Uma Aplicação Web para Predição de Desempenho de Alunos

Márcio Pereira[1,4], **Alana Oliveira**[2,4], **Mario M. Teixeira**[3,4], **Carlos de Salles S. Neto** [3,4]

[1] Bacharelado Interdisciplinar em Ciência e Tecnologia – BICT

[2] Coordenação de Engenharia da Computação – ECP

[3]Departamento de Informática – DEINF

[4]Universidade Federal do Maranhão
Av. dos Portugueses, 1966 - Campus do Bacanga
CEP: 65080-805 – São Luís - MA

marcio_goncalves@outlook.com, alana@ecp.ufma.br, {mario,csalles}@deinf.ufma.br

Abstract. *This article describes the development of a web application for students' academic performance. The article deals with the main phases of Data Mining (data selection, attribute selection, predictive model evaluation). The predictive model receives as input behavioral data entry and provides a student performance estimate using three prediction techniques in a user-friendly web interface.*

Resumo. *Este artigo descreve a criação de uma aplicação web para prever desempenho acadêmico de alunos. O artigo aborda as principais fases de Mineração de Dados (obtenção dos dados, seleção dos atributos, avaliação dos modelos preditivos). O modelo preditivo recebe como entrada aspectos comportamentais a fim de fornecer a previsão de aproveitamento usando três técnicas de predição diferentes, em uma interface web de fácil uso.*

1. Introdução

Para [Romero and Ventura 2007], a tomada de decisão nos processos de aprendizagem envolvem observar o comportamento do aluno, analisar os dados históricos e estimar a eficácia das estratégias pedagógicas adotadas. No entanto, para [Romero and Ventura 2007], quando os alunos trabalham em ambientes virtuais, esse monitoramento informal não é possível e os educadores devem procurar outras maneiras de obter essa informação.

O ambiente tradicional de sala de aula é, ainda, o modelo de sistema educativo mais utilizado no mundo. Dentro das salas de aula convencionais, os professores tentam melhorar o desempenho do aluno monitorando o processo de aprendizagem e analisando sua performance através de registros em papel e observação individual. Além disso, os educadores podem usar outras informações on-line (páginas da web on-line e páginas de conteúdo do curso), bancos de dados multimídias, etc., mas analisar e fazer o monitoramento desse volume de interações entre professor, alunos e conteúdo é uma importante e complexa tarefa do educador no processo de ensino-aprendizagem.

A mineração de dados pode ajudar nesse processo de aprendizado. Assim, é possível compreender de forma mais eficaz e adequada os alunos, como eles aprendem, o

contexto na qual a aprendizagem ocorre, além de outros fatores que influenciam a aprendizagem [Baker et al. 2011]. Por exemplo, é possível predizer o desempenho de um aluno baseado em alguns atributos comportamentais e usar os resultados para personalizar o ambiente e os métodos de ensino para aquele determinado aluno.

Diante disso, o presente estudo objetiva descrever a criação de uma aplicação web para predizer o desempenho de alunos, através de Mineração de Dados Educacionais por meio de três algoritmos de classificação. Com essa aplicação, pretende-se predizer o desempenho acadêmico do aluno no decorrer do curso, mediante seu nível de participação e interação.

Este trabalho está estruturado da seguinte forma: a seção 2 aborda as noções que fundamentam o trabalho, abordando conceitos de Mineração de Dados Educacionais, bem como a importância e técnicas da mineração de dados educacionais. A seção 3 descreve o procedimento metodológico utilizado. A seção 4 apresenta os resultados. Finaliza-se com a seção 5 ao apresentar a conclusão.

2. Mineração de dados Educacionais

Para [Baker et al. 2011], o termo Mineração de dados, é conhecido como Descoberta de Conhecimentos em Bancos de Dados, ou KDD (do inglês, "Knowledge Discovery in Databases"). O termo faz referência a disciplina que tem como objetivo encontrar novas informações através da análise de grande volume de dados. A mineração de dados é uma das ferramentas tecnológicas mais promissoras na atualidade, ela visa permitir a coleta e o armazenamento de uma grande quantidade de informações que, ao serem analisadas, podem se tornar uma valiosa fonte para as instituições, bem como para a comunidade científica [Rabelo et al. 2017].

A Mineração de Dados Educacionais (do inglês, "Educational Data Mining", ou EDM) utiliza métodos e técnicas de aprendizado de máquina tradicional e estatístico para explorar e analisar dados originados no contexto educacional. O objetivo principal dessa abordagem é analisar as diferentes variáveis envolvidas no processo de ensino aprendizagem e utilizá-las para elaborar modelos preditivos, a fim de fazer a classificação de alunos quanto ao seu desempenho [Fernandes et al. 2019].

Segundo [Baker et al. 2011], a EDM é definida por como uma área de pesquisa que tem como principal objetivo o desenvolvimento de métodos para explorar conjuntos de dados coletados em ambientes educacionais. Com tais métodos visa-se, por exemplo, entender melhor o estudante no seu processo de aprendizagem, analisando sua interação com o ambiente [Costa et al. 2013].

Existem muitos métodos utilizados em EDM que são originalmente da área de mineração de dados [Witten et al. 2016]. Contudo, para [Baker et al. 2010], muitas vezes é necessário modificar estes métodos, por causa da necessidade de considerar a hierarquia da informação.

Neste trabalho são utilizados métodos preditivos supervisionados para predizer desempenho estudantil. A meta da predição é desenvolver modelos que deduzam aspectos específicos dos dados, conhecidos como variáveis preditivas, através da análise e fusão dos diversos aspectos encontrados nos dados, chamados de variáveis preditoras [Baker et al. 2011]. A taxonomia proposta por [Baker et al. 2011], indica três tipos de

predição: classificação, regressão e estimação de densidade. Para esse trabalho foi utilizado três dos algoritmos de classificação dos mais conhecidos: Naive Bayes, Árvore de Decisão (C5.0) e Máquina de vetores de suporte (do inglês, "Support Vector Machines", ou simplesmente SVM). A escolha foi baseada na larga aplicação dos mesmos, disponibilidade para uso e pelo fato de que cada algoritmo representa uma abordagem teórica diferente.

3. Metodologia

O processo de Descoberta de Conhecimento a partir de dados estruturados — KDD (Knowledge Discovery in Databases) é composto por cinco fases: Seleção de Dados, Pré-Processamento, Transformação, Mineração e Interpretação / Avaliação. A figura 1 mostra como essas fases foram incorporadas na aplicação web desenvolvida neste trabalho.

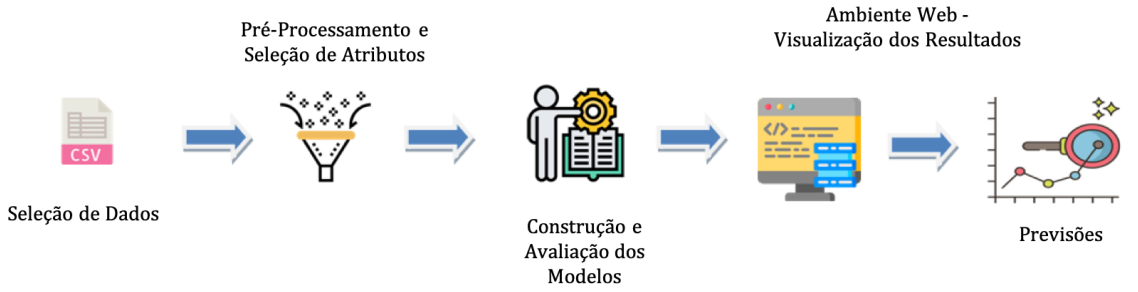

Figura 1. Processo

3.1. Seleção de Dados

Os dados utilizados neste trabalho, foram obtidos na plataforma Kaggle [1] no formato de um único arquivo CSV. Os dados foram coletados em 2016 usando uma ferramenta de rastreamento de atividade de aluno, chamada Experience API (ou xAPI) [Amrieh et al. 2016]. A xAPI é uma tecnologia de aprendizado que possibilita a coleta sobre uma ampla variedade de experiência que uma pessoa tem (on-line e off-line).

O conjunto de dados consiste em 480 registros e 16 atributos. As características são classificadas em três categorias principais:

- Características demográficas (como gênero e nacionalidade).
- Informações acadêmicas (escolaridade, turma e semestre).
- Características comportamentais (como a mão levantada na sala de aula, visualizações de materiais e satisfação escolar).

3.2. Pré-Processamento e Seleção dos Atributos

O pré-processamento é a parte do processo em que é feita a remoção de estruturas indesejadas das fontes de dados. Nesta etapa de limpeza busca-se eliminar inconsistências e valores errados ou incompletos, para que estes não influenciem no resultado dos algoritmos, bem como para reduzir a dimensão dos dados [Gonçalves et al. 2018]. Utilizando bibliotecas do software R, não se constatou a necessidade de retirar nenhum atributo. Sendo assim, a base de dados permaneceu com o mesmo número registros e atributos.

[1] kaggle.com

Nesta etapa, também é feita a seleção de atributos. Ou seja, são escolhidas quais informações, dentre as bases de dados existentes, devem ser efetivamente consideradas para o estudo [Rodrigues et al. 2017]. Inicialmente, conjunto de dados continha informações sobre sexo, nacionalidade, local de nascimento, pai responsável pelo estudante, etc. Optou-se por uma seleção manual dos quais, foram selecionados 7 atributos considerados mais representativo para compor os indicadores de previsão de desempenho. Os indicadores selecionados são mostrados na Tabela 1.

Tabela 1. Atributos selecionados na fase de mineração

Atributos Selecionados	**Informação**
Semestre	F = Primeiro e S = Segundo
Disciplina	Disciplina do Curso
NumVezesMaoLev	Quantidade de vezes em que o aluno levanta a mão na sala de aula (0 a 100)
NumContVisit	Quantidade de vezes em que o aluno visita o conteúdo do curso (0 a 100)
NumVisualiAnuncios	Quantidade de vezes em que o aluno verifica novos anúncios (0 a 100)
NumDiscussaoForum	Quantidade de vezes em que o aluno participa de fórum de discussão (0 a 100)
DiasAusencia	Quantidade de dias de ausência para cada aluno Above 7 (acima de 7) Under 7 (abaixo de 7)

3.3. Construção e Avaliação dos Modelos

Na etapa de construção dos modelos, fez se o emprego de ferramentas computacionais do R , utilizou-se três algoritmos de classificação o Naive Bayes, Árvore de Decisão(C5.0) e o SVM.

Na fase de avaliação são abordados os resultados referentes à construção dos modelos preditivos e suas respectivas validações. É de extrema importância validar os modelos de mineração entendendo suas qualidades e características antes de implantá-los em um ambiente de produção [Chapman et al. 2000].

Uma das técnicas mais utilizados na avalição de modelos é a validação cruzadas (cross-validation), que visa dividir um dataset em conjuntos de treino e teste, usando o conjunto de treino para treinar o modelo e o conjunto de teste para avaliar quão bom o modelo generaliza para dados que ele ainda não conhece. Neste trabalho foram utilizadas técnicas de validação cruzadas por 10 vezes (10 fold cross validation).

Para compreender os erros gerados pelo classificador foi feita a construção da matriz de erros denominada matriz de confusão para cada modelo. Com a matriz de confusão é possível obter uma métrica sobre o desempenho dos algoritmos como mostra a Tabela 2. De forma geral, o algoritmo Árvore de decisão obteve os melhores resultados, tendo a maior taxa de acerto (acurácia) de 96.88%.

Tabela 2. Desempenho dos três algoritmos

Algoritmos	Acurácia	Sensibilidade	Especificidade
Naive Bayes	0.7188	0.7189	0.8483
Árvore de Decisão	0.9688	0.9711	0.9839
SVM	0.7604	0.7629	0.8708

3.4. Criação do Ambiente Web

Esta fase teve como objetivo a criação de um ambiente para visualização dos dados e a criação de dashboards para exibir as predições feitas pelo modelo. Para criação do painel de controle, usou-se o Shiny. O Shiny é um pacote do software R que facilita a criação de aplicativo da Web interativos diretamente do R como mostra a Figura 2. Além disso, é possível hospedar aplicativos independentes em uma página Web ou integrá-los aos documentos do R Markdown ou simplesmente para criar dashboards.

A figura 2 exibe o layout do Ambiente Web de predição de desempenho, no lado esquerdo é a área de menu. Nela, é possível selecionar uma página inicial onde será exibida informações sobre as etapas de mineração, selecionar a página de previsão,obter informações dos dados, e também selecionar informações adicionais sobre os algoritmos utilizados.

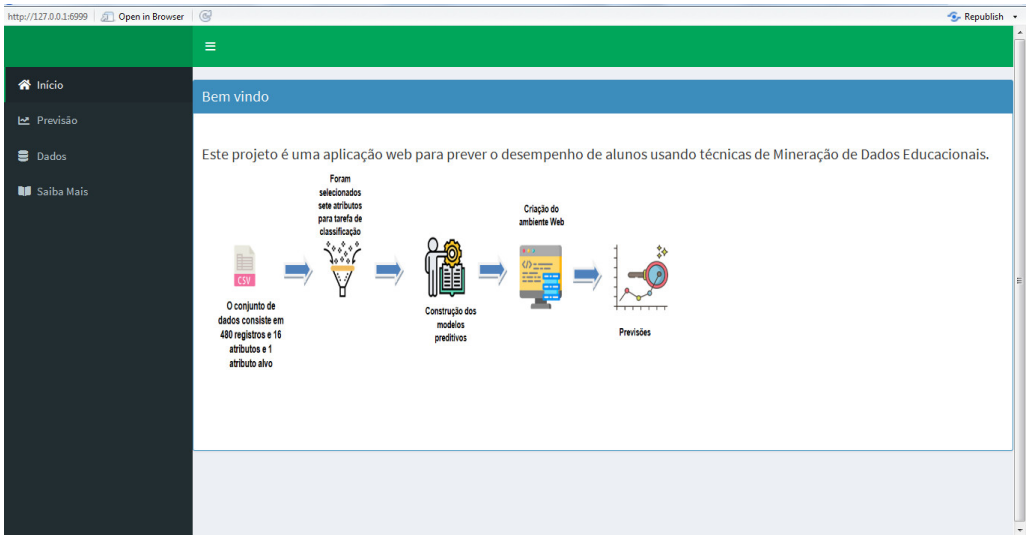

Figura 2. O layout do Ambiente Web

4. Resultados e Discussão

A figura 3 apresenta a interface relacionada à predição de desempenho estudantil.Através dela, o utilizador pode realizar a predição de um novo caso, selecionando valores para os sete indicadores ou atributos do aluno obtido na fase de mineração. O sistema, então, mostra a predição de desempenho do aluno nos três modelos de aprendizado de máquina construídos.

As figuras 4, 5 e 6 exibem uma previsão de desempenho de um aluno X nos três modelos. Após o usuário informar os valores de cada um dos atributos, é mostrado

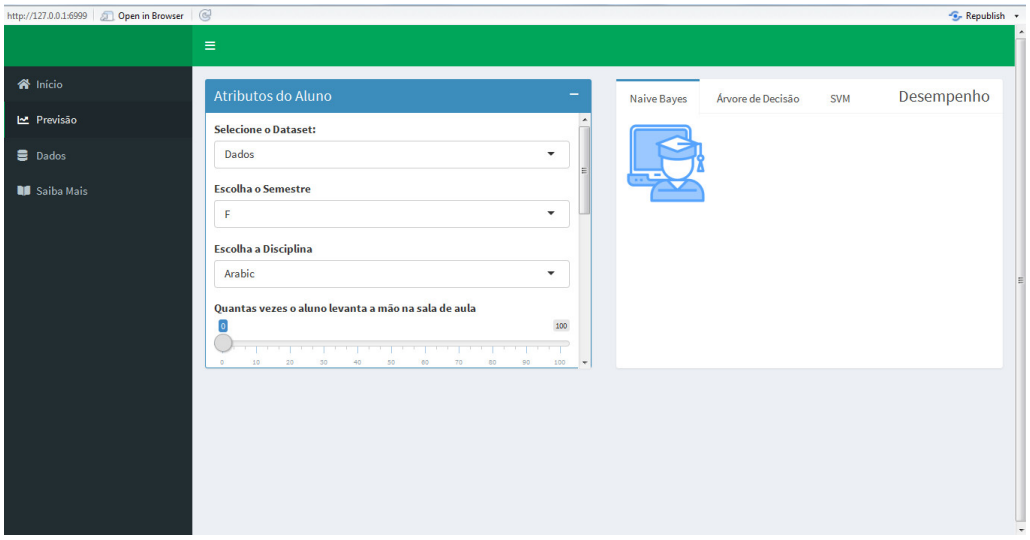

Figura 3. Página de previsão de desempenho

ao lado uma caixa contendo informação sobre o desempenho do aluno e a acurácia do algoritmo. A caixa na cor verde representa um aluno com desempenho Nível Alto, a caixa na cor amarela corresponde um aluno com desempenho Nível Médio e a caixa na cor vermelha corresponde um aluno com desempenho Nível Baixo. Estes resultados de previsões mostradas na figura são produzidos a partir dos modelos criados durante a fase de mineração.

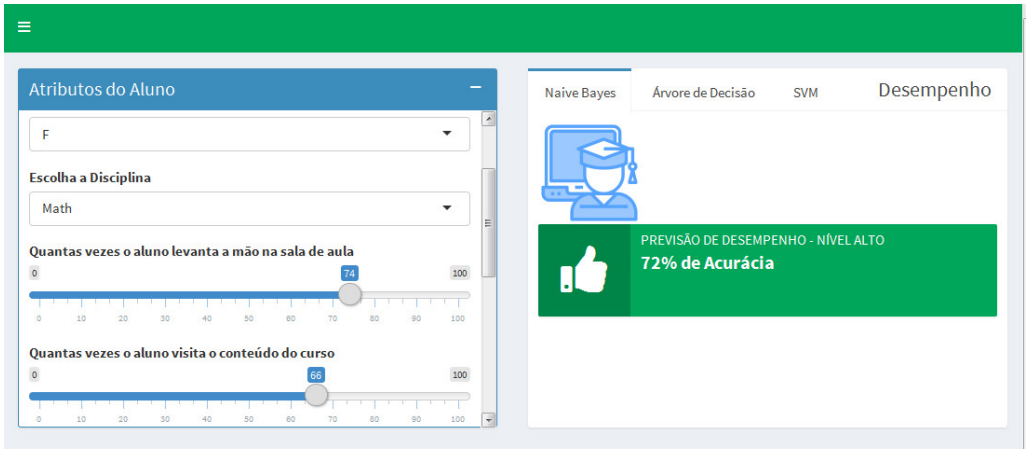

Figura 4. Resultado do modelo Naive Bayes

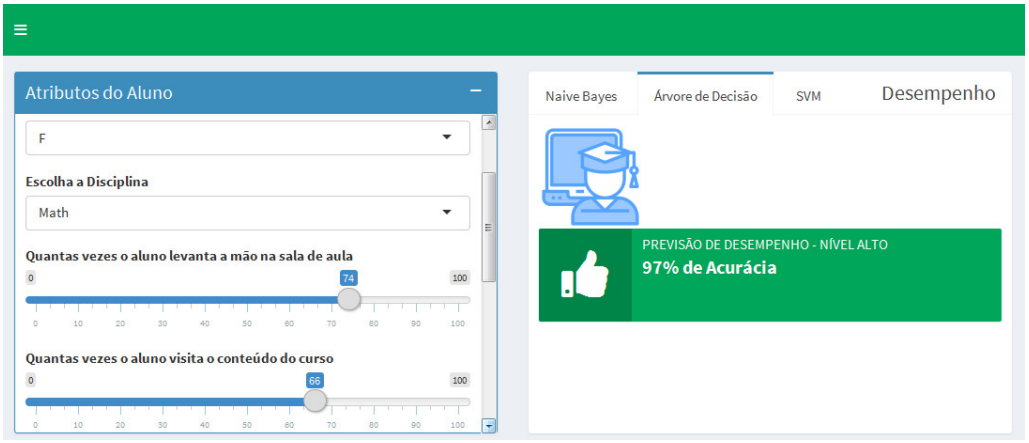

Figura 5. Resultado do modelo Árvore de Decisão

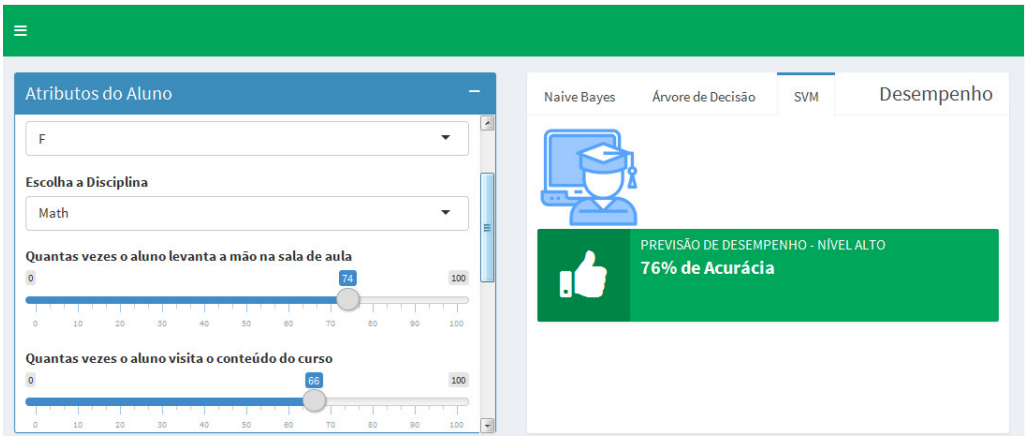

Figura 6. Resultado do modelo SVM

5. Conclusão

Este trabalho buscou criar uma aplicação web que pudesse mostrar de forma dinâmica as previsões de desempenho acadêmico de um estudante no decorrer de um curso baseado em alguns atributos comportamentais.

Os métodos de mineração de dados educacionais mostraram-se satisfatórios. Mesmo considerando uma base dados relativamente pequena, os modelos preditivos gerados obtiveram resultados satisfatórios em termo de acurácia, com destaque para o algoritmo Árvore de Decisão que obteve os melhor resultado. Com relação à aplicação criada, ela contribuiu para facilitar o acompanhamento e a visualização dos resultados obtidos pelos algoritmos.

Para trabalhos futuros, pretende-se adaptar o ambiente web para receber qualquer tipo de base de dados educacionais fornecidos pelo usuário e também permitir o educador ou usuário selecionar os atributos que julgar necessário para medir o desempenho dos alunos, podendo assim fazer uma análise mais individual do estudante. Uma outra pretensão é utilizar outros algoritmos de classificação disponível com objetivo de realizar testes comparativos do desempenho de cada um, verificando as vantagens e desvantagem de cada um deles, considerado a base de dados.

Referências

Amrieh, E. A., Hamtini, T., and Aljarah, I. (2016). Mining educational data to predict student's academic performance using ensemble methods. *International Journal of Database Theory and Application*, 9(8):119–136.

Baker, R. et al. (2010). Data mining for education. *International encyclopedia of education*, 7(3):112–118.

Baker, R., Isotani, S., and Carvalho, A. (2011). Mineraçao de dados educacionais: Oportunidades para o brasil. *Brazilian Journal of Computers in Education*, 19(02):03.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., Wirth, R., et al. (2000). Crisp-dm 1.0: Step-by-step data mining guide. *SPSS inc*, 16.

Costa, E., Baker, R. S., Amorim, L., Magalhães, J., and Marinho, T. (2013). Mineração de dados educacionais: conceitos, técnicas, ferramentas e aplicações. *Jornada de Atualização em Informática na Educação*, 1(1):1–29.

Fernandes, E., Holanda, M., Victorino, M., Borges, V., Carvalho, R., and Van Erven, G. (2019). Educational data mining: Predictive analysis of academic performance of public school students in the capital of brazil. *Journal of Business Research*, 94:335–343.

Gonçalves, T. C., da Silva, J. C., and Cortes, O. A. C. (2018). Técnicas de mineração de dados: um estudo de caso da evasão no ensino superior do instituto federal do maranhão. *Revista Brasileira de Computação Aplicada*, 10(3):11–20.

Rabelo, H., Burlamaqui, A., Valentim, R., de Souza Rabelo, D. S., and Medeiros, S. (2017). Utilização de técnicas de mineração de dados educacionais para predição de desempenho de alunos de ead em ambientes virtuais de aprendizagem. In *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*, volume 28, page 1527.

Rodrigues, R., Gomes, A. S., and Adeodato, P. (2017). Uma abordagem de mineração de dados educacionais para previsão de desempenho a partir de padrões comportamentais de autorregulação da aprendizagem. In *Anais dos Workshops do Congresso Brasileiro de Informática na Educação*, volume 6, page 13.

Romero, C. and Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. *Expert systems with applications*, 33(1):135–146.

Witten, I. H., Frank, E., Hall, M. A., and Pal, C. J. (2016). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.