Universidade Federal do Parana
Setor de Ciências Exatas
Departamento de Estatística
Programa de Especialização em Data Science e Big Data

Guilherme Matos Barbosa

ÁRVORE DE DECISÃO APLICADO A EVASÃO ESCOLAR

Curitiba
2022

Guilherme Matos Barbosa

ÁRVORE DE DECISÃO APLICADO A EVASÃO ESCOLAR

Monografia apresentada ao Programa de Especialização em *Data Science* e *Big Data* da Universidade Federal do Paraná como requisito parcial para a obtenção do grau de especialista.

Orientador: Prof. Dr. Wagner Hugo Bonat

Curitiba
2022

Especialização em Data Science e Big Data
Universidade Federal do Paraná
dsbd.leg.ufpr.br

ÁRVORE DE DECISÃO APLICADO A EVASÃO ESCOLAR

Guilherme Matos Barbosa[1]
Wagner Hugo Bonat[2]

Resumo

A evasão escolar pode ser definida como a descontinuação de um estudante no seu ensino, podendo ocorrer de diferentes formas e, esse fenômeno, está cada vez mais presente no cenário do ensino superior. Torna-se importante então buscar maneiras de reduzir a taxa de evasão. As técnicas de aprendizagem de máquina permitem identificar padrões e gerar modelos computacionais que podem predizer se um aluno será um evasor. Este trabalho usou modelos de classificação de árvore de decisão e *randon forest* para predizer quais alunos estão em risco de evasão, assim como identificar os atributos mais determinantes utilizando dados do SIGA (Sistema de Gestão Acadêmica) da Universidade Federal do Paraná, que contempla informações sobre matrícula, cursos, professores, disciplinas, frequências, notas, entre outras. As variáveis mais significativas foram índice de rendimento acadêmico, carga curricular e setor de estudo. O melhor resultado de predição foi obtido pelo algoritmo *randon forest* com uma acurácia de 0,78 e AUC de 0,87.
Palavras-chave: Evasão escolar. Mineração de dados. Aprendizagem de máquina.

Abstract

School dropout can be defined as the discontinuation of a student's education, which can occur in different ways, and this phenomenon is increasingly present in the higher education scenario. It is therefore important to look for ways to reduce the dropout rate. Machine learning techniques allow us to identify patterns and generate computational models that can predict whether a student will be a truant. This work used decision tree and randon forest classification models to predict which students are at risk of dropping out, as well as to identify the most determinant attributes using data from SIGA (Academic Management System) of the Federal University of Paraná, which includes information about enrollment, courses, professors, subjects, frequencies, grades, among others. The most significant variables were academic performance index, course load, and study sector. The best prediction result was obtained by the randon forest algorithm with an accuracy of 0.78 and AUC of 0.87.
***Keywrods**: School dropout. Data mining. Machine learning.*

1 Introdução

A evasão escolar é caracterizada pela descontinuação da trajetória de estudos de um indivíduo e está cada vez mais presente nas instituições de ensino. Esse fenômeno é complexo e multifatorial e afeta negativamente as instituições [1]. Diante disso, as instituições estão cada vez mais focadas em combater e prevenir que os alunos evadam [2].

Os dados sobre a evasão não tem uma emissão oficial, isso se deve à difícil caracterização do aluno que evade, pois ele pode deixar de frequentar um curso mas permanecer na instituição, pode abandonar na totalidade mas ainda ter vínculo com outra, e, pode também, apenas deixar de frequentar uma disciplina [3].

A evasão é algo que gera prejuízos tanto na esfera pública quanto privada. No setor público reflete como uma perda de receita investida na formação de mão de obra, que não se tornará qualificada, tem o agravante social, pois os alunos evadidos indiretamente impedem outros alunos de adentrarem no sistema ao ocuparem vagas mas não concluírem o curso e nas instituições privadas acarreta principalmente em um menor lucro, além disso, há desgaste emocional do aluno evadido em ambas as esferas [4].

Contudo, estudos vêm sendo realizados como uma medida de prevenção à evasão, a partir da utilização de dados fornecidos pelas próprias instituições. Esses dados são aplicados a técnicas de mineração de dados e aprendizagem de máquina (AM) uma vez que essas ferramentas conseguem transformar um grande volume de dados em informações relevantes [2].

Através das técnicas da AM é possível identificar padrões que podem ser utilizados para gerar um modelo computacional propiciando assim classificar se um aluno é provável concluinte ou provável evasor baseado em suas características e nas do curso em que está inserido. Além disso, algumas permitem realizar uma avaliação interna do modelo facilitando identificar os atributos mais determinantes na classificação como, por exemplo, árvores de decisão e *random forest* [2].

Diante do exposto, este trabalho teve como objetivo utilizar árvore de decisão e *random forest* para classificar

[1]Aluno do programa de Especialização em Data Science & Big Data, guilhermematos@ufpr.br.
[2]Professor do Departamento de Estatística - DEST/UFPR, wbonat@ufpr.br.

alunos com potencial de evasão além de quantificar e qualificar os atributos mais determinantes utilizando os dados coletados do Sistema de Gestão Acadêmica da Universidade Federal do Paraná (SIGA/UFPR).

2 Metodologia

Este capítulo apresenta os procedimentos metodológicos da pesquisa com o intuito de atingir os objetivos do trabalho.

A figura 1 é um esquema com todas as etapas do processo realizado que se inicia com a caracterização de um problema ou de uma ideia, neste trabalho trata-se da evasão no ensino superior seguida da aplicação das técnicas de mineração e avaliação dos resultados.

Figura 1: Etapas do Processo.

Com o objetivo bem definido, obtém-se o conjunto de dados, com eles selecionados, limpos e transformados é realizada então a mineração de dados, onde são aplicadas as técnicas de mineração e aprendizagem de máquina. Na etapa de avaliação dos resultados é realizado a interpretação do modelo extraído como, por exemplo, se o modelo de classificação gerado consegue prever com elevada acurácia a probabilidade de um estudante evadir da instituição ou analisar se ainda é preciso melhorar de alguma forma.

2.1 O conjunto de dados

Os dados utilizados são da base de dados do SIGA (Sistema Integrado de Gestão Acadêmica) trata-se de um sistema desenvolvido pela Universidade Federal do Paraná, esses dados tem um volume crescente e incluem: matrícula, cursos, professores, disciplinas, frequências, notas, situação do aluno no curso, caso tenha evadido qual forma de evasão entre outras informações e possibilita realizar diversas investigações em busca de padrões e informações relacionadas com a condição do discente.

As informações foram coletadas através de uma consulta em SQL (*Structured Query Language*) diretamente da base de dados, as informações sensíveis não foram selecionadas impossibilitando assim a identificação de qualquer aluno.

Os dados analisados neste estudo são referentes a 6.901 registros, sendo 3861 formados e 3040 evadidos. Cada registro é a última informação de cada aluno em cada curso, dessa maneira cada estudante possui no máximo um vínculo com cada curso.

2.2 Limpeza e preparo dos dados

Primeiramente, na variável que indica qual a forma de evasão foi identificado classificações de alunos que fugiam do objetivo da pesquisa ou que teriam pouca interferência no resultado, como, por exemplo, alunos com as seguintes formas de evasão: "Desistência Vestibular/PROVAR", "Falecimento", "Mudança de Campus/Habilitação interna", "Não Confirmação de Vaga", "Término de intercâmbio" e para a maioria dos que estavam como "Término de Registro Temporário".

Depois foi realizada uma etapa de seleção de variáveis, em que foram removidas aquelas que não faziam sentido para o contexto, que apresentavam muitos valores nulos ou inválidos, ou apresentavam baixíssima variabilidade entre os valores.

Em seguida para corrigir problemas relacionados a dados faltantes ou *outliers* as variáveis foram categorizadas tendo como base o cálculo do InfoValue (IV) estatística que permite categorizar e avaliar o poder de discriminação de uma variável, de modo que cada categoria de cada variável tenha no mínimo 5% de representatividade em relação ao todo.

Outro ponto foi a criação de uma nova variável para saber a idade do aluno ao ingressar no curso através da data de nascimento do aluno e da data de entrada no curso.

Desta forma, após a limpeza do conjunto de dados, para modelagem foram consideradas 15 variáveis conforme a tabela 1.

Tabela 1: Dicionário do conjunto de dados.

Variável	Descrição
status	Situação do aluno no curso
sexo	Sexo do aluno.
racaCor	Raça/Cor do aluno.
cota	Descrição da cota do aluno.
turno	Turno do curso.
periIngresso	Período de Ingresso.
formaIngresso	Forma de Ingresso.
setor	Setor que o aluno pertence.
idaIngresso	Idade de Ingresso no Curso.
paisNascimento	País de Nascimento.
estaNascimento	Estado de Nascimento.
ira	Índice de Rendimento Acadêmico no 1° semestre.
repNota	Quantidade de Reprovações por Nota no 1° semestre.
repFrequencia	Quantidade de Reprovações por Frequência no 1° semestre.
chCurriculo	Carga Curricular cursada no 1° semestre.

2.3 Modelos empregados

2.3.1 Árvore de Decisão

A árvore de decisão são embasados em uma ação de divisão e conquista, sua estrutura representa um número de possíveis caminhos de decisão e um resultado para cada caminho. Um atributo é selecionado para servir de raiz da árvore e os ramos são criados a partir de cada valor do atributo selecionado avaliando uma regra predefinida, uma condição ou operação matemática, as folhas da árvore representam o valor previsto. O trajeto desde a raiz até à folha corresponde a uma regra de classificação [5][3].

São exemplos de algoritmos de árvore de decisão CHAID, CTree, C4.5, CART e Hoeffding Tree. Esses diferentes algoritmos são parecidos entre si, a principal diferença é como as variáveis são selecionadas e o critério de particionamento e de parada para o crescimento da árvore [6].

2.3.2 *Random Forest*

O método *random forest*, em sua abordagem quanto um classificador, gera um modelo que é basicamente um agrupamento de árvores de decisão, onde cada árvore possui características próprias [2].

O algoritmo consiste na construção de um número grande de árvores de decisão não correlacionadas, usadas no final para a construção de um modelo final. Tem a vantagem de através do modelo identificar as variáveis mais importantes, ser flexível e com boa acurácia [7].

2.4 Métricas

A *matriz de confusão* é um teste para comparar qual classificação o algoritmo previu com a classificação real da instância, fornece um detalhamento do desempenho do modelo. Os termos utilizados na composição de uma matriz de confusão são:

- Verdadeiro Positivo (TP): número de exemplos positivos classificados corretamente;
- Falso Negativo (FN): número de exemplos negativos classificados incorretamente;
- Falso Positivo (FP): número de exemplos positivos classificados incorretamente;
- Verdadeiro Negativo (TN): número de exemplos negativos classificados corretamente.

A *acurácia* representa a taxa de acertos em relação ao total de amostras.

$$\text{Acurácia: } A = \frac{TP + TN}{TP + FP + TN + FN}.$$

A *sensibilidade* representa a taxa de verdadeiros positivos corretamente classificados.

$$\text{Sensibilidade: } R = \frac{TP}{TP + FN}.$$

A *especificidade* representa a taxa de falsos positivos corretamente classificados.

$$\text{Especificidade: } E = \frac{TN}{TN + FP}.$$

A *precisão* representa a taxa de objetos positivos classificados corretamente, podendo ser um falso positivo ou um verdadeiro positivo.

$$\text{Precisão: } P = \frac{TP}{TP + FP}.$$

O *F1* é a média harmônica entre precisão e a sensibilidade.

$$\text{F1: } F = 2 \cdot \frac{P \cdot R}{P + R}.$$

A curva *ROC* (*Receiver Operating Characteristic*) é um gráfico que mostra o desempenho dos modelos, representamos a sensibilidade no eixo y e 1 - especificidade no eixo x.

Para resumir a qualidade mensurada pela curva, é comum a utilização da métrica *AUC* (*Area Under the Curve*), um algoritmo que calcula a área sob a curva ROC, podendo assim comparar os classificadores utilizando um único escalar, a área abaixo da curva *ROC*.

3 Resultados e Discussões

Foram desenvolvidos modelos para estimar a propensão a evasão dos alunos e determinar as variáveis mais determinantes utilizando as técnicas de modelagem de árvore de decisão (CART, CTree) e *random forest*.

No processo de modelagem, após a composição e tratamento do conjunto de dados, o mesmo foi dividido em duas partes, sendo 70% da base para treino e 30% para teste.

A base de treino é submetida ao classificador para treinamento do modelo, que é calibrado conforme os dados apresentados. A estratégia definida de treinamento foi com validação cruzada 10-fold e 3 repetições e a métrica de AUC para avaliação dos resultados.

Após esta etapa, apresentam-se os exemplos da base de teste para o modelo, que deverá realizar a predição de suas classes.

O ponto de corte foi de 0.8, assim gerou-se a matriz de confusão, para este estudo o número de verdadeiro positivo é os discentes que realmente evadiram, e ao realizar a predição foi classificado como evadido. O número de verdadeiro negativo é os discentes que não evadiram, e ao realizar a predição foi classificado como formado.

Através dos resultados da matriz de confusão foi possível obter as métricas dos modelos criados. Na tabela 2 é apresentado os resultados, ordenados por AUC.

Tabela 2: Comparação dos modelos do experimento.

	Random Forest	CTree	CART
AUC	0,87	0,84	0,80
Sensibilidade	0,66	0,96	0,93
Especificidade	0,87	0,47	0,49
Acurácia	0,78	0,73	0,72
Precisão	0,80	0,69	0,68
F1-Score	0,72	0,80	0,79

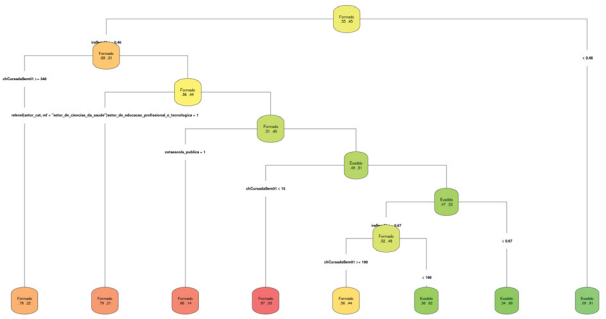

Figura 2: Árvore gerada do modelo CART.

A AUC que foi a métrica utilizada para encontrar o melhor modelo nessa fase de comparação todos obtiveram um valor expressivo entre 0,80 até 0,87, onde o *random forest* foi o melhor.

A métrica sensibilidade que neste trabalho é de grande valia, tendo em vista que é mais problemático deixar de acompanhar os alunos com probabilidade de evadir do que acompanhar os com probabilidade de formar obteve-se valores acima de 0,9 para os modelos CTree e CART.

Analisando os resultados dos modelos nas predições realizadas, percebe-se que, foi alcançado bons resultados nas métricas calculadas, indicando que os modelos provavelmente apresentarão bons resultados preditivos em dados não vistos.

Os resultados das árvores de decisão foram bons, mas inferiores aos do *random forest* . Por ser um algoritmo mais simples, a construção de uma única árvore de decisão apresenta um custo computacional menor, mas um resultado inferior.

O modelo construído pelo *random forest* foi o que obteve o melhor desempenho e geral. Este classificador é o que permite a escolha de um critério mais alto de aceitação. Mesmo com um resultado para sensibilidade um pouco mais baixo, o método se mostrou mais estável e confiável considerando todas as outras métricas, que mostram eficácia tanto em suas previsões de evasão quanto de sucesso.

A grande vantagem destas técnicas é que o modelo criado permite a avaliação direta das características mais importantes para a classificação de um estudante. A fim de discutir sobre os fatores que, potencialmente, mais influenciam na distinção entre alunos evadidos e formados, foi gerado árvores de decisão, por exemplo, a figura 2, e também utilizando o *random forest* foi gerado gráfico da importância das variáveis para o modelo como mostra a figura 3.

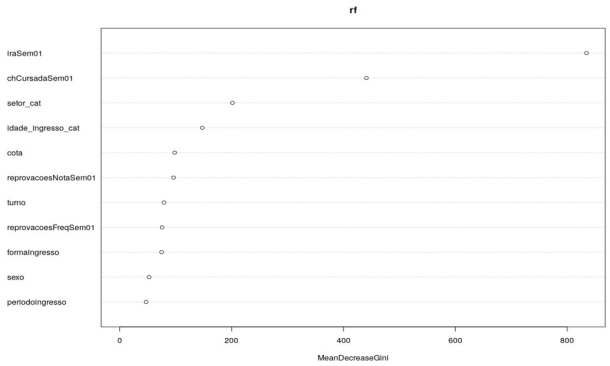

Figura 3: Gráfico da importância das variáveis para o modelo *random forest*.

Através das figuras 2 e 3 pode-se analisar que as variáveis mais importantes para os modelos foram iguais tendo o índice de rendimento acadêmico, a carga horária cursada e o setor como as principais variáveis.

4 Conclusões

A evasão tem impacto em toda a comunidade escolar e não é fácil de ser combatida. As vagas desocupadas no sistema de ensino geram um desperdício de recursos, sendo assim, identificar padrões de um aluno que possivelmente não será um concluinte pode ser uma forma de enfrentar essa problemática.

As tecnologias de aprendizado de máquina estão crescendo consideravelmente nos últimos anos, principalmente no meio educacional. Nesse sentido, o principal objetivo tem sido gerar conhecimento a partir dos milhares de registros dos bancos de dados de sistemas e dar subsídios para pesquisas, que, no que lhe concerne, podem ser voltadas para as políticas de combate à evasão.

Neste trabalho com os modelos testados foi possível verificar quais as principais causas de evasão e também as variáveis que mais contribuem e apresentam maior risco para tal evento.

Também foi possível comprovar que as técnicas de AM podem ser utilizadas de forma satisfatória para a

mineração de dados, mostrando que a árvore de decisão (CART, CTree) e *random forest* são classificadores que possibilitam que a tarefa de identificar os alunos passíveis de evasão possa ser realizada de forma automática.

Agradecimentos

Agradeço ao Prof. Wagner Hugo Bonat pela dedicação e compromisso para realização desse trabalho. A minha família, principalmente minha namorada Danielle por todo suporte e ajuda para realização desse e outros trabalhos. E por fim, aos docentes, por viabilizarem essa oportunidade de qualificação profissional.

Referências

[1] G. L. Santos and P. B. V. Guimarães. *Governo Digital - uma abordagem interdisciplinar na gestão da educacão superior*. Editora Motres, 2019.

[2] L. A. Teodoro and M. A. A. Kappel. *Aplicação de Técnicas de Aprendizado de Máquina Para Predição de Risco de Evasão Escolar em Instituições Públicas de Ensino Superior no Brasil*. Revista Brasileira de Informatica na Educação- RBIE, 2020.

[3] L. R. S. Assis. *Perfil de Evasão no Ensino Superior Brasileiro: uma Abordagem de Mineração de Dados*. Mestrado Profissional em Computação Aplicada - Universidade de Brasília, 2017.

[4] I. L. Hoffmann, R. C. M. Nunes, and F. Martins. *As informações do Censo da Educação Superior na implementação da gestão do conhecimento organizacional sobre evasão*, volume 26. 2019.

[5] A. M. Souza. *Machine learning e a evasão escolar - Análise preditiva no suporte à tomada de decisão*. Mestrado em Sistemas de Informação e Gestão do Conhecimento - Universidade FUMEC, 2020.

[6] T. Escovedo and A. Koshiyama. *Introdução a Data Science - Algoritmos de Machine Learning e Métodos de Análise*. Casa do Código, 2021.

[7] R. L. de Carvalho A. T. R. da Silva L. C. C. P. Soares, R. A. Ronzani. *Aplicação de técnicas de aprendizado de máquina em um contexto ccadêmico com foco na identificação dos alunos evadidos e não evadidos*. Revista Humanidades e Inovação, 2020.