Relatório
Tratamento de dados, viéses e privacidade

1nd Jadson Crislan Santos Costa
Instituto de Computação
Universidade Estadual de Campinas
UNICAMP
Campinas SP, Brasil
k255164@dac.unicamp.br

2rd Kamila de Almeida Benevides
Instituto de Computação
Universidade Estadual de Campinas
UNICAMP
Campinas SP, Brasil
k254880@dac.unicamp.br

3st Karla Gabriele Florentino da Silva
Instituto de Computação
Universidade Estadual de Campinas
UNICAMP
Campinas SP, Brasil
k272454@dac.unicamp.br

I. Introdução

Identificar precocemente a evasão de estudantes permite que a instituição de ensino promova ações para retenção e sucesso dos alunos em uma trajetória acadêmica bem-sucedida. (HEIJMANS; ARAÚJO; MENDES, 2014) classificam a evasão como um fenômeno complexo, multifacetado e multicausal, atrelado a fatores pessoais, sociais e institucionais, que podem resultar na saída provisória ou definitiva. (STEIMBACH, 2012) em sua dissertação adota o termo abandono escolar, pois considera evasão um ato solitário, levando a responsabilizar o indivíduo pelo seu afastamento. Enquanto diversas razões podem ser motivadoras para o ato, inclusive o abandono por parte da instituição.

(BORJA; MARTINS, 2014) afirmam que a evasão escolar provoca graves consequências sociais, acadêmicas e econômicas. No cenário universitário, este é problema recorrente, além da perda do aluno no meio acadêmico, os gastos gerados são elevados. No Brasil, segundo o relatório do Banco Mundial publicado em 2017, o custo médio anual para manter um aluno em universidades federais, foi de R$40.900.

Neste trabalho, pretende-se desenvolver um modelo de classificação supervisionado de aprendizado de máquina ético para a tarefa de predição de evasão e sucesso acadêmico sobre os dados de 4.424 alunos do ensino superior disponibilizados por (REALINHO VALENTIM; BAPTISTA, 2021). O conjunto de dados é tabular e possui 36 variáveis que incluem informações sobre a nacionalidade, trajetória acadêmica e fatores socioeconômicos.

II. Trabalhos relacionados

Nos últimos anos, diferentes modelos de *machine learning* (ML) tem sido usados por pesquisadores para predição de desempenho acadêmico (MDUMA; KALEGELE; MACHUVE, 2019). No cenário universitário do Brasil, Costa et al. (2017) fizeram *fine-tuning* em quatro modelos (Support Vector Machine (SVM), Decision Tree (DT), Neural Network (NN) e Naive Bayes (NB)) para identificação de insucesso em cursos introdutórios de programação de alunos de uma universidade pública brasileira. Os resultados do estudo revelaram que o modelo SVM superou os demais ao prever com 92% e 83% de acurácia as reprovações de alunos dos cursos à distância e presencial, respectivamente. Silva, Almeida e Ramalho (2020) utilizaram variáveis de discentes, docentes, turma e curso para identificar o risco de reprovação dos estudantes de todos os cursos da Universidade Federal da Paraíba na disciplina de Cálculo Diferencial e Integral I. Os autores compararam e ranquearem as dez variáveis que mais influenciaram no desempenho de modelos NB, SVM, DT, K-Nearest Neighbors, Logistic Regression e Penalized Methods.

Em um estudo recente, Bonifro et al. (2020) realizaram experimentos com dados reais pseudonimizados de 15.000 estudantes de onze escolas de uma grande universidade. Os autores apontam a existência de uma concentração de abandono durante o primeiro ano de curso, reforçando a importância de ações preventivas desde o início. As métricas utilizadas para avaliação dos modelos foram acurácia, especificidade e sensibilidade. Por fim, Bonifro et al. (2020) afirmam que a ferramenta tirou partido de dados sensíveis e que não é possível provar que é modelo é justo quanto ao gênero dos estudantes.

Karimi-Haghighi et al. (2021) analisaram o problema de previsão do risco de abandono dos estudos e desempenho insuficiente pela perspectiva de um algorítmico *fairness*. Um conjunto de dados anônimo constituído de informações de 881 estudantes de Engenharia da Computação de uma universidade europeia foi utilizado. Os autores avaliaram resultados discriminatórios em termos da acurácia (AUC) e taxas de erro (Taxa Generalizada de Falsos Positivos e Taxa Generalizada de Falsos Negativos). Foram obtidos modelos ML calibrados com AUC de 0,77 e 0,78 utilizando apenas informações no momento da inscrição.

Singh e Alhulail (2022) empregaram um modelo de regressão logística em quatro etapas para análise de evasão numa amostra de 1.723 estudantes-professores de escolas públicas de formação de professores da Etiópia. O primeiro modelo aprendeu somente com variáveis pessoais, o segundo modelo com variáveis acadêmicas, o terceiro combinou variáveis acadêmicas e pessoais e no quarto modelo variáveis socioeconômicas foram inclusas. O estudo revelou que utilizar somente informações pessoais ou acadêmicas não é suficiente para explicar os dados, com valores R^2 iguais a 14.9% e 20.6% respectivamente. Enquanto variáveis como ”postsecondary

aspirations", "academic performance", "negative interactions with teachers", e "academic reason for choosing institution" tiveram maior influência nas predições.

III. Tratamento dos dados

Os dados que serão utilizados foram tratados anteriormente pelos criadores do banco de dados, onde foram tratados anomalias, outliers e valores em faltantes.

A. Vieses

Diferentes formas de viés podem representar uma ameaça à validade em diversas situações. No contexto dos dados educacionais, os vieses discutidos nesse projeto serão: Viés Social, Viés de representação e Viés Temporal.

Em relação ao viés social, esse pode se manifestar de algumas maneiras, por exemplo, professores ou orientadores podem inadvertidamente avaliar o desempenho dos alunos com base em estereótipos sociais, como raça, cor, idade e histórico escolar anterior. Além disso, a análise dos dados socioeconômicos e da escolaridade dos pais pode revelar desigualdades sociais. Se o modelo de machine learning não considerar esses fatores e não for treinado para lidar com o viés social, pode subestimar o potencial de alunos de origens socioeconômicas menos privilegiadas, aumentando assim o risco de evasão para esses alunos. Para tentar evitar esse tipo de conclusão do modelo, e equipe buscará em apresentar a saída do modelo de forma a explicar os fatores que levaram ou poderá levar o aluno se evadir.

O viés de representação pode ocorrer quando a amostra de dados coletados não é representativa da população estudantil total. Isso pode acontecer se certos grupos demográficos, como estudantes de origens socioeconômicas mais baixas, forem sub ou super-representados nos dados. Nos dados educacionais usados para esse projeto, onde a maioria dos alunos tem nacionalidade de Portugal.

O viés temporal pode afetar a validade do modelo de machine learning se os dados forem coletados em momentos diferentes ou sob condições diferentes. Em relação aos dados educacionais que serão utilizados, os dados coletados referem-se a registos de estudantes inscritos entre os anos letivos 2008/2009 e 2018/2019, esse espaço de tempo abrange um período relativamente longo e pode incluir mudanças substanciais nas políticas educacionais, estruturas curriculares e condições socioeconômicas que afetam os alunos, isso pode afetar significativamente o viés temporal na análise de dados educacionais. Outro ponto relacionado ao viés temporal é que na base de dados contém a taxa de desemprego, índice de inflação e o produto interno bruto do país na data da coleta, isso pode ser usado para relacionar a evasão com a situação socioeconômica.

B. Privacidade

Os dados educacionais que serão utilizados neste projeto contém muitos atributos que poderiam passar por processos que certifiquem uma melhor privacidade para as pessoas cujos dados estão no banco de dados. A base de dados é composta por dados provenientes de diversas fontes, tanto internas quanto externas à instituição educacional, como não foram coletadas diretamente dos alunos, não existe algum tipo de termo de consentimento deles. Por esse lado, a base de dados podem conter informações que um determinado aluno não queira informar, por exemplo, sua nota, dados dos pais e dívidas escolares.

Algumas metodologias que poderiam ser utilizadas para certificação de privacidade são: (i) Realizar uma avaliação de riscos à privacidade, isso envolve identificar como os dados dos alunos serão coletados, armazenados, usados e quais riscos à privacidade podem surgir. Além disso, excluir a chance de um aluno ser identificado. A principal perda associada à avaliação de riscos é o tempo e os recursos necessários; (ii) Desenvolver políticas de privacidade claras e obter o consentimento dos alunos para o uso de seus dados são práticas essenciais.

Referências

BONIFRO, F. D. et al. Student dropout prediction. In: SPRINGER. *Artificial Intelligence in Education: 21st International Conference, AIED 2020, Ifrane, Morocco, July 6–10, 2020, Proceedings, Part I 21*. [S.l.], 2020. p. 129–140.

BORJA, I. M. F. de S.; MARTINS, A. M. de O. Evasão escolar: desigualdade e exclusão social. *Revista Liberato*, v. 15, n. 23, p. 93–102, 2014.

COSTA, E. B. et al. Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses. *Computers in Human Behavior*, v. 73, p. 247–256, 2017. ISSN 0747-5632. Disponível em: ⟨https://www.sciencedirect.com/science/article/pii/S0747563217300596⟩.

HEIJMANS, R. D.; ARAÚJO, A. C. de; MENDES, J. de S. Evasão na educação: estudos, políticas e propostas de enfrentamento. 2014.

KARIMI-HAGHIGHI, M. et al. Predicting early dropout: Calibration and algorithmic fairness considerations. *arXiv preprint arXiv:2103.09068*, 2021.

MDUMA, N.; KALEGELE, K.; MACHUVE, D. A survey of machine learning approaches and techniques for student dropout prediction. Data Science Journal, 2019.

REALINHO VALENTIM, V. M. M. M. J.; BAPTISTA, L. *Predict students' dropout and academic success*. 2021. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5MC89.

SILVA, A. Ferreira da; ALMEIDA, A. T. Cavalcanti de; RAMALHO, H. M. de B. Predição do risco de reprovação no ensino superior usando algoritmos de machine learning. *Teoria e Prática em Administração*, v. 10, n. 2, 2020.

SINGH, H. P.; ALHULAIL, H. N. Predicting student-teachers dropout risk and early identification: A four-step logistic regression approach. *IEEE Access*, IEEE, v. 10, p. 6470–6482, 2022.

STEIMBACH, A. A. Juventude, escola e trabalho: razões da permanência e do abandono no curso técnico em agropecuária integrado. *Curitiba, PR*, 2012.