Técnicas de aprendizagem de máquina para classificação de estudantes com base no seu perfil e percepção sobre a universidade

Machine learning techniques for classifying students based on their profile and perception of the university

Gabriel Carlos Venturini[1], Bruno Samways dos Santos[2]

RESUMO

O objetivo deste trabalho é analisar o desempenho de algoritmos de aprendizado de máquina para classificar os estudantes entre "promotor" e "detrator", identificando os atributos mais importantes para os modelos. Foi utilizado um conjunto de dados composto por 198 estudantes de uma universidade pública federal do Paraná, com informações sobre sua graduação, dados pessoais, rotina e percepções no âmbito universitário. A partir de técnicas de balanceamento no conjunto de dados como *Oversampling*, *Undersampling* e os dois métodos combinados, em conjunto com modelos de aprendizagem de máquina como *Random Forest*, *Multilayer Perceptron*, *XGBoost* e *Support Vector Machine*, classificou-se o estudante entre promotor e detrator, sendo promotor os estudantes que avaliam a universidade com notas de 8 a 10, e os detratores de 0 a 7. Para avaliar o poder preditivo dos métodos, foramutilizadas as métricas de avaliação como acurácia, *f-score*, precisão e *recall*, principalmente para classe dosdetratores. O melhor modelo para classificar os estudantes com confiabilidade aceitável foi a combinação detodos os modelos em um sistema de *voting*, com uma acurácia geral de 90%.

PALAVRAS-CHAVE: aprendizagem de máquina; satisfação; universidade.

ABSTRACT

The objective of this work is to analyze the performance of machine learning algorithms to classify students between "promoter" and "detractor", identifying the most important attributes for the models. A data set composed of 198 students from a federal public university in Paraná was used, with information about their graduation, personal data, routine and perceptions within the university context. Using data set balancing techniques such as Oversampling, Undersampling and the two methods combined, together with machine learning models such as Random Forest, Multilayer Perceptron, XGBoost and Support Vector Machine, the student was classified between promoter and detractor , being promoters of students who evaluate the university with grades from 8 to 10, and detractors from 0 to 7. To assess the predictive power of the methods, evaluation metrics such as accuracy, f-score,precision, and recall were used, mainly for the class of detractors. The best model to classify students withacceptable reliability was the combination of all models in a voting system, with an overall accuracy of 90%.

KEYWORDS: machine learning; satisfaction; university.

INTRODUÇÃO

Segundo Weerasinghe et al. (2017), a satisfação dos estudantes pode ser influenciada por vários fatores, incluindo o tamanho das turmas, a qualidade dos recursos e instalações, a interação com professores e colegas, e a eficácia dos serviços e suporte administrativo. Nesse sentido, é importante que a universidade ofereça serviços como orientação acadêmica, monitorias e atividades extracurriculares que possam contribuir para o desenvolvimento dos estudantes.

[1] Universidade Tecnológica Federal do Paraná, Londrina, Paraná, Brasil. E-mail: gabriel.vent@hotmail.com. ID Lattes: 5816844827056769.
[2] Docente no curso Engenharia de Produção. Universidade Tecnológica Federal do Paraná, Londrina, Paraná, Brasil. E-mail: brunosantos@utfpr.edu.br. ID Lattes: 5500192844287607.

Complementando a importância da satisfação do estudante, na teoria da expectativa de valor proposta por Wigfield e Eccles (2000), é destacado a importância da relação entre as expectativas dos estudantes no panorâma da universidade e o valor que eles atribuem à sua experiência educacional. Segundo essa teoria, quando os estudantes têm altas expectativas em relação ao ensino superior e percebem que suas experiências atendem ou excedem suas expectativas, isso leva a uma maior satisfação estudantil. Porém caso esta experiência não atenda sua expectativa, pode gerar uma maior insatisfação do estudante, resultando em altas taxas de evasão além de prejudicar a reputação da instituição.

Porém o modo como a satisfação do estudante é avaliado também é relevante, Delen (2011) afirma que a maioria das pesquisas sobre satisfação estudantil se baseia em métodos qualitativos, como entrevistas e grupos focais, que fornecem uma compreensão aprofundada, mas não permitem generalizações amplas. Sendo assim, é destacado a necessidade de estudos quantitativos, que envolvam a coleta de dados por meio de questionários estruturados e análise estatística, para obter informações mais abrangentese representativas sobre a satisfação dos estudantes.

Atualmente, é notável a crescente disponibilidade de dados, entre eles dados educacionais, algumas instituições de ensino têm usado a mineração de dados para identificar padrões e tendências em dados de estudantes (KITCHIN, 2014). A aplicação de técnicas de mineração de dados e aprendizado de máquina no ambiente acadêmico permite a analisar variáveis como notas, frequência, histórico escolar, percepções e informações sociodemográficas dos alunos. Isso possibilita a identificação de padrões dos estudantes, avaliação da eficácia de programas educacionais e a implementação de estratégias pedagógicas mais efetivas (ROMERO; VENTURA, 2010).

O processo que envolve a descoberta e extração de informações úteise relevantes a partir de conjuntos de dados é conhecido como *Knowledge Discovery in Databases* (Descoberta de Conhecimento em Bancos de Dados), ou KDD. Esta metodologia vem sendo amplamente utilizada em diversas áreas, como finanças, marketing, saúde, ciência e educação, devidoà sua capacidade de fornecer informações valiosas para a tomada de decisões e a soluçãode problemas (WITTEN et al., 2016).

A aprendizagem de máquina também desempenha um papel importante neste processo. Envolve a construção de algoritmos e modelos que permitem aos sistemas aprenderem a partir dos dados e melhorarem seu desempenho ao longo do tempo, e entre todas as tecnologias digitais que ganham força no mundo, o aprendizado de máquina é a que mais se destaca, sendo talvez uma das mais atraentes (GRANVILLE, 2019).

Em uma aplicação prática de Carlotto e Câmara (2022) foi avaliado o poder preditivo de algumas variáveis como as sociodemográficas, acadêmicas, e fatores estressantes, concluindo que a satisfação com o curso desempenha um importante papel para a continuidade dos estudos e que deve ser considerada pelos estudantes e pelas instituições de ensino superior.

Com o objetivo de analisar o desempenho de algoritmos de aprendizado de máquina para classificar os estudantes, e identificar os atributos mais importantes para os modelos, surge o questionamento: como é possível predizer a satisfação de estudantes por meio de dados que retratam seus comportamentos e percepções de suas experiências universitárias, utilizando técnicas de aprendizado de máquina?

METODOLOGIA

Os dados secundários utilizados nesta pesquisa foram obtidos a partir de um instrumento de coleta elaborado na pesquisa de Homma Junior (2022), contendo cinco seções principais, sendo elas: Termo de Consentimento Livre e Esclarecido (TCLE); informações gerais; perfil sociodemográfico; informações sobre rotina de estudo e bem-estar físico e mental; avaliação da estrutura e experiência de aprendizagem. O estudo foi aprovado pelo CEP (Comitê de Ética e Pesquisa) sob o CAAE: 57956421.6.0000.0177 e número do parecer: 5.458.136. Para mais informações sobre o instrumento utilizado, acessar Homma Junior (2022).

Os estudantes foram separados em promotor, detrator e neutro. Promotores avaliam a universidade com nota 8 ou mais sendo a classe 0, e detratores com nota 7 ou menos, classe 1.

O conjunto de treino e teste foi dividido em 50/50, para proporcionar maior quantidade da classe minoritária para o treinamento, oportunizando ao modelo mais exemplos para identificação de padrões ocultos.

Tabela 1 – Balanceamento da classe de interesse com neutro e com diluição do neutro

Classe	Vl. Abs.	%
Promotor	154	80,21%
Detrator	38	19,79%

Fonte: Autor (2022).

O primeiro balanceamento utiliza a técnica *SMOTE,* uma técnica de *Oversampling* para adicionar instâncias sintéticas da classe minoritária até atingir uma proporção especificada da classe majoritária que melhora a acurácia de classificadores para a classe minoritária (CHAWLA et al., 2002), sendo a proporção utilizada de 0,66.

Para o segundo balanceamento, é utilizado a técnica *RandomUnderSampler,* uma técnica de *Undersampling*, que retira instâncias da classe majoritária até a classe minoritária atingir uma proporção especificada da classe majoritária, proporcionando um conjunto mais compacto e consequentemente menor custo de processamento (CHAWLA et al., 2002), sendo a proporção utilizada de 0,33.

Enquanto para o terceiro balanceamento foi utilizado a técnica *SMOTE*, em conjunto com o *RandomUnderSampler,* obtendo uma combinação de balanceamentos válidos como desenvolvido por (CHAWLA et al., 2002).

Tabela 2 – Balanceamento dos conjuntos de treino

Conjunto de treino *SMOTE*			Conjunto de treino *RandomUnderSampler*			Conjunto de treino misto		
Classe	Vl. Abs.	%	Classe	Vl. Abs.	%	Classe	Vl. Abs.	%
Promotor	77	60,63%	Promotor	57	75,00%	Promotor	37	59,68%
Detrator	50	39,37%	Detrator	19	25,00%	Detrator	25	40,32%

Fonte: Autor (2022).

RESULTADOS

Entre todos os 12 modelos criados, cinco obtiveram resultados mais promissores, sendo eles os modelos de 1 a 5:

1. RF utilizando um conjunto de dados misto;
2. MLP utilizando um conjunto de dados com *oversampling*;
3. XGB utilizando um conjunto de dados misto;
4. XGB utilizando um conjunto de dados com *oversampling*;
5. SVC utilizando um conjunto de dados com *oversampling*.

Tabela 3 – Resultados dos modelos

	Modelo 1		Modelo 2		Modelo 3		Modelo 4		Modelo 5	
	Promotor	Detrator	Promotor	Detrator	Promotor	Detrator	Promotor	Detrator	Promotor	Detrator
Precisão	0,92	0,62	0,94	0,47	0,93	0,61	0,93	0,56	0,91	0,60
Recall	0,90	0,68	0,78	0,79	0,88	0,74	0,86	0,74	0,90	0,63
f-score	0,91	0,65	0,85	0,59	0,91	0,67	0,89	0,64	0,90	0,62
Acurácia	0,85		0,78		0,85		0,83		0,84	

Fonte: Autor (2023).

É possível observar que entre os modelos selecionados, não há a presença de nenhum modelo com o conjunto de dados com *undersampling*, como esperado, tiveram dificuldades para classificar os detratores, indicando uma possível generalização excessiva do modelo (*underfitting)*.

Porém é necessário analisar se o modelo pode estar sofrendo *overfitting*, para isso é realizado a predição do próprio conjunto de treinamento. Sendo esperado um aumento na performance, quatro dos cinco modelos que mais performaram estão possivelmente sofrendo um ajuste excessivo (*overfitting*), visto que performaram com perfeição no conjunto de treino, e uma queda considerável no conjunto de testes, enquanto o modelo 3 apresentou apenas um pequeno aumento ao ser avaliado no conjunto de treinos comparado ao conjunto de testes.

Tabela 4 – Performance no conjunto de treino para verificar o *overfitting*

	Modelo 1		Modelo 2		Modelo 3		Modelo 4		Modelo 5	
	Promotor	Detrator	Promotor	Detrator	Promotor	Detrator	Promotor	Detrator	Promotor	Detrator
Precisão	1,00	1,00	1,00	1,00	0,85	0,86	1,00	1,00	1,00	1,00
Recall	1,00	1,00	1,00	1,00	0,92	0,76	1,00	1,00	1,00	1,00
f-score	1,00	1,00	1,00	1,00	0,88	0,81	1,00	1,00	1,00	1,00
Acurácia	1,00		1,00		0,85		1,00		1,00	

Fonte: Autor (2022).

A partir da combinação dos 12 modelos, obtemos o modelo *ensemble voting*, em que os resultados obtidos se demonstraram melhores, demonstrando principalmente um aumento na sensibilidade para os detratores (classe 1), sem a perda de desempenho no conjunto dos promotores, classe 0, como consta na tabela 16:

Tabela 5 – *Modelo de v*otação majoritária (*voting*)

	Promotor	Detrator
Precisão	0,94	0,74
Recall	0,94	0,74
f-score	0,94	0,74
Acurácia	0,90	

Fonte: Autor (2022).

Para os classificadores *Random Forest* e *XGBoost*, é possível utilizar o método *feature_importances* para elencar quais variáveis foram utilizadas pelos modelos para realizar a classificação, atribuindo um peso a cada uma delas, e a qual categoria do perfil do estudante tal variável pertence.

Tabela 6 – Importância das categorias das variáveis

Categoria	Contribuição	Qtd. Variáveis	Média
Notas	44,84%	12	3,74%
Rotina	24,50%	17	1,44%
Pessoal	17,54%	16	1,10%
Graduação	13,13%	11	1,19%

Fonte: Autor (2022).

CONCLUSÃO

Por utilizar diversos modelos para realizar a predição das instâncias, o método *ensemble* de votação majoritária (*voting*), foi o modelo que melhor desempenhou entre todos os outros. Ao analisar os modelos individualmente, o *XGBoost* treinado com conjunto de dados misto (modelo 3) foi o que desempenhou melhor.

As variáveis relacionadas a categoria de notas foram as que mais demonstraram importância para os modelos, envolvendo tanto notas para professores, quanto para estruturas da universidade. Ou seja, todas são variáveis que a universidade pode atuar para tentar fazer a diferença para o estudante, priorizando estruturas e aspectos específicos no ambiente universitário visando maior satisfação do estudante.

Sendo assim, é possível direcionar esforços e recursos para que os problemas levantados na universidade possam utilizar deste trabalho para direcionar e priorizar aspectos e áreas, tanto quanto para agir diretamente nos pontos mais críticos, sempre visando a maior percepção da satisfação pelo estudante.

Conflito de interesse

Não há conflito de interesse.

REFERÊNCIAS

CHAWLA, N. V. et al. SMOTE: Synthetic minority over-sampling technique. **The Journal of Artificial Intelligence Research**, v. 16, p. 321–357, 2002. https://doi.org/10.1613/jair.953

CARLOTTO, M. S.; GONÇALVES C MARA, S. As Intenções de abandonar o curso universitário: um estudo de predição e mediação. **Revista educação em questão**, v. 60, n. 65, 2022. https://doi.org/10.21680/1981-1802.2022v60n65id29277.

DELEN, D. A comparative analysis of machine learning techniques for student retention management. **Decision Support Systems**, v. 49, n. 4, p. 498–506, 2010. https://doi.org/10.1016/j.dss.2010.06.003

Wigfield, A., & Eccles, J. S. (2000). Expectancy–value theory of achievement motivation. **Contemporary Educational Psychology**, 25(1), 68–81. https://doi.org/10.1006/ceps.1999.1015

WEERASINGHE, I. M. S.; LALITHA, R.; FERNANDO, S. Students' satisfaction in higher education literature review. **American journal of educational research**, v. 5, n. 5, p. 533–539, 2017. DOI: 10.12691/education-5-5-9

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: a publication of the IEEE Systems, Man, and Cybernetics Society**, v. 40, n. 6, p. 601–618, 2010. https://doi.org/10.1109/TSMCC.2010.2053532

KITCHIN, R. **The data revolution: Big data, open data, data infrastructures and their consequences**. Christchurch, New Zealand: Sage Publications, 2014. https://doi.org/10.1111/jors.12293

GRANVILLE, Lisandro Z. Machine Learning: Desafios para um Brasil competitivo. **SBC**, 2019.

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). **Data Mining: Practical Machine Learning Tools and Techniques**. Morgan Kaufmann. https://doi.org/10.1016/C2009-0-19715-5

HOMMA JUNIOR, E. **Análise do perfil de estudantes de graduação baseada nas suas experiências e percepções sobre a universidade: uma abordagem por técnicas de clusterização**. [s.l.] Universidade Tecnológica Federal do Paraná, 21 nov. 2022.