Aplicação do modelo de regressão logística na analítica de aprendizagem

Olga L. Anglas R. Tarumoto[1], Thais Tomé Carvo[2], Mário H. Tarumoto[3]

1 Introdução

Na área da Educação o uso das Tecnologias Digitais de Informação e Comunicação (TDIC) desempenharam um papel muito importante na forma de comunicarmos, aprendermos e vivermos, constituindo-se assim uma forte aliada ao processo de ensino – aprendizagem na Educação a Distância (EaD), no entanto não é mais um privilégio de cursos ministrados nesta modalidade, pois está sendo utilizado também no ensino presencial e no ensino híbrido, tanto de área pública como privada. Desta forma não se pode pensar em ensino, aprendizagem e avaliação sem o uso destas tecnologias independente da modalidade de ensino. Esses cursos se desenvolvem nos Ambientes Virtuais de Aprendizagem (AVA) disponibilizando aos estudantes diversas ferramentas pedagógicas configurando-se como uma alternativa ao processo de ensino – aprendizagem virtual ou semi-presencial.

Várias instituições de ensino já estão implementando sistemas de avaliação da qualidade dos cursos. Diante disso, os professores são levados a adotar inovações em termos de ferramentas e estratégias que lhes possibilitam identificar os estudantes com maior dificuldade de aprendizagem e construir novas formas de apoiar seu processo de aquisição de conhecimentos (MACFADYEN E DAWSON, 2010). Neste contexto, surge uma nova solução tecnológica que promete auxiliar a universidade como um todo a avaliar e monitorar os estudantes: a analítica. Nascida no meio empresarial, a analítica de negócios/analítica empresarial (*business analytics*), também denominada inteligência de negócios/inteligência empresarial (*business intelligence*), surgiu como um processo que integra metodologias, *hardware* e *software* para coletar, armazenar e analisar grandes quantidades de dados, a fim de possibilitar uma tomada de decisão nas organizações mais rápida e fundamentada (BALLARD *et al*., 2006). Esta ideia está sendo implementada também no campo educacional. Ferguson (2013) destaca três expressões que representam áreas de pesquisa com interesses similares e que se sobrepõem: "mineração de dados educacionais" (*education data mining*), "analítica acadêmica" (*academic analytics*) e "analítica de aprendizagem" (*learning analytics*). O objetivo deste trabalho, foi o de aplicar o modelo de regressão logística para analisar os dados obtidos no AVA, configurando se como parte da analítica de aprendizagem, cujo interesse é estimar a probabilidade de aprovação em uma determinada disciplina ministrada a distância.

2 Alguns fundamento teóricos

2.1 Analítica de Aprendizagem

Não há uma definição única e consensual sobre o tema. Para a *Society for Learning Analytics Research* (SoLAR), a Analítica de Aprendizagem é considerada como uma metodologia que permite a "medição, coleta, análise e divulgação de dados sobre os estudantes e seus contextos, com o propósito de compreender e otimizar a aprendizagem e os ambientes em que ela ocorre". Segundo (ELIAS, 2011), é um campo emergente onde ferramentas analíticas são usadas para melhorar a aprendizagem e a educação. Segundo Chatti *et al*. (2012), é um campo multidisciplinar que envolve

[1] Depto de Estatística – FCT/Unesp. email: *olga.anglas@unesp.br*.
[2] Curso de Estatística – FCT/Unesp. email: *thaiscarvo1996@gmail.com*.
[3] Depto de Estatística – FCT/Unesp. email: *mario.tarumoto@unesp.br*.

aprendizado de máquina, inteligência artificial, recuperação de dados, estatísticas e visualização. Além disso, é um campo em que várias áreas convergem para a Aprendizagem Aprimorada pela Tecnologia. Estas áreas são: Analítica Acadêmica, Pesquisa-ação, Mineração de Dados Educacionais, Sistemas de Recomendação e Aprendizagem Adaptativa.

Ainda de acordo com Chatti *et al.* (2012), o processo da Analítica de Aprendizagem é um ciclo iterativo e é geralmente realizado em três etapas principais: 1. Coleta de dados e pré-processamento, 2. Análise e ação e 3. Pós-processamento.

2.2 Ambiente Virtual de Aprendizagem (AVA)

O Ambiente Virtual de Aprendizagem (AVA) é uma sala virtual que permite a utilização de mídias, linguagens e recursos que apresentam conteúdos e interações necessários para a aprendizagem do estudante proporcionando a interação e comunicação entre seus atores (estudante, tutor, professor). Como produto dessas interações, esses ambientes armazenam uma grande quantidade de dados, que na sua forma bruta não fornecem nenhuma informação, porém se transformados adequadamente produzem informações significativas para análises e tomada de decisões úteis para cada um dos seus atores como também a nível institucional. Segundo Cônsolo (2016), o AVA é constituído pela parte tecnológica e pela parte pedagógica e embora cada fabricante crie seu ambiente com algumas especificidades, eles basicamente contêm ferramentas similares em seu funcionamento.

2.3 Técnicas Estatísticas

Entre as várias possibilidade de análise de dados, neste caso de Analítica de Dados, uma delas é a aplicação do Modelo de Regressão Logística, considerando que neste trabalho, o interesse é o de predizer a possibilidade de aprovação em uma disciplina. A análise de regressão é uma técnica estatística que tem como objetivo descrever a relação entre uma variável resposta e um conjunto de variáveis explicativas, através de um modelo que tenha um bom ajuste. Na regressão logística, a variável resposta é dicotômica ou binária, isto é, aquela que apresenta duas possibilidades de resposta (sucesso ou fracasso).

Hosmer e Lemeshow (2000) dizem que há pelo menos duas razões para utilização do modelo logístico na análise de variáveis respostas dicotômicas: de um ponto de vista matemático, é de extremamente flexível e fácil de ser utilizado e do ponto de vista biológico permite uma interpretação das covariáveis de forma bastante rica e direta.

Um modelo de regressão logística simples é usado para o caso de regressão com uma variável explicativa. Neste caso, a variável resposta é dicotômica, ou seja, é atribuído valor 1 para o acontecimento de interesse (“sucesso”) e valor 0 para o acontecimento complementar (“fracasso”), com probabilidades $\pi(x) = P(Y = 1|X)$ e $1 - \pi(x) = P(Y = 0|X)$, respectivamente. Na regressão logística múltipla, tem-se diferentes escalas e várias variáveis independentes. Diferente do caso simples, onde há apenas uma variável independente. Hosmer e Lemeshow (2000), considera um conjunto de p variáveis independentes pelo vetor $x' = (x_1, x_2, \dots, x_p)$, o logito é expresso por : $g(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$. E o modelo de regressão logística fica: $\pi(x) = \frac{e^{g(x)}}{1+e^{g(x)}}$.

Uma das informações importantes em um modelo de regressão logística é a razão de chances (*odds ratio*). A interpretação dos parâmetros de um modelo de regressão logística é realizada através da comparação entre a probabilidade de sucesso e a probabilidade de fracasso, usando a função odds ratio OR (razão de chances). Assim, quando a variável independente é dicotômica, ela é codificada como 0 ou 1. A razão das chances (“Odds Ratio”), denotada por Ψ será:

$$\Psi = \frac{\frac{\pi(1)}{1-\pi(1)}}{\frac{\pi(0)}{1-\pi(0)}}$$

Então, no caso do modelo de regressão logística simples, a razão de chances, para a variável independente dicotômica será:

$$\Psi = \frac{\left(\frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}\right) / \left(\frac{1}{1+e^{\beta_0+\beta_1}}\right)}{\left(\frac{e^{\beta_0}}{1+e^{\beta_0}}\right) / \left(\frac{1}{1+e^{\beta_0}}\right)} = \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}} = e^{\beta_1}$$

A razão de chances é uma medida de associação muito utilizada em muitas áreas. É um parâmetro de grande interesse no modelo de regressão logística devido sua fácil interpretação. Ela definida como a chance de ocorrência de um evento entre indivíduos que têm um fator de risco, comparadas com indivíduos não expostos, sujeitos ao evento.

3 APLICAÇÃO

3.1 Descrição e organização do banco de dados.

Para ilustrar o uso das Técnicas Estatísticas aplicadas na Analítica de Aprendizagem foram utilizados os dados provenientes do Programa Rede São Paulo de Formação (REDEFOR/Unesp) referente ao curso de especialização de Educação Especial na Perspectiva Inclusiva (EI) ministrado pela Unesp na modalidade a distância. O público alvo foram educadores do Ensino Fundamental II e Ensino Médio, que pertencem ao quadro do Magistério da SEE/SP: supervisores, diretores, professores-coordenadores e professores. O curso de EI foi realizado no período de fevereiro de 2014 a junho de 2015 com a participação de 999 estudantes que foram divididos em 29 turmas, sendo que cada turma ficou sobre a responsabilidade de um tutor denominado de Tutor-*Online*. Neste período foram ofertadas oito disciplinas (D01 até D08) com carga horária de 444 horas (mais o período de elaboração do Trabalho Acadêmico). Neste curso foi utilizado o AVA-*Moodle* (**M**odular **O**bject **O**riented **D**istance **LE**arning)/Unesp (versão 2.5). O foco neste trabalho foi na disciplina D01 – Diversidade e Cultura Inclusiva com carga horária de 50 horas, ofertada no período de 12/03/14 ao 15/04/14 com duração de 5 semanas.

No início da disciplina D01 foi apresentado uma Agenda de abertura na qual os estudantes tiveram que acessar o vídeo de abertura da disciplina. Em caso de dúvidas no decorrer da disciplina, os estudantes poderiam fazer uso de um Fórum chamado de "Fale com o Tutor". Foi solicitado também aos estudantes a leitura do Manual do Cursista. Neste trabalho foi adotado o termo estudante no lugar de cursista. Enfatizamos que durante todo o desenvolvimento da disciplina D01, o Fórum "Fale com o tutor" foi usado assiduamente pelos estudantes. Na aplicação, trabalhou-se com 10 turmas escolhidas aleatoriamente, as quais foram selecionadas usando o comando *sample* do software R. Assim as turmas que fizeram parte deste estudo foram: Turma 01 (T01) – 35 estudantes, Turma 02 (T02) – 34 estudantes, Turma 03 (T03) – 32 estudantes, Turma 05 (T05) – 35 estudantes, Turma 06 (T06) – 35 estudantes, Turma 15 (T15) – 35 estudantes, Turma 18 (T18) – 35 estudantes, Turma 25 (T25) – 35 estudantes, Turma 27 (T27) – 33 estudantes e Turma 29 (T29) – 29 estudantes, totalizando 336 estudantes.

A variável de interesse são os eventos/ações dos estudantes no Ambiente Virtual de Aprendizagem (AVA) *Moodle*. Os logs do Moodle registram quando algum usuário cria/acrescenta, visualiza, atualiza/edita, entre outros recursos. Somente os administradores do ambiente tem permissão para acessar as informações sobre os logs registrados, enquanto que os professores e

tutores tem acesso apenas aos logs do curso, dos recursos/atividades e no perfil do usuário/alunos. Por fim, os alunos podem visualizar apenas os próprios logs disponíveis a ele em seu perfil, para isso é necessário permitir mostrar relatório das atividades nas configurações do curso. Assim, o Moodle gera relatórios que informam a navegação dos seus usuários. Desta forma, do AVA – Moodle/Unesp da disciplina D01 do curso EI, retiramos todas as informações referentes aos logs de acesso e ações realizados dentro do referido ambiente (dados brutos). O *log* do Moodle é composto pelos seguintes atributos: hora, nome completo, usuário afetado, contexto do evento, componente , nome do evento, descrição, origem e endereço IP. Por outro lado, o *log* do Moodle é composto pelos seguintes componentes (arquivo, chat, diário, escolha, fórum, página, pesquisa, questionário, sistema, tarefa, URL). Neste trabalho num primeiro momento separamos os logs brutos das 10 turmas amostradas e fizemos uma limpeza nos dados, isto é, retirar todo indivíduo que não era estudante (isto é, professor, tutor, visitante, etc.), após isso separamos todos os eventos/ações que não traziam nenhuma informação. Na sequência toda essa informação das (10 turmas amostradas) foram colocadas em um único arquivo Excel contendo os dados limpos e acrescentando a informação das notas obtidas de um outro relatório referente as notas obtidas na D01. De tal forma que este último arquivo foi utilizado para realizar as análises estatísticas, em particular a previsão do desempenho dos estudantes. No final, o banco de dados possui 47 variáveis, com as contagens do número de acesso no ambiente, retiradas de eventos/ações praticadas pelos estudantes no ambiente virtual, com 10 turmas e 336 observações. Dentre estes 336 estudantes, foi observado que 307 foram reprovados e 29 aprovados, o que configura a variável resposta do trabalho.

Para a construção do modelo foi utilizado o método de seleção de variáveis Stepwise. Após 24 iterações e AIC 113,05 o modelo selecionou 29 variáveis.

Tabela 14: Tabela com ID respectivos nomes dos parâmetros

ID	Nome do parâmetro
	Intercepto
V1	total_acessos
V2	assign_submission.statement.accepted
V3	assign_submit
V4	assign_view
V5	assign_view.confirm.submit.assignment.form
V6	assign_view.submit.assignment.form
V7	autoattend_view.own
V8	autoattendep_view.own
V9	blog_view
V10	chat_report
V11	chat_view
V12	course_view
V13	course_view.section
V14	data_add
V15	data_record.delete
V16	data_update
V17	data_view
7V18	folder_view
V19	forum_add.post
V20	forum_delete.post
V21	forum_search

ID	Nome_do parâmetro
V22	forum_unsubscribe
V23	forum_update.post
V24	forum_view.discussion
V25	forum_view.forum
V26	forum_view.forums
V27	page_view
V28	user_view
V29	user_view.all

Fonte: Autoria própria (2018)

A função logito é dada por:

$$g(x) = -9{,}735 - 19{,}985(V1) + 21{,}954(V1) + 20{,}367(V3) + 20{,}136(V4) + 26{,}262(V5) + 19{,}554(V6) + 20{,}432(V7) + 19{,}264(V8) + 11{,}051(V9) + 20{,}475(V10) + 19{,}887(V11) + 20{,}042(V12) + 19{,}950(V13) + 24{,}143(V14) + 16{,}922(V15) + 19{,}268(V16) + 19{,}998(V17) + 20{,}233(V18) + 20{,}014(V19) + 18{,}266(V20) + 23{,}031(V21) + 18{,}870(V22) + 22{,}350(V23) + 19{,}964(V24) + 19{,}944(V25) + 15{,}830(V26) + 18{,}940(V27) + 19{,}896(V28) + 20{,}577(V29)$$

A probabilidade de um estudante reprovar é dada por:

$$\pi(x) = \frac{e^{g(x)}}{1 + e^{g(x)}}$$

Considerando um ponto de corte de 0,6 e as probabilidades estimadas pelo modelo, temos a seguinte tabela de classificação:

Tabela 16: Tabela de classificação (ponte de corte=0,6)

	Observado		
Classificado	Reprovado (Y=0)	Aprovado (Y=1)	Total
Reprovado(Y=0)	27	2	29
Aprovado(Y=1)	1	306	307
Total	28	308	336

Fonte: Autoria própria (2018)

Podemos perceber que o modelo pode ser considerado um bom modelo, pois entre os reprovados, acertou 27 dos 29 e em relação aos aprovados o modelo acertou 306 de um total de 307. A sensibilidade e a especificidade neste ponto de corte são respectivamente 99,35% e 96,4%.

A taxa global de classificação é $\frac{27+306}{336} = 0{,}9910$, ou seja, 99,1% é explicado corretamente pelo modelo quando o corte é igual a 0,6.

Em seguida, para a validação do modelo utilizamos duas turmas não amostradas, o critério de escolha dessas turmas foi através da menor e da maior média, ao analisar a média das demais turmas, encontramos a turma 08 e a turma 17, com média 7.3 e 9.26, respectivamente.

Na turma 08, o modelo acertou 4 dos 8 estudantes que foram reprovados e 25 aprovados de um total de 27 estudantes. Neste caso, temos uma sensibilidade de 86,20% e especificidade de 66,6 %. Na turma 17, não possui nenhum estudante reprovado, a turma conta com 35 estudantes dos quais 32 o modelo classificou como aprovado. Neste caso, temos apenas a sensibilidade de 100%.

5 CONSIDERAÇÕES FINAIS

A questão do monitoramento de logs de sistemas educacionais é algo que as instituições precisam levar a sério, deve ter rígido controle sobre as permissões concedidas para os envolvidos no processo educacional, para que não haja nenhum conflito, como por exemplo, um tutor excluir algum recurso. Somente pela avaliação dos logs é possível recuperar todas as informações do aluno no curso e assim avaliar quais recursos/atividades ele acessou, para isso é necessário que haja comprometimento com as informações.

O uso da regressão logística pode nos auxiliar nesse estudo, encontrando modelos que possam prever a probabilidade de um determinado evento acontecer. O modelo conseguiu explicar corretamente 99,1% com um ponto de corte de 0,6.

A Analitica de Aprendizagem é uma área de pesquisa recente que busca informações dentro de um Ambiente Virtual de Aprendizagem, que podem ser utilizadas para melhorar o processo de ensino-aprendizagem. Os dados gerados pela interação dos estudantes dentro do AVA pode ser utilizado para melhorar o desempenho individual e/ou coletivo dos estudantes, como por exempo identificando alunos em risco de reprova ou evasão.

Por outro lado é importante levar em consideração que a Analitica de Aprendizagem é utilizada para construir modelo com base na obervação do passado e presente, para que possa estimar o futuro do comportamento dos estudantes.

REFERÊNCIAS

BALLARD, C. et at. **Improving business** performance **insight with business intelligence and business process management.** S.l.: IBM, 2006.

CHATTI, M. A.; DYCKHOFF, A. L.; SCHROEDER, U.; THÜS, H. **A reference model for learning analytics.** International Journal of Technology Enhanced Learning, v. 4, n. 5, p. 318-331,2012.

CÔNSOLO, Adriane. **O que é Ambiente Virtual de** Aprendizagem**?**. 2016. Disponível em: <www.coachead.com.br/ambiente-virtual-de-aprendizagem >. Acesso em: 12 ago.2018.

ELIAS, T. **Learning analytics:** definitions, processes and potential. 2011. Não publicado. Disponível em: <http://learninganalytics.net/LearningAnalyticsDefinitionsProcessesPotential.pdf> Acesso em: 25 fev. 2018.

FERGUSON, R. Learning Analytics for open and distance education. **CEMCA EdTech Notes**, s/n, p. 1-8, 2013. Disponível em: <http://cemca.org.in/ckfinder/userfiles/files/EdTech%20Notes_LA_Rebecca_15%20May.pdf>.Acesso em: 15 mar. 2018

HOSMER, D. W.;LEMESHOW, S. **Applied Logistic Regression**. 2nd ed. New York: John Wiley, 2000. 375p.

MACFADYEN, L. P.; DAWSON, S. Mining LMS data to develop an ‘‘early warning system” for educators: a proof of concept. **Computers & Education**, v. 54, p. 588–599, 2010.