Refas
Revista Fatec Zona Sul
1ª. Edição Outubro de 2014

Análise e Desenvolvimento de Sistemas

DEFINIÇÃO DE MODELOS DE APRENDIZADO DE MÁQUINA PARA PREDIÇÃO DE EVASÃO DE ALUNOS DO CURSO TÉCNICO

DEFINITION OF MACHINE LEARNING MODELS TO PREDICT STUDENTS WITHDRAWAL FROM THE TECHNICAL COURSE

DEFINICIÓN DE MODELOS DE MACHINE LEARNING PARA PREDECIR LA RETIRADA DE ESTUDIANTES DEL CURSO TÉCNICO

João Carlos Lima Silva[1]

Artigo recebido em janeiro de 2022
Artigo aceito em julho de 2022

RESUMO

A evasão no contexto escolar é um problema que afeta diretamente a qualidade da educação. Em cursos técnicos de nível médio as consequências são ainda mais graves. O uso de aprendizado de máquina permite realizar a previsão de comportamentos baseado em experiências anteriores. Através de dados socioeconômicos de alunos e do desempenho deles no processo seletivo, foi possível predizer com uma acurácia média de 75,1% se um aluno deixaria o curso. Ao ter acesso à essa informação, é possível realizar um trabalho preventivo de acompanhamento desses alunos com a finalidade de evitar que ele se evada do curso.

Palavras-chave: Evasão. Inteligência Artificial. Gestão Escolar.

ABSTRACT

Dropout in the school context is a problem that directly affects the quality of education. In technical courses the consequences are even more serious. Using machine learning allows you to predict behavior based on past experience. Through socioeconomic data of students and their performance in the selection exam, it was possible to predict with an average accuracy of 75.1% if a student would dropout. By having access to this information, it is possible to carry out preventive follow-up work with these students in order to prevent them from evading the course.

Keywords: Dropout. Artificial intelligence. School management.

[1]Diretor e professor da Etec Irmã Agostina. E-mail: joaocarloslima@me.com. Orcid: 0000-0002-2861-1474.

RESUMEN

La deserción en el contexto escolar es un problema que afecta directamente la calidad de la educación. En los cursos técnicos de nivel medio, las consecuencias son aún más graves. El uso del aprendizaje automático permite predecir comportamientos en base a experiencias previas. A través de datos socioeconómicos de los estudiantes y su desempeño en el proceso de selección, fue posible predecir con una precisión promedio de 75.1% si un estudiante abandonaría el curso. Al tener acceso a esta información, es posible realizar un trabajo preventivo de seguimiento de estos estudiantes para evitar que abandonen el curso.

Palabras clave: Evasión. Inteligencia artificial. gestión escolar.

1 INTRODUÇÃO

O orçamento para a educação no estado de São Paulo para o ano de 2019 foi de R$32 bilhões (Lei 16.923 de 07/01/2019 Alesp). Um aluno que deixa de frequentar o curso acaba por causar um grande desperdício do dinheiro público investido em infraestrutura, instalações e principalmente recursos humanos. Investimento esse que poderia ser aplicado na formação de outros alunos. Além disso, a perda mais significativa é a não formação do profissional que deixou as aulas. Uma vez que um aluno evade essa vaga não pode ser ocupada por outro aluno. Reduzir a evasão escolar no âmbito do ensino médio profissionalizante estadual é um grande desafio que deve ser encarado.

Dados do Centro Estadual de Educação Profissional Paula Souza, responsável pela administração da ETECs (Escolas Técnicas Estaduais) e FATECs (Faculdades de Tecnologia) do Estado de São Paulo, apontam que o maior índice de evasão de alunos ocorre nos períodos iniciais dos cursos (CPS, 2019). Isso agrava ainda mais o problema, uma vez que essa perda se reflete nos módulos seguintes.

Um coordenador de curso de ETEC é responsável por aproximadamente 180 alunos e tem uma carga horária de 12 horas semanais em média. Entre as suas atribuições estão: acompanhamento dos registros de aula, frequência de alunos e professores, cumprimento de carga horária, projetos interdisciplinares e acompanhamento dos alunos. Devido à essa realidade o coordenador de curso não consegue realizar um acompanhamento individualizado de todos os alunos. É necessário que o coordenador foque seus esforços nos alunos com maior necessidade. Porém, nem sempre os alunos que desistem do curso sinalizam antecipadamente as dificuldades e quando o coordenador toma ciência da situação já é tarde demais.

O Aprendizado de Máquina é uma subárea da Inteligência Artificial que permite a criação de modelos treinados com capacidade de identificar classes a partir de instâncias previamente determinadas (RUSSELL e NORVIG, 2010). As aplicações modernas de Aprendizado de Máquina realizam diversos tipos de previsões de comportamento humano baseado em um histórico anterior, como por exemplo, prever se um cliente irá comprar determinado produto de acordo com as páginas que ele visitou em um site.

A aplicação do Aprendizado de Máquina no problema da evasão pode permitir que a comunidade escolar tenha uma sinalização antecipada dos alunos com maior chance de abandonar o curso. Dessa forma, é possível que seja realizado um trabalho de prevenção e suporte para que o aluno consiga concluir o curso de forma satisfatória.

Este trabalho de pesquisa tem por objetivo geral definir qual é o modelo de aprendizagem de máquina que oferecer maior acurácia na predição da evasão de alunos no contexto dos cursos técnicos de nível médio.

Definir um modelo preciso de previsão permite que as decisões tomadas no âmbito da gestão escolar sejam mais efetivas na redução da evasão.

O artigo está organizado da seguinte maneira: Na seção 2 são apresentados os conceitos fundamentais de Evasão Escolar e Aprendizado de Máquina. Na Seção 3 apresentam-se os trabalhos relacionados; na Seção 4 descrevem-se a metodologia utilizada para criação e avaliação de modelos de classificação de alunos evadidos; os resultados são apresentados na Seção 5 e as considerações finais e trabalhos futuros na Seção 6.

2 REFERENCIAL TEÓRICO

Para compreensão mais aprofundada do problema, faz necessário compreender dois conceitos: Evasão Escolar e Aprendizado de Máquina.

2.1 Evasão Escolar no Ensino Médio

Em 2018 o Brasil teve quase dois milhões de matrículas no Ensino Profissional Técnico de Nível Médio. Dessas matrículas, 24% ocorreram no Estado de São Paulo, que é o estado da Federação com o maior número de vagas ocupadas na modalidade (INEP 2019).

O problema da evasão escolar acontece em todos os níveis de ensino. Porém, no ensino médio de nível técnico em São Paulo ele é ainda mais grave por dois fatores: 1) a universalidade do ensino (grande oferta em todo estado) e 2) não obrigatoriedade.

Os pesquisadores distinguem três dimensões que devem ser investigadas no contexto da evasão escolar: 1) níveis de escolaridade (médio, técnico ou superior); 2) tipo de evasão (descontinuidade, retorno, definitiva); 3) razões que motivam a evasão (trabalho, problemas de saúde, problemas sociais, quebra de expectativa com o curso etc.) (Jordan et al. 1996).

A compreensão geral é de que alunos com renda mais baixa possuem um menor índice de rendimento pois muitos deles precisam trabalhar para sustento próprio e da família, cansados do Ensino, muitos adolescentes desistem dos estudos sem completar o curso (Sousa et al. 2011).

O Conselho Europeu em 2004 propôs como solução para redução de exclusão e evasão de alunos, a prévia identificação e o encaminhamento adequado daqueles que estão em situação de risco (COUNCIL, 2004).

A presente pesquisa está inserida no contexto do ensino técnico de nível médio, investigando evasões definitivas, ou seja, de alunos que não concluem o curso independente da motivação.

2.2 Aprendizado de Máquina

A área de Inteligência artificial possui algumas subáreas. Uma delas é a área de aprendizagem de máquina que consiste na capacidade de um programa se adaptar a novas circunstâncias e detectar e extrapolar padrões (Russel 2010).

Um algoritmo de aprendizado de máquina utiliza dados do passado para prever o futuro. Eles podem ser utilizados para predição de situações, classificação e regressão. Mitchell (2017) define aprendizado de máquina como sendo a capacidade de melhorar o desempenho na realização de alguma tarefa por meio da experiência.

O aprendizado de máquina pode ser supervisionado (quando são informados para o algoritmos resultados prévios, que ele utiliza para a classificação) ou não supervisionados (quando esses dados não são apresentados para o algoritmo).

A presente pesquisa utiliza o aprendizado supervisionado para classificação. São apresentados dados de situações passadas de evasão de alunos para treinar os modelos que irão predizer novos casos de evasão.

3 MÉTODO

A partir do artigo publicado por (BRITO, ALMEIDA, *et al.*, 2014) pretende-se implementar melhorias nas técnicas de inteligência artificial com o objetivo de obter modelos melhores para predição de evasão de alunos. Essa nova pesquisa será aplicada no contexto de uma escola técnica estadual que oferece quatro cursos: Técnico em Administração, Técnico em Informática, Técnico em Química e Técnico e Nutrição e Dietética.

3.1 Conjunto de Dados

O artigo original não disponibilizou os dados utilizados para treinar e testar o modelo. Para realização dos experimentos foram levantados dados reais de alunos ingressantes e concluintes dos cursos técnicos.

Para acesso aos cursos os alunos realizam um processo seletivo composto por uma prova de múltipla escolha. Essa prova avalia cinco competências gerais. O resultado dessa avaliação de competências foi utilizado no *dataset*.

Jordan *et al.* (1996) compreendem que questões sociais influenciam na evasão de alunos. Portanto foram incluídos dados socioeconômicos dos ingressantes na base de dados. Esses dados foram obtidos através do cadastro do aluno na ficha de matrícula.

A **Erro! Fonte de referência não encontrada.** apresenta os doze atributos do conjunto de dados utilizados na presente pesquisa, bem como os valores mínimos e máximos de cada atributo.

Tabela 1 – Atributos do conjunto de dados

Atributo	Mín.	Máx.	Observações
Renda	0	15	Renda do aluno em salário-mínimo
Afrodescendente	0	1	Valor lógico indicativo de afrodescendência
Escola Pública	0	1	Valor que indica se o aluno estudou em escola pública
Curso	1	4	Id do curso do aluno
Casado	0	1	Valor lógico indicativo do estado civil do aluno
Idade	14	62	Idade do aluno quando iniciou o curso
N1 a N5	0	1	Porcentagem de acertos de cada competência avaliada
Concluinte	**0**	**1**	**Valor lógico que indica se o aluno concluiu o curso**

Fonte: autor

Após o levantamento e normalização dos dados foram obtidos 852 registros[2]. Porém, alguns desses alunos não realizaram o processo seletivo por serem oriundos de transferências de outras escolas, sendo assim não possuíam os dados completos e foram removidos do *dataset*. O conjunto de dados final possui 789 registros de alunos.

3.2 Algoritmos e Classificação

Seguindo a metodologia do artigo original foram utilizados cinco algoritmos de aprendizado de máquina, com o objetivo de encontrar o que apresenta melhor acurácia na predição da evasão. Os algoritmos estão listados a seguir juntamente com sua classificação: *Naive Bayes* (NB), Método do Vizinho mais próximo (KNN), Máquina de Vetor de Suporte (SVM), Árvore de Decisão (*Random Forest*), Redes Neurais Artificiais (MLP).

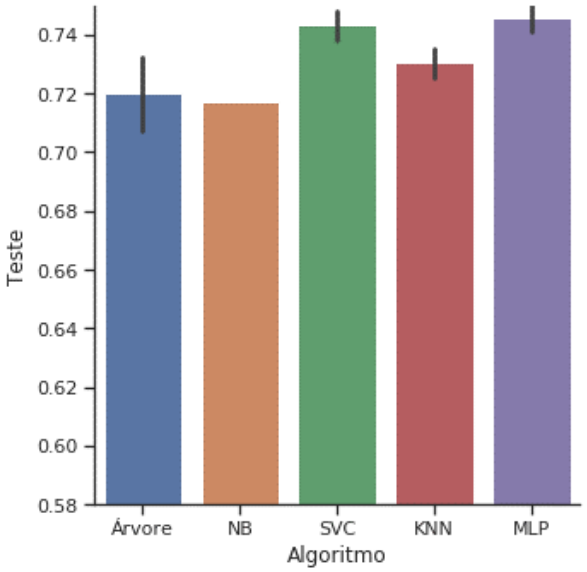
O artigo original executou os experimentos no aplicativo Weka[3], porém a presente pesquisa optou por utilizar a biblioteca *scikit-learn* do Python, pois essa estratégia oferece mais recursos e um maior controle sobre todo o processo de treino e teste dos modelos (JOVIC, BRKIC e BOGUNOVIC, 2014).

Para realização do treino e teste foi utilizada validação cruzada com o conjunto de dados sendo dividido em 10 partes. O critério de escolha do melhor desempenho foi a acurácia média medida nos testes. O resultado da execução inicial dos testes apresentou um desempenho próximo do obtido no artigo original de 2014.

[2] *Dataset* disponível em https://bit.ly/2Yu6aPd no formato CSV.
[3] https://www.cs.waikato.ac.nz/ml/weka/.

A acurácia média e o desvio padrão estão representados no Gráfico 1. Esse é o *baseline* do presente artigo, que tem como objetivo superar essa acurácia através da alteração de parâmetros dos algoritmos. A acurácia média do algoritmo da *baseline* foi de 73,3%.

Gráfico 1 – Acurácia dos algoritmos da baseline



Fonte: autor

Classificadores como o *Support Vector Machine* assumem que os dados estão centralizados em zero e apresentam uma variação de mesma ordem. Se um campo tiver uma variação maior que outro, ele pode dominar a função objetivo e tornar o classificador menos eficiente (PEDREGOSA, VAROQUAUX, *et al.*, 2011). No conjunto de dados utilizado temos o campo idade e renda que apresentam essa característica, Gráfico 1.

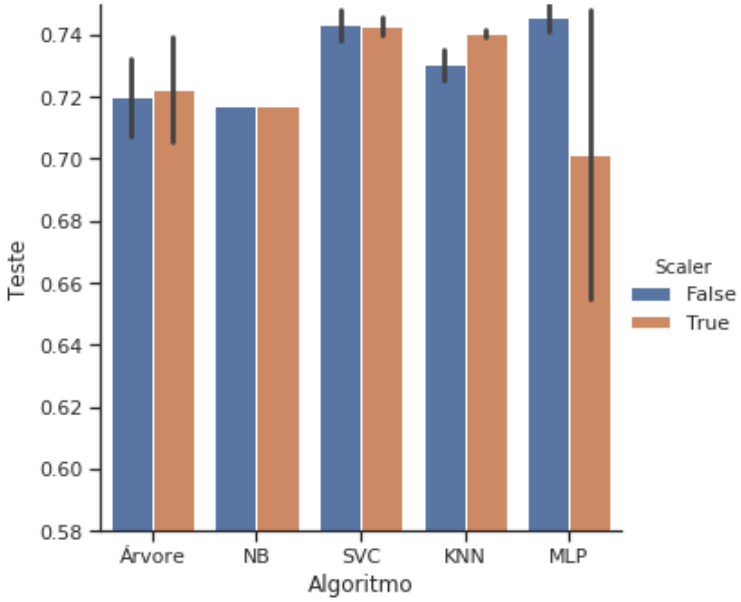
Portanto, a primeira alteração realizada foi escalar os dados utilizando o *StandartScaler*. Essa biblioteca padroniza os dados através da média e do desvio padrão utilizando a Fórmula 1.

$$z = \frac{x-u}{s} \quad (1)$$

Em que *z* é o dado escalado, *x* é o dado original, *u* é a média dos dados de treino e *s* é o desvio padrão (BUITINCK, LOUPPE, *et al.*, 2013).

Com os dados escalados, o desempenho médio apresentou uma melhora nos algoritmos da Árvore de Decisão e KNN. Na estratégia do *Naive Bayes* o desempenho foi o mesmo e tivemos um desempenho médio menor nos outros dois algoritmos. O resultado e comparação com a *baseline* estão representados no Gráfico 2.

Gráfico 2 – Comparação de desempenho com dados escalados e não escalados



Fonte: autor

A acurácia média dos algoritmos executados com os dados escalados foi de 72,6%. O desempenho 1% menor foi puxado pela baixa acurácia do MLP, que por sua vez apresentou uma variação significativa.

3.3 Análise de Desempenho por Cursos

Para encontrar um modelo de predição de evasão mais preciso, foram separados os dados por curso e realizando os testes e treinos de cada um dos algoritmos para cada conjunto de registro.

Com esse novo teste, foi selecionado o algoritmo que apresentou o melhor desempenho para cada curso, pois entendeu-se que os cursos possuem características diferentes que influenciam na sua conclusão. Por exemplo, o curso técnico em química exige uma base maior de exatas, sendo assim o desempenho nas competências de exatas do processo seletivo pode ter um peso maior na conclusão do curso.

Os testes realizados utilizaram as mesmas configurações dos algoritmos da *baseline* com os dados escalados. A Tabela 2 mostra os algoritmos selecionados para cada curso e a acurácia média testada.

Tabela 2 – Melhores algoritmos por curso

Cód. Curso	Curso	Algoritmo	Acurácia
1	Administração	Árvore de Decisão	82,3%
2	Informática	SVC	72,1%
3	Nutrição e Dietética	Árvore de Decisão	74,4%
4	Química	SVC	66,5%

Fonte: autor

Tendo selecionado o algoritmo com melhor performance na configuração da *baseline*, foram exploradas as alterações nos hiperparâmetros de cada algoritmo de aprendizagem de máquina.

Existem inúmeras combinações de hiperparâmetros, testar todas elas de forma manual seriam muito dispendioso. Por isso foi utilizado o *RandomizedSearchCV* que realiza uma busca aleatória para um conjunto de parâmetros especificados e retorna o melhor conjunto de parâmetros para o modelo.

A busca aleatória tem desempenho próximo da busca exaustiva e com um custo computacional bem menor (PEDREGOSA, VAROQUAUX, *et al.*, 2011). Dessa forma foi possível explorar mais possibilidade do que as propostas inicialmente pelo artigo que definiu a *baseline*.

Para o curso de administração, foi selecionado o algoritmo Árvore de Decisão. Foram testadas aleatoriamente 40 combinações de 6 parâmetros e o desempenho foi de 82,4%.

O modelo do curso de informática foi o SVC. Ele foi testado com alteração de quatro parâmetros em um espaço de 20 combinações. O desempenho do classificador SVC para o curso de informática foi de 74,8%.

No curso de nutrição foi selecionado para a análise o algoritmo de Árvore de Decisão. O *RandomizedSearchCV* recebeu 6 parâmetros e testou 40 combinações diferentes. O desempenho do melhor classificador para o curso de nutrição foi de 74,9%. Finalmente, para o curso de química foi utilizado o algoritmo SCV com variação de 4 parâmetros, gerou-se 80 combinações e obteve-se uma acurácia média de 68,7%.

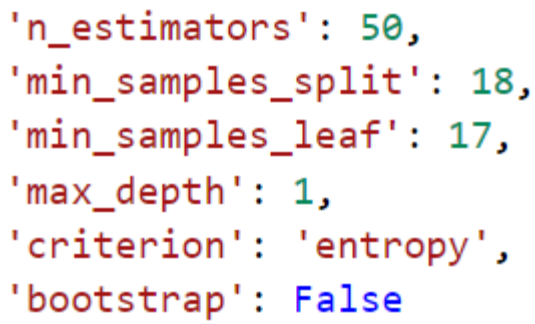
4 RESULTADOS E DISCUSSÃO

Após análise exploratória foram encontrados modelos de predição baseados em aprendizagem de máquina capazes de identificar alunos que não irão concluir o curso. Esses

modelos apresentam uma acurácia melhor do que os modelos apresentados inicialmente por BRITO *et al*. (2014).

No curso de administração a acurácia teve um pequeno aumento, chegando à média de 82,4% quando utilizado o algoritmo da Árvore de Decisão (*RandomForestClassifier*) com os parâmetros listadas na Figura 1.

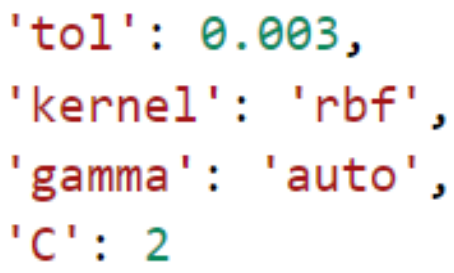
Figura 1 – Parâmetros do melhor modelo para o curso de administração



Fonte: autor

Para o curso de informática, ao utilizar o classificador SVC com as configurações listadas na Figura 2, obtivemos uma melhoria de 3,6% em relação ao *baseline*, alcançando uma acurácia média de 74,8%.

Figura 2 – Parâmetros do melhor modelo para o curso de informática

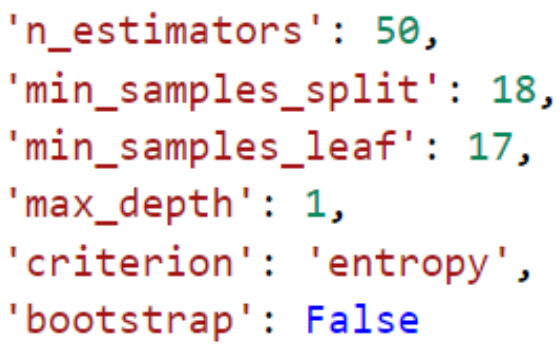


Fonte: autor

O algoritmo do curso de nutrição foi o da Árvore de Decisão (*RandomForestClassifier*), que trouxe uma acurácia média de 74,9% com a configuração de parâmetros exibidos na Figura 3. Este curso teve uma pequena melhoria na acurácia com o uso do *RandomizedSearchCV*.

Isso se deu por conta dos parâmetros selecionados inicialmente já apresentarem um bom desempenho no espaço de testes. Mesmo assim, com a escala dos dados, o resultado foi bem melhor que a média da *baseline*.

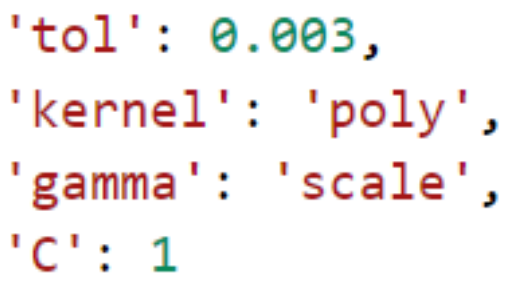
Figura 3 – Parâmetros do melhor modelo para o curso de nutrição



Fonte: autor

Finalmente, o algoritmo utilizado para a classificação dos alunos do curso de química foi o SVC com os parâmetros exibidos na Figura 4. Esse modelo teve acurácia média de 68,7%, que representa uma melhoria de 3,2% em relação ao teste inicial para os dados deste curso (66,5%).

Figura 4 – Parâmetros do melhor modelo para o curso de química



Fonte: autor

O uso de bibliotecas de busca de modelos como o *RandomizedSearchCV* permitiu uma exploração mais ampla das combinações de parâmetros para os modelos selecionais. Dessa forma foi possível melhorar o desempenho dos classificadores, obtendo uma melhor acurácia na predição dos alunos que não irão concluir o curso.

5 CONSIDERAÇÕES FINAIS

O presente trabalho construiu, testou e validou diversos modelos de classificação baseados em aprendizado de máquina. Esses modelos podem ser utilizados para predição de alunos que irão evadir o curso.

Além disso, partindo de uma proposta inicial (*baseline*) e alterando diversos parâmetros, foi possível melhorar de forma significativa o desempenho desses algoritmos, medido através da acurácia média.

Como contribuição a presente pesquisa apresentou os hiperparâmetros utilizados na elaboração desses modelos. Esses hiperparâmetros foram dimensionados para o classificador que melhor se adaptou aos dados de cada curso.

Uma limitação do trabalho foi ter aplicado a busca de hiperparâmetros apenas nos algoritmos que apresentaram melhor performance em cada curso. Como essa performance foi resultado da definição de parâmetros específicos, é possível que outros modelos apresentem um desempenho melhor com outros parâmetros.

Como sugestões de trabalhos futuros que podem complementar essa pesquisa, sugere-se investigar a interferência de cada atributo listado na Tabela 1 no resultado. Recomenda-se também a busca de hiperparâmetros em algoritmos diferentes dos testados nesta pesquisa.

6 REFERÊNCIAS

BRITO, D. M. et al. **Predição de desempenho de alunos do primeiro período baseado nas notas de ingresso utilizando métodos de aprendizagem de máquina**. Simpósio Brasileiro de Informática na Educação, 2014. 882-890.

BUITINCK, L. et al. **API design for machine learning software: experiences from the scikit-learn**. ECML PKDD Workshop: Languages for Data Mining and Machine Learning, p. 108-122, 2013.

CPS. **Mapeamento das Escolas Técnicas**. São Paulo: Centro Paula Souza - Secretaria do Desenvolvimento Econômico do Estado de São Paulo, v. 44, 2019.

INEP. **Sinopse Estatística da Educação Básica 2018**. Brasília: INEP - Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, 2019.

JORDAN, W. A.; LARA, J.; MCPARTLAND, J. M. **Exploring the causes of early dropout among race-ethnic and gender groups**. Youth and Society, p. 62-94, 1996.

JOVIC, A.; BRKIC, K.; BOGUNOVIC, N. **An overview of free software tools for general data mining**. 7th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), p. 1112-1117, 2014.

PEDREGOSA, F. et al. **Scikit-learn: Machine Learning in Python**. Journal of Machine Learning Research, v. 12, p. 2825-2830, 2011.

RUSSELL, S.; NORVIG, P. **Artificial Intelligence: A Modern Approach**. [S.l.]: Prentice Hall, 2010.