ESTUDO SOBRE EVASÃO NOS CURSOS DE GRADUAÇÃO DE UMA INSTITUIÇÃO DE ENSINO SUPERIOR PRIVADA: APLICAÇÃO DE REGRESSÃO LOGÍSTICA[1]

Ricardo Ferreira Vitelli
Universidade do Vale do Rio dos Sinos
vitelli@unisinos.br

Cleonice Silveira Rocha
Universidade do Vale do Rio dos Sinos
nice@unisinos.br

Rosangela Fritsch
Universidade do Vale do Rio dos Sinos
rosangelaf@unisinos.br

Resumo: O artigo estuda a evasão de alunos nos cursos de graduação em uma instituição de ensino superior privada. Configura-se como uma pesquisa de natureza quantitativa que utiliza na construção das variáveis a técnica de regressão logística para a análise dos dados. Os dados foram coletados nas bases de dados da instituição. Com essa técnica é possível determinar características de perfil tem o aluno potencial evadido, para que se possa gerir o fenômeno, desenvolvendo ações de combate e prevenção da evasão. Ao final do estudo identificou-se que a evasão está associada fundamentalmente aos fatores: indefinição na escolha profissional, desempenho acadêmico e condição financeira dos discentes.

Palavras-chave: evasão; regressão logística; ensino superior.

INTRODUÇÃO

Segundo Ristoff (1999) a discussão nacional sobre evasão surge no âmbito da crise de modelo e da crise gerencial e deve ser tratada no contexto da avaliação institucional. A evasão é um fenômeno complexo, associado à satisfação de expectativas de pessoas, e esta, por sua vez, a fatores e variáveis objetivas e subjetivas. É reflexo de múltiplas causas que precisam ser compreendidas no contexto socioeconômico, político e cultural e de inadequações do sistema educacional. Relaciona-se com a perda de alunos que iniciam, mas não concluem seus cursos e configura-se como desperdício social, acadêmico e econômico. É um dos problemas que afligem as instituições de ensino em geral. Nas instituições privadas, constitui-se em uma importante perda de receita. A evasão se caracteriza por ser um processo de exclusão determinado por fatores e variáveis internos e externos às IES. Sob a perspectiva de um fenômeno institucional, pode ser reflexo de uma política incipiente de permanência do aluno.

A educação com qualidade social e a democratização da gestão implicam a garantia do direito à educação para todos, por meio de políticas públicas, materializadas em programas e

1 Este texto compõe a produção do edital 38/2010, Programa Observatório de Educação INEP/CAPES, Núcleo em Rede, Projeto nº 44, Indicadores de Qualidade e Gestão Democrática.

ações articuladas, com acompanhamento e avaliação da sociedade, tendo em vista e melhoria dos processos de organização e gestão dos sistemas e das instituições educativas. Implicam ainda, processos de avaliação capazes de assegurar a construção da qualidade social inerente ao processo educativo, de modo a favorecer o desenvolvimento e a apreensão de saberes científicos, artísticos, tecnológicos, sociais e históricos, compreendendo as necessidades do mundo do trabalho, os elementos materiais e a subjetividade humana (ASSUMPÇÃO, 2010, p.233-234).

A preocupação com este fenômeno é crescente, pois o ensino superior brasileiro tem apresentado índices de evasão elevados em seus cursos de graduação. Este processo é percebido tanto em Instituições de âmbito público quanto particular e muitos fatores contribuem para a concretização deste fato.

Pesquisadores apontam a evasão como um dos principais problemas do sistema educacional brasileiro. Souza (1999) afirma que são modestas as pesquisas no Brasil sobre o fenômeno. Esse fato preocupa pesquisadores e tem levado-os a tentar descobrir as principais causas da evasão, propondo alternativas para elevar o número de estudantes que concluem seus cursos.

Biazus (2004) também destaca que é importante verificar e levantar as causas motivadoras da evasão, com o intuito de minimizar o número dos acadêmicos que abandonam o ensino superior, o que poderia levar o curso a realizar uma avaliação constante, e, em especial, nas suas inter-relações com a comunidade, tendo em vista a busca da qualidade do ensino-aprendizagem e da sua responsabilidade com a sociedade de forma a otimizar os investimentos empreendidos. Além destes fatos é preocupante a consequência social deste fenômeno, no sentido de contribuir para um processo de exclusão e de criação de um ambiente que interfere na sustentabilidade das instituições de ensino superior privadas, alvo desse estudo.

A partir do conceito da Evasão como saída definitiva do aluno de seu curso de origem sem concluí-lo (MEC/SESu) foram identificados dois indicadores na revisão de literatura: Evasão anual média: percentagem de alunos matriculados em uma IES/curso que, não tendo se formado, não se matriculou no ano/semestre seguinte. Evasão total: número de alunos que, tendo entrado num determinado curso/IES, não obteve o diploma ao final de um período de anos.

O presente estudo tem por objetivo avaliar e identificar, a partir de um conjunto de informações disponíveis em banco de dados, variáveis que podem contribuir para a evasão. A partir do conhecimento das variáveis que interferem na evasão o estudo também pretende estabelecer um perfil de aluno evadido para que se possa agir de forma pró-ativa com esse público. Além disso, o estudo busca conhecer as possíveis interações existentes entre duas ou mais variáveis e como elas aumentam ou diminuem a chance de um aluno se evadir de um curso de graduação da Instituição.

Outro aspecto importante é a construção de um modelo matemático que possa predizer, com a máxima exatidão possível, a probabilidade de um aluno se evadir de um curso de graduação, a partir de um determinado conjunto de variáveis que delineia seu perfil.

Foi definido que o estudo avaliaria um grupo de ingressantes, em um determinado período, em todos os cursos de graduação da Instituição, acompanhando-os ao longo de um período

de cinco anos. A escolha do período está relacionada ao fato de que, nesse tempo, já poderiam estar formados nos cursos pesquisados. Assim sendo, todos os ingressantes, por todas as formas de ingresso, em todos os cursos de graduação formaram o público-alvo desse estudo.

Do público-alvo da pesquisa foi estabelecido um conjunto de variáveis, passíveis de serem encontradas nos bancos de dados da Instituição, composto da seguinte forma: sexo; idade; estado civil; local de residência; média de desempenho nas atividades, média de desempenho no vestibular; quantidade de atividades matriculadas; percentual de atividades reprovadas; percentual de atividades aprovadas; percentual de atividades canceladas; percentual de atividades sem frequência (desistência); percentual de atividades desistentes; percentual de créditos concluídos; inadimplente; três semestres contínuos sem matrícula; média de créditos matriculados por semestre; ajuda financeira; transferência interna; forma de ingresso; área (Curso) e tamanho do curso (em créditos). A composição desse conjunto de variáveis foi determinada a partir da possível relevância das variáveis na construção do processo de evasão resultante da revisão literatura.

Com esses dados pode-se constituir uma equação de regressão múltipla onde a variável resposta (dependente) é se evadir (sim/não) e as variáveis independentes são as citadas anteriormente. Com esses dados se constituiu uma análise multivariada de dados. A análise multivariada pode ser entendida como um processo onde se estabelece uma combinação linear de variáveis com pesos empiricamente determinados. As variáveis são especificadas pelo pesquisador, sendo os pesos determinados pela técnica utilizada para se analisar os resultados da coleta das variáveis. No processo de análise multivariada de dados a variável definida como resposta (dependente) passa a ser uma combinação linear das demais variáveis (independentes).

Na análise multivariada temos um conjunto de técnicas para análise de dados. A opção por uma técnica, em detrimento das demais, está relacionada a fatores tais como: o nível de mensuração das variáveis e o objetivo do estudo, entre outros. A escolha de uma técnica multivariada depende também do nível de mensuração das variáveis. No caso desse objeto de estudo a variável resposta da pesquisa é a evasão, mensurada da seguinte forma: o aluno se evade ou o aluno não se evade. Além de ser um nível de mensuração nominal é dicotômico sendo, portanto, uma variável que se enquadra na possibilidade de uso de uma análise discriminante múltipla.

Segundo Hair *et all* (2005) a Análise Discriminante Múltipla (*MDA – multiple discriminant analysis*) é a técnica multivariada adequada quando a variável dependente é dicotômica. A análise discriminante é aplicável em situações nas quais a amostra total pode ser dividida em grupos baseados em uma variável dependente e seu objetivo é entender diferenças entre os perfis dos grupos; determinar quais variáveis independentes explicam o máximo de diferenças nos perfis e estabelecer procedimentos para classificar indivíduos em grupos, com base em seus escores.

Uma alternativa de análise é a técnica de Análise de Regressão Logística. Algumas razões que justificam o uso da Análise de Regressão Logística em detrimento da Análise Discriminante são: a Análise Discriminante depende estritamente de se atenderem as suposições de

normalidade multivariada e de iguais matrizes de variância-covariância nos grupos, a Regressão Logística não depende desta suposição; a Regressão Logística é muito mais robusta quando tais pressupostos não são atendidos. Além disso, a Regressão Logística apresenta uma gama maior de diagnóstico dos resultados.

Em Corrar *et all* (2007) vemos que a Regressão Logística estima os parâmetros com o apoio do método de máxima verossimilhança e não com o dos mínimos quadrados, usado na Análise Discriminante. Com a máxima verossimilhança buscam-se coeficientes que nos permitam estimar a maior probabilidade possível de um evento ocorrer ou de certa característica se fazer presente. Este fato é importante na medida em que isto vem ao encontro do objetivo deste estudo. Além disso, a Regressão Logística é mais indicada quando existe a presença de variáveis independentes no modelo com nível de mensuração nominal, o que ocorre neste estudo. Por estas razões justifica-se o uso da Regressão Logística neste estudo, em detrimento da Análise Discriminante.

Para a efetivação da Regressão Logística alguns passos precisam ser estabelecidos. Inicialmente a decisão pela variável a ser definida como a variável resposta do modelo. A partir de um estudo piloto realizado com um Curso podemos perceber que a quarta matrícula apresenta um ponto de corte no período de tempo muito importante, pois neste momento alunos com três semestres sem matrícula tendem a ter uma chance muito grande de se evadirem definitivamente. Desta forma, estabeleceu-se que o aluno com três semestres seguidos sem matrícula fosse a variável resposta.

A segunda etapa do processo consiste em criar variáveis *Dummy* para as variáveis nominais com mais de duas categorias de resposta. Para Hill *et all* (1999) as variáveis *dummies*, também designadas como variáveis binárias, são variáveis explicativas que podem tomar um de dois valores. Essas variáveis constituem instrumento poderoso para representar características qualitativas de dados. Como no presente estudo existem variáveis com estas características, o uso deste tipo de recurso passa a ser fundamental.

Assim sendo, uma variável nominal com duas categorias de resposta do tipo aluno inadimplente seria representada da seguinte forma: **Não** = 1 e **Sim** = 0. Para o caso das variáveis com mais possibilidades de resposta como *estado civil* que tem 5 categorias de resposta as variáveis *dummies* assumem valores de 1 quando da presença de uma categoria no modelo e 0 quando da ausência das demais categorias no modelo.

Para o desenvolvimento da análise foi utilizado o *software* SPSS®. O método escolhido para o ajuste das variáveis foi o *stepwise*. Nesse método as variáveis são introduzidas no modelo uma a uma. Após a inclusão de cada variável o modelo é avaliado se melhora sua capacidade preditiva e, passo a passo, são incluídas novas variáveis até que se encontre uma combinação ótima de variáveis. Para Hair *et all* (2005) o método *stepwise* permite ao pesquisador examinar a contribuição de cada variável independente para o modelo de regressão. Cada variável é considerada para inclusão antes do desenvolvimento da equação. A variável independente com maior contribuição é acrescentada em um primeiro momento.

Por estas considerações então fica definida que a opção será pela técnica de Regressão Logística, utilizando o método *stepwise*, a partir do conjunto de variáveis estabelecido e tendo como variável resposta (dependente) a não realização de três matrículas seguidas (Sim/Não) que o caracteriza como potencial evadido.

A avaliação da qualidade do ajuste da Regressão Logística passa pela análise de uma série de testes e indicadores que contribuem para que se possa decidir a este respeito. A seguir é feita uma análise conjunta de seus resultados. Não existe uma orientação sobre qual é o mais importante, pois avaliam situações e concluem a partir de distintas visões. Por isso, é importante avaliar todos conjuntamente e não esquecer que uma amostra muito grande possibilita maior sensibilidade dos testes aplicados, como é o caso deste estudo.

Os indicadores log likelihood value, Cox & Snell R^2 e Nagelkerke R^2.

Uma das principais medidas de avaliação da Regressão Logística é o *log likelihood value* (-2LL). Este indicador mostra a capacidade de o modelo estimar a probabilidade associada à ocorrência de determinado evento. No estudo em questão o evento está associado à evasão, quanto menor o valor deste indicador, maior o poder preditivo do modelo (definir um aluno como sendo evadido quando ele realmente é, e vice-versa).

O teste *Cox & Snell R Square* serve para comparar o desempenho de modelos concorrentes. Entre duas equações logísticas igualmente válidas, deve-se preferir a que apresente o *Cox & Snell R Square* mais elevado. Este indicador baseia-se no *Likelihood value* e quanto maior o seu valor melhor a qualidade do ajuste. Nagelkerke propôs um ajuste neste índice para que ele pudesse chegar a 1, sua finalidade é a mesma do *Cox & Snell R Square*, porém assume a "ideia" do coeficiente de explicação do ajuste pela regressão linear múltipla (assume valores de 0 a 1).

Nenhum indicador entre os estabelecidos é considerado como mais importante no momento da escolha pelo melhor ajuste. Para Corrar *et all* (2007) como não são conflitantes entre si, recomenda-se utilizá-los em conjunto, com a devida prudência.

Ao analisarmos os resultados apresentados na tabela 1, verificamos que todos os indicadores apresentados indicam que o quinto passo de ajuste do modelo de regressão apresenta melhores resultados, sendo assim, em um primeiro momento, parece ser a melhor escolha.

Tabela 1: Resultados das medidas da qualidade de ajuste do modelo nos cinco passos de construção do modelo logístico.

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	2100,752(a)	0,543	0,744
2	1982,077(a)	0,557	0,763
3	1882,764(a)	0,568	0,778
4	1753,885(b)	0,581	0,796
5	1690,795(c)	0,588	0,805

Fonte: Pesquisa do autor.

a Estimation terminated at iteration number 6 because parameter estimates changed

by less than 0,001.
b Estimation terminated at iteration number 7 because parameter estimates changed by less than 0,001.
c Estimation terminated at iteration number 20 because maximum iterations has been reached. Final solution cannot be found.

O teste Hosmer e Lemeshow.

Este indicador é obtido através de um teste Qui-quadrado que consiste em dividir o número de observações em cerca de 10 classes e, em seguida, comparar as frequências preditas com as observadas. Em função disso, a finalidade deste teste é verificar se existem diferenças significativas entre as classificações realizadas pelo modelo e a realidade observada. A certo nível de significância, o teste busca aceitar a hipótese de que não existam diferenças entre os valores preditos e observados. Caso exista diferença entre os valores, então o modelo não seria capaz de produzir estimativas e classificações muito confiáveis. Como se pode observar na tabela 2, em todas as etapas de ajuste, o modelo sempre aceita a hipótese de que não exista diferença significativa entre os valores observados e esperados.

A eficiência se baseia no fato de que, ao comparar os valores observados e os valores preditos pelo modelo o teste não encontrou diferenças significativas. Com esse resultado o modelo proposto seria indicado, pois não mostra diferença entre o valor observado e o estimado pelo modelo. Um aspecto que comprova este fato é apresentado no quadro 3 onde se percebe que existe pouca diferença entre o valor observado e o valor esperado, demonstrando uma boa qualidade de ajuste.

Tabela 2: Resultados do teste Hosmer e Lemeshow nos cinco passos de ajuste do modelo logístico.

Step	Chi-square	df	Sig.
1	26,520	8	0,001
2	150,238	8	0,000
3	54,439	8	0,000
4	22,541	8	0,004
5	27,305	8	0,001

Fonte: Pesquisa do Autor

Tabela 3: Tabela do teste de Hosmer e Lameshow no quinto passo de ajuste do modelo logístico.

Passo		Evadido 3 = Não		Evadido 3 = Sim		Total
		Observed	Expected	Observed	Expected	Observed
Step 5	1	400	399,507	0	0,493	400
	2	398	394,798	2	5,202	400
	3	366	355,001	34	44,999	400
	4	150	178,011	250	221,989	400
	5	49	58,040	351	341,960	400
	6	44	28,651	356	371,349	400

	7	20	16,402	380	383,598	400
	8	10	10,588	390	389,412	400
	9	8	5,900	395	397,100	403
	10	4	2,101	390	391,899	394

Fonte: Pesquisa do Autor

O teste Wald.

A estatística Wald tem por finalidade aferir o grau de significância de cada coeficiente da equação logística, inclusive a constante. Em outras palavras tem como objetivo verificar se cada parâmetro estimado é significativamente diferente de zero. Essa estatística segue uma distribuição Qui-quadrado e quando a variável dependente tem um único grau de liberdade pode ser calculada elevando-se ao quadrado a razão entre o coeficiente que está sendo testado e o respectivo erro-padrão. Na tabela 4 apresentamos os resultados do SPSS® apenas para o quinto passo do modelo de ajuste.

Tabela 4: Resultados do teste Wald no quinto passo de ajuste do modelo logístico.

Passo / Variáveis		B	S.E.	Wald	df	Sig.
Step 5	Área			152,886	5	0,000
	Área(1)	-0,188	0,234	0,643	1	0,423
	Área(2)	-1,195	0,190	39,763	1	0,000
	Área(3)	0,267	0,226	1,394	1	0,238
	Área(4)	0,740	0,240	9,466	1	0,002
	Área(5)	1,728	0,227	57,974	1	0,000
	Qtd.Discip	-0,203	0,008	577,650	1	0,000
	Créd.Concl	-0,038	0,003	126,551	1	0,000
	MédiaDiscip	0,717	0,067	116,201	1	0,000
	Inadimplente(1)	20,597	2949,588	0,000	1	0,994
	Constant	-17,554	2949,588	0,000	1	0,995

Fonte: Pesquisa do Autor

Onde:

B - simboliza o coeficiente da variável incluída no modelo. Este coeficiente pode ser positivo quando a variável aumenta então aumenta a probabilidade de o aluno se evadir. Para o caso da variável ser dicotômica então depende do seu valor estabelecido, se 1 aumenta a probabilidade, caso contrário diminui.

S.E. - o erro-padrão associado ao coeficiente de cada variável.

Wald - representa o valor do teste para cada coeficiente e a constante do modelo.

df - indica os graus de liberdade do teste

sig - é o nível de significância do teste. Para os casos em que o valor for inferior a 0,05 é porque o coeficiente é significativamente diferente de zero e faz parte da equação de regressão.

A definição sobre o modelo final de regressão logística passa pela análise da qualidade do modelo final, realizada anteriormente, e da composição do modelo com o melhor conjun-

to possível de variáveis. O modelo inicialmente proposto apresentou o seguinte conjunto de variáveis independentes: Idade (em anos); Estado civil (com uso de variável *dummy*); Cidade (com uso de variável *dummy*); Média de desempenho no vestibular (0 a 10); Forma de ingresso (com uso de variável *dummy*); Transferência interna (Sim ou Não); Percentual de atividades aprovadas; Percentual de atividades reprovadas; Percentual de atividades com cancelamento; Percentual de atividades sem frequência; Percentual de atividades desistentes; Quantidade de semestres sem matrícula; Quantidade de créditos do programa acadêmico; Média de desempenho nas atividades acadêmicas; Recebe algum tipo de ajuda financeira (Sim ou Não); Três semestres seguidos sem matrícula (Sim ou Não) – Variável resposta.

Além dessas também foram incluídas as variáveis: Área do curso (com uso de variável *dummy*); Quantidade de atividades matriculadas; Percentual de créditos concluídos; Média de atividades matriculadas por semestre e Inadimplente (Sim ou Não), que acabaram sendo significativas na construção do modelo.

Avaliando os resultados iniciais observamos na tabela 5 resumo inicial dos dados que foram efetivamente processados no modelo logístico. De um total de 4.435 casos 3.997 (90,17%) foram utilizados sendo que 437 casos foram eliminados por não terem informação em pelo menos uma variável.

Tabela 5: Resumo dos dados processados no modelo logístico.

Unweighted Cases(a)		N	Percent
Selected Cases	Included in Analysis	3997	90,1
	Missing Cases	437	9,9
	Total	4434	100,0
Unselected Cases		0	0,0
Total		4434	100,0

Fonte: Pesquisa do Autor

a If weight is in effect, see classification table for the total number of cases.

A tabela 6 apresenta as diversas etapas de ajuste do modelo, e como ele classifica os indivíduos como sendo evadidos ou não. Na parte onde aparece o termo *observed* temos os valores observados da variável pesquisada (evadido em três semestres) que correspondem aos valores reais. Em *predicted* temos como o modelo classifica os indivíduos em evadidos ou não. Neste caso temos que, por exemplo, no primeiro passo existiam 2.548 alunos como evadidos e o modelo previu 2.409 tendo uma margem de acerto de 94,5% dos casos. Da mesma forma ele tem 82,7% de chance de acertar os não evadidos e 90,3% de chance de modo geral.

Estes valores vão se alterando, conforme cada novo passo (ajuste) vai sendo feito e novas variáveis são incluídas e excluídas pelo método *stepwise*, melhorando a predição do modelo. Ao passo que na última etapa a chance de acerto para os evadidos é de 96% e 92,5% geral, mostrando que a opção pelo modelo delineado pelo passo cinco é, em um primeiro momento, melhor que os demais. Os resultados apresentados na tabela 7 indicam a composição final do modelo, assim como as respectivas variáveis que o integram.

Tabela 6: Capacidade preditiva do modelo nos cinco passos do ajuste.

Observed			Predicted		
			Evadido 3		Percentage
			Não	Sim	Correct
Step 1	Evadido 3	Não	1199	250	82,7
		Sim	139	2409	94,5
	Overall Percentage				90,3
Step 2	Evadido 3	Não	1211	238	83,6
		Sim	123	2425	95,2
	Overall Percentage				91,0
Step 3	Evadido 3	Não	1224	225	84,5
		Sim	110	2438	95,7
	Overall Percentage				91,6
Step 4	Evadido 3	Não	1243	206	85,8
		Sim	109	2439	95,7
	Overall Percentage				92,1
Step 5	Evadido 3	Não	1251	198	86,3
		Sim	102	2446	96,0
	Overall Percentage				92,5

Fonte: Pesquisa do Autor

a The cut value is 0,500

Tabela 7: Modelo de regressão logística após cinco etapas de ajuste.

Passo / variáveis		B	Exp(B)	95,0% C.I.for EXP(B)	
				Lower	Upper
Step 5	Área				
	Área(1)	-0,188	0,829	0,524	1,312
	Área(2)	-1,195	0,303	0,209	0,439
	Área(3)	0,267	1,306	0,838	2,036
	Área(4)	0,740	2,096	1,308	3,358
	Área(5)	1,728	5,628	3,607	8,780
	Qtd.Discip	-0,203	0,816	0,803	0,830
	Créd.Concl	-0,038	0,962	0,956	0,969
	MédiaDiscip	0,717	2,049	1,798	2,334
	Inadimplente(1)	20,597	881258508,487	0,000	.
	Constant	-17,554	0,000		

Fonte: Pesquisa do Autor

A regressão logística mostra que no quinto passo de ajuste do modelo as variáveis que o compõe são: área de residência; quantidade de atividades matriculadas no período; percentual de créditos já concluídos; a média de atividades acadêmicas matriculadas por semestre e a inadimplência.

Na apresentação do modelo final algumas variáveis foram incluídas mesmo não tendo significância na predição. Para melhor entender este fato será feita outra escolha de variáveis, retirando estas não significantes do modelo logístico inicialmente proposto.

A variável quantidade de atividades matriculadas, apesar de significativa tem colinearidade com as variáveis percentuais de aprovação, reprovação etc, uma vez que estas últimas são oriundas de uma divisão pela variável quantidade de atividades. Este fato confunde o modelo, pois as variáveis acabam tendo correlação perfeita entre si dando a ideia de que podem ser a mesma variável. Assim sendo, para não retirar o conjunto de variáveis medidas em percentual será retirada a variável quantidade de atividades matriculadas no período.

Após novos ajustes no modelo logístico os resultados mostram que o aproveitamento da amostra foi igual ao modelo anterior, conforme a tabela 8. O tamanho final foi de 3.997 casos, representado 90,1% da amostra selecionada.

Tabela 8: Resumo final dos dados processados no modelo logístico.

Unweighted Cases(a)		n	Percent
Selected Cases	Included in Analysis	3997	90,1
	Missing Cases	437	9,9
	Total	4434	100,0
Unselected Cases		0	0,0
Total		4434	100,0

Fonte: Pesquisa do Autor

a If weight is in effect, see classification table for the total number of cases.

Por outro lado os indicadores de qualidade de ajuste do modelo são menores do que os obtidos no modelo anterior. Porém, como descrito mais adiante eliminou alguns problemas encontrados anteriormente. No segundo bloco de variáveis o método realizou oito etapas de ajuste (tabela 9), sendo a última a de melhor qualidade de predição.

Tabela 9: Resultados das medidas da qualidade de ajuste do modelo nos oito passos de ajuste do modelo logístico.

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	3751,085(a)	0,310	0,425
2	3087,515(b)	0,416	0,569
3	2997,922(b)	0,429	0,587
4	2918,250(b)	0,440	0,602
5	2864,284(b)	0,447	0,613
6	2816,022(b)	0,454	0,622
7	2810,624(b)	0,455	0,623
8	2806,718(b)	0,455	0,624

Fonte: Pesquisa do Autor

a Estimation terminated at iteration number 5 because parameter estimates changed by less than 0,001.

b Estimation terminated at iteration number 6 because parameter estimates changed by less than 0,001.

O indicador que avalia a ausência de diferença entre os valores reais e os estimados pelo modelo (Hosmer e Lemeshow) indica, diferentemente do modelo anterior, a presença de indícios para se rejeitar a hipótese da não diferença entre os valores (tabela 10). Este resultado é obtido a partir de uma significância de 1%, a partir do sétimo passo. Considerando um nível de significância de 5% as hipóteses não são rejeitadas, até mesmo porque na tabela 11, assim como no modelo proposto anteriormente, apresenta valores esperados e observados muito próximos. Como o tamanho de amostra é bastante grande o teste passa a ter mais sensibilidade em rejeitar a hipótese nula, por isso, é recomendável utilizar uma significância de 5% para a tomada de decisão. Através da última etapa de ajuste (*step* 8) podemos ver que a comparação entre o valor observado e o esperado indica uma boa qualidade de ajuste.

Tabela 10: Tabela do teste de Hosmer e Lameshow no oitavo passo de ajuste do modelo logístico.

Step	Chi-square	df	Sig.
1	62,175	8	0,000
2	45,650	8	0,000
3	31,698	8	0,000
4	18,390	8	0,018
5	26,227	8	0,001
6	25,951	8	0,001
7	19,654	8	0,012
8	16,368	8	0,037

Fonte: Pesquisa do Autor

Tabela 11: Tabela do teste de Hosmer e Lameshow no oitavo passo de ajuste do modelo logístico.

		Evadido 3 = N		Evadido 3 = S		Total
		Observed	Expected	Observed	Expected	Observed
Step 8	1	394	391,017	6	8,983	400
	2	343	354,004	57	45,996	400
	3	279	268,854	121	131,146	400
	4	194	178,421	206	221,579	400
	5	101	110,906	299	289,094	400
	6	56	68,363	344	331,637	400
	7	34	40,382	366	359,618	400
	8	30	22,404	370	377,596	400
	9	13	11,179	387	388,821	400
	10	5	3,470	392	393,530	397

Fonte: Pesquisa do Autor

Por outro lado o teste Wald mostra (tabela 12) resultados mais consistentes para a significância dos coeficientes do modelo. Neste ajuste das variáveis todos os coeficientes se mostraram significativos. Além deste fato alterou-se o quadro de variáveis, delineando um perfil diferenciado a ser acompanhado.

Tabela 12: Resultados do teste Wald no oitavo passo do modelo.

		B	S.E.	Wald	df	Sig.
Step 8	Area			291,567	5	0,000
	Area(1)	0,417	0,203	4,240	1	0,039
	Area(2)	-1,062	0,159	44,645	1	0,000
	Area(3)	0,932	0,189	24,267	1	0,000
	Area(4)	1,221	0,186	43,261	1	0,000
	Area(5)	2,021	0,188	115,062	1	0,000
	Sexo(1)	0,286	0,110	6,721	1	0,010
	Desemp.Vestibular	0,120	0,061	3,889	1	0,049
	TRI(1)	-0,803	0,119	45,599	1	0,000
	Créd.Concl	-0,060	0,003	542,973	1	0,000
	MédiaDesemp	-0,393	0,053	54,818	1	0,000
	MédiaDiscip	-0,350	0,041	71,603	1	0,000
	AjudaFinanc(1)	0,952	0,101	88,001	1	0,000
	Constant	5,213	0,399	170,246	1	0,000

Fonte: Pesquisa do Autor

Alguns exemplos de variáveis mostram que receber ajuda financeira tende a diminuir a evasão, por outro lado realizar transferência interna entre cursos tende a aumentar a chance de evasão. Outras variáveis como média de desempenho em atividades ou mesmo média de matrículas por atividade, quando aumentam, tendem a reduzir a probabilidade de evasão. Com relação à capacidade preditiva do modelo (tabela 13) podemos ver que o novo modelo de regressão tem um pouco menos de precisão 91,1% ao detectar o aluno como evadido, mas continua ainda muito significativa. Em função do modelo atual não mostrar incoerências entre a presença ou não dos coeficientes do modelo, pois todos são significativos, a opção pelo modelo atual indica um resultado mais confiável, apesar das evidências apresentadas.

Tabela 13: Capacidade preditiva do modelo no último passo de ajuste.

Observed			Predicted		
			Evadido 3		Percentage Correct
			Não	Sim	
Step 8	Evadido 3	Não	1069	380	73,8
		Sim	227	2321	91,1
	Overall Percentage				84,8

Fonte: Pesquisa do Autor

Na tabela 14 temos então as informações sobre Exp(B) que identifica o aumento ou a queda na probabilidade de o aluno se evadir em função de determinada característica.

Segundo Corrar *et all* (2007) é importante afirmar que o efeito dos coeficientes sobre a razão de chance é sempre de natureza multiplicativa, e não aditiva, como ocorre em um modelo de regressão linear. Por essa razão, quando se obtém um coeficiente igual a zero o efeito sobre a variável dependente também é nulo. Também cabe destacar que quando o valor da constante for positivo produz um resultado superior a um, portanto contribui para elevar a razão de chance e o contrário quando for negativo.

Por exemplo, para a variável média de desempenho em atividades acadêmicas (Média-Desemp) o valor de B é negativo, indicando que quanto melhor o desempenho do aluno em suas atividades menores as chances de o aluno se evadir do curso. O valor de Exp(B) para esta variável é de 0,675, desta forma a cada 1 grau que se aumente na média de desempenho do aluno diminui em 32,5% de chance de o aluno se evadir do curso (**32,5% = 0,675-1 = -0,325 = -32,5%**), supondo que as demais variáveis permaneçam constantes. O intervalo de confiança para esta estimativa (**95,0% C.I.for EXP(B)**) mostra que existe 95% de chance de que a diminuição fique entre 25,1% e 39,2%.

Já para o caso de uma variável dicotômica – receber ajuda financeira (AjudaFinanc) o Exp(B) é de 0,952, então se o aluno não recebe ajuda financeira (pelo modelo a variável ajuda financeira assume o valor 1 quando a resposta for não, portanto presente na equação) aumenta a probabilidade de se evadir do curso. Em Exp(B) vemos que esta chance aumenta em 159%.

Tabela 14: Modelo de regressão logística após oito etapas de ajuste.

				95,0% C.I.for EXP(B)	
		B	Exp(B)	Lower	Upper
Step 8	Área				
	Área(1)	0,417	1,518	1,020	2,258
	Área(2)	-1,062	0,346	0,253	0,472
	Área(3)	0,932	2,540	1,753	3,680
	Área(4)	1,221	3,390	2,356	4,877
	Área(5)	2,021	7,548	5,217	10,921
	Sexo(1)	0,286	1,330	1,072	1,651
	Desemp.Vestibular	0,120	1,128	1,001	1,271
	TRI(1)	-0,803	0,448	0,355	0,565
	Créd.Concl	-0,060	0,942	0,937	0,947
	MédiaDesemp	-0,393	0,675	0,608	0,749
	MédiaDiscip	-0,350	0,705	0,650	0,764
	AjudaFinanc(1)	0,952	2,590	2,123	3,159
	Constant	5,213	183,570		

Fonte: Pesquisa do Autor

PREVISÕES COM O MODELO DE REGRESSÃO LOGÍSTICA

A partir do conjunto de dados estabelecido para se ajustar um modelo de regressão logística algumas considerações importantes são feitas. Em primeiro lugar as conclusões apresentadas pelo modelo dependem muito do conjunto de variáveis que se optou por incluir inicialmente no modelo. Partindo da suposição de que o modelo escolhido possa representar a realidade observada os resultados de cada variável são pertinentes a realidade de cada grupo observado, isto é, recomenda-se uma atualização do modelo em um período máximo de um ano. Ao final das etapas de ajuste o modelo de regressão logística pode realizar previsão com base em probabilidade de chance de acerto conhecida. O modelo logístico final proposto é apresentado através da fórmula 1.

Y=5,12+0,42a1-1,06a2+0,93a3+1,22a4+2,02a5+0,29sx+0,12dv-0,80tri-0,06cc-0,39mdp-0,35mdi+0.95af (1)

Onde:

Y é a variável resposta (o aluno se evade sim ou não)

5,12 é a constante do modelo

a1 quando o curso for da área 1 (demais igual a zero)

a2 quando o curso for da área 2 (demais igual a zero)

a3 quando o curso for da área 3 (demais igual a zero)

a4 quando o curso for da área 4 (demais igual a zero)

a5 quando o curso for da área 5 (demais igual a zero)

quando o curso for da área 6 (zero para todas)

sx variável sexo (masculino = 0 e feminino = 1)

dv média de desempenho no vestibular

tri aluno já realizou transferência interna (Não = 1 e Sim = 0)

cc quantidade de créditos já concluídos

mdp Média de desempenho nas atividades acadêmicas

mdi Média de atividades acadêmicas matriculadas por semestre

af Recebe algum tipo de ajuda financeira (Não = 1 e Sim = 0)

Dessa forma a probabilidade de o aluno se evadir do curso de graduação acaba sendo medida não pela fórmula de ajuste, pois ela é construída a partir de um modelo *logit* (logarítmico natural). Mas a probabilidade pode ser obtida pela fórmula 2. Nesta expressão, após substituirmos os valores das variáveis é possível determinar a probabilidade de evasão, como a variável resposta é dicotômica então o modelo estabelece que, se a probabilidade calculada for superior a 0,50 o aluno se evade, caso contrário não.

$$P(evadir) = \frac{1}{1+e^{-y}} \qquad (2)$$

CONSIDERAÇÕES SOBRE OS RESULTADOS DA ANÁLISE MULTIVARIADA

Os passos adotados até a conclusão final do modelo de regressão logística avaliaram as contribuições das variáveis a partir de um conjunto de indicadores de qualidade de ajuste. Além deste fato, coube também ao pesquisador, avaliar os resultados obtidos e buscar novos caminhos para a solução dos problemas encontrados.

O modelo final de ajuste incluiu variáveis oriundas das mais diversas áreas. Tanto variáveis de escolha por curso a desempenho no vestibular passando por outras variáveis de natureza diferenciada. Este fato alerta para o indício de que existe uma complexidade de fatores que interferem no fenômeno da evasão. Estes fatores podem ser muito diversos de um aluno para outro. Enquanto que para alguns a indecisão pela escolha passa a ter importância fundamental, para outros as lacunas decorrentes de sua formação anterior dificultam sua trajetória no curso.

O modelo final apresenta, a partir da composição de um conjunto de variáveis, a probabilidade de ocorrência de um evento (aluno se evadir). Esta probabilidade, quando inferior a 50% é um indicativo de que talvez o aluno não vá se evadir, por outro lado, quando superior a este valor, mostra indícios de que ele pode se evadir. As colocações sobre a ocorrência da variável resposta não são conclusivas, pois trabalham com probabilidades e não certezas.

CONSIDERAÇÕES FINAIS

A decisão por determinado conjunto de variáveis interfere diretamente nos resultados obtidos. Além deste fato, o período de corte também influencia nos resultados obtidos. Outro fator relevante neste processo é que este modelo não é estanque. Uma vez alteradas algumas condições o modelo passa automaticamente por transformações. Por este motivo, não é recomendável que ele seja adotado por um período de tempo muito longo. Pelo menos uma vez por ano cabe realizar um novo ajuste no modelo para verificar a necessidade de inclusão de novas variáveis ou de alteração nos pesos dos coeficientes.

A construção do modelo seguiu passos sedimentados em concepções teóricas e avaliou somente a concepção de evasão a partir de variáveis incluídas em banco de dados, desta forma não se completa por si só. Nesse processo é importante buscar fatores subjetivos que não podem ser mensurados através de um modelo matemático. Assim sendo, a realização de uma pesquisa qualitativa pode contribuir muito na complementação dos resultados apresentados nesse estudo. A complementaridade das informações pode ser feita a partir de investigações qualitativas já existentes na Instituição que avaliam aspectos subjetivos (pesquisa com alunos com perfil para evasão, perfil de alunos que permanecem nos cursos, mapeamento e diagnóstico dos pontos de interação dos alunos com a Instituição, avaliação institucional dos serviços, entre outras).

As variáveis que compõem o modelo logístico mostram que alterações nos resultados das mesmas podem fazer com que um aluno passe de perfil de evadido para não evadido. Por outro lado, algumas variáveis não podem ser alteradas em função de sua natureza como sexo ou mesmo média de ingresso no vestibular. Mas no caso da média do vestibular pode ser disponibilizado um acompanhamento para estes alunos, contribuindo também para a melhoria da

variável média de desempenho em atividades acadêmicas. Outras variáveis como média de compra na matrícula associada à quantidade de créditos já concluídos devem exigir esforços maiores da instituição para melhorar esses indicadores. Avaliando apenas as variáveis passíveis de serem alvo de ações institucionais alguns agrupamentos foram estabelecidos. O primeiro grupo envolve as variáveis: média de desempenho nas atividades acadêmicas e média de desempenho no vestibular – desempenho acadêmico. As variáveis: área do curso e transferência interna – escolha profissional. E a quantidade de créditos contratados e recebimento de ajuda financeira – suporte financeiro. A variável tempo de Curso também se mostrou significativa, principalmente para alunos ingressantes, onde a evasão é maior. Quanto mais tempo o aluno permanece no Curso menor tende a ser a probabilidade de evasão. Através da figura 1 podemos resumir essas variáveis que interferem na evasão, passíveis de serem alvo de ações para amenizar os problemas.

Figura 1: Representação sintética das variáveis que interferem na evasão.

Esse processo implica na necessidade de um conjunto políticas institucionais que sejam implementadas para se mapear e identificar ações que possam contribuir para atenuar o fenômeno da e

vasão. Nas demais etapas desse estudo, a partir de uma investigação mais ampla, pode-se propor com mais segurança um conjunto de ações a serem desenvolvidas.

REFERÊNCIAS

ASSUMPÇÃO, Eracilda da. Planejamento e Avaliação – uma interlocução necessária. In: WERLE, Flávia Obino Corrêa (org.) **Avaliação em Larga Escala**: foco na escola. São Leopoldo: Oikos; Brasília, Liber Livro, 2010.

BIAZUS, Cleber Augusto. **Sistema de fatores que influenciam o aluno a evadir-se dos cursos de graduação na UFSM e na UFSC**: **um estudo no curso de Ciências Contábeis**. Florianópolis, Universidade Federal de Santa Catarina, 2004. Tese. (Doutorado em Engenharia de Produção).

CORRAR, Luiz J.; PAULO, Edílson; FILHO, José Maria Dias. **Análise Multivariada para os cursos de Administração, Ciências Contábeis e Economia**. São Paulo. Editora Atlas, 2007.

HAIR, Joseph F.; ANDERSON, Rolph E.; TATHAM, Ronald L.; BLACK, William B. **Análise Multivariada de Dados**. 5ª edição. Porto Alegre. Editora Bookman, 2005.

HILL, Carter.; GRIFFITHS, William.; JUDGE, George. **Econometria**. São Paulo. Editora Saraiva, 1999.

LEVIN, Jack. **Estatística Aplicada a Ciências Humanas**. 2ª edição. São Paulo. Editora Harbra Ltda, 1997.

RISTOFF, Dilvo I. **Universidade em Foco**: reflexões sobre a educação superior. Florianópolis, Insular, 1999.

SOUZA, Irineu Manoel de. **Causas da Evasão nos cursos de graduação da Universidade Federal de Santa Catarina.** Florianópolis, Universidade Federal de Santa Catarina, 1999. Dissertação (Mestrado em Administração).