VI SEMANA UNIVERSITÁRIA DA URCA
XXIV SEMANA DE INICIAÇÃO CIENTÍFICA DA URCA
13 a 17 de Dezembro de 2021
Tema: "Centenário de Paulo Freire: contribuição da divulgação científica e tecnológica em defesa da vida, da cidadania e da educação"

RISCO DE REPROVAÇÃO NO ENSINO SUPERIOR: UMA APLICAÇÃO DOS ALGORITMOS DE MACHINE LEARNING PARA A UNIVERSIDADE REGIONAL DO CARIRI

Pedro Willame Pereira da Silva[1], Andréa Ferreira da Silva[2], Jayane Freires Ferreira[3] José Alexandre Quieroga do Nascimento[4]

Resumo: A reprovação no ensino superior tem se caracterizado como um problema nas Instituições de Ensino Superior (IES) no Brasil, em paralelo ao aumento da oferta de vagas na educação superior. Assim, este estudo tem como objetivo avaliar o risco de reprovação dos discentes matriculados nas disciplinas de pesquisa da Universidade Regional do Cariri (URCA), Unidade Descentralizada de Iguatu, utilizando-se microdados cedidos pela URCA, do período de 2010 a 2019. Aplicaram-se algoritmos de *Machine Learning* (ML) para a análise dos dados e estimação das previsões, selecionando como o modelo preditivo o *Elastic Net*. Resultou-se que das 775 observações que compõem a base de teste, quanto ao *status* (reprovados e aprovados) houve uma taxa de predição correta de 91,47% da *Accuracy*. Ainda, cerca de 23% dos discentes (178) foram efetivamente previstos como reprovados (*Sensitivity*). Os resultados obtidos através deste estudo, qualificam os algoritmos de ML como ferramentas efetivas no auxílio aos gestores educacionais na implementação de estratégias que tenham como finalidade reduzir os índices de reprovação no ensino superior.

Palavras-chave: *Machine Learning*. Economia. Políticas Públicas. Reprovação.

1. Introdução

Problemas como a reprovação, a retenção e a evasão incidem sobre uma realidade presente na formação de muitos discentes nos mais diversos cursos de formação nas Instituições de Ensino Superior (IES) no Brasil. Caracterizam-se não só como dificuldades estruturais do contexto educacional, mas acarreta também um aumento nos custos de financiamento do funcionamento nas universidades públicas, diminuindo os níveis de produtividade e eficiência destas. Diversas pesquisas abordam essa problemática, dialogando sobre a necessidade de interferência em diferentes níveis, partindo desde o núcleo familiar, social e econômico, até ao que remete aos insumos escolares e

1 Graduando em Ciências Econômicas pela Universidade Regional do Cariri (URCA), email: pedrowill2@hotmail.com
2 Doutora em Economia Aplicada pela Universidade Federal da Paraíba (UFPB). Docente do curso de Ciências Econômicas da Universidade Regional do Cariri (URCA) email: andrea.economia@yahoo.com.br
3 Graduanda em Ciências Econômicas pela Universidade Regional do Cariri (URCA), email: jayfreires2014@gmail.com
4 Graduando em Ciências Econômicas pela Universidade Regional do Cariri (URCA), email: jose.queirogan@gmail.com

diretrizes da educação (SILVA-FILHO et al., 2007; SANTOS-BAGGI; LOPES, 2011).

Nesse sentido, esta pesquisa tem por objetivo avaliar de maneira precoce, os discentes que possuam um determinado risco de reprovação no que tange às disciplinas de pesquisa científica, sendo estas adotadas por serem as disciplinas com maiores taxas de reprovação e desistência, além de ter caráter indispensável ao processo de escrita do trabalho de conclusão de curso. Então, para atingir este objetivo, será utilizada uma estratégia semelhante a que foi adotada na tese de Silva (2019), onde em um dos seus ensaios sobre economia aplicada, analisou-se a reprovação no ensino superior através da aplicação dos algoritmos de *Machine Learning* (ML) para a Universidade Federal da Paraíba (UFPB).

2. Objetivo

Avaliar o risco de reprovação dos discentes matriculados nas disciplinas de pesquisa na Universidade Regional do Cariri (URCA), Unidade Descentralizada de Iguatu (UDI). Descrevendo os algoritmos de *Machine Learning* aplicados e analisando o modelo de melhor performance preditiva.

3. Metodologia

A variável de resposta a ser estimada nos algoritmos de classificação de ML, será uma variável binária e assume 1 quando o discente apresenta o *status* de matrícula reprovado, e 0, caso contrário, ou seja, quando for aprovado. Os dados apresentados neste artigo foram coletados a partir de microdados oriundos do Departamento de Tecnologia da Informação (DTI) da URCA/UDI, contendo informações sobre os discentes que ingressaram nos quatro cursos que compõem a UDI: Ciências Econômicas; Direito; Educação Física e Enfermagem, no período de 2010 a 2019.

Com a finalidade de atingir o objetivo apresentado no estudo foram aplicados uma série de algoritmos de ML e utilizados processos de seleção e avaliação do algoritmo de melhor performance preditiva. É notório que a adoção de um maior conjunto de variáveis acarretaria em uma melhor análise, no entanto, observou-se uma grande quantidade de *missings* na base, principalmente com relação a variáveis a nível de professor e mais variáveis descrevendo os discentes. A base de dados, que está dividida por curso, contém características dos discentes, além de informações sobre o desempenho acadêmico nas disciplinas, sobretudo nas com maior número de reprovações, sendo estas as cadeiras relacionadas com a escrita do trabalho de conclusão de curso, e pesquisa científica, ressaltando-se que os estudantes com dados presentes neste estudo tiveram sua identidade preservada

4. Resultados

A partir da análise da tabela 1, observa-se uma similaridade na performance dos algoritmos. Segundo Prati, Batista e Monard (2008) e Kuhn e Johnson (2013), o método mais comumente utilizado para combinar a *Sensitivity* e *Specificity*, indispensáveis às análises das variáveis, é a curva ROC, este analisando o comportamento que estima a área sob a curva ROC, a *AUC ROC*, e a *Accuracy* foram os fundamentos, que conjuntamente determinaram a seleção dos algoritmos com maior capacidade de predição: os métodos baseados em árvores de decisão: *Bagging* e *Random Forest*, o *Penalized Methods LASSO*, *Penalized Methods Elastic Net*, Regressão Logística*, KNN* e *Naïve Bayes Classifier*.

Tabela 1 – Estimação dos algoritmos de *Machine Learning* para prever o risco de reprovação dos discentes matriculados nas disciplinas de pesquisa na URCA/UDI, nos anos de 2010 e 2019.

Algoritmos		**Critérios para avaliação do desempenho (%)**				
		Accuracy	***Sensitivity***	***Specificity***	***AUC ROC***	***IC 95% (AUC)***
Regressão Linear		79,14	56,52	80,57	77,19	(72,20 – 82,19)
Penalized Methods	***Ridge***[1]	79,66	55,43	81,19	77,40	(72,46 – 82,35)
	Lasso[2]	92,25	17,39	96,98	76,86	(72,40 – 81,33)
	Elastic Net[3]	91,47	22,82	95,81	76,97	(72,52 – 81,44)
Regressão Logística		81,92	50,00	83,93	77,00	(71,82 – 82,19)
K-Nearest Neighbors (KNN)[4]		93,60	2,17	99,38	76,55	(72,22 – 81,40)
Naive Bayes Classifier		93,35	15,21	98,28	56,75	(53,05 – 60,46)
Support Vector Machines (SVM)		78,95	56,37	80,26	77,10	(72,32 – 81,40)
Decision Tree Based Methods	***C. trees***	78,76	54,17	81,74	74,58	(72,38 – 80,78)
	Bagging	80,89	53,26	82,63	75,95	(71,04 – 80,86)
	Random Forest	81,92	53,26	83,73	77,60	(72,66 – 82,55)
	Boosting	77,83	52,72	78,84	74,51	(69,49 – 79,43)

Fonte: Elaboração própria a partir dos microdados do DTI/URCA.
Nota: [1]λ_{Ridge} = 0, 01 e MSE=0,0516; [2] λ_{Lasso} = 0, 01 e MSE=0,0527; [3]λ_{EN} = 0, 01, $\alpha = 0{,}5$ e MSE=0,0519; [4]*K* = 300 (Parâmetros *tunning*).

De acordo com James et al. (2013), os algoritmos que pertencem ao grupo dos métodos lineares com penalidade procuram penalizar os coeficientes estimados a fim de limitar a variância em troca de aumentar de forma não significativa o viés, indo de acordo ao que fala Hastie et al. (2009), que corroborando essa ideia, infere que estes métodos são os mais adequados no processo de predição. Por meio dessa análise, os modelos finais selecionados

com o intuito de prever o risco de reprovação dos alunos que se matricularão nas disciplinas de pesquisa da URCA-UDI são: *LASSO* e *Elastic Net.*

A partir dos dados acerca dos algoritmos utilizados na Tabela 2 e adotando um nível de significância de 5%, há evidências de que não se rejeite a hipótese nula onde as distribuições das predições sejam equivalentes de acordo com o *p-valor* no modelo: *Elastic Net*, considerando o teste de McNemar, se deduz que há uma diferença no resultado do teste Elastic Net quando comparado aos outros modelos, tornando-o assim, o de performance mais confiável dentro dos parâmetros observados. Com isso, com relação ao processo de avaliação (Tabela 1), a *AUC ROC*, que foi abordada como ferramenta de otimização e para selecionar os modelos no decorrer do aprendizado; corroborando com o teste de McNemar que rejeitou a hipótese nula de um dos dois modelos previamente selecionados, apresentou resultado de 76,97%. Em relação a *Accuracy*, o modelo obteve uma taxa de predição de 91,47% dos discentes da URCA-UDI entre os anos de 2010 e 2019 (base de teste), se enquadrando na situação de reprovado e aprovado. Ainda por meio da interpretação dos resultados, é viável afirmar que, com uma Taxa de Falso Positivo (TFP) (1 - *Specificity*) em torno de 4,19% [1- 95,81% (*Specificity*)] no modelo, é possível 18 predizer que a uma estimativa de 23% (*Sensitivity* ou a Taxa de Verdadeiro Positivo) dos discentes que se matricularam nas disciplinas de pesquisa na URCA-UDI irão reprovar entre os anos de 2010 e 2019, assim, dos 775 estudantes da base de teste, 178 foram corretamente preditos.

Tabela 2 – Testes de McNemar para os algoritmos de *Machine Learning* - Teste Individual

Algoritmos		**Teste McNemar**	
		Teste Chi-squared	*p-valor*
Regressão Linear		*181,31*	0,0000
Penalized Methods	***Ridge***	170,87	0,0000
	Lasso	8,0083	0,0046
	Elastic Net	0,6136	0,4334
Regressão Logística		124,89	0,0000
K-Nearest Neighbors (KNN)		64,646	0,0000
Naive Bayes Classifier		27,01	0,0000
Support Vector Machines (SVM)		84,496	0,0000
Decision Tree Based Methods	***C. trees***	138,47	0,0000
	Bagging	147,57	0,0000
	Random Forest	133,03	0,0000
	Boosting	142,99	0,0000

Fonte: Elaboração própria a partir dos microdados do DTI/URCA.

5. Conclusão

Por meio dos resultados obtidos durante o desenvolvimento desta pesquisa, não estando estes isentos de limitações, é razoável indicar que esse seja um instrumento viável ao que remete ao fornecimento de um aparato de informações mais precisas, visando auxiliar as ações dos gestores educacionais na busca pela diminuição nos índices de reprovação nos cursos de ensino superior. Desta forma, as IES podem dar um maior auxílio aos gestores e professores, ao passo que as ações teriam uma maior taxa de produtividade, uma vez que serão previamente direcionadas aos problemas de maior degradação da educação superior. Assim, acarretando em uma redução da retenção, evasão e os custos, e ao mesmo tempo, elevar os índices de diplomação e elevar a produtividade e qualificação do ensino superior no Brasil.

6. Referências

BARROSO, Marta F.; FALCÃO, Eliane BM. **Evasão universitária: o caso do Instituto de Física da UFRJ**. IX Encontro Nacional de Pesquisa em Ensino de Física, v. 9, p. 1-14, 2004.

DA SILVA, Andréa Ferreira; DE ALMEIDA, Aléssio Tony Cavalcanti; DE BRITO RAMALHO, Hilton Martins. **PREDIÇÃO DO RISCO DE REPROVAÇÃO NO ENSINO SUPERIOR USANDO ALGORITMOS DE MACHINE LEARNING.**

JAMES, G., WITTEN, D., HASTIE, T., & TIBSHIRANI, R. (2013). **An introduction to statistical learning** (Vol. 112, pp. 3-7). New York: springer.

KUHN, M., & JOHNSON, K. (2013). **Applied predictive modeling (Vol. 26).** New York: Springer.

PRATI, R. C., BATISTA, G. E. A. P. A., & MONARD, M. C. (2008). **Curvas ROC para avaliação de classificadores.** Revista IEEE América Latina, 6(2), 215-221.

SILVA, Andrea Ferreira da. **Ensaios sobre economia aplicada: doações eleitorais, compras públicas, análise de políticas afirmativas e reprovação no ensino superior. 2019**.

SILVA FILHO, Roberto Leal Lobo et al. **A evasão no ensino superior brasileiro. Cadernos de pesquisa**, v. 37, n. 132, p. 641-659, 2007