial Neural Network e Support Vector Machine para seu estudo, pois segundo o autor estas técnicas apresentam boa eficácia em diferentes ambientes e têm sido usados para identificar alunos com falhas acadêmicas (Costa et al., 2017). Os resultados obtidos com os algoritmos foram parecidos, porém o Support Vector Machine foi o que atingiu melhor eficácia na pesquisa (Costa et al., 2017).

Gamie et. al. (2020) usou algoritmos de ML em conjunto para analisar os fatores que impactam o desempenho dos alunos no último ano de curso e propor um modelo preditivo com taxas de precisão aprimoradas em comparação com outros no mesmo conjunto de dados. Os autores também avaliaram as características mais significativas que podem afetar no desempenho dos alunos. O modelo de Gamie et al. (2020) seguiu as seguintes etapas: inicializar grupos de recursos com base em fontes de dados; gerar combinações possíveis de grupos de repetição a fim de detectar a melhor combinação; Aplicar Support Vector Machine, Decision Tree e Neural Networks em cada partição (combinação de características); Aplicar Random Forest como técnica de ensacamento em cada partição (combinação de recursos); Aplicar XgBoost e AdaBoost com Decision Tree como aprendiz base de cada partição

Revista ARACÊ
ISSN: 2358-2472

(combinação de recursos); selecionar o melhor classificador junto com a melhor combinação de grupos; Aumentar o Support Vector Machine linear e não-linear e o Random Forest e salvar os resultados e; comparar a precisão da classificação (Gamie et al., 2020). Nesta pesquisa, concluiu-se que dados demográficos não afetam a precisão dos resultados finais, portanto, foram excluídos para não gerar esforço computacional desnecessário (Gamie et. al., 2020).

Poucos autores das pesquisas selecionadas usaram apenas um algoritmo para o estudo, é o caso de Beaulac e Rosenthal (2019), que utilizaram Random Forest para prever se o aluno concluirá seu curso de graduação e para prever quais cursos atraem mais os alunos. Os autores reiteram que RF são fáceis de usar, rápidos para treinar e superam os modelos de Regressão Linear em precisão de previsões. Beaulac e Rosenthal (2019) ajustaram os classificadores de RF e os compararam com dois modelos de regressão logística para prever se um aluno conclui seu curso ou não (Beaulac & Rosenthal, 2019).

Outro autor que se aproveitou de apenas um algoritmo foi Chui et al. (2020), que propôs um modelo reduzido de avaliação com o Support Vector Machine para prever alunos em riscos e possíveis alunos a abandonarem seus cursos. Após, o autor realizou uma comparação com pesquisas relacionadas, assim, ele concluiu que seu modelo pode ser adotado, pois reduz o tempo de treinamento quando o conjunto de dados fornecido é grande (Chui et al., 2020).

Nikolic et al. (2020) apresentou um sistema baseado em Natural Language Processing (NLP) para avaliar textos livres expressos por alunos em pesquisa de opiniões na língua Sérvia e fazer com que, uma instituição superior consiga verificar quais aspectos os alunos estão satisfeitos ou insatisfeitos. O autor também utilizou os algoritmos K-Nearest Neighbors (KNN), Naive Bayes e Support Vector Machine para realizar o aprendizado de máquina e o que obteve melhor avaliação foi o Support Vector Machine (Nikolic et al., 2020).

Support Vector Machine, Naive Bayes, Regressão Logística, JRip, J48, Multilayer Perceptron e Random Forest foi utilizado por Mia et al. (2019) para prever o número de alunos registrados em um semestre para ajudar no pré-planejamento de uma universidade privada de Bangladesh. Mia et al. (2019) comparou os classificadores e o Support Vector Machine obteve melhor precisão, já Random Forest obteve menor precisão. O autor pretende considerar mais atributos e aumentar o número de universidades avaliadas no futuro (Mia et al., 2019).

Suharjito (2019) explorou os algoritmos K-Nearest Neighbor, Naive Bayes e Decision Tree para analisar e encontrar uma melhor solução de modelagem na identificação de preditores de abandono de alunos em uma universidade de Jacarta. Ao combinar os algoritmos com método

Revista ARACÊ
ISSN: 2358-2472

Ensemble Classifier e testado várias vezes a precisão teve melhor desempenho. Dentre os algoritmos o que obteve melhor resultado foi o K-Nearest Neighbor (Suharjito, 2019).

3.2 COLETA DE DADOS

Já na questão referente à coleta de dados, a maioria das pesquisas utilizou somente a coleta nos bancos de dados dos sistemas acadêmicos com representação de 26 artigos, as pesquisas que utilizaram coleta de dados dos sistemas acadêmicos juntamente com um ou mais questionários somam 4 artigos, e as pesquisas que levaram em conta apenas questionários somam 2 artigos.

Vidhya e Vadivu (2020) utilizaram um conjunto de questionários abrangendo todos os aspectos dos fatores de aprendizagem dos alunos (dados pessoais dos alunos, padrão de aprendizagem, comportamento, fatores emocionais, inteligência múltipla e habilidades cognitivas). A partir dos resultados obtidos nos questionários os autores aplicaram os algoritmos de aprendizado de máquina e assim, conseguiram classificar os alunos nas categorias “Excelente”, “Bom”, “Médio” e “Ruim”. Com os resultados, é possível tomar medidas para melhorar o resultado dos alunos e também a reputação da instituição (Vidhya & Vadivu, 2020).

Muñiz et al. (2019) utilizou numa primeira etapa os dados extraídos do sistema institucional onde ele extraiu: dados de identificação, sexo, local de nascimento, nacionalidade, deficiência, tamanho da família, qualificação dos pais e ocupações atuais, nota média do ensino médio, pontuação do exame de admissão universitária, idade quando admitido, data da primeira matrícula, prioridades indicadas no aplicativo de admissão do curso, área do conhecimento correspondente ao curso do aluno, número de créditos inscritos, créditos passados e pontuação média, bolsa, situação acadêmica atual e destino de transferência, quando houver. Numa segunda etapa foi realizado um questionário relacionado a estado civil, nível de renda, tipo de moradia durante o curso, motivação para escolha do curso e universidade, participação em atividades de boas-vindas para calouros e opinião deles, tempo gasto com estudo, trabalho e trabalhos domésticos, avaliação dos requisitos do programa de satisfação com pontuações, avaliação das relações pessoais, intenção de abandono e razões, satisfação com a universidade e, se caso o aluno desistiu, a situação atual e satisfação com os resultados de sua decisão (Miñiz et al., 2019).

Delen (2010) empregou variáveis relacionadas ao nível acadêmico, características financeiras e demográficas dos alunos (Delen, 2010). Já Adekitan e Salau (2019) utilizaram as notas do ensino médio, nível de participação das aulas, assiduidade, notas intermediárias, relatórios de laboratórios, notas de tarefas de casa, pontuação de seminários, conclusão de tarefas e notas gerais para prever a evasão escolar no ensino superior (Adekitan & Salau, 2019).

Revista ARACÊ
ISSN: 2358-2472

Silva et al. (2020) usufruiu de dados disponibilizados pela Plataforma Lattes do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) para analisar os dados dos docentes como tempo de graduação, publicações no ano, doutorado e dedicação exclusiva, juntamente, com os dados institucionais, onde foram retirados os dados do aluno como nota no vestibular e nota de matemática no vestibular. Silva et al. (2020) afirma que as disciplinas de cálculo estão relacionadas com a deficiência do conhecimento geral vinda da educação básica, por isso o uso das notas de matemática podem influenciar de forma direta no desempenho do aluno no ensino superior (Silva et al., 2020).

Segundo Sultana et al. (2017), quando comparado entre alguns algoritmos de ML (Regressão Logística, Decision Tree, Naive Bayes e Neural Networks), Neural Networks obteve melhor desempenho na previsão de desempenho dos alunos, porém utilizando apenas características cognitivas, por isso, os autores avaliaram características cognitivas e não cognitivas para a previsão de abandono escolar. Para Sultana et al. (2017), características não cognitivas podem ajudar a aumentar a precisão da predição de abandono. Os autores coletaram dados através de questionários de autoavaliação, questionário de apoio social, questionário de habilidades e liderança e questionário de autoavaliação estudantil para analisar as características não cognitivas dos estudantes do primeiro ano da faculdade de engenharia elétrica e ciência da computação. Os dados coletados dos questionários foram pré-processados juntamente com fontes secundárias e utilizados os algoritmos de Machine Learning (Decision Tree, Regressão Logística, Naive bayes e Neural Networks) para classificação dos dados. Sultana et al. (2017) verificou que algumas características não cognitivas ajudam na previsão e precisão e algumas outras características não, dependendo do curso analisado, e que quando características não cognitivas são acopladas com características cognitivas resultam em previsões e precisões melhores. O objetivo da pesquisa foi utilizar os algoritmos de ML para gerar dados e conseguir informar aos alunos com desempenho ruim a tempo de melhorarem seus desempenhos e reduzir assim o abandono escolar (Sultana et al., 2017).

Gamie et al. (2019) fez a análise de um conjunto de dimensões e cada dimensão incluiu um conjunto de variáveis. A análise de previsão é aplicada para cada dimensão, e por fim, é realizada uma comparação entre as dimensões. Três fatores foram a base das variáveis: atividades dos alunos, estilo de ensino e a categorização de conteúdo.

Suharjito (2019) utilizou dados demográficos e de desempenho acadêmico (média cumulativa de notas, avaliação interna, avaliação externa, atividades extracurriculares, histórico do ensino médio e atividades sociais para prever o aluno com risco de abandono. Segundo Suharjito (2019), nos primeiros dois anos o gênero também influencia na qualidade do aprendizado, assim como características como

Revista ARACÊ
ISSN: 2358-2472

idade, restrições financeiras, ausência do aluno, influência dos pais, oportunidade de emprego e estado civil (Suharjito, 2019).

4 CONSIDERAÇÕES FINAIS

O artigo teve como objetivo fazer uma revisão sistemática com o intuito de avaliar pesquisas anteriores que utilizaram técnicas de Machine Learning na gestão universitária.

A pesquisa aferiu que dentro da Gestão Universitária a maior parte dos artigos representam um envolvimento com risco de evasão escolar, desempenho acadêmico do aluno, conseguir prever os resultados dos alunos e analisar a evasão a fim de evitar desistência, abandono ou atrito de cursos pelo aluno. Estes resultados mostram que a preocupação das instituições de ensino superior com a evasão escolar é grande em todos os continentes.

Também se observou que os algoritmos mais utilizados nas pesquisas foram Support Vector Machine, Decision Tree, Random Forest, Neural Networks e Regressão Logística (Logistic Regression). O algoritmo que se destaca é Support Vector Machine, apresentando melhor precisão em boa parte das pesquisas. Já o Decision Tree foi considerado o mais transparente. Outro que trouxe bons resultados foi Random Forest, que pode ser fácil de usar e treinar. Apesar de Naive Bayes não estar entre os algoritmos mais utilizados nas pesquisas, ele obteve os melhores resultados no trabalho de Gamie et al. (2019).

Random Forest, Support Vector Machine, Naive Bayes, Decision Tree e XGBoost foram usados individualmente em algumas pesquisas. Apesar de alguns autores preferirem aplicar apenas um algoritmo, no geral, os estudos mostram que as utilizações de um conjunto de algoritmos selecionados trazem uma melhor previsão dos resultados.

O Brasil está representado nas pesquisas com 3 artigos, destacando o artigo de Costa et al. (2017) o qual tem um grande número de citações (113). Porém, os artigos mais aceitos limitam- se à língua inglesa.

Por fim, este artigo ajudou a analisar os algoritmos mais utilizados e os que obtêm melhor desempenho nas pesquisas com Machine Learning, os objetivos das pesquisas e averiguar, além de outros países, o que o Brasil tem desenvolvido no tema.

Pretende-se aprofundar o estudo nos algoritmos de ML mais utilizados nas pesquisas avaliadas e analisar como eles se desempenham na evasão escolar no ensino superior brasileiro, analisar os fatores que mais impactam na evasão escolar as instituições federais brasileiras, e a partir disso, utilizar os algoritmos que mostram eficácia nas pesquisas relacionadas para o treinamento e testes da base de

Revista ARACÊ
ISSN: 2358-2472

dados de uma instituição federal do Brasil e propor um modelo de previsão de evasão escolar para as IFES brasileiras.

Revista ARACÊ
ISSN: 2358-2472

REFERÊNCIAS

ADEJO, O. W.; CONNOLLY, T. Predicting Student Academic Performance Using Multi-Model Heterogeneous Ensemble Approach. Journal of Applied Research in Higher Education, 2018.

Adekitan, Aderibigbe Israel; SALAU, Odunayo. The impact of engineering students' performance in the first three years on their graduation result using educational data mining. Heliyon, 2019.

ANDIFES. Associação Nacional dos Dirigentes das Instituições Federais de Ensino Superior. Diplomação, retenção e evasão nos cursos de graduação em instituições de ensino superior públicas. Resumo do Relatório apresentado a ANDIFES, ABRUEM e SESu/MEC pela Comissão Especial. p. 55-65, 1996.

BEAULAC, Cédric; ROSENTHAL, Jeffrey S. Predicting university students' academic success and major using random forests. Research in Higher Education, 2019.

CHUI, K.; FUNG, D.; LYTRAS, M.; LAM, T. Predicting at-risk university students in a virtual learning environment via a machine learning algorithm. Computer in Human Behavior, 2020.

COSTA, Evandro B.; FONSECA, Baldoino; SANTANA, Marcelo A.; DE ARAÚJO, Fabrísia F.; REGO, Joilson. Evaluating the effectiveness of educational data mining techniques for early prediction of student's academic failure in introductory programming courses. Computers in Human Behavior, p. 247-256, 2017.

DELEN, Dursun. A comparative analysis of machine learning techniques for student retention management. Decision Support Systems, 2010.

DEO, R. C.; YASEEN, Z. M.; AL-ANSARI, N.; NGUYEN-HUY, T.; LANGLANDS, T.A.M.; GALLIGAN, L. Modern Artificial Intelligence Model Development for Undergraduate Student Performance Prediction: An Investigation on Engineering Mathematics Courses. IEEE Access, 2020.

EZZ, M.; ELSHENAWY, A. Adaptive Recommendation System Using Machine Learning Algorithms for Predicting Student's best Academic Program. Education and Information Technologies, 2019.

FERNÁNDEZ-MARTÍN, T.; SOLÍS-SALAZAR, M.; MARÍA TERESA, M. T.; MOREIRA-MORA, T. E. Un Análisis Multinomial y Predictivo de los Factores Asociados a la Deserción Universitaria. Revista Eletrónica Educare/Educare Eletronic Journal, 2018.

FERNÁNDEZ, Diego Buenaño; GIL, David; MORA, Sergio Luján. Application of machine learning in predicting performance for computer engineering students: a case study. Sustainability, 2019.

FREITAS, Francisco A. da S.; VASCONCELOS, Francisco F. X.; PEIXOTO, Solon A.; HASSAN, Mohammad Mehedi; DEWAN, M. Ali Akber; ALBUQUERQUE, Victor Hugo C. de; REBOUÇAS FILHO, Pedro P. IoT Systems for school dropout prediction using machine learning techniques based on socioeconomic data. Electronics, 2020.

GAMIE, E.; EL-SEOUD, M. S. A.; SALAMA, M. A. Comparative Analysis for Boosting Classifier in the Context of Higher Education. International Journal of Emerging Technologies in Learning, 2020.

Revista ARACÊ
ISSN: 2358-2472

GAMIE, Eslam Abou; EL-SEOUD, M. Samir; SALAMA, Mostafa A.; HUSSEIN, Walid. Multi-dimensional analysis to predict student's grades in higher education. International Journal of Emerging Technologies in Learning, v. 14, n. 02, 2019.

GOMES, O. F.; GOMIDE, T. R.; GOMES, M. A. N.; ARAÚJO, D. C.; MARTINS, S.; FARONI, W. Sentidos e implicações da gestão universitária para os gestores universitários. Revista Gestão Universitária na América Latina - GUAL, 2013.

HAI-TAO, P.; MING-QU, F.; HONG-BIN, Z.; BI-ZHEN, Y.; JIN-JIAO, L.; CHUN-FANG, L.; YAN-ZE, Z.; RUI, S. Predicting Academic Performance of Students in Chinese-foreign Cooperation in Running Schools with Graph Convolutional Network. Neural Computing and Applications, 2020.

HUNG, H. C.; LIU, I.-F.; LIANG, C.-T.; SU, Y.-S. Applying Educational Data Mining to Explore Student's Learning Patterns in the Flipped Learning Approach for Coding Education. Symmetry, 2020. Hussain, M.; Zhu, W.; Zhang, W.; Abidi, S.M.R. (2018). Student Engagement Predictions in an e-Learning System and Their Impact on Student Course Assessment Scores. Hindawi - Computational Intelligence and Neuroscience.

IATRELLIS, O.; SAVVAS, I. K.; FITSILIS, P.; GEROGIANNIS, V. C. A Two-phase Machine Learning Approach for Predicting Student Outcomes. Education and Information Technologies, 2020.

INEP. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. Censo da educação superior 2019. Acessado em: https://download.inep.gov.br/educacao_superior/censo_superior/documentos/2020/Press_Kit_Censo_Superior_2019.pdf. Último acesso em 20 jan. 2021.

INEP. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. Metodologia de indicadores e trajetória de curso. Acessado em: https://download.inep.gov.br/informacoes_estatisticas/indicadores_educacionais/2017/metodologia_indicadores_trajetoria_curso.pdf. 2017. Último acesso em 20 jan. 2021.

LAU, L. K. Institutional factors affecting students retention. Education, p. 126-137, 2003.

MANZANARES, M. C. S.; SÁNCHEZ, R. M.; GONZÁLEZ, A. A.; LLAMAZARES, M. C. E.; DEUS, M. A. Q. Detección del Alumno en Riesgo en Titulaciones de Ciências de la Salut: Aplicación de Técnicas de Learning Analytics. European Journal of Investigation in Health, Psychology and Education, 2018.

MIA, M.; BISWAS, A.; SATTER, A.; HABIB, T. Registration status prediction of students using machine learning in the context of private university of Bangladesh. International Journal of Innovative Technology and Exploring Engineering, 2019.

MOHER, D.; LIBERATI, A.; TETZLAFF, J.; ALTMAN, D. G. The PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. Disponível em: <www.prisma-statement.org>.

MUÑIZ, Luis Rodriguez; BERNARDO, Ana B.; ESTEBAN, María; DIAZ, Irene. Dropout and transfer paths: what are the risk profiles when analyzing university persistence with machine learning techniques? Plos One, 2019.

Revista ARACÊ
ISSN: 2358-2472

MUSSO, M.; HERNÁNDEZ, C.; CASCALLAR, E. Predicting Key Educational Outcomes in Academic Trajectories: a Machine-learning Approach. Higher Education, 2020.

NIKOLIC, Nikola; GRLJEVIC, Olivera; KOVACEVIC, Aleksandar. Aspect-based sentiment analysis of reviews in the domain of Higher. Electronic Library, p. 44-64, 2020.

PILLAI, B. Raveendran. Deep Regressor: Cross Subject Academic Performance Prediction System for University Level Students. International Journal of Innovative Technology and Exploring Engineering, 2019.

SCHLICKMANN, Raphael. Administração universitária: desvendando o campo científico no Brasil. Universidade Federal de Santa Catarina. Tese de doutorado, 2013.

SHIRATORI, N. Modeling dropout behavior patterns using bayesian networks in small-scale private university. In: Proceedings of the 2017 6th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI), Hamamatsu, Japan, 9-13 jul. 2017. p. 170-173.

SILVA FILHO, Roberto Leal Lobo e; MOTEJUNAS, Paulo Roberto; HIPÓLITO, Oscar; LOBO, Maria Beatriz de Carvalho Melo. A evasão no ensino superior brasileiro. Cadernos de Pesquisa, v. 37, n. 132, p. 641-659, set.-dez. 2007.

SILVA, Andréa Ferreira da; ALMEIDA, Aléssio Tony Cavalcanti de; RAMALHO, Hilton Martins de Brito. Predição do risco de reprovação no ensino superior usando algoritmos de machine learning. Teoria e Prática em Administração, jul.-dez. 2020.

SUHARJITO, Nindhia Hutagaol. Predictive modelling of student dropout using ensemble classifier method in higher education. Advances in Science, Technology and Systems Journal, v. 4, n. 4, p. 206-211, 2019.

SULTANA, Sara; KHAN, Sharifullah; ABBAS, Muhammad A. Predicting performance of electrical engineering students using cognitive and non-cognitive features for identification of potential dropouts. International Journal of Electrical Engineering Education, v. 54, n. 2, p. 105-118, 2017.
TSAI, S.-C.; CHEN, C.-H.; SHIAO, Y.-T.; CIOU, J.-S.; WU, T.-N. Precision Education with Statistical Learning and Deep Learning: a Case Study in Taiwan. International Journal of Educational Technology in Higher Education, 2020.

VIDHYA, R.; VADIVU, G. Towards developing an ensemble based two-level student classification model (ESCM) using advanced learning patterns and analytics. Journal of Ambient Intelligence and Humanized Computing, 2020.

WOTAIFI, T.; AL-SHAMERY, E. Mining of Completion Rate of Higher Education Based on Fuzzy Feature Selection Model and Machine Learning Techniques. International Journal of Recent Technology and Engineering, 2019.

XIAO, M.; YI, H. Research on Adaptive Learning Prediction Based on XAPI. International Journal of Information and Education Technology, 2020.

Revista ARACÊ
ISSN: 2358-2472

YANG, Y.; HOOSHYAR, D.; PEDASTE, M.; WANG, M.; HUANG, Y.-M.; LIM, H. Predicting Course Achievement of University Students Based on their Procrastination Behaviour on Moodle. Soft Computing, 2020.

YANG, Y.; HOOSHYAR, D.; PEDASTE, M.; WANG, M.; HUANG, Y.-M.; LIM, H. Prediction of Student's Procrastination Behaviour Through their Submission Behavioural Pattern in Online Learning. Journal of Ambient Intelligence and Humanized Computing, 2020.