Revista Brasileira de Informática na Educação – RBIE
Brazilian Journal of Computers in Education
(ISSN online: 2317-6121; print: 1414-5685)
http://br-ie.org/pub/index.php/rbie

Submission: 08/Oct/2017; 1st round notif.: 24/Nov/2017; New version: 18/Jan/2018; 2nd round notif.: 29/Jan/2018
Camera ready: 08/Jun/2018; Edition review: 29/Oct/2018; Available online: 12/Nov/2018; Published: 01/Jan/2019

Predição de Sucesso Acadêmico de Estudantes: Uma Análise sobre a Demanda por uma Abordagem baseada em *Transfer Learning*

Title: Student Success Prediction: An Analysis of the Demand for a Transfer Learning Approach

Daniel A. Guimarães de los Reyes
Escola de Engenharia e TI
Centro Universitário Ritter dos Reis
daniel.a.g.reyes@gmail.com

Everton André Thomas
Escola de Engenharia e TI
Centro Universitário Ritter dos Reis
everton.a.thomas@gmail.com

Lilian Landvoigt da Rosa
Mestrado em Design
Centro Universitário Ritter dos Reis
land.lilian@gmail.com

Wilson P. Gavião Neto
Mestrado em Design e Escola de Engenharia e TI
Centro Universitário Ritter dos Reis
wilson_gaviao@uniritter.edu.br

Resumo

Interações de estudantes com Ambientes Virtuais de Aprendizagem (AVA) geram logs que permitem reconstruir cada atividade realizada. A análise destes dados tem proporcionando uma melhor compreensão do comportamento de estudantes e dos processos de ensino e aprendizagem. Neste contexto, inúmeros trabalhos têm relatado resultados promissores na tarefa de predição de desempenho de estudantes, permitindo que ações proativas possam ser tomadas no sentido de evitar insucessos acadêmicos. Usualmente, técnicas de mineração de dados empregadas na construção de modelos preditivos utilizam registros históricos (passados) de dados, assumindo-se, desta forma, a premissa de que o preditor construído irá realizar predições em contextos futuros que sejam similares aos contextos (passados) que foram utilizados na sua concepção. Ainda que seja razoável assumir que a diversidade de contextos educacionais existentes se reflita nos dados gerados, poucos são os trabalhos que discutem o impacto de tal premissa na área de Mineração de Dados Educacionais (MDE), o que resulta em modelos que podem apresentar desempenho insatisfatório quando utilizados em condições educacionais não previstas. Este trabalho propõe uma análise empírica no sentido de verificar indícios de diferenças entre dados provenientes de contextos educacionais distintos na tarefa de predição de insucesso acadêmico de estudantes. Emprega-se dados de logs de mais de 3.000 estudantes de ensino superior na modalidade EAD. A metodologia adotada é baseada na própria abordagem de classificação supervisionada, comumente utilizada em tarefas de predição, sendo que busca-se, especificamente, verificar se contextos educacionais distintos são de fato separáveis em termos dos dados que geram. Ainda que o cenário de dados envolva atividades comuns a estudantes de uma mesma disciplina, os experimentos indicam uma acurácia de até 83% na separação de dados provenientes de períodos letivos distintos. Embora empíricos, os resultados indicam uma direção similar àquela apontada por outros trabalhos, contribuindo sobre a necessidade da utilização de técnicas de transfer learning e/ou adaptação de domínio no projeto dos modelos preditivos voltados a prevenção de insucessos acadêmicos.

Palavras-chave: *Predição de Sucesso Acadêmico; Learning Analytics; Mineração de Dados Educacionais; Transfer Learning; Covariate Shift.*

Cite as: De Los Reyes, D. A. G., Thomas, E. A., Rosa. L. L., Gavião Neto, W. P.. (2019). Student Success Prediction: An Analysis of the Demand for a Transfer Learning Approach (Predição de sucesso acadêmico de estudantes: uma análise sobre a demanda por uma abordagem baseada em transfer learning). Brazilian Journal of Computers in Education (Revista Brasileira de Informática na Educação – RBIE), 27(1), 01-24. DOI: 10.5753/RBIE.2019.27.01.01.

Abstract

Student interactions with Learning Management Systems (LMS) generate logs, which are usually stored, allowing to recover each student activity. Analysis of these data with data mining and/or learning analytics techniques have been provided a better understanding of student behavior and teaching-learning processes. In this context, a number of studies have been reporting promising results in the task of predicting student performance, which allows proactive actions to avoid academic failures. Usually, data mining techniques estimate predictive models by using (past) historical data, assuming the premise that the estimated predictor will make predictions in future contexts that are similar to the (past) contexts which were used in its design. Although it is reasonable to assume that the diversity of existing educational contexts is reflected in the data, few studies discuss the impact of the aforementioned premise in the area of Educational Data Mining (EDM), resulting in models that may perform poorly when used under unforeseen educational conditions. This paper proposes an empirical analysis to verify evidences of differences between data from different educational contexts in the task of predicting students' academic failure. Logs of more than 3,000 distance higher education students are used, and the adopted methodology is based on the supervised classification approach, commonly used in prediction tasks. Specifically, we aim to verify if distinct educational contexts are in fact separable in terms of the data they generate. Although data scenarios involve activities common to students in the same subject, the experiments indicate an accuracy of up to 83% in the separation of data from different academic terms. Although empirical, our results indicate a similar direction to that pointed out by other studies, contributing about the need of using transfer learning and/or domain adaptation techniques in the design of predictive models that aim to support proactive actions to prevent student failures.

Keywords: *Academic sucess prediction; Educational Data Mining; Transfer Learning; Covariate Shift.*

1 Introdução

O desenvolvimento das Tecnologias da Informação e Comunicação (TICs) e seu crescente uso na educação têm representado uma oportunidade para o enriquecimento dos processos de ensino e aprendizagem. Neste contexto, a educação a distância (EAD) encontra ferramentas que permitem o acompanhamento do comportamento de estudantes em ambientes digitais. Por meio dos rastros gerados a partir das interações dos usuários com Ambientes Virtuais de Aprendizagem (AVA), é gerado um volume grande de dados que permite a reconstrução de cada atividade realizada. A análise e apresentação destes dados educacionais, por meio de técnicas de mineração e/ou visualização de dados, tem prestado apoio a atividades de ensino, proporcionando a estudantes, professores, pesquisadores e gestores uma melhor compreensão de comportamentos e de processos de ensino e aprendizagem, possibilitando intervenções mais precisas no aprimoramento de resultados acadêmicos (Dawson et al., 2014; Siemens & Long, 2011).

Learning Analytics (LA) (Ferguson, 2012) e Mineração de Dados Educacionais (MDE) (Peña-Ayala, 2013) são denominações sob as quais se desenvolve a literatura que discute avanços na área de educação pelo uso da tecnologia. Ainda que utilizem métodos, técnicas e abordagens por vezes distintas, ambas têm contribuído com o objetivo comum de melhorar a qualidade da análise de dados educacionais em larga escala, para apoiar tanto a pesquisa básica quanto a prática na educação (Siemens & Baker, 2012). Neste contexto, destaca-se o uso da mineração de dados em tarefas preditivas (Peña-Ayala, 2014; Gašević et al., 2016), com resultados promissores em prestar suporte à intervenções proativas que visam, por exemplo, melhorar o desempenho acadêmico de estudantes (Bousbia & Belamri, 2014).

Técnicas ou algoritmos da área de *machine learning* são usualmente empregados em tarefas preditivas, como para predizer insucessos ou evasões de estudantes (Thammasiri et al., 2014; Hu et al., 2014; You, 2016) ou ainda identificar precocemente estudantes que enfrentam dificuldades de aprendizado (Baker et al., 2011). Contudo, como discutido na Seção 2.2, técnicas preditivas de *machine learning* geralmente buscam aprender padrões sobre dados históricos, com a finalidade de realizar inferências sobre eventos futuros, assumindo-se, assim, que contextos passados, utilizados na construção de um modelo preditivo, possuem características semelhantes ao contexto para o qual é realizada a predição (Pan & Yang, 2010; Lu et al., 2015). Na área educacional, embora muitos trabalhos reportem alta acurácia ao empregar algoritmos de *machine learning* na predição de sucesso acadêmico de estudantes, pouco discute-se sobre a generalização ou portabilidade destes modelos preditivos para contextos distintos, como diferentes cursos, disciplinas ou mesmo outras instituições de ensino (Gašević et al., 2016). Neste cenário, a diversidade de contextos gerada pela variabilidade da carga de atividades demandada a diferentes turmas de estudantes também constitui-se em um desafio, como destacado por Jayaprakash et al. (2014). Ainda que dentro do contexto de uma mesma disciplina, professores com diferentes níveis de exigência, disponibilizando materiais de forma mais frequente, são exemplos de características que podem se refletir na quantidade de interações realizadas por estudantes em AVA e, consequentemente, confundir a habilidade de um modelo preditivo capturar padrões de comportamento de estudantes similares (Jayaprakash et al., 2014).

Deste modo, embora técnicas de *machine learning* estejam sendo amplamente adotadas na descoberta de padrões de atividades e predição de resultados educacionais (Peña-Ayala, 2014;

Gašević et al., 2016), raros são os trabalhos na área de MDE e LA, como detalhado na Seção 3, que discutem o pressuposto de que os dados de comportamentos passados de estudantes, utilizados para construir um modelo preditivo, possuem características similares aos dados de comportamentos de estudantes para os quais deseja-se realizar predições. Como consequência da existência de diferenças significativas entre dados provenientes de diferentes contextos, se faz recomendável estudos que apontem técnicas capazes de compatibilizar tais dados/contextos, como têm-se discutido em inúmeros trabalhos sob denominações de *Transfer learning*, *Domain adaptation*, *Dataset shift* ou *Covariate shift* (Quionero-Candela et al., 2009; Sun et al., 2016; Moreno-Torres et al., 2012; Pan & Yang, 2010; Lu et al., 2015).

Este trabalho apresenta uma análise empírica sobre *logs* de interações de mais de 3.000 estudantes com AVA, na modalidade de ensino superior a distância, com o objetivo de verificar indícios de divergências entre distribuições de dados que são gerados em contextos educacionais envolvidos em tarefas de predição de insucesso acadêmico. A metodologia adotada é a própria abordagem de classificação supervisionada, comumente utilizada em tarefas de predição, buscando-se, especificamente, verificar se contextos educacionais que comumente são assumidos como sendo similares são, de fato, separáveis em termos dos dados que geram. Como principais contribuições destacam-se:

- Uma análise realizada no cenário de predição de períodos letivos correntes (isto é, enquanto as disciplinas estão em andamento), ao contrário de muitos trabalhos na literatura que concentram-se na construção de preditores para a avaliação do desempenho de estudantes após o término do período letivo analisado (Hu et al., 2014);
- Utiliza-se apenas dados de *logs* de interação com AVA, minimizando-se riscos de comprometimento de privacidade ao empregar-se dados de perfil sócio-econômico e/ou demográfico de estudantes (You, 2016);
- Os dados empregados não envolvem registros prévios ou intermediários sobre o desempenho de estudantes, como notas provenientes de avaliações. Ainda que esta informação se mostre como uma das melhores preditoras de resultados acadêmicos (Manhães et al., 2011; Thammasiri et al., 2014), há cenários importantes e críticos que não dispõe de tal informação, como nas primeiras semanas letivas de estudantes ingressantes em cursos superiores. Este cenário têm sido foco de estudos sobre evasão em que aponta-se, por exemplo, que um percentual significativo de estudantes que evadem não realizam se quer a entrega da primeira tarefa (Simpson, 2004);

O conteúdo do presente artigo está organizado como segue. A Seção 2 revisa a tarefa de predição no contexto da MDE. Na Seção 2.2, detalha-se o processo de classificação supervisionada, usual em tarefas de predição, bem como as ideias de *Transfer Learning* e *Covariate Shift*, ambas associadas a abordagens que buscam conciliar contextos diferentes de dados em uma mesma tarefa de predição. Seção 3 discute a literatura no sentido de sustentar a contribuição deste trabalho. A abordagem experimental adotada por este trabalho é apresentada na Seção 4 e, em seguida, a Seção 5 detalha e discute os resultados alcançados. Por fim, as conclusões são apresentadas na Seção 6.

2 Mineração de dados educacionais: a tarefa de predição

A Mineração de dados (MD) tem sido usada em diversas áreas para identificar padrões de comportamento e encontrar *insights* que gerem melhorias em produtos e serviços. Na área da educação, a Mineração de Dados Educacionais (MDE) é considerada uma disciplina voltada ao desenvolvimento de métodos para explorar dados oriundos de ambientes educacionais e utilizá-los para compreender melhor os processos de ensino-aprendizagem (Baker et al., 2011). De forma similar, inúmeros trabalhos também estão sendo propostos sob a denominação de área conhecida como *Learning Analytics* (Peña-Ayala, 2013), que envolve medir, coletar, analisar e comunicar sobre dados de estudantes e seus contextos, com a finalidade de compreender e otimizar a aprendizagem e os ambientes em que esta ocorre (Chatti et al., 2012).

A compreensão de como a MD ocorre inicia-se por estabelecer-se tarefas básicas de um sistema que tem por objetivo automatizar a descoberta de conhecimento sobre uma base dados. Neste sentido, pode-se distinguir dois tipos básicos de tarefas (Fayyad et al., 1996): *predição* e *descrição*, que ainda dividem-se em tarefas de *classificação, regressão, agrupamento* ou *clustering*, *associação* e *sumarização*. Dentre estas, este trabalho concentra-se na tarefa de predição, uma das mais empregadas no contexto educacional (Peña-Ayala, 2014).

2.1 Tarefas preditivas

Uma tarefa preditiva caracteriza-se pela busca de um padrão de comportamento que seja capaz de prever o comportamento de uma futura entidade (Fayyad et al., 1996). Como exemplo no contexto educacional, pode-se citar uma tarefa de predição de insucesso acadêmico de estudantes em uma dada disciplina, em que busca-se estimar, sobre dados de edições passadas da disciplina, um padrão de comportamento de estudantes que tenha resultado em insucesso. Deste modo, uma vez que este padrão de comportamento seja identificado em um estudante que esteja cursando uma edição corrente da disciplina, pode-se inferir a predição de insucesso para ele, possibilitando que ações proativas sejam tomadas de forma a evitar a confirmação do insucesso. De um ponto de vista prático, a Figura 1(a) ilustra, em cada linha, um estudante e seu comportamento em termos de totalizações de categorias de interações com um AVA, realizadas ao longo de um período letivo. Assumindo-se que estes dados dizem respeito a uma disciplina já finalizada, em que já se conhece a situação final dos estudantes (Aprovado ou Reprovado), busca-se então descobrir um comportamento recorrente que esteja associado com o desempenho acadêmico de Aprovação ou Reprovação. Esta etapa denomina-se de *Treinamento*, e tem como resultado um *modelo preditivo* ou um simplesmente um *preditor*. Na Figura 1(a), as colunas ***forum, resource*** e ***assign*** são denominadas de variáveis de entrada, e a coluna **Final** é denominada de variável de saída ou variável alvo, sendo que os valores que aparecem nesta variável de saída são denominados de *classes*. Isto é, no caso da variável **Final** mostrada na Figura 1(a), observa-se duas classes de estudantes: *Rep* (reprovados) e *Apr* (aprovados). Neste cenário, espera-se então que o preditor estimado seja resultado da descoberta de um padrão que relacione valores de variáveis de entrada (***forum, resource*** e ***assign***) com valores da variável de saída (**Final**).

Finalmente, emprega-se o preditor na etapa denominada de predição, que caracteriza-se pela presença de valores de variáveis de entrada e pela ausência dos valores da variável de saída, conforme ilustrado na Figura 1(b). Neste cenário, a partir dos dados de comportamento corrente

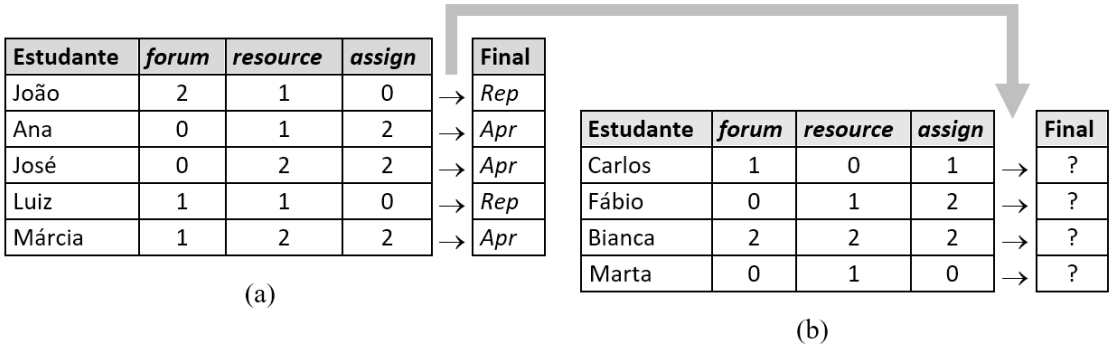


Figura 1: **(a) Etapa de treinamento do modelo preditivo:** associa quantificações históricas de interações (fórum, resource e assign) com o status final na disciplina (Aprovado ou Reprovado). **(b) Etapa de predição:** modelo preditivo infere um status final a partir de quantificações de interações do período letivo corrente.

de um estudante (isto é, totalizações de interações em ***forum, resource*** e ***assign***), é papel do preditor inferir os valores (Rep ou Apr) da variável de saída para tal estudante.

Tarefas de mineração de dados de cunho preditivo ainda se distinguem em **classificação** e **regressão** (Fayyad et al., 1996). O tipo do dado que consta na variável de saída é o determinante. No caso de uma tarefa de classificação, a variável de saída apresenta dados categóricos ou discretos, como exemplificado pela variável **Final** na Figura 1(a), em que os estudantes estão categorizados em *Apr* ou *Rep*. Sendo assim, a tarefa preditiva ilustrada na Figura 1 trata de uma tarefa de classificação. Por outro lado, se a variável **Final** fosse representada em termos de valores numéricos e contínuos, como uma nota numérica e real entre 0 e 10, a tarefa de predição seria tratada como uma tarefa de regressão.

2.2 Classificação supervisionada

Classificação supervisionada é um processo usual pelo qual se implementa uma solução para uma tarefa preditiva, e consiste de duas fases (Han et al., 2011), como detalhado na Seção 2.1. Na fase de treinamento, emprega-se um algoritmo para se estimar, ou aprender, um modelo preditivo/classificador sobre um conjunto de dados. Este algoritmo, usualmente oriundo da área de *machine learning* (Faceli et al., 2011), tem por objetivo estabelecer uma associação entre tuplas de entrada com seus respectivos rótulos de classe, como ilustrado na Figura 1. Na segunda fase, o modelo/-classificador estimado é usado para classificar tuplas cujo rótulo de classe é desconhecido, tendo como resultado a inferência de um rótulo para cada nova tupla de entrada.

Modelos preditivos com frequência aplicam funções de aprendizado supervisionado para estimar valores desconhecidos, ou futuros, de variáveis dependentes com base nas características de variáveis independentes relacionadas (Peña-Ayala, 2013). De forma geométrica, a Figura 2 ilustra o processo de classificação supervisionada para um problema de classificação binária (duas classes), tendo duas variáveis de entrada e representadas nos eixos x e y. De especial importância é o fato do modelo preditivo ser uma fronteira de separação, que quando empregada na fase de predição deve ser capaz de manter uma separação dos elementos de cada classe com acurácia similar ou superior àquela alcançada na fase de treinamento. Ainda com base na Figura 2, é importante considerar as características das distribuições dos dois conjuntos de dados representados

pelas Figuras 2(a) e 2(b), comumente denominados de domínio de dados *origem*, ou de treino, e domínio de dados *alvo*, ou de teste, respectivamente, e o pressuposto de que ambos conjuntos devam manter as mesmas características estatísticas sob pena do modelo preditivo estimado na fase de treinamento não separar adequadamente os elementos das classes na fase de predição que ocorre no domínio de dados alvo.

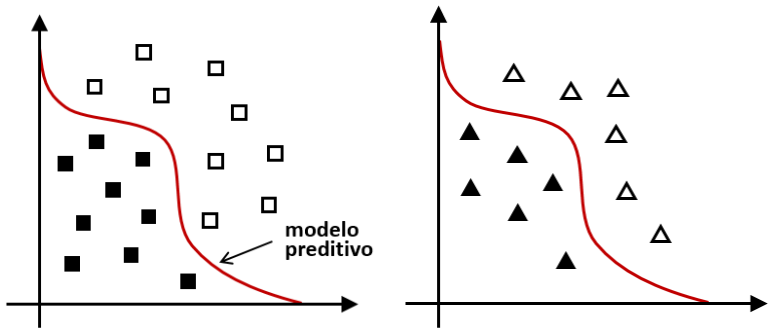


Figura 2: Classificação supervisionada envolvendo elementos de duas classes, □ e ■, e duas variáveis de entrada (eixos x e y). **(a) Fase de treinamento:** modelo preditivo (fronteira de separação) é estimado. **(b) Fase de predição ou teste:** novos elementos, △ e ▲, são classificados conforme a fronteira definida pelo modelo preditivo.

Na prática, como forma de avaliar a acurácia de um modelo preditivo, um conjunto de tuplas já rotuladas é reservada para teste. Este conjunto de tuplas de teste é mantido a parte do conjunto de treinamento e não faz parte da construção do modelo preditor. Assim, a análise da acurácia se dá pela comparação do rótulo inferido pelo modelo preditor com o real rótulo de classe de cada tupla que compõe o conjunto de teste.

2.3 *Transfer learning* e *covariate shift*

Como exposto anteriormente, técnicas preditivas, como a classificação supervisionada, buscam aprender padrões sobre bases históricas de dados com a finalidade de inferir algo sobre eventos futuros. Contudo, tradicionalmente tais técnicas assumem que os dados utilizados para treinar um modelo preditivo possuem as mesmas características do conjunto de dados alvo, para o qual é realizada a predição (Pan & Yang, 2010; Lu et al., 2015). Entretanto, quando a distribuição dos dados (futuros) alvo altera-se em relação aos dados de treinamento, os modelos preditivos frequentemente precisam ser reconstruídos, demandando a coleta de novos dados de treinamento (Lu et al., 2015). Em muitas situações práticas este processo pode recair em um cenário de insuficiência de novos dados para treinamento, ou em um contexto oneroso devido a inviabilidade de reconstrução dos modelos a cada novo cenário de predição ou simplesmente pela natureza irregular do fenômeno a ser predito.

Domain adaptation, covariate shift, dataset shift e *transfer learning* são termos usados na literatura como forma de abordar o problema descrito acima (Moreno-Torres et al., 2012; Sugiyama et al., 2008; Lu et al., 2015). Com base na ideia de transferir/aproveitar o aprendizado/conhecimento já adquirido por um modelo preditivo, a literatura de *transfer learning* talvez seja a mais abrangente em termos da discussão sobre os diferentes cenários do problema, propondo abordagens que permitam que domínios, tarefas e distribuições de dados usados em treinamento e teste

sejam compatibilizados.

Especificamente, o cenário em que os dados de treinamento e os dados de teste seguem diferentes distribuições de probabilidade caracteriza um problema de *covariate shift*, podendo levar um modelo preditivo a apresentar um desempenho insatisfatório em termos de assertividade (Quionero-Candela et al., 2009; Zadrozny, 2004; Sun et al., 2016).

3 Trabalhos relacionados

Antever resultados acadêmicos de estudantes é um desafio que tem sido alvo de inúmeros trabalhos (Macfadyen & Dawson, 2010; Dawson et al., 2014; Gašević et al., 2016; Peña-Ayala, 2014), sobretudo na condição atual em que as tecnologias de informação e comunicação permitem armazenar dados de diversas atividades que ocorrem em ambientes educacionais. Em especial, os *logs* de atividades de estudantes em AVA têm sido frequentemente utilizados para tal propósito, sendo considerados de fácil acesso, escaláveis, não-intrusivos e indicadores de eventual sucesso ou fracasso (Macfadyen & Dawson, 2010).

A Tabela 1 caracteriza a literatura de MDE quando o objetivo é a predição de desempenho acadêmico. Visando um cenário em que busca-se minimizar riscos relacionados ao comprometimento de informações privadas de estudantes (Chatti et al., 2012; Fortenbacher et al., 2013), a seleção de trabalhos relacionados tem foco em estudos que envolvem dados provenientes de AVA, ainda que o emprego de informações socioeconômicas, demográficas e de perfil do acadêmico possibilite um incremento na assertividade dos modelos preditivos. Na Tabela 1, diferencia-se, ainda, a literatura em termos de 3 aspectos: (i) variáveis/atividades que são utilizadas como entrada na construção de modelos preditivos e que são consideradas mais relevantes ou resultam em modelos mais assertivos; (ii) predições realizadas para um cenário de disciplinas em andamento, visando oferecer suporte a tomadas de decisão em tempo oportuno (Hu et al., 2014); (iii) aplicação de técnicas de *transfer learning* sobre dados educacionais com vistas a compatibilizar dados provenientes de contextos educacionais distintos, ou ainda trabalhos que discutam os desafios envolvidos na tentativa de se generalizar modelos preditivos para uso em diferentes contextos.

O uso de registros de interação de estudantes com AVA para predição de sucesso acadêmico, durante o decorrer do curso, é discutido por Barber & Sharkey (2012). Contudo, assim como Márquez-Vera et al. (2013) e Kampff (2009), é proposta a utilização de informações pessoais de estudantes como variáveis de entrada. Ainda que o emprego destas informações possibilite um incremento na assertividade dos modelos preditivos, a coleta de tais dados usualmente depende da aplicação de questionários, ao mesmo tempo que, como mencionado anteriormente, pode ter seu uso sujeito a restrições por representar riscos a violação de informações privadas dos estudantes (Chatti et al., 2012; Fortenbacher et al., 2013).

Diversos dados/atividades são exploradas como variáveis de entrada na tarefa de predição de desempenho acadêmico. Romero et al. (2013) explora o conteúdo de fóruns de discussão em AVA, enquanto Gottardo et al. (2014) quantifica interações em ambientes virtuais em 3 categorias: interações entre estudantes, entre estudante-professor e entre estudante-sistema. Por outro lado, Gašević et al. (2016) discute a diversidade de condições instrucionais nas quais os dados são gerados, destacando a grande influência deste fator na assertividade de preditores de desempenho

Tabela 1: Sumário de trabalhos relacionados.

Referência	Tipo de dado		Variáveis mais relevantes	Transfer learning?	Período corrente?
	Logs	Outros			
Macfadyen & Dawson (2010)	X		Mensagens enviadas Posts em fóruns de discussão Tarefas finalizadas	NÃO	SIM
Barber & Sharkey (2012)	X	X	Notas parciais Créditos cursados	NÃO	SIM
Er (2012)	X		Frequência por semana Notas parciais e finais	NÃO	SIM
Romero et al. (2013)	X		Posts em fóruns de discussão Número de palavras Avaliação de posts por instrutores Métricas de redes sociais	NÃO	NÃO
Hu et al. (2014)	X		Tempo gasto online Leituras de material de apoio	NÃO	SIM
Agudo-Peregrina et al. (2014)	X		Interações: estudante-professor e entre estudantes relacionadas às avaliações relacionadas participação ativa	NÃO	NÃO
Gottardo et al. (2014)	X		Mensagens recebidos do professor Nota parcial (questões corretas) Participação em chats	NÃO	NÃO
Jayaprakash et al. (2014)	X	X	Notas parciais Nota média (GPA)	NÃO	SIM
Lara et al. (2014)	X		Acessos (login) Acessos aos recursos	NÃO	SIM
You (2016)	X		Estudo regular Data de submissão das tarefas Acessos Nota parcial	NÃO	SIM
Gašević et al. (2016)	X	X	Acessos (login) Acesso aos recursos	NÃO	SIM
Cechinel et al. (2015)	X		Interações de estudantes por semana Interações de professores e tutores	NÃO	SIM
Hoang et al. (2016)		X	Nota final nas disciplinas	SIM	NÃO
Boyer & Veeramachaneni (2015)	X		Qtde média de submissões por tarefa Tempo gasto em material de apoio	SIM	SIM

acadêmico. Embora estes trabalhos discutam importantes fatores no comportamento virtual de estudantes, as análises realizadas envolvem dados de atividades de estudantes que abrangem o final de períodos letivos, e podem não refletir uma realidade parcial que permite acompanhar o progresso do estudante ao longo do período letivo. A importância da mensuração das interações com AVA para o acompanhamento do progresso e engajamento de estudantes é ressaltada, tanto na predição de desempenho acadêmico (Er, 2012; Hu et al., 2014; Lara et al., 2014; You, 2016; Cechinel et al., 2015) quanto na visualização de dados educacionais (Duval, 2011; Santos et al., 2012). Dados temporais têm destaque na construção de modelos preditivos (Lara et al., 2014; Cechinel et al., 2015; Rigo et al., 2014), permitindo identificar variações nos resultados da predição enquanto disciplinas estão em andamento, conforme o progresso do estudante no curso/disciplina.

Embora diversos trabalhos no campo da mineração de dados e/ou *machine learning* sugiram o emprego de técnicas para compatibilizar domínios/contextos de dados em tarefas preditivas (Pan & Yang, 2010; Sun et al., 2016; Daume III & Marcu, 2006), poucos são os estudos que abordam o assunto na área educacional (Lagus, 2016). Neste sentido, Gašević et al. (2016) destaca que, embora alguns trabalhos resultem em bons níveis de assertividade, outros divergem, e que a causa da divergência está no fato de ambos desconsiderem a diversidade de condições instrucionais nas quais os dados são gerados.

Jayaprakash et al. (2014) propõe generalizar um modelo preditivo treinado sobre os dados de uma instituição de ensino e aplicando-o a dados de outras instituições. Embora conclua que haja retenção de boa parte do poder preditivo do modelo entre diferentes instituições, aponta a necessidade da existência de um processo de registro de notas parciais de estudantes ao longo do período letivo, ao mesmo tempo que não relativiza resultados como função de diferentes condições instrucionais internas as instituições de ensino, como sugerido por Gašević et al. (2016).

Em lugar de analisar modelos preditivos entre instituições de ensino, Hoang et al. (2016) também expõe o problema de adaptação de contextos educacionais, porém entre dois cursos de graduação distintos, embora correlatos: Ciência da Computação e Engenharia da Computação. Através da aplicação de técnicas de *transfer learning*, discute-se um modelo preditivo treinado sobre um curso **A** e a classificação/predição sobre os estudantes de um outro curso **B**. Os experimentos não consideram, contudo, dados de interações com AVA, apenas notas finais de estudantes em disciplinas a partir do segundo ano de curso, sendo que o foco da tarefa preditiva está sobre o fato do estudante concluir o curso de graduação. *Transfer learning* também é empregado por Voß et al. (2015), buscando predizer o desempenho de estudantes alemães com base em dados de estudantes norte-americanos, porém sobre dados de interações em Sistemas Tutores Inteligentes, e não sobre interações de AVA.

Utilizando dados obtidos a partir de ambientes online voltados ao ensino em massa, conhecidos como MOOCs (*Massive Open Online Courses*), Boyer & Veeramachaneni (2015) comparam resultados de dois cenários de predição de evasão: com e sem o emprego de técnicas de *transfer learning*. A comparação é realizada sobre dados de três ofertas consecutivas do mesmo curso (2012/1, 2012/2 e 2013/1), sendo duas ofertas usadas para treino e um para teste, em uma abordagem de validação cruzada. Ao mesmo tempo que o trabalho conclui que os resultados não são conclusivos a respeito do benefício do emprego da técnica de *transfer learning* utilizada, apenas assume-se que as três ofertas do curso podem representar contextos/dados diferentes, não sendo apresentada uma caracterização, por exemplo, estatística sobre a existência de reais diferenças

nos dados das três ofertas do curso. Logo, os resultados inconclusivos sobre a necessidade de se compatibilizar dados de treino e teste (por *transfer learning*) podem ter causa no fato de que as três ofertas do curso não carregam diferenças significativas em termos do comportamento dos estudantes envolvidos em cada oferta, sobretudo quando trata-se de uma grande quantidade de estudantes em cada oferta (entre 29.000 e 155.000).

Em resumo, apesar de existirem trabalhos que discutam a necessidade de compatibilização de contextos de dados de treino e teste em tarefas preditivas na área educacional, ou mesmo que já empreguem técnicas de *transfer learning* para tal propósito, verifica-se uma carência de trabalhos na conjuntura:

- que combina a utilização de, apenas, dados gerados por interações de estudantes com AVA em instituições de ensino superior e um cenário de dados de predição para períodos letivos correntes;
- e que quantifique diferenças entre (dados de) contextos de treino e teste, na forma de *covariate shift*, em cenários típicos de predição de sucesso acadêmico de estudantes, de modo a justificar, quantitativamente, a necessidade de se avaliar os benefícios do uso de técnicas de *transfer learning* na busca de modelos preditivos escaláveis e/ou portáveis através de diferentes contextos educacionais/instrucionais.

4 Estudo Proposto

Estudar a diversidade de contextos educacionais, e/ou instrucionais, e os efeitos que produzem sobre interações de estudantes com AVA pode ser útil na identificação dos impactos causados na assertividade de predições na área de mineração de dados educacionais, bem como prestar suporte na busca de um modelo preditivo mais genérico e capaz de ser utilizado em diferentes contextos educacionais. Generalizar modelos preditivos sobre resultados educacionais vem sendo apontado como um desafio e tem sido alvo de estudos recentes (Gašević et al., 2016; Jayaprakash et al., 2014), inclusive na prática da indústria de softwares educacionais (Essa & Ayad, 2012).

Mesmo sendo oriundos de uma mesma disciplina, é possível que dados históricos, utilizados na construção de modelos preditivos, apresentem características estatísticas diferentes quando comparados a dados gerados posteriormente, como discutido na Seção 2.2. Estas diferenças podem estar associadas a diversidade de contextos instrucionais comumente presentes em ambientes educacionais, como uma mesma disciplina ser ministrada por diferentes tutores/professores, refletindo, por exemplo, diferentes frequências de postagem de conteúdos, ou mesmo turmas sendo compostas por estudantes com diferentes perfis de respostas a atividades de ensino. Neste cenário, propõe-se investigar empiricamente a existência do fenômeno de *covariate shift*, como descrito na Seção 2.3, e a consequente necessidade de utilização de técnicas de *transfer learning*.

Embora a literatura apresente diferentes maneiras para se detectar *covariate shift* (Raza et al., 2015), adota-se neste trabalho uma abordagem popularmente utilizada na prática de *machine learning* (Martin, 2014 (accessed November 1, 2016)) e baseada na ideia de discriminar/classificar dados como sendo de treino versus teste (Bickel et al., 2007), empregando-se as próprias técnicas de classificação supervisionada para distinguir se uma instância provêm do domínio de dados de

origem/treino ou do domínio alvo/teste. Deste modo, em lugar de predizer o sucesso acadêmico de estudantes, busca-se predizer o contexto instrucional do qual o estudante provêm e, consequentemente, analisar se um modelo preditor é capaz de diferenciar e classificar tais contextos pelas características de comportamento de seus estudantes.

A Figura 3 ilustra o experimento proposto. À esquerda, cada estudante está associado ao período letivo (2013/1 ou 2013/2) em que cursou uma dada disciplina, constituindo a fase de treinamento. À direita, ilustra-se a fase de teste, em que busca-se predizer para cada estudante o período letivo do qual é proveniente. Nesta tarefa, quanto maior a assertividade, mais clara é a existência de *covariate shift* e a consequente necessidade de se analisar a compatibilidade entre dados de comportamento de estudantes quando da tarefa de se predizer sucessos acadêmicos de estudantes em 2013/2 tendo por base os estudantes do período letivo anterior, 2013/1.

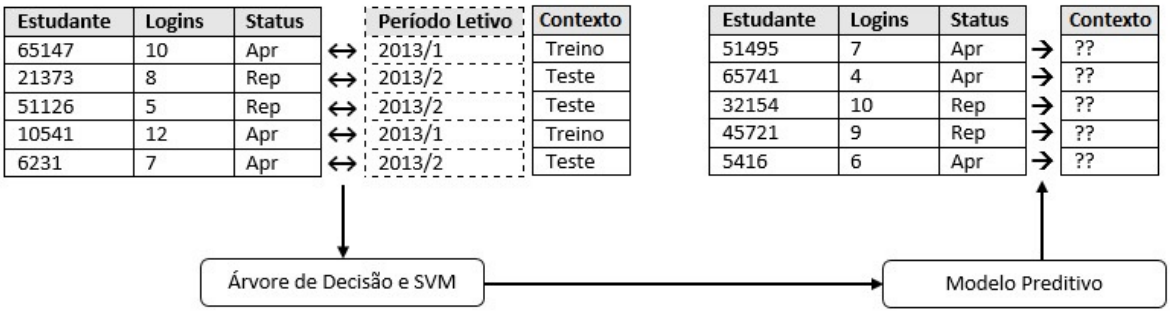


Figura 3: Experimento proposto visando distinguir contextos instrucionais supostamente refletidos nas atividades dos estudantes. Algoritmos de classificação supervisionada, Árvore de Decisão e *Support Vector Machines* (SVM), estimam modelos preditivos, à esquerda, na tentativa de distinguir/predizer os contextos instrucionais dos estudantes à direita, em que contextos instrucionais são definidos em termos de períodos letivos (2013/1 vs 2013/2).

É importante destacar que, neste trabalho, tendo por base a abordagem de *machine learning* na construção de preditores (isto é, aprender padrões sobre dados do passado para prever o futuro), assume-se ou delimita-se a definição de contextos instrucionais ao mínimo de duas variáveis: períodos letivos e disciplinas, desconsiderando-se outros fatores que podem caracterizar diferenças em condições instrucionais, como diferentes perfis de professores ou diferentes perfis de estudantes sendo agrupados em diferentes turmas de uma mesma disciplina. Neste *setup*, em um cenário prático de predição de sucesso acadêmico, as turmas de estudantes de um período letivo corrente constituem o domínio de dados alvo/teste, isto é, estudantes para os quais deseja-se predizer um status final de sucesso/insucesso na disciplina, e as turmas de estudantes de períodos letivos anteriores, da mesma disciplina, constituem o domínio (origem) de dados de treinamento do modelo preditivo, como detalhado nas Seções 2.1 e 2.2.

A seguir, o estudo proposto é detalhado em conformidade com a metodologia clássica de descoberta de conhecimento em bases de dados - KDD, como proposta por Fayyad et al. (1996).

4.1 Seleção de Dados

Para este estudo, selecionou-se um conjunto de dados formado por mais de 1 milhão de *logs* de interações de estudantes com o ambiente virtual Moodle[1], no contexto de cursos totalmente EAD

[1] https://moodle.org/

Tabela 2: Sumário dos dados utilizados nos experimentos.

Disciplina	Estudantes	Turmas	Aprovações
Lógica	421	6	42%
Processo Administrativo	926	12	72%
Raciocínio Lógico	1003	14	52%
Matemática para Administração	735	21	35%
Total	3085	53	

de uma universidade do sul do Brasil. Esses dados foram providos através de projeto em parceria com a empresa GVDasa - Inteligência Educacional[2], sendo previamente processados para preservar a identidade dos estudantes e sua utilização aprovada por comitê de ética para fins de pesquisa acadêmica. A Tabela 2 sumariza o cenário dos dados empregados nos experimentos. As disciplinas foram selecionadas por apresentarem três cenários distintos em termos de balanceamento de aprovações/reprovações: número de aprovações bastante maior que o de reprovações, a situação inversa e por fim um contexto em que a quantidade de aprovações e reprovações se equiparam. Cada disciplina possui duração de nove semanas (ou módulos), sendo as duas últimas semanas reservadas para avaliações presenciais. Neste trabalho, entende-se como insucesso acadêmico o estudante que não alcançou o status de aprovado em uma disciplina, seja por desistência ou por insuficiência de rendimento. O processo de avaliação é composto por dois graus, 1 e 2. Com peso de 30% do grau final, o grau 1 decorre de atividades virtuais realizadas ao longo do período letivo, sendo que 70% do grau final resulta de avaliação presencial realizada na 8ª semana. O status de sucesso ocorre quando o grau final é maior ou igual a 6,0. Como forma de recuperação do grau 2, o estudante pode ainda realizar uma segunda avaliação presencial.

4.2 Pré-processamento

Ambientes virtuais de aprendizagem como o Moodle dispõe de diversas categorias de atividades e recursos através das quais os estudantes interagem com o sistema e desempenham suas tarefas. No cenário de dados de inúmeras turmas de estudantes como histórico de uma disciplina, é de se esperar que algumas categorias de atividades/recursos estejam propostas em algumas turmas e em outras não, ainda que seja usual as instituições de ensino seguirem um mesmo padrão pedagógico de atividades para disciplinas à distância. Por exemplo, "chat" pode não ser uma atividade sempre presente em turmas de uma determinada disciplina. Logo, analisar *logs* provenientes de "chat" envolvendo turmas que ora oferecem a atividade e ora não, pode implicar em distorções na análise. Com isso, neste trabalho, descartou-se categorias de atividades/recursos que foram empregadas esporadicamente em algumas turmas do histórico das disciplinas. As atividades mantidas, e comuns a todas as turmas do histórico de uma disciplina, são detalhadas nas próximas seções.

4.3 Transformação

A etapa de transformação ocorre pela quantificação de cada categoria de *log* para cada estudante, como ilustrado na Figura 1(a) e detalhado na Seção 2.1. Desta forma, os valores das variáveis

[2] http://gvdasa.com.br/

de entrada empregadas neste trabalho resultam do acúmulo (quantidade de vezes ou tempo) de interações que os estudantes realizam em dado período de tempo, sendo cada variável definida pela categoria da interação. A Tabela 3 sumariza as variáveis empregadas neste trabalho. Cabe destacar que tal conjunto de variáveis resultou da busca por abranger a natureza das variáveis comumente empregadas na literatura relacionada e brevemente descritas na coluna "Variáveis mais relevantes" da Tabela 1, tendo ainda como limitador, conforme mencionado na seção 4.2, variáveis longitudinais comuns ao histórico das disciplinas.

Tabela 3: Sumário das variáveis de entrada computadas a partir dos registros de *logs* de estudantes no AVA Moodle.

Variável	**Descrição**
Login_Quantidade	Quantidade acumulada de logins
Forum_Quantidade_Post	Quantidade acumulada de postagens em fóruns de discussão
Forum_Quantidade_Visualizações	Quantidade acumulada de visualizações de discussões de fóruns
Forum_TempoUso	Tempo acumulado em atividades de fóruns de discussão
Chat_Quantidade_Mensagens	Quantidade acumulada de postagens em chats
Chat_TempoUso	Tempo acumulado em atividades de chat
Assignment_Post_Quantidade	Quantidade acumulada de submissões de tarefas
Assignment_View_Quantidade	Quantidade acumulada de visualizações de tarefas
Resource_View_Quantidade	Quantidade acumulada de visualizações de material de apoio
Resource_View_Tempo	Tempo acumulado em visualizações de material de apoio
Quiz_Quantidade	Quantidade acumulada de respostas a questionários
Quiz_TempoUso	Tempo acumulado em atividades de questionários
Log_Post_Quantidade	Total acumulado de postagens
Log_View_Quantidade	Tempo acumulado de visualizações
TempoUsoTotal	Tempo acumulado total em atividades no AVA
TempoUsoMadrugada	Tempo acumulado em atividades entre 0:00 e 06:00
TempoUsoManha	Tempo acumulado em atividades no turno da manhã
TempoUsoTarde	Tempo acumulado em atividades no turno da tarde
TempoUsoNoite	Tempo acumulado em atividades no turno da noite
Dias_Semana_Distintos	Quantidade acumulada de dias distintos em que fez login

Como forma de compatibilizar unidades e dispersões distintas entre as variáveis, cada variável tem valores normalizados por meio de *z-scores* (Han et al., 2011), desta maneira tem-se dados normalizadas com média zero e desvio padrão um.

4.4 Mineração

De modo a permitir intervenções proativas por parte de instituições de ensino e que possam evitar insucessos acadêmicos, este trabalho tem como foco cenários de predição que prestem suporte a sistemas conhecidos como *Early Warning Systems* (Hu et al., 2014; Jayaprakash et al., 2014), cujas predições são realizadas para um período letivo corrente, como ilustrado na Figura 1(b). Neste sentido, são analisados dois cenários: semana 3 e semana 6, representando um cenário prático em que predições são realizadas para a semana 4 e 7, respectivamente. Assim, cada disciplina tem variáveis de entrada que acumulam atividades de estudantes até a semana 3, em um cenário mais precoce, e até a semana 6, configurando um cenário mais tardio e supostamente com dados mais ricos.

Como ilustrado na Figura 3, objetiva-se a predição/classificação dos contextos instrucionais nos quais os estudantes cursaram determinada disciplina. Especificamente, os rótulos de con-

textos instrucionais a serem classificados/distinguidos são dois, e correspondem aos contextos de dados que seriam empregados na fase de treino (passado) e na fase de teste (presente) em uma tarefa de predição de sucesso/insucesso acadêmico. Para isso, dentro do histórico de turmas de uma disciplina, cada estudante é rotulado conforme o período letivo em que cursou tal disciplina. Deste modo, a coluna **Contexto**, ilustrada na Figura 3, constitui a variável alvo do processo de classificação supervisionada.

Com a nova coluna **Contexto** que indica se o estudante é proveniente de treino ou de teste, utiliza-se algoritmos de *machine learning* para que estes aprendam as relações entre as variáveis de entrada e a variável alvo/saída. Neste trabalho, emprega-se dois algoritmos comumente utilizados em problemas de classificação, Árvores de Decisão e *Support Vector Machine* (SVM). Na área de MDE, árvores de decisão são usuais (Peña-Ayala, 2014) e justificáveis pela qualidade de permitir uma melhor interpretação dos modelos preditivos estimados (Faceli et al., 2011). SVM também é usual na MDE, e justifica-se por frenquentemente alcançar maiores taxas de assertividade em estudos comparativos (Costa et al., 2017; Thammasiri et al., 2014) em diversas áreas de aplicação (Liu, 2011). As implementações disponíveis na biblioteca de software Scikit-learn[3] (Pedregosa et al., 2011) foram empregadas para ambos algoritmos neste trabalho, sendo que o ajuste de parâmetros se deu pela abordagem de busca exaustiva (*grid search*), como implementado no pacote de seleção de modelos[4].

Em resumo, a tarefa de mineração (classificação supervisionada) proposta neste trabalho constitui-se de 16 experimentos: 4 disciplinas $\times$ 2 cenários de predição (semanas 3 e 6) $\times$ 2 algoritmos (Árvores de Decisão e SVM).

4.5 Validação

Algoritmos de *machine learning*, como Árvores de Decisão e SVM, podem apresentar resultados tendenciosos quando utilizados diretamente sobre cenários de dados desbalanceados (He & Garcia, 2009), isto é, quando na fase de treinamento, há uma desproporção entre entradas das classes envolvidas na tarefa de classificação. Na prática, este cenário é recorrente em tarefas de predição de sucesso/insucesso acadêmico, uma vez que, historicamente, é comum haver uma quantidade significativamente menor de estudantes na classe de insucesso (Márquez-Vera et al., 2013).

De forma a minimizar a influência de desbalanceamento de dados, considerando ainda a disponibilidade não homogênea do histórico de dados/estudantes ao longo dos períodos letivos, procedeu-se com o agrupamento mostrado na Tabela 4 para a formação dos conjuntos (de estudantes) de treino e teste.

Estando os estudantes rotulados como pertencentes a contextos de treino e teste, adota-se o método usual de validação cruzada em 10 partes (*10-fold-cross-validation*) como forma de validar os modelos preditivos (Faceli et al., 2011). Cada *fold* foi balanceado com estudantes provenientes dos contextos de treino e de teste, empregando-se amostragem aleatória sobre a classe de treino, uma vez que, como observa-se na Tabela 4, há mais estudantes no contexto de treino do que no contexto de teste.

[3] http://scikit-learn.org
[4] http://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html

Tabela 4: Distribuição de estudantes por períodos letivos. Agrupamento de períodos letivos, verde e azul, delimitam os contextos instrucionais de treino e teste em cada disciplina.

PERÍODOS LETIVOS	**DISCIPLINAS**			
Ano/semestre/bimestre	**Lógica**	**Matemática para Administração**	**Processo Administrativo**	**Raciocínio Lógico**
2012/01/01	119	41	253	
2012/01/02		140		226
2012/02/01	84	49	165	
2012/02/02		131		185
2013/01/01	69	58	156	
2013/01/02		124	36	281
2013/02/01	81	72	77	
2013/02/02		72	70	243
2014/01/01	68	48	91	
2014/01/02			78	68
Treino	**272**	**419**	**574**	**692**
Teste	**149**	**316**	**352**	**311**

Neste *setup* de validação, cabe ainda destacar que, para cada *fold*, manteve-se o desbalanceamento original das disciplinas em termos de estudantes de sucesso e insucesso. Desta forma, os contextos de treino e teste constituem-se das mesmas taxas de estudantes com insucesso/sucesso, minimizando-se conclusões triviais a respeito da possível causa de diferenças, entre contextos de treino e teste, estar no fato dos contextos apresentarem diferentes razões entre estudantes de sucesso e insucesso. A Figura 4 ilustra a distribuição de estudantes em cada *fold* do processo de validação cruzada.

Treino	**50%**	*sucesso* *insucesso*	*x%* *y%*
Teste	**50%**	*sucesso* *insucesso*	*x%* *y%*

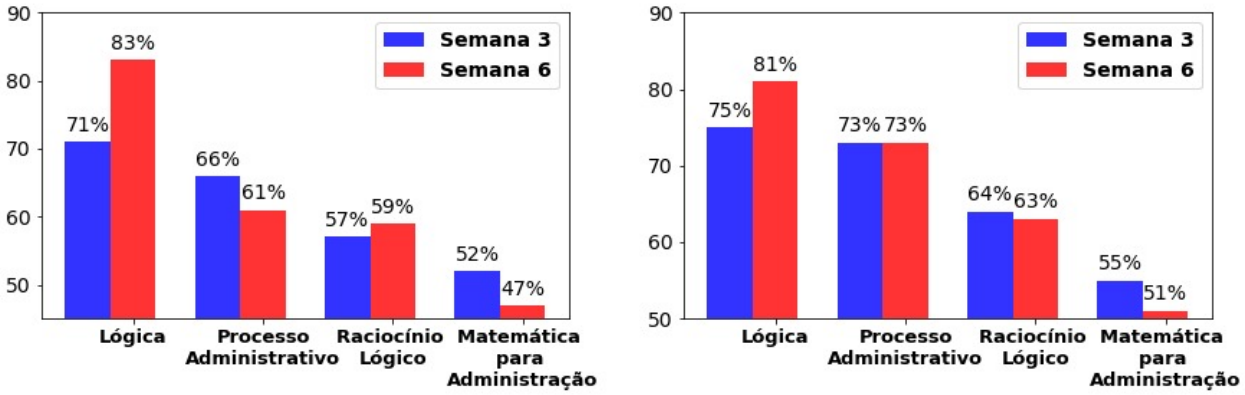
Figura 4: Distribuição de estudantes em cada *fold* do processo de validação cruzada. Os percentuais *x* e *y* correspondem as proporções de sucesso e insucesso em cada disciplina..

5 Resultados

Nesta seção são apresentados os resultados de assertividade na tarefa de classificar/distinguir os contextos instrucionais (treino versus teste) definidos pelo agrupamento de estudantes mostrado na Tabela 4, considerando apenas duas informações na delimitação do que se entende por contexto instrucional: períodos letivos e disciplinas. Cabe destacar que a Tabela 4 constitui um típico cenário de classificação supervisionada, composto por dados de treino e teste, em que buscaria-se predizer o sucesso/insucesso acadêmico dos estudantes que compõe o conjunto de teste. Tal tarefa de predição pode ser negativamente afetada em caso da existência do fenômeno de *covariate shift*, isto é, em caso da existência de diferenças estatísticas entre os dados empregados no treinamento do modelo preditivo e os dados provenientes do contexto para o qual objetiva-se realizar predições/inferências. Deste modo, a assertividade na tarefa de distinguir contextos de (dados de)

treino versus teste deve ser analisada como indicativo da necessidade de empregar-se técnicas de compatibilização de tais contextos, antes de proceder-se para a tarefa preditiva propriamente dita, como discutido na Seção 2.2.

Um vez que trata-se de uma tarefa de classificação/predição sobre um cenário balanceado de dados, adota-se tão somente a acurácia média (entre os *folds* da validação cruzada) como métrica de avaliação de resultados. A Figura 5 mostra a acurácia alcançada pelos algoritmos SVM e Árvore de Decisão em cada cenário de predição. A Tabela 5 estratifica os valores de acurácia em termos de estudantes que obtiveram sucesso e insucesso. Desta maneira, pode-se observar:



(a) (b)

Figura 5: Acurácia média na distinção de contextos instrucionais que caracterizam cenários de treino e teste (ver Tabela 4) em uma tarefa típica de predição de sucesso acadêmico de estudantes. Acurácia como função de 4 disciplinas, 2 instantes de predição (semana 3 e 6) e 2 algoritmos: (a) Árvore de decisão e (b) SVM.

Disciplina	**Classificador**	**Classe**	**Acurácia**	
			Semana 3	Semana 6
Processo Administrativo	Árvore de Decisão	Suc	0.69	0.63
		Insuc	0.61	0.57
	SVM	Suc	0.75	**0.75**
		Insuc	**0.68**	0.66
Lógica	Árvore de Decisão	Suc	0.83	**0.92**
		Insuc	0.65	**0.78**
	SVM	Suc	0.82	0.92
		Insuc	0.71	0.74
Matemática para Administração	Árvore de Decisão	Suc	0.56	0.44
		Insuc	0.49	0.44
	SVM	Suc	**0.58**	0.50
		Insuc	**0.54**	0.52
Raciocínio Lógico	Árvore de Decisão	Suc	0.58	0.63
		Insuc	0.56	0.55
	SVM	Suc	0.65	**0.66**
		Insuc	**0.63**	0.60

Tabela 5: Acurácia média de cada classificador em função da variável sucesso acadêmico. Em negrito destacam-se os melhores resultados para cada disciplina de acordo com o sucesso ou insucesso de cada estudante. .

- Os algoritmos de Árvore de Decisão e SVM produziram resultados proporcionalmente similares quando a análise volta-se às disciplinas, indicando resultados mais consistentes com

os contextos representados pelas disciplinas e menos afetados por virtudes de um algoritmo em relação ao outro.

- Dentre as 4 disciplinas, 3 resultaram em valores médios de acurácia acima de 65%, culminando com valores que ultrapassam 80% na disciplina de lógica, indicando que um tratamento dos efeitos de *covariate shift* pode ser benéfico em uma tarefa de predição de sucesso acadêmico neste cenário de dados.

- Dentre as 4 disciplinas, matemática para administração apresentou os menores valores de acurácia, indicando possíveis contextos instrucionais mais estáveis ao longo do histórico de períodos letivos analisados.

- Por outro lado, observa-se que os valores de acurácia são inversamente proporcionais a quantidade de turmas em cada disciplina (ver Tabela 2). Os valores de acurácia são maiores na disciplina Lógica (apenas 6 turmas), decrescendo para a disciplina de Processo Administrativo (12 turmas), passando por Raciocínio Lógico (14 turmas) e coincidindo os piores valores de acurácia com a maior quantidade de turmas (21 turmas) na disciplina de Matemática para Administração. Empiricamente, a justificativa pode estar no fato de que a maior quantidade de turmas implica em uma diversidade de contextos educacionais presente tanto no conjunto de dados de treino quanto no de teste, resultando em maiores dificuldades na tarefa de distinguir o conjunto de turmas provenientes dos períodos letivos de treino do conjunto de turmas provenientes dos períodos letivos de teste.

- Na Tabela 5, destacam-se as altas acurácias dentre estudantes de sucesso na disciplina de lógica, indicando uma clara diferença de comportamento de estudantes de sucesso entre os contextos instrucionais analisados. Seja qual for a causa, este resultado indica que alcançar sucesso, por parte de estudantes, no contexto de treino (períodos letivos de 2012/1 a 2013/1) mostrou-se ser algo claramente separável de se alcançar sucesso no contexto instrucional de teste (períodos letivos de 2013/2 a 2014/1), ainda que ambos contextos estejam quantificados/representados pelo mesmo conjunto de variáveis.

Adicionalmente, e tendo por base as disciplinas com maior acurácia, a Tabela 6 mostra as (três) variáveis de entrada que melhor distinguem os contextos de treino e teste segundo o critério *Information Gain* (Quinlan, 1986), usual também na MDE (Ramaswami & Bhaskaran, 2009; Romero et al., 2013; Baradwaj & Pal, 2011). Observa-se que as interações sobre questionários se mostram relevantes na disciplina de Lógica, sendo as interações com material de apoio também importantes na disciplina de Processo Administrativo. Ainda que com uma acurácia resultante menor, interações de cunho social, envolvendo fóruns e chats, predominam no contexto da disciplina de Raciocínio Lógico.

5.1 Discussão de resultados

É razoável considerar que estudantes que participam de uma mesma turma, em uma disciplina, frequentam o mesmo espaço virtual no AVA e são expostos ao mesmo conjunto de atividades de ensino, ainda que possam reagir de forma distinta a um mesmo conjunto de atividades. Desta forma, uma turma em uma disciplina parece ser um ponto de partida razoável, e factível, na

Tabela 6: Seleção das três melhores variáveis segundo *Information Gain*.

Disciplina	Semana 3	Semana 6
Lógica	Quiz_Quantidade	Quiz_Quantidade
	Log_Post_Quantidade	Quiz_TempoUso
	Quiz_TempoUso	Log_Post_Quantidade
Processo Administrativo	Quiz_Quantidade	Resource_View_Quantidade
	Resource_View_Quantidade	Resource_View_Tempo
	Resource_View_Tempo	Quiz_Quantidade
Raciocínio Lógico	Forum_Quantidade_Post	Forum_Quantidade_Post
	Forum_Quantidade_Visualizações	Assignment_View_Quantidade
	Chat_TempoUso	Chat_TempoUso

delimitação de contextos instrucionais que possam ser caracterizados quantitativamente em termos de dados gerados pelos estudantes envolvidos.

Na prática, não são raras as ocasiões em que a organização de atividades de uma disciplina sofre alterações ao longo de períodos letivos, ou que turmas sejam conduzidas por diferentes educadores/tutores, fazendo com que turmas de estudantes que cursaram a mesma disciplina, porém em períodos letivos diferentes, talvez não possam ser diretamente comparadas ou agrupadas sob o mesmo cenário de atividades ou mesmas condições instrucionais. Assim, embora turmas possam ser distintas em termos de seu contexto instrucional, na prática de tarefas preditivas na MDE diversas turmas são agrupadas para o treinamento de modelos preditivos, sem um controle de variáveis que podem determinar diferenças significativas de condições instrucionais entre as turmas que compõe o histórico de uma disciplina. Gašević et al. (2016) aponta que estas circunstâncias podem ajudar a explicar divergências de resultados na literatura de predição de sucesso de acadêmico.

Da mesma forma, como mostrado na Tabela 4, neste trabalho diversas turmas são agrupadas, porém para testar a hipótese de que dois grupos de turmas, pertencentes ao histórico de uma disciplina, são, na verdade, dois contextos instrucionais distintos, tendo o período letivo como único critério para tal divisão: (i) turmas de períodos letivos mais antigos, representando um contexto instrucional histórico sobre o qual um modelo preditivo seria treinado, e (ii) turmas de estudantes de períodos letivos mais recentes, representando um contexto instrucional de teste em que o modelo preditivo seria usado para produzir inferências.

Assim, os resultados de acurácia mostrados na Seção 5 devem ser analisados a luz dos agrupamentos de turmas formados e necessários em tarefas preditivas. Neste sentido, a similaridade observada entre contextos de treino e teste na disciplina de matemática para administração, verificada pela baixa acurácia alcançada na distinção/classificação de tais contextos, deve ser também analisada com base na maior quantidade de turmas que compõe o histórico da disciplina. Neste cenário, não deve-se descartar uma análise mais profunda que possa revelar que turmas com características similares estejam presentes em ambos contextos de treino e teste, tornando-os similares nesta situação, mas que em um outro cenário de agrupamento de turmas possa indicar contextos divergentes. Por outro lado, apesar do agrupamento de turmas poder influenciar negativamente na distinção de contextos instrucionais, os demais resultados, como verificados para a disciplina de lógica, confirmam a principal hipótese deste trabalho a respeito da existência do fenômeno de *covariate shift* em cenários de predição de sucesso acadêmico de estudantes.

A conclusão acima, pela demanda em analisar fenômenos como *covariate shift*, está em acordo com outros resultados apontados na literatura. Pela análise de diferentes níveis de assertividade na predição de sucesso em 9 disciplinas, Gašević et al. (2016) concluiu pela importância crítica de considerar condições instrucionais no desenvolvimento de modelos preditivos, sobretudo na análise de características de dados de interações com AVA e consequências ao fundi-los de forma descontextualizadas na busca por desejáveis modelos preditivos mais genéricos/portáveis. Jayaprakash et al. (2014) também destaca a necessidade de compatibilizar-se dados previamente a construção de modelos preditivos, ainda que não reporte resultados como função de diferentes contextos instrucionais. Por outro lado, os resultados obtidos na disciplina de matemática para administração, indicando uma possível similaridade entre os contextos de treino e teste, estão em acordo com as conclusões de Boyer & Veeramachaneni (2015). Embora realize analises em um cenário de dados massivos e tenha foco em evasão, Boyer & Veeramachaneni (2015) não conclui pela melhoria de assertividade ao empregar *transfer learning* na maior parte dos cenários testados, remetendo para uma possível inexistência de diferenças significativas entre dados de treino e teste, como verificado neste trabalho para a disciplina de matemática para administração.

6 Conclusão

Este trabalho apresentou uma análise empírica com o objetivo de verificar indícios de divergências entre distribuições de dados que são gerados em contextos educacionais envolvidos em tarefas de predição de insucesso acadêmico. Em tarefas de predição, usualmente emprega-se técnicas de *machine learning* que são capazes de aprender padrões sobre dados históricos (de comportamento de estudantes) a fim de utilizar-se o padrão aprendido sobre dados (de comportamentos de estudantes) provenientes de períodos (letivos) correntes. Assumindo-se, desta forma, que dados (comportamentos) do passado e do presente seguem uma mesma distribuição de dados. Contudo, raros são os trabalhos na literatura de Mineração de Dados Educacionais e/ou *Learning Analytics* que avaliam as consequências de tal suposição em tarefas preditivas.

Sobre *logs* de interações com AVA de mais de 3.000 estudantes, na modalidade de ensino superior a distância, verificou-se indícios de divergências, na forma de *covariate shift*, entre dados de estudantes provenientes de períodos letivos distintos em 3 disciplinas, dentre 4 avaliadas. A partir dos resultados do estudo proposto, contribui-se com uma elucidação de possíveis causas para resultados divergentes na literatura de predição de sucesso acadêmico de estudantes (Gašević et al., 2016), ao mesmo tempo que recomenda-se uma análise prévia, embora nem sempre necessária, no sentido de compatibilizar-se contextos históricos de dados visando melhorias de assertividade dos modelos preditivos.

Como continuação natural deste trabalho, proceder-se-á com a aplicação de técnicas de *transfer learning*, como a técnica proposta por Sun et al. (2016), a fim de avaliar a possível melhora de assertividade de predição para os contextos em que verificou-se uma maior separabilidade de dados de comportamento de estudantes. Com isso, espera-se contribuir na busca por modelos preditivos escaláveis e/ou portáveis através de diferentes contextos educacionais/instrucionais, os quais são de grande interesse da indústria de softwares educacionais (Essa & Ayad, 2012).

Acknowledgements

Agradecemos à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pela bolsa de estudos concedida para Lilian Landvoigt da Rosa.

References

Agudo-Peregrina, Á. F., Iglesias-Pradas, S., Conde-González, M. Á., & Hernández-García, Á. (2014). Can we predict success from log data in vles? classification of interactions for learning analytics and their relation with performance in vle-supported f2f and online learning. *Computers in human behavior*, *31*, 542–550. [GS Search] doi: 10.1016/j.chb.2013.05.031

Baker, R., Isotani, S., & Carvalho, A. (2011). Mineração de dados educacionais: Oportunidades para o brasil. *Revista Brasileira de Informática na Educação*, *19*(02), 03. [GS Search] doi: 10.5753/RBIE.2011.19.02.03

Baradwaj, B. K., & Pal, S. (2011). Mining educational data to analyze students' performance. *International Journal of Advanced Computer Science and Applications*, *2*(6). Retrieved from https://arxiv.org/ftp/arxiv/papers/1201/1201.3417.pdf [GS Search]

Barber, R., & Sharkey, M. (2012). Course correction: Using analytics to predict course success. In *Proceedings of the 2nd international conference on learning analytics and knowledge* (pp. 259–262). [GS Search] doi: 10.1145/2330601.2330664

Bickel, S., Brückner, M., & Scheffer, T. (2007). Discriminative learning for differing training and test distributions. In *Proceedings of the 24th international conference on machine learning* (pp. 81–88). [GS Search] doi: 10.1145/1273496.1273507

Bousbia, N., & Belamri, I. (2014). Which contribution does edm provide to computer-based learning environments? In *Educational data mining* (pp. 3–28). Springer. [GS Search] doi: 10.1007/978-3-319-02738-8_1

Boyer, S., & Veeramachaneni, K. (2015). Transfer learning for predictive models in massive open online courses. In *International conference on artificial intelligence in education* (pp. 54–63). [GS Search] doi: 10.1007/978-3-319-19773-9_6

Cechinel, C., Araujo, R. M., & Detoni, D. (2015). Modelling and prediction of distance learning students failure by using the count of interactions. *Brazilian Journal of Computers in Education*, *23*(03), 1. [GS Search] doi: 10.5753/RBIE.2015.23.03.1

Chatti, M. A., Dyckhoff, A. L., Schroeder, U., & Thüs, H. (2012). A reference model for learning analytics. *International Journal of Technology Enhanced Learning*, *4*(5-6), 318–331. [GS Search] doi: 10.1504/IJTEL.2012.051815

Costa, E. B., Fonseca, B., Santana, M. A., de Araújo, F. F., & Rego, J. (2017). Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic

failure in introductory programming courses. *Computers in Human Behavior*, *73*, 247–256. [GS Search] doi: 10.1016/j.chb.2017.01.047

Daume III, H., & Marcu, D. (2006). Domain adaptation for statistical classifiers. *Journal of Artificial Intelligence Research*, *26*, 101–126. [GS Search] doi: 10.1613/jair.1872

Dawson, S., Gašević, D., Siemens, G., & Joksimovic, S. (2014). Current state and future trends: A citation network analysis of the learning analytics field. In *Proceedings of the fourth international conference on learning analytics and knowledge* (pp. 231–240). [GS Search] doi: 10.1145/2567574.2567585

Duval, E. (2011). Attention please!: learning analytics for visualization and recommendation. In *Proceedings of the 1st international conference on learning analytics and knowledge* (pp. 9–17). [GS Search] doi: 10.1145/2090116.2090118

Er, E. (2012). Identifying at-risk students using machine learning techniques: A case study with is 100. *International Journal of Machine Learning and Computing*, *2*(4), 476. Retrieved from http://www.ijmlc.org/papers/171-L003.pdf [GS Search]

Essa, A., & Ayad, H. (2012). Improving student success using predictive models and data visualisations. *Research in Learning Technology*, *20*(sup1), 19191. [GS Search] doi: 10.3402/rlt.v20i0.19191

Faceli, K., Lorena, A. C., Gama, J., & Carvalho, A. (2011). Inteligência artificial: Uma abordagem de aprendizado de máquina. *Rio de Janeiro: LTC*. [GS Search]

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine*, *17*(3), 37. [GS Search] doi: 10.1609/aimag.v17i3.1230

Ferguson, R. (2012). Learning analytics: drivers, developments and challenges. *International Journal of Technology Enhanced Learning*, *4*(5-6), 304–317. [GS Search] doi: 10.1504/IJTEL.2012.051816

Fortenbacher, A., Beuster, L., Elkina, M., Kappe, L., Merceron, A., Pursian, A., ... Wenzlaff, B. (2013). Lemo: A learning analytics application focussing on user path analysis and interactive visualization. In *Intelligent data acquisition and advanced computing systems (idaacs), 2013 ieee 7th international conference on* (Vol. 2, pp. 748–753). [GS Search] doi: 10.1109/IDAACS.2013.6663025

Gašević, D., Dawson, S., Rogers, T., & Gasevic, D. (2016). Learning analytics should not promote one size fits all: The effects of instructional conditions in predicting academic success. *The Internet and Higher Education*, *28*, 68–84. [GS Search] doi: 10.1016/j.iheduc.2015.10.002

Gottardo, E., Kaestner, C. A. A., & Noronha, R. V. (2014). Estimativa de desempenho acadêmico de estudantes: Análise da aplicação de técnicas de mineração de dados em cursos a distância. *Revista Brasileira de Informática na Educação*, *22*(1). Retrieved from https://pdfs.semanticscholar.org/67e4/2719d9ea7785af18436058c154d750c3d78c.pdf [GS Search]

Han, J., Pei, J., & Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier. [GS Search]

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, *21*(9), 1263–1284. [GS Search] doi: 10.1109/TKDE.2008.239

Hoang, N. D., Chau, V. T. N., & Phung, N. H. (2016). Combining transfer learning and co-training for student classification in an academic credit system. In *Computing & communication technologies, research, innovation, and vision for the future (rivf), 2016 ieee rivf international conference on* (pp. 55–60). [GS Search] doi: 10.1109/RIVF.2016.7800269

Hu, Y.-H., Lo, C.-L., & Shih, S.-P. (2014). Developing early warning systems to predict students' online learning performance. *Computers in Human Behavior*, *36*, 469–478. [GS Search] doi: 10.1016/j.chb.2014.04.002

Jayaprakash, S. M., Moody, E. W., Lauría, E. J., Regan, J. R., & Baron, J. D. (2014). Early alert of academically at-risk students: An open source analytics initiative. *Journal of Learning Analytics*, *1*(1), 6–47. [GS Search] doi: 10.18608/jla.2014.11.3

Kampff, A. J. C. (2009). Mineração de dados educacionais para geração de alertas em ambientes virtuais de aprendizagem como apoio à prática docente. Retrieved from http://www.lume.ufrgs.br/handle/10183/19032 [GS Search]

Lagus, J. (2016). *Course outcome prediction with transfer learning methods* (Master's thesis, University of Helsinki, Helsinki, Finland). [GS Search] doi: 10138/165915

Lara, J. A., Lizcano, D., Martínez, M. A., Pazos, J., & Riera, T. (2014). A system for knowledge discovery in e-learning environments within the european higher education area–application to student data from open university of madrid, udima. *Computers & Education*, *72*, 23–36. [GS Search] doi: 10.1016/j.compedu.2013.10.009

Liu, B. (2011). *Web data mining: Exploring hyperlinks, contents, and usage data*. Springer Berlin Heidelberg. Retrieved from https://books.google.com.br/books?id=jnCi0Cq1YVkC [GS Search]

Lu, J., Behbood, V., Hao, P., Zuo, H., Xue, S., & Zhang, G. (2015). Transfer learning using computational intelligence: a survey. *Knowledge-Based Systems*, *80*, 14–23. [GS Search] doi: 10.1016/j.knosys.2015.01.010

Macfadyen, L. P., & Dawson, S. (2010). Mining lms data to develop an "early warning system" for educators: A proof of concept. *Computers & education*, *54*(2), 588–599. [GS Search] doi: 10.1016/j.compedu.2009.09.008

Manhães, L. M. B., Da Cruz, S. M. S., Costa, R. J. M., Zavaleta, J., & Zimbrão, G. (2011). Previsão de estudantes com risco de evasão utilizando técnicas de mineração de dados. In *Brazilian symposium on computers in education (simpósio brasileiro de informática na educação-sbie)* (Vol. 1). Retrieved from http://br-ie.org/pub/index.php/sbie/article/view/1585 [GS Search]

Márquez-Vera, C., Cano, A., Romero, C., & Ventura, S. (2013). Predicting student failure at school using genetic programming and different data mining approaches with high dimensional and imbalanced data. *Applied intelligence*, *38*(3), 315–330. [GS Search] doi: 10.1007/s10489-012-0374-8

Martin, F. (2014 (accessed November 1, 2016)). A simple machine learning method to detect covariate shift [Computer software manual]. Retrieved from https://blog.bigml.com/2014/01/03/simple-machine-learning-to-detect-covariate-shift/

Moreno-Torres, J., Raeder, T., Alaiz-Rodríguez, R., Chawla, N., & Herrera, F. (2012). A unifying view on dataset shift in classification. *Pattern Recognition*, *45*(1), 521–530. [GS Search] doi: 10.1016/j.patcog.2011.06.019

Pan, S. J., & Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 1345–1359. [GS Search] doi: 10.1109/TKDE.2009.191

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... others (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, *12*(Oct), 2825–2830. Retrieved from https://dl.acm.org/citation.cfm?id=2078195 [GS Search]

Peña-Ayala, A. (2013). *Educational data mining: Applications and trends* (Vol. 524). Springer. [GS Search]

Peña-Ayala, A. (2014). Educational data mining: A survey and a data mining-based analysis of recent works. *Expert systems with applications*, *41*(4), 1432–1462. [GS Search] doi: 10.1016/j.eswa.2013.08.042

Quinlan, J. R. (1986, mar). Induction of decision trees. *Machine Learning*, *1*(1), 81–106. [GS Search] doi: https://doi.org/10.1007/BF00116251

Quionero-Candela, J., Sugiyama, M., Schwaighofer, A., & Lawrence, N. D. (2009). *Dataset shift in machine learning*. The MIT Press. [GS Search]

Ramaswami, M., & Bhaskaran, R. (2009). A study on feature selection techniques in educational data mining. *Journal of Computing*, *1*(1), 7–11. Retrieved from https://arxiv.org/abs/0912.3924 [GS Search]

Raza, H., Prasad, G., & Li, Y. (2015). Ewma model based shift-detection methods for detecting covariate shifts in non-stationary environments. *Pattern Recognition*, *48*(3), 659–669. [GS Search] doi: 10.1016/j.patcog.2014.07.028

Rigo, S. J., Cambruzzi, W., Barbosa, J. L., & Cazella, S. C. (2014). Educational data mining and learning analytics applications in dropout: opportunities and challenges. *Brazilian Journal of Computers in Education*, *22*(01), 132. [GS Search] doi: 10.5753/rbie.2014.22.01.132

Romero, C., López, M.-I., Luna, J.-M., & Ventura, S. (2013). Predicting students' final performance from participation in on-line discussion forums. *Computers & Education*, *68*, 458–472. [GS Search] doi: 10.1016/j.compedu.2013.06.009

Santos, J. L., Govaerts, S., Verbert, K., & Duval, E. (2012). Goal-oriented visualizations of activity tracking: a case study with engineering students. In *Proceedings of the 2nd international conference on learning analytics and knowledge* (pp. 143–152). [GS Search] doi: 10.1145/2330601.2330639

Siemens, G., & Baker, R. S. (2012). Learning analytics and educational data mining: towards communication and collaboration. In *Proceedings of the 2nd international conference on learning analytics and knowledge* (pp. 252–254). [GS Search] doi: 10.1145/2330601.2330661

Siemens, G., & Long, P. (2011). Penetrating the fog: Analytics in learning and education. *EDUCAUSE review*, *46*(5), 30. Retrieved from https://er.educause.edu/articles/2011/9/penetrating-the-fog-analytics-in-learning-and-education [GS Search]

Simpson, O. (2004). The impact on retention of interventions to support distance learning students. *Open Learning: The Journal of Open, Distance and e-Learning*, *19*(1), 79–95. [GS Search] doi: 10.1080/0268051042000177863

Sugiyama, M., Nakajima, S., Kashima, H., & Buenau, P. V. (2008). Direct importance estimation with model selection and its application to covariate shift adaptation. In *Advances in neural information processing systems* (pp. 1433–1440). [GS Search] doi: 10.1007/s10463-008-0197-x

Sun, B., Feng, J., & Saenko, K. (2016). Return of frustratingly easy domain adaptation. In *Aaai conference on artificial intelligence* (Vol. 6, p. 8). Retrieved from https://arxiv.org/abs/1511.05547 [GS Search]

Thammasiri, D., Delen, D., Meesad, P., & Kasap, N. (2014). A critical assessment of imbalanced class distribution problem: The case of predicting freshmen student attrition. *Expert Systems with Applications*, *41*(2), 321–330. [GS Search] doi: 10.1016/j.eswa.2013.07.046

Voß, L., Schatten, C., Mazziotti, C., & Schmidt-Thieme, L. (2015). A transfer learning approach for applying matrix factorization to small its datasets. *International Educational Data Mining Society*. Retrieved from https://files.eric.ed.gov/fulltext/ED560791.pdf [GS Search]

You, J. W. (2016). Identifying significant indicators using lms data to predict course achievement in online learning. *The Internet and Higher Education*, *29*, 23–30. [GS Search] doi: 10.1016/j.iheduc.2015.11.003

Zadrozny, B. (2004). Learning and evaluating classifiers under sample selection bias. In *Proceedings of the 21th international conference on machine learning* (p. 114). [GS Search] doi: 10.1145/1015330.1015425