Classificação dos candidatos ao vestibular da FECILCAM via técnicas estatísticas multivariadas

Tatiane C. da Silva
Faculdade Estadual de Ciências e Letras de Campo Mourão - Departamento de Matemática
87303100, Campo Mourão, PR
E-mail: taticazarin@yahoo.com.br

Gislaine A. Periçaro
Faculdade Estadual de Ciências e Letras de Campo Mourão - Departamento de Matemática
87303100, Campo Mourão, PR
E-mail: gpericaro@gmail.com

***Resumo**: A presente pesquisa buscou verificar a validação da aplicação de técnicas estatísticas multivariadas na classificação dos candidatos ao vestibular da Faculdade Estadual de Ciências e Letras de Campo Mourão, FECILCAM-PR, como aprovados ou reprovados, baseada em 19 variáveis sócio-educacionais. As informações contidas no questionário sócio-educacional respondido pelos candidatos serviram como banco de dados à aplicação das técnicas de Análise Fatorial e Regressão Logística, possibilitando verificar a validade da utilização de técnicas de simplificação e classificação em seu desempenho final.*

1. Introdução

Atualmente, um dos principais fatores que sustentam o desenvolvimento tecnológico e profissional tem sido a educação, que atua de forma direta na qualificação pessoal e social. Por isso esta é uma questão que gera grande preocupação nos mais diversos setores da sociedade, nos quais a consolidação de seus objetivos é ponderada por meio dos processos metodológicos empregados. Para que a qualidade seja atingida é necessário que todos os setores da educação almejem um ideal comum, proporcionando a validação dos objetivos propostos ou a busca deste patamar, e uma forma de estruturar essa análise é buscar informações relevantes, muitas vezes ditas insignificantes.

Esse direcionamento atua na descoberta de conhecimento em banco de dados, ou prospecção de conhecimento (*Knowledge Discovery in Databases – KDD*) e de acordo com CARVALHO (1999) *apud* MARTINHAGO (2005) esse é um processo multidisciplinar, que "combina técnicas, algoritmos e definições de todas as áreas com o objetivo principal de extrair conhecimento a partir de grandes bases de dados". Atua na descoberta de conhecimentos, desenvolvendo e validando técnicas, ferramentas e métodos que buscam extrair padrões até então implícitos no banco de dados. Uma das principais etapas desse processo, que trabalha diretamente na manipulação numérica é denominada Mineração de Dados, ou *Data Mining,* definida por BERRY E LINOFF (1997 *apud* ANDRADE *et al,* 2005) como a exploração e análise de grandes quantidades de dados, de maneira automática ou semi-automática, com o objetivo de descobrir padrões e regras relevantes utilizando algoritmos com eficiência computacional aceitável.

Interligado a essa análise e visando estruturar um banco de dados que forneça informações relevantes, é conhecido que algumas Instituições de Ensino Superior – IES, fornecem aos seus candidatos o preenchimento de um questionário sócio-educacional. As informações contidas nesses questionários podem estabelecer relações entre as variáveis sócio-educacionais e o desempenho dos candidatos nas provas de seleção, podendo auxiliar os administradores das IES na tomada de decisões, visando à melhoria da qualidade do ensino. Para PANIZZI (2004) *apud* MARTINHAGO (2005) os "órgãos governamentais não devem apenas se preocupar com o ingresso dos jovens no ensino superior, mas principalmente com a permanência destes nas instituições". Dessa forma, percebe-se a importância de delinear o perfil dos candidatos ao vestibular, de forma a auxiliar na elaboração de projetos que atendam

às necessidades dos acadêmicos e, consequentemente, forneçam subsídios à permanência desses na Instituição. Tratando-se de um estudo voltado a inúmeras características, surge a necessidade de utilizar métodos estatísticos multivariados que garantam um conhecimento geral da estrutura numérica das variáveis e a mineração de dados, que fornece técnicas de análise que possibilitam estruturar os dados, destacando suas dependências.

A estatística multivariada é definida por CUADRAS (1981) *apud* TRIVELLONI (1998) como "uma parte da estatística e da análise de dados que estuda, interpreta e elabora o material estatístico sobre a base de um conjunto de $n > 1$ variáveis, que podem ser do tipo quantitativo, qualitativo ou uma mescla de ambos". Além de aperfeiçoar os resultados populacionais, a técnica multivariada possui outras características, relacionadas à combinação linear ou não entre as variáveis, classificações e interdependência entre as mesmas. (PLA, 1986 *apud* ALVES, 2005). Dessa forma, é possível estabelecer padrões e relações na análise proposta, possibilitando a simplificação e generalização dos resultados.

A presente pesquisa visou utilizar os métodos que compreendem as técnicas exploratórias de sintetização dos dados, destacando-se: Análise de Componentes Principais, Análise Fatorial e a Regressão Logística, com o objetivo de verificar a validade da aplicação da padronização dos dados no desempenho dos candidatos ao vestibular da FECILCAM baseada nas variáveis sócio-educacionais dos mesmos. As técnicas utilizadas encontram-se descritas a seguir.

A Análise de Componentes Principais tem como objetivo principal explicar a estrutura da variância e da covariância do vetor aleatório original, por meio de combinações lineares entre as variáveis observadas, sendo essas combinações denominadas Componentes Principais. (MARQUES, 2006). Tem como característica principal tornar as variáveis não-correlacionadas entre si, além de classificar as variâncias explicativas, ou seja, expõe as componentes em ordem decrescente, obedecendo à variância máxima determinada na explicação do fenômeno.

A Análise Fatorial visa à simplificação dos dados, mantendo a variabilidade, com a menor perda possível de informações. Esse método multivariado busca a explicação numérica, possibilitando estimar a relação entre as variáveis em questão. ANDREOLI (1998) diz que a Análise Fatorial é uma técnica de análise multivariada que tem como objetivo examinar a interdependência entre as variáveis e a sua principal característica é a capacidade de redução de dados. Percebe-se então a essência da estatística multivariada quando empregada essa técnica: utilizar um banco de dados relacionados a diversas variáveis, buscando explicar o desenvolvimento dos dados e, conseqüentemente, a generalização dos resultados. Os valores numéricos obtidos possibilitam encontrar o valor correspondente a cada elemento amostral. Tais valores, denominados escores, podem também ser utilizados em análise de variância e regressão, já que estabelecem a dependência na estrutura numérica. (ZANELLA, 2006).

A Regressão Logística é um método, ou uma abordagem de modelagem matemática, que objetiva descrever a relação entre uma variável resposta - dependente - e uma ou mais variáveis explicativas – independentes, ou relacionando variáveis quantitativas e qualitativas. De acordo com MARQUES (2006) a principal característica que define a regressão logística é o fato de a variável resposta ser dicotômica ou binária (0,1), enquanto que na regressão linear são consideradas apenas variáveis contínuas. Ainda de acordo com este autor as razões para a escolha da regressão logística são: a extrema flexibilidade e facilidade de uso, além de proporcionar interpretações significativas.

2. Metodologia

A presente pesquisa engloba um estudo associado à aplicação da análise multivariada no desempenho de 1157 candidatos ao vestibular de Verão 2007, ingressos no ano de 2008, da Faculdade Estadual de Ciências e Letras de Campo Mourão – FECILCAM, que oferece à comunidade nove cursos: Administração, Ciências Contábeis, Ciências Econômicas, Engenharia de Produção Agroindustrial, Geografia, Letras, Matemática, Pedagogia e Turismo e

Meio Ambiente. São realizados, por ano, dois vestibulares do tipo vocacionado, denominados Vestibular de Inverno e Verão, realizados em junho e dezembro, respectivamente, do ano que antecede o ingresso dos aprovados na instituição. No total são oferecidas 265 vagas por vestibular.

Das 30 questões que compunham o questionário formulado pela Instituição, foram selecionadas 19, sendo caracterizadas por: estado civil; estado de residência; zona de localização da residência; renda mensal; instrução do pai; instrução da mãe; tipo de moradia; participação financeira na família; característica do ensino fundamental; tempo de conclusão do ensino médio; característica do ensino médio; turno em que cursou o ensino médio; tipo de formação; participação em cursinho pré-vestibular; curso superior; meio de informação; cor; sexo e idade. A análise dos dados, e consequentemente, dos resultados, foi auxiliada por alguns recursos e programas computacionais, tais como os *Softwares Excel, Statistica* e *Minitab.*

3. Resultados e Discussão

A Análise Fatorial permite estabelecer as variáveis que melhor explicam a variabilidade dos dados e, portanto, influenciam na resposta do candidato no questionário. Com a matriz inicial de dados, que continha as respostas dos candidatos, ditas observáveis, foi aplicada a análise fatorial, tendo como característica principal a simplificação, ou redução de dados a fatores que permitem obter um número menor de variáveis alternativas, não correlacionadas e que sintetizem as informações referentes ao fenômeno observado em uma variância explicada. A variância explicada pelos fatores, por meio da Análise de Componentes Principais, é dada no quadro 1.1.

nr	Autovalores	Variância	Autovalores Acumulados	% Variância Explicada
1	2,509132	13,20596	2,509132	13,20596
2	1,954406	10,28635	4,463538	23,49231
3	1,279559	6,734521	5,743097	30,22683
4	1,191842	6,272854	6,934939	36,49968
5	1,095764	5,767177	8,030703	42,26686
6	1,071632	5,640167	9,102335	47,90703
7	1,064159	5,600835	10,16649	53,50786
8	1,052851	5,541319	11,21934	59,04918
9	0,971842	5,114958	12,19119	64,16414
10	0,95553	5,029105	13,14672	69,19324
11	0,859184	4,522021	14,0059	73,71526

Quadro 1.1 - Autovalores e % Variância Explicada

Na interpretação dos resultados, o fato de existir correlação linear entre as variáveis, permite o agrupamento em fatores. Utilizando os métodos de escolha do número de fatores tem-se que, de acordo com Critério de Kaiser, no qual são considerados os autovalores maiores ou iguais a 1, o número de fatores seria 8, o que corresponde a apenas 59,05% da variabilidade total dos dados. Consideramos, então, a proporção da variância explicada em relação à total, sendo utilizados os autovalores superiores a 0,85, totalizando 11 fatores que explicam 73,71% da variância total.

A fim de identificar quais variáveis melhor carregam cada fator, realizamos a rotação Varimax. Nesse processo os fatores são translacionados próximos de variáveis que o carregam com maior intensidade, consequentemente, apontando as variáveis com maior contribuição, enquanto as demais se tornam numericamente próximas à zero. De acordo com diversos autores, o carregamento é considerado significativo na determinação dos fatores quando possui valores superiores a 0,7, em módulo.

Tabela 1.1 - Peso dos fatores após a rotação Varimax

	Fator 1	Fator 2	Fator 3	Fator 4	Fator 5	Fator 6	Fator 7	Fator 8	Fator 9	Fator 10	Fator 11
Var 1	0,124	-0,654	-0,167	0,196	-0,048	0,057	0,011	0,166	0,150	0,093	-0,113
Var 2	-0,094	-0,039	-0,015	-0,015	0,032	-0,034	-0,010	-0,030	-0,035	0,963	-0,011
Var 3	0,180	0,160	0,014	-0,068	-0,031	0,088	0,056	0,055	-0,865	0,047	0,045
Var 4	-0,559	-0,180	0,073	-0,030	0,283	-0,022	-0,197	0,226	-0,023	-0,155	0,051
Var 5	-0,765	0,106	-0,015	0,011	0,067	0,121	0,045	-0,040	0,115	0,084	0,018
Var 6	-0,773	0,161	0,013	-0,066	0,127	-0,038	0,037	0,001	0,046	0,098	0,058
Var 7	0,042	0,018	-0,015	0,024	0,022	0,034	-0,019	-0,934	0,039	0,029	0,001
Var 8	-0,062	0,143	-0,744	0,088	0,108	0,046	-0,063	0,154	-0,014	0,087	-0,034
Var 9	-0,151	0,011	-0,036	-0,024	0,847	-0,020	-0,018	-0,037	-0,004	0,023	0,033
Var 10	-0,119	0,792	-0,124	0,063	0,043	0,071	-0,051	0,074	-0,072	0,018	-0,165
Var 11	-0,126	0,021	0,018	0,074	0,850	0,037	0,026	0,017	0,040	0,015	0,032
Var 12	0,197	-0,476	0,346	0,356	-0,137	0,011	0,101	0,009	0,066	-0,007	-0,047
Var 13	0,020	0,070	-0,038	0,859	0,077	0,029	-0,011	-0,026	0,030	-0,015	0,085
Var 14	0,096	0,056	0,048	-0,085	-0,065	0,043	0,076	0,002	0,029	0,013	-0,945
Var 15	0,116	0,686	-0,014	0,346	-0,122	0,019	0,056	0,039	0,067	0,022	-0,007
Var 16	0,004	0,043	0,047	-0,006	-0,009	0,040	-0,958	-0,016	0,040	0,011	0,071
Var 17	0,330	0,191	0,078	-0,232	0,050	0,609	0,142	0,104	0,397	0,089	0,209
Var 18	0,137	-0,112	-0,713	-0,053	-0,130	-0,009	0,155	-0,215	0,031	-0,082	0,088
Var 19	0,198	0,047	0,075	-0,129	0,003	-0,819	0,107	0,085	0,207	0,079	0,132

Percebe-se que um fator pode ser explicado por mais de uma variável, e por outro lado há variáveis que não carregam nenhum dos fatores. Dessa forma, a relação do fator e do número de variáveis torna-se hipotética. O carregamento de cada fator associado às variáveis pode também ser exemplificado na representação gráfica entre as variáveis.

Como o carregamento dos fatores é identificado pelas variáveis, pode-se denominar cada fator, segundo sua maior explicação. Os resultados desse novo processo, especificados a seguir, representam as variáveis contidas no questionário com respostas mais correlacionadas:

Fator	**Denominação**
1	Formação dos pais
2	Tempo de conclusão do ensino médio
3	Contribuição familiar
4	Tipo de formação escolar
5	Caracterização da formação escolar
6	Idade
7	Meios de informação utilizados
8	Moradia
9	Zona de localização da residência
10	Estado em que reside
11	Participação em cursinho

Quadro 1.3 – Denominação dos fatores

Utilizando a análise Fatorial e a rotação Varimax determinou-se uma nova estrutura dos dados, agora reduzidos dimensionalmente, que revelam os dados referentes às observações individuais. Os coeficientes dos escores fatoriais explicitam a contribuição de cada variável na formação de cada fator. A nova matriz encontrada apresenta valores normalizados e não-observáveis, que se torna a nova base de dados para a aplicação das demais técnicas multivariadas, e traz os valores individuais dos candidatos na formação dos 11 fatores.

Tais variáveis, agora dispostas em fatores, podem delinear o perfil dos candidatos ao vestibular da Instituição quando ponderados separadamente. Como o objetivo do trabalho é verificar a validade na classificação dos candidatos, tais dados serão utilizados na técnica de Regressão Logística, modelando o resultado final a partir das variáveis e do desempenho dos candidatos. A fim de comparar os resultados, a Regressão também foi aplicada aos dados brutos, porém os resultados obtidos com a análise fatorial foram significativamente melhores.

A Regressão Logística tem por objetivo "saber quais variáveis independentes influenciam ao resultado (variável dependente) e usá-las numa função para prever o resultado de um indivíduo à custa das variáveis independentes." (REGRESSÃO..., *on-line*). Para isso, o resultado final do candidato caracterizou a variável dependente, que foi nomeada 0 aos reprovados e 1 aos aprovados. Os parâmetros $(\beta_0, ..., \beta_{11})$ obtidos pelo modelo de regressão foram estimados por meio do algoritmo de quase Newton, com o auxílio do *Software Statistica*, que determinam a função *logit*. O modelo é dado por:

$$\hat{\pi}(x) = \frac{e^{-1,1382 + 0,2663 x_1 - 0,3211 x_2 + 0,1237 x_3 + ... - 0,0015 x_{10} - 0,1757 x_{11}}}{1 + e^{-1,1382 + 0,2663 x_1 - 0,3211 x_2 + 0,1237 x_3 + ... - 0,0015 x_{10} - 0,1757 x_{11}}} \quad (1.1)$$

Para determinar o resultado final de um elemento quando relacionado à amostra, basta substituir os escores fatoriais no modelo acima definido e associá-lo à variável binária considerada. A verificação dos erros e acertos na fase de treinamento para o modelo estimado é realizada de acordo com a análise da matriz de confusão, apresentada a seguir, que evidencia a melhor classificação para os candidatos reprovados.

		Classificação prevista		Percentual de acerto
		π_1	π_2	
Classificação real	π_1	50	253	16,50%
	π_2	22	832	97,42%

Quadro 1.4 - Matriz de Confusão – Regressão Logística.

Os resultados obtidos numa análise completa da Regressão possibilitam identificar a probabilidade de significância de cada um dos 11 fatores, conforme exposto.

Fatores	**Coef (β)**	***p***
1	0,266392	1,15
2	-0,321184	0,000
3	0,123754	0,077
4	0,120651	0,081
5	0,307832	0,000
6	-0,0272234	0,690
7	0,0488565	0,478
8	0,390821	0,000
9	0,0894627	0,154
10	-0,0015404	0,982
11	-0,175702	0,008

Quadro 1.5 – Coeficientes e probabilidade de significância dos coeficientes aplicada aos 11 fatores.

De acordo com os dados referentes à probabilidade de significância dos coeficientes (valor-*p*), temos que os fatores 1, 3, 4, 6, 7, 9 e 10 obtiveram uma maior probabilidade, possibilitando extingui-los do modelo de predição (considerando um nível de significância de 5%). Desconsiderando estes fatores, a regressão logística foi aplicada novamente e as probabilidades encontradas para os novos coeficientes, aproximaram-se do desejado (Quadro 1.6), confirmando a representação dos fatores na análise numérica, sendo obtidas como probabilidade de classificação correta: 98,48% aos candidatos reprovados e 7,92% para os candidatos aprovados.

Fatores	**Coef (β)**	***p***
2	-0,322353	0,000
5	0,308496	0,000
8	0,391807	0,000
11	-0,169905	0,009

Quadro 1.6 – Coeficientes e probabilidade de significância dos coeficientes aplicada aos 4 fatores.

Entretanto, quando comparamos a técnica de modelagem utilizada na classificação, pode-se perceber que há uma pequena melhora na classificação dos reprovados, para o modelo gerado a partir dos 4 fatores. Isso indica que os fatores desconsiderados não alteram de forma significativa o resultado final da classificação e poderiam ser utilizados no modelo de regressão. Esta semelhança entre os resultados se justifica pelo fato de que ao aplicar a análise fatorial às 19 variáveis originais, obtivemos novas variáveis, os 11 fatores não correlacionados entre si, utilizados na regressão logística.

4. Considerações Finais

A estatística multivariada atua como uma área de grande importância, seja pelo desenvolvimento dos métodos e softwares computacionais, seja pelo seu amplo meio de aplicação em diversas áreas do conhecimento. Aliada à pesquisa operacional e a outros ramos, possibilita o grande objetivo do estudo estatístico: analisar dados ou fenômenos interpretá-los algebricamente, e consequentemente, fornecendo resultados relevantes a conclusões futuras.

Reconhecendo a importância da aplicação das técnicas de *Data Mining* na análise de dados, e da estatística multivariada como ferramenta foi possível estruturar a relação entre o desempenho e as variáveis sócio-educacionais dos candidatos ao vestibular de verão 2007 da FECILCAM. Mesmo sabendo que tais variáveis não tenham caráter informativo ou definam o resultado de um candidato, tornou-se possível verificar o comportamento das mesmas, tornando visível o processo de análise e padronização quando se tem um banco de dados multivariados. Com o auxílio das técnicas de Análise Fatorial e Regressão Logística foi realizada a redução do banco de dados, e a conseqüente formulação do algoritmo de classificação. Tais técnicas tinham como objetivo, reduzir a estrutura dos dados para aplicá-los na classificação dos candidatos, ou seja, em seu desempenho final. Comparando os coeficientes significativos o modelo mostrou-se melhor na classificação dos reprovados. Dessa forma, percebe-se a importância da manipulação e limpeza dos dados antes do processo final de análise, a fim de que os dados já tenham uma representatividade total, alcançando resultados significativamente melhores na análise multivariada.

Referências

1. A. V. “Wangenheim. Reconhecimento de Padrões”. Artigo disponível em: <http://ww.inf.ufsc.br/~ patrec/estatisticas.html>
2. A. Zanella. “Identificação de fatores que influenciam na qualidade do ensino de matemática, através da análise multivariada”. Dissertação de Mestrado, Santa Maria, 2006.
3. C. A. P. Trivelloni; N. Hochheim. “Avaliação de imóves com técnicas de análise multivariada”. In: Congresso Brasileiro de Cadastro técnico Multifinalitário – UFSC. Florianópolis, 1998.
4. D. F. Andrade, *et al.* “Estatística e Redes Neurais em Mineração de dados”. Dissertação de Mestrado, UFSC [ca. 2003].
5. J. M. Marques. Notas de aula da disciplina de Análise Multivariada Aplicada à Pesquisa, do curso de Mestrado em Métodos Numéricos em Engenharia, da Universidade Federal do Paraná. Curitiba, 2006.
6. S. A. Mingoti. Análise de Dados através de métodos de estatística multivariada: uma abordagem aplicada. Belo Horizonte: UFMG, 2005.
7. S. B. Andreoli. “Estrutura fatorial do questionário de morbidade psiquiátrica do adulto aplicado em amostras representativas de três cidades brasileiras (Brasília, São Paulo e Porto Alegre)”. Dissertação de mestrado, UNIFESP.
8. S. Martinhago. “Descoberta de conhecimento sobre o processo seletivo da UFPR”. Dissertação de mestrado, Curitiba, 2005.
9. V. Alves. “Avaliação de imóveis baseada em métodos estatísticos multivariados”. Dissertação de Mestrado, UFPR, 2005.