# USO DE TÉCNICAS DE APRENDIZAGEM DE MÁQUINA PARA ANÁLISE DA EVASÃO DE CURSOS DE ENSINO SUPERIOR : APLICAÇÃO EM UMA UNIVERSIDADE NO INTERIOR DO ESTADO DE SÃO PAULO

Barbara Celi Braga Camargo [1]

Leandro Guarino de Vasconcelos [2,3]

## RESUMO

A evasão escolar no ensino superior é um problema frequente e que causa diversas perdas em diferentes âmbitos. Com a possibilidade de traçar o caminho anterior à evasão é possível que a instituição possa agir para evitar o abandono por parte dos alunos de seus cursos. As novas tecnologias podem ser usadas a nosso favor na análise das similaridades entre os alunos que evadem, visto que a análise de uma quantidade grande de dados de maneira manual é inviável. A utilização de aprendizagem de máquina em análises educacionais já foi reportada em alguns trabalhos. Neste projeto analisamos os fatores que levam os estudantes a evasão do ensino superior, usando dados referentes aos cursos de uma universidade no interior do estado de São Paulo no decorrer dos últimos dez anos. Utilizando a aprendizagem de máquina foi possível identificar o perfil de alunos evadidos com precisão de 94% usando diversos classificadores a partir de um conjunto de atributos pré-selecionados.

**Palavras-chave:** Evasão escolar, Ensino superior, Aprendizado de máquina, Atributos, Classificadores.

## INTRODUÇÃO

A evasão escolar é um fenômeno social complexo, definido como interrupção no ciclo de estudos (GAIOSO, 2005). É um problema que vem preocupando as instituições de ensino, sejam públicas ou particulares. O abandono do aluno sem a finalização dos

[1] Graduanda do Curso de Análise e Desenvolvimento de Sistema da FATEC- SP, bcbc.fisica@gmail.com;

[2] Professor Doutor da FATEC- SP, le.guarino@gmail.com ;

[3] Projeto de Iniciação Científica financiado pelo Centro Paula Souza-SP.

seus estudos representa uma perda social, de recursos e de tempo de todos os envolvidos no processo de ensino (LOBO, 2011).

Para termos um panorama da situação, o trabalho de Silva Filho (2007) apresenta uma análise no período entre 2000 e 2005, no conjunto formado por todas as Instituições de Ensino Superior (IES) do Brasil. Segundo os autores, a evasão média foi de 22% e atingiu 12% nas públicas e 26% nas particulares.

O aumento quantitativo do número de vagas foi considerável nos últimos anos, mas a sua concentração se deu em maior parte instituições privadas, o que mostra que o acesso à universidade pública ainda é um privilégio de poucos. Apesar do aumento no acesso ao nível superior, o número de profissionais formados ainda é muito baixo quando comparados a outros países. No trabalho de Silva Filho e Hipólito (2009) é apontado que somente 8% da população adulta tem formação superior no Brasil, enquanto outros países apresentaram um percentual maior ainda na década de 1990: Coreia, 32%; Espanha, 28%; Rússia, 55% e Chile, 13%. Ademais, podemos destacar dois fatores para o baixo número de pessoas com formação superior no país, primeiramente a dificuldade no acesso, devido a questões econômicas e educacionais visto que o vestibular é um meio de seleção. Uma vez dentro da universidade os estudantes encontram um segundo problema, o de permanecer até o fim do curso.

Os resultados obtidos por Ferretti e Madeira (1992) apontam que nos últimos dez anos os estudantes relataram a necessidade de trabalhar para conseguir manter-se no ensino superior e ajudar na receita familiar, sendo esse um novo perfil do estudante brasileiro. As classes menos favorecida da sociedade passaram a estudar em Instituições de Ensino Superior da rede privada, já que as instituições públicas, apresentam uma forte concorrência para o ingresso bem como seus horários de cursos que se alternam durante o dia, impossibilitam os alunos com o perfil de trabalhador, de cursá-las (SOUZA et al., 2017).

Entretando, não é apenas as questões financeiras que colaboram para a evasão. Um dos pontos enfatizados por Silva Filho e outros (2007) é o fato de se minimizar frequentemente as razões da evasão, apontando em geral a falta de recursos financeiros do aluno como a principal causa para a interrupção de seus estudos. É importante que se

priorize também a compreensão das questões de ordem acadêmica, como as expectativas do aluno em relação ao curso ou à instituição que podem encorajá-lo ou desestimulá-lo a priorizar a conclusão do seu curso (BAGGI, 2010).

O estudo elaborado por Silva Filho e Lobo (2007) argumenta que para medirmos a evasão seria necessário acompanhar o histórico escolar de cada aluno, pois assim poderíamos identificar quando ele abandonou ou se transferiu de curso. Os autores salientam que os pesquisadores dessa área precisam estudar dados referentes à evasão nas diferentes Instituições de Ensino Superior, tais como, total de matrículas e número de ingressantes e de concluintes (BAGGI, 2010).

As novas tecnologias podem ser usadas a nosso favor na análise das similaridades entre os alunos que evadem de seus cursos, visto que a análise de uma quantidade grande de dados de maneira manual é inviável. A utilização de aprendizagem de máquina em análises educacionais já foi reportada em alguns trabalhos. A aprendizagem de Máquina (AM) ou em inglês *Machine Learning* (ML) é uma sub-área da Inteligência Artificial que estuda métodos computacionais para adquirir novos conhecimentos e novas habilidades. Existem vários métodos de aprendizagem de máquina, como por exemplo, a aprendizagem por hábito, por instrução, por dedução, por analogia e por indução (FRANK et al., 2016).

As técnicas de AM podem ser divididas em duas classes em geral: aprendizagem supervisionada e aprendizagem não supervisionada. A aprendizagem supervisionada consiste no algoritmo de aprendizagem (indutor) receber um conjunto de exemplos, e cada exemplo ser formado por um conjunto de atributos de entrada e atributos de saída, ou seja, o indutor recebe um conjunto de exemplo para treinamento para os quais os rótulos da classe já são conhecidos. Já na aprendizagem não supervisionada, o indutor recebe um conjunto de exemplos formados por conjuntos de atributos de entrada somente, ou seja, o indutor analisa os exemplos fornecidos para tentar agrupá-los de alguma maneira (FRANK et al., 2016). Essas técnicas são utilizadas para buscar e detectar padrões em uma grande quantidade de dados, o que auxilia no entendimento dos mesmos.

Este trabalho teve como objetivo principal encotrar similiaridade que levam os estudantes à evasão do ensino superior, usando dados referentes aos cursos de uma universidade no interior do estado de São Paulo no decorrer dos últimos dez anos.

Como dito, a evasão escolar é um problema frequente no ensino superior, o que causa diversas perdas como a de mão de obra qualificada e o financiamento governamental. Com a possibilidade de traçar o caminho anterior à evasão é possível que a instituição possa se adequar para acolher e diminuir a quantidade de alunos que abandonam seus cursos.

## METODOLOGIA

Para este projeto propomos a utilização de técnicas de aprendizagem de máquina para a caracterização dos alunos que tendem a evadir dos cursos de ensino superior. A Aprendizagem de Máquina (AM) é uma sub-área da Inteligência Artificial que visa usar métodos computacionais para encontrar padrões em uma grande quantidade de dados.

O tema de evasão acadêmica e AM foi estudado por Amorim (2008). Em seu trabalho foi explorado os principais aspectos que podem levar um aluno a trancar ou abandonar seu curso. Nesse artigo, foram implementadas três fases principais para a criação de um sistema de previsão, sendo elas: seleção de atributos, levantamento dos dados e escolha dos classificadores, testando a acurácia dos mesmos e, por último, mostrou as estatísticas referentes à evasão de cada curso. Utilizamos uma metodologia similar a este trabalho.

A metodologia deste projeto consiste na pesquisa bibliográfica referentes aos conceitos que foram utilizados na pesquisa, como linguagem de máquina e evasão escolar. Em conjunto utilizamos dados sobre os alunos dos cursos superiores, coletados nos últimos dez anos em uma universidade do interior de São Paulo, e aplicamos num software que utiliza aprendizado de máquina para a obtenção dos resultados, o *WEKA* (*Waikato Environment for Knowledge Analysis*).

O *WEKA* procede à análise computacional e estatística dos dados fornecidos recorrendo a técnicas de mineração de dados tentando, indutivamente, a partir dos

padrões encontrados gerar hipóteses para soluções e elaborar inclusive teorias sobre os dados em questão (FRANK et al., 2016).

*Quadro 1: Atributos selecionados referente a cada aluno para o estudo de evasão.*

*Fonte: Elaboração Própria.*

| **Atributos Selecionados** |
|---|
| Idade |
| Gênero |
| Raça |
| Cidade |
| Trabalha |
| Curso |
| Noturno/Diurno |
| Período |
| Quantidade de disciplinas com reprovação |
| Nota no vestibular |
| Classificação no vestibular |
| Média de faltas por semestre |
| Entrou por transferência? |
| Aluno já fez curso na FATEC antes? |
| Quantidade de reprovações em PI |
| índice de rendimento acadêmico |
| Quantidade de trancamentos |
| Já desistiu de outros cursos na FATEC? |
| Quantidade de reprovações por falta |
| Média de disciplinas cursadas por semestre. |
| |
| Classe: EVADIDO/MATRICULADO/FORMADO |

Na utilização do *WEKA* para a classificação dos alunos é necessário inserir dados referentes a estes alunos. Visto que a FATEC de Guaratinguetá tem em seu banco de dados várias informações sobre os alunos da instituição nos últimos dez anos, selecionamos diversos atributos que podem ajudar a identificar os fatores que levam a evasão, apresentamos no Quadro 1 os atributos selecionados para o nosso estudo.

Uma vez de posse desses dados, criamos uma tabela em formato CSV e a inserimos no *WEKA*. Iniciamos os testes com diversos classificadores e avaliamos a precisão. O método que exploramos foi o método supervisionado chamado árvore de decisão. As árvores na computação são estruturas de dados formadas por um conjunto de elementos que armazenam informações chamadas nós. Os nós representam regiões onde são realizados testes lógicos para a separação dos dados - o primeiro nó é chamado de nó raiz e é o nó principal da árvore de decisão (SATO et al., 2013).

## RESULTADOS E DISCUSSÃO

Diversos atributos e diferentes algoritmos foram testados, a fim de encontrar os resultados mais precisos para a solução do problema, que no nosso caso é identificar alunos em potencial de evasão. Na Figura 1, são apresentados os melhores resultados dos algoritmos, com destaque em amarelo, azul e verde, iniciando pelo melhor resultado.

| Classificador | Precisão | Evadidos | Formados |
|---|---|---|---|
| .meta.Bagging | 94.6711 % | 888 | 551 |
| .meta.RandomCommittee | 94.4737 % | 879 | 557 |
| .meta.ViaRegression | 94.8026 % | 879 | 562 |
| .trees.RandomForest | 94.3421 % | 874 | 560 |
| .trees.REPTree | 94.0789 % | 879 | 551 |
| .trees.J48 | 94.5395 % | 874 | 563 |
| .trees.LMT | 94.0789 % | 867 | 563 |
| .rules.JRip | 94.2105 % | 881 | 551 |

*Figura 1 : Resultados referente aos testes utilizando como atributo percentual de rendimento dos alunos. Fonte: Elaboração Própria.*

Os resultados apresentados na Figura 1 levam em consideração várias características dos alunos, como idade, gênero, cidade, quantidade de disciplinas cursadas no semestre, quantidade de reprovações, quantidade de reprovações em projetos interdisciplinares, entre outros. Neste primeiro caso, foi utilizado também o percentual de rendimento dos alunos. Pode-se ver que a precisão utilizando os classificadores indicados é acima de 94%. Ou seja, é possível identificar o perfil de alunos evadidos com precisão de 94%.

O classificador *REPTree* teve um bom percentual de acerto no número de alunos que evadiram o curso. Como o principal objetivo do nosso trabalho era identificar os alunos que podem abandonar o curso, exploramos a árvore de decisão obtida pelo classificador, Figura 2.

```
REPTree
============

pr < 59.59
|   pr < 51.31 : EVADIDO  (1259/5) [638/4]
|   pr >= 51.31
|   |   qtde_disciplinas_reprovacao < 15.5 : EVADIDO  (159/10) [74/4]
|   |   qtde_disciplinas_reprovacao >= 15.5
|   |   |   qtde_trancamentos < 0.5
|   |   |   |   classificacao_vestibular < 7 : EVADIDO  (2/0) [2/1]
|   |   |   |   classificacao_vestibular >= 7
|   |   |   |   |   idade < 31.5 : FORMADO  (16/0) [13/3]
|   |   |   |   |   idade >= 31.5
|   |   |   |   |   |   afrodescendencia = N : FORMADO  (4/0) [1.67/0.67]
|   |   |   |   |   |   afrodescendencia = S : EVADIDO  (2/0) [1.33/0]
|   |   |   qtde_trancamentos >= 0.5 : EVADIDO  (23/6) [10/4]
pr >= 59.59
|   qtde_trancamentos < 1.5
|   |   media_disciplinas_semestre < 10.41
|   |   |   media_disciplinas_semestre < 5.15 : EVADIDO  (14/0) [10/2]
|   |   |   media_disciplinas_semestre >= 5.15
|   |   |   |   cidade = LORENA : FORMADO  (208/5) [107/7]
|   |   |   |   cidade = GUARATINGUETA
|   |   |   |   |   ja_fez_fatec = N
|   |   |   |   |   |   media_disciplinas_semestre < 7.54 : FORMADO  (128/10) [54/3]
|   |   |   |   |   |   media_disciplinas_semestre >= 7.54
|   |   |   |   |   |   |   pr < 73.31
|   |   |   |   |   |   |   |   qtde_disciplinas_reprovacao < 0.5 : EVADIDO  (2/0) [3/0]
|   |   |   |   |   |   |   |   qtde_disciplinas_reprovacao >= 0.5 : FORMADO  (76/2) [34/0]
|   |   |   |   |   |   |   pr >= 73.31 : FORMADO  (356/2) [174/3]
|   |   |   |   |   ja_fez_fatec = S : FORMADO  (17/6) [13/2]
|   |   |   |   cidade = TAUBATE : FORMADO  (10/0) [4/0]
|   |   |   |   cidade = APARECIDA
```

*Figura 2: Parte da árvore de decisão gerada a partir do classificador REPTree. Fonte: Elaboração Própria.*

O algoritmo *REPTree*, constrói árvores de decisão para classificação ou regressão com base no ganho de informação/variância e poda esta árvore usando uma poda guiada por erro (WITTEN et al., 2011). No caso apresentado na Figura 2, vimos os atributos utilizados na classificação. Notamos que o índice de rendimento é o primeiro indicador para evasão. A quantidade de disciplinas reprovadas, cursadas e o trancamento do curso também foram utilizados na classificação. Um outro atributo importante na análise foi a cidade em que o aluno residia.

Um algoritmo pode ter boa precisão, contudo pode não apresentar uma boa precisão na análise do dado que queremos avaliar. Para analisarmos de maneira mais precisa os resultados é necessário avaliar a matriz de confusão gerada pelo sistema, Figura 3.

```
=== Confusion Matrix ===

   a   b   <-- classified as
 879  29 |   a = EVADIDO
  61 551 |   b = FORMADO
```

*Figura 3: Matriz de confusão gerada a partir do classificador REPTree. Fonte: Elaboração Própria.*

Na matriz de confusão gerada pelo *REPTree*, vemos que 879 evadidos foram classificados corretamente enquanto 29 foram classificados incorretamente. No caso de alunos formados, 551 alunos foram classificados de maneira correta e apenas 61, de forma incorreta.

Realizando testes sem o atributo percentual de rendimento, foram alcançados resultados excelentes, mas com percentual abaixo do primeiro caso apresentado. A Figura 4 mostra os resultados referentes a esse segundo conjunto de casos.

| Classificador | Precisão | Evadidos | Formados |
|---|---|---|---|
| .meta.AdaBoostM1 | 85.6579 % | 810 | 492 |
| .meta.Bagging | 88.6184 % | 826 | 521 |
| .meta.FilteredClassifier | 88.75 % | 850 | 499 |
| .meta.ViaRegression | 89.2763 % | 835 | 522 |
| .trees.RandomForest | 88.0263 % | 823 | 515 |
| .trees.REPTree | 88.0921 % | 849 | 490 |
| .trees.J48 | 89.2763 % | 833 | 524 |
| .rules.JRip | 87.8947 % | 826 | 510 |
| | | | |

*Figura 4 : Resultados referente aos testes não utilizando como atributo percentual de rendimento dos alunos. Fonte: Elaboração Própria.*

Mesmo sem a utilização do percentual de rendimento, conseguimos uma precisão de mais de 88% na maioria dos classificadores testados. Esse resultado mostra que, mesmo sem todos os atributos selecionados no Quadro 1, podemos encontrar resultados excelentes para a previsão de evasão de alunos, podendo extrapolar esse resultado para outras instituições de ensino.

Para a medição da acurácia do nosso método, aplicamos a classificação para todos os semestres entre 2014 e 2018. Na Figura 5, vemos os valores de acurácia de cada semestre em porcentagem. Notamos que apenas no segundo semestre de 2014 temos uma acurácia abaixo de 50% de acerto, em todos os outros casos analisados os valores de acurácia são acima de 60%. Neste caso, mesmo o valor com 50% de acerto, temos um resultado favorável, visto que nosso objetivo é que a universidade saiba os alunos em pontencial desistência e possa agir antes do abandono.

| Ano | Acurácia final |
|---|---|
| 2014-2 | 40,54 |
| 2015-1 | 60,82 |
| 2015-2 | 82,52 |
| 2016-1 | 70,23 |
| 2016-2 | 68,60 |
| 2017-1 | 76,47 |
| 2017-2 | 75,79 |
| 2018-1 | 74,23 |
| 2018-2 | 68,00 |

*Figura 5: Acurácia em porcentagem da previsão por semestre entre os anos de 2014 e 2018. Fonte: Elaboração Própria.*

Com os resultados obtidos, é possível aplicar a árvore de decisão nos dados de alunos atualmente matriculados, a fim de prever quais alunos irão evadir do curso e, consequentemente, ajudar a faculdade a diminuir o número de evasão.

## CONSIDERAÇÕES FINAIS

Neste trabalho analisamos os fatores que podem levar estudantes do ensino superior a abandonar seus cursos. Para isso utilizamos pesquisas anteriores e aplicamos métodos de aprendizagem de máquina num conjunto de dados obtidos por uma universidade no interior do estado de São Paulo nos últimos dez anos. Utilizando essa metodologia conseguimos resultados muito bons, chegando a precisão de mais de 94% de acerto na previsão com diferentes classificadores. Retirando alguns atributos fizemos outro teste, onde conseguimos uma precisão de 88% mesmo com a diminuição de dados. Nossa análise mostra que este modelo pode ser aplicável a diversas instituições mesmo com quantidade reduzida de dados sobre os alunos. Em nosso trabalho futuro iremos implementar esse método em uma plataforma interativa capaz de prever em tempo real alunos que têm possibilidade de abandonar seu curso.

## REFERÊNCIAS

AMORIM, J. V. M.. Técnicas de Aprendizado de Máquina Aplicada na Previsão de Evasão Acadêmica. **Simpósio brasileiro de informática na educação**, 2008.

BAGGI, C. A. S. . Evasão e avaliação institucional: uma discussão bibliográfica Pontifícia Universidade Católica de Campinas. **Dissertação (Mestrado em Educação) - Pontifícia Universidade Católica de Campinas.** Campinas. 2010.

FRANK, E., HALL, M. A. e WITTEN, I. H.. The WEKA Workbench. **Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques"**, Morgan Kaufmann, Fourth Edition, 2016.

GAIOSO, N. P. L.. O fenômeno da evasão escolar na educação superior no Brasil. **Dissertação (Mestrado em Educação) – Programa de Pós-Graduação em Educação da Universidade Católica de Brasília**, Brasília, 2005.

LOBO, M. B. C. M.. Panorama da evasão no ensino superior brasileiro: aspectos gerais das causas e soluções. **Instituto Lobo / Lobo & Associados Consultoria**. 2011.

SATO, L. Y.; SHIMABUKURO, Y. E.; KUPLICH, T. M.; GOMES, V. C. F. Análise comparativa de algoritmos de árvore de decisão do sistema WEKA para classificação do uso e cobertura da terra. **Simpósio brasileiro de sensoriamento remoto**, 16. (SBSR), 2013.

SILVA FILHO, R. L. L. . A evasão no ensino superior brasileiro. **Cadernos de Pesquisa, São Paulo**, v. 37, n. 132, p. 641-659, 2007.

SILVA FILHO, R. L. L.; HIPÓLITO, O. Financiamento e expansão do ensino superior. Disponível em: **http://www.jornaldaciência.org.br/Detalhe.jsp?id=62770**. Acesso em: 16 abr. 2009.

SOUZA, C., da SILVA, C. & GESSINGER, R. . Um estudo sobre a evasão no ensino superior do Brasil nos últimos dez anos. ***Congresso CLABES***,2017.

WITTEN et al. **Data Mining: Practical Machine Learning Tools and Techniques,** Morgan Kaufmann, Burlington, MA, 2011.