APLICAÇÃO DE TÉCNICAS MULTIVARIADAS PARA ANÁLISE DOS ESCORES DE CLASSIFICAÇÃO DOS CANDIDATOS AO CONCURSO VESTIBULAR 2007 DA UFSM

Paulo Roberto Machado Calil
Programa de Pós-Graduação em Estatística e Modelagem Quantitativa
Universidade Federal de Santa Maria
Santa Maria – RS

RESUMO

O presente artigo surge da necessidade de uma ferramenta estatística que auxilie a Universidade Federal de Santa Maria (UFSM), mais especificamente, a Comissão Permanente do Vestibular (COPERVES), na análise da metodologia de classificação dos candidatos ao processo seletivo Vestibular. Para isso, uma das técnicas mais conhecidas para classificação de dados é a Análise Discriminante de Fisher, a qual neste trabalho fará uso dos escores individuais dos candidatos por disciplina com o objetivo de avaliar um índice de discriminação entre candidatos através de técnicas de análise multivariada. O poder de discriminação entre os candidatos pode ser entendido como um índice, o qual é o valor da função discriminante calculado para cada candidato, baseado em suas notas de provas objetivas e redação. A partir dessas notas, foram obtidas duas funções discriminantes e três funções de classificação para os três grupos (não-selecionados, selecionados e classificados), e uma função discriminante e duas funções de classificação para dois grupos (selecionados e classificados). O valor da função de classificação determina a alocação do candidato em uma das três possíveis classes. Conclui-se que a Análise Discriminante de Fisher possui uma excelente capacidade de discriminação quando os grupos são compostos por amostras de proporções grandes e medianas em relação ao número de parâmetros estimados. Assim, cursos com média e alta densidade de candidatos por vaga apresentam bons resultados, enquanto que os grupos formados por amostras pequenas, ou cursos com poucos candidatos, resultam em baixa acurácia na estimação dos parâmetros das funções discriminantes, e conseqüentemente prejudicando a acurácia final de classificação.

Palavras-chave: Análise Discriminante de Fisher, Índice Discriminante, Funções de Classificação, Processo Seletivo Vestibular.

ABSTRACT

The present work arises from the necessity of a statistic key which helps the COPERVES in the analysis of the classification methodology of the candidates to the Vestibular. The Discriminant Analysis of Fisher is a well knowed technique to data classification, which in this work will employ the individual scores of the candidates for matter with the aim to value the discrimination rate between candidates through of multivarieds techniques. The capacity of discrimination between the candidates can be understood like an index, which is the value of the discriminant function calculated for each candidate, based on their objective tests and composition marks. According to these marks, were obtained two discriminant functions and three classification functions for three groups (no selecteds, selecteds and successfuls), and a discriminant function and two classification functions for two groups (selecteds and successfuls). The value of the classification function will determine the position of the candidate in one of the three possible situations. In this way, the Discriminant Analysis shows an excellent capacity to discriminate when the groups are samples of medium and large dimensions with reference to the number the respected parameters, and low capacity to classify when the groups are small samples because show a low accuracy in the respected parameters of the discriminant functions, going down the final classification accuracy.

Keywords: Fisher Discriminant Analysis, Discrimination Index, Classification Functions, Vestibular Selective Process.

1 INTRODUÇÃO

No Brasil, a educação tem sido motivo de preocupação quanto à formação dos seus professores e estudantes. Barbosa et al. (1995) afirma que “a insatisfação diante deste quadro tem levado líderes e estudiosos do problema a buscarem estratégias capazes de melhorar o desempenho das instituições educacionais”.

Existem duas formas de seleção para o ingresso no ensino superior em todo o mundo: (i) durante os primeiros anos dos cursos superiores; ou (ii) na entrada dos cursos superiores.

O método de seleção dentro da universidade é empregado na Bélgica e na Argentina, por exemplo. Nesse processo os estudantes são admitidos diretamente para o curso.

O método de seleção na entrada da universidade é um sistema adotado em vários países como Brasil, Estados Unidos, Japão, China, Grécia, Inglaterra, Canadá e Alemanha. Esse processo pode ser realizado de diversas maneiras, como: prova de habilitação em matérias, avaliação de personalidade, critérios por tempo de espera e sorteio de vagas.

O presente Artigo surge da necessidade de uma ferramenta estatística que auxilie a Universidade Federal de Santa Maria/(COPERVES), na análise da metodologia de classificação dos candidatos ao processo seletivo Vestibular. Uma das técnicas mais conhecidas para a classificação de dados é a Análise Discriminante de Fisher, a qual fará uso dos escores individuais dos candidatos por disciplina, com o objetivo de avaliar o índice de discriminação entre candidatos através das técnicas de análise multivariada.

Segundo os autores Santos e Milioni (2005), a Análise Discriminante de Fisher “consiste em separar classes de objetos e prever a pertinência do novo objeto a uma classe”.

2 METODOLOGIA

2.1 Análise Discriminante

Hair et al. (2005) afirmam que a Análise Discriminante tem como princípio básico, a estimativa de relação entre uma variável dependente não-métrica (categórica) e um conjunto de variáveis independentes métricas, conforme a demonstração geral:

$$Y_1 = X_1 + X_2 + ... + X_n \quad (1)$$

(não-métrica) (métricas)

A análise identifica o grupo ao qual o indivíduo pertence. Apresentando a categoria que estão os vestibulandos não-selecionados, selecionados e classificados no Vestibular 2007.

A Análise Discriminante é a determinação de uma variável estatística a partir da combinação linear de duas ou mais variáveis independentes, que individualizam os grupos estudados de maneira eficiente. A discriminação é adquirida estabelecendo os pesos das variáveis para maximizar a variância entre grupos e minimizar a variância dos grupos. Essa combinação linear é conhecida como função discriminante, representada da seguinte forma:

$$Z_{jk} = a + W_1X_{1k} + W_2X_{2k} + ... + W_nX_{nk} \quad (2)$$

onde:
Z_{jk} = escore Z discriminante da função discriminante j para o indivíduo k;
a = intercepto;
W_i = peso discriminante para a variável independente i;
X_{ik} = variável independente i para o indivíduo k;
i =1,2,...,n.

O teste de significância para a função discriminante se dá através da distância entre os centróides (médias) dos grupos, comparando as distribuições dos escores discriminantes dos mesmos. Na Figura 1, há duas distribuições de escores discriminantes. No primeiro diagrama a sobreposição nas distribuições é pequena, significando que a função discriminante separa

bem os grupos (A e B), enquanto que no segundo a sobreposição é grande, significando que a função discriminante não é um bom discriminador entre o grupo A e B.

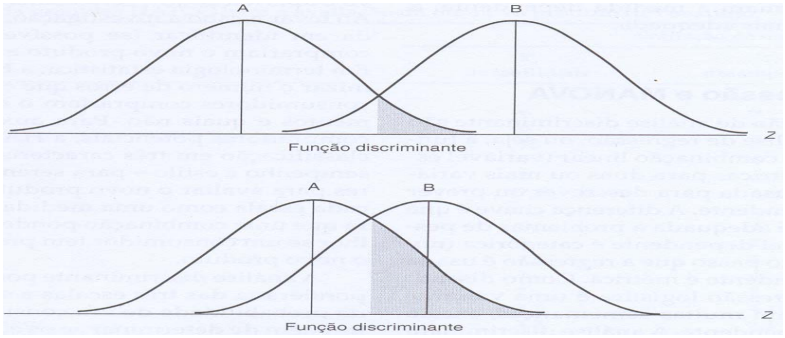

Figura 1 - Representação univariada de escores *Z* discriminantes.

Na Figura 2, há uma representação geométrica da função discriminante de dois grupos. Esse tipo de ilustração gráfica possibilita entender melhor a natureza da Análise Discriminante, mostrando o que realmente acontece com uma função. Supondo-se que existam dois grupos, A e B, e duas medidas, V_1 e V_2, para cada membro dos dois grupos, os pontos pequenos representam as medidas das variáveis do grupo B, enquanto que os pontos grandes representam o grupo A.

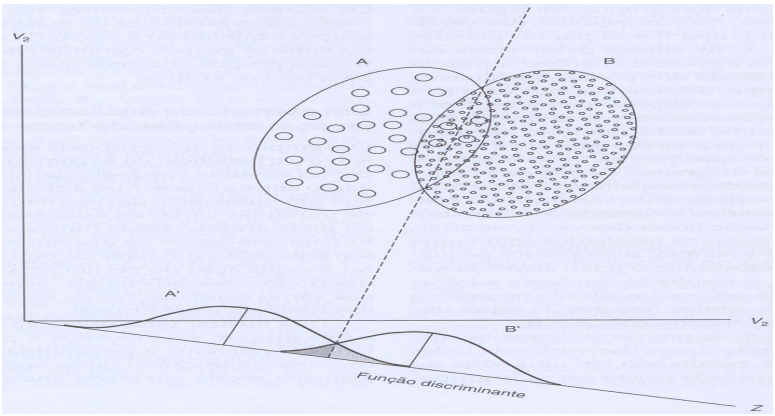

Figura 2 - Ilustração gráfica de análise discriminante de dois grupos.

2.2 O Processo de Decisão para Análise Discriminante

Hair et al. entendem (2005) que a aplicação da Análise Discriminante pode ser demonstrada a partir da construção de um modelo de seis estágios, os quais são descritos nesta literatura através de um fluxograma. Este fluxograma pode ser encontrado na bibliografia citada e não serão reproduzidos neste trabalho. O primeiro passo da análise é definir os objetivos. A seguir abordar questões específicas de planejamento e certificar-se de que as suposições inerentes estão sendo atendidas. A partir dessas premissas, realizar a dedução da função discriminante e determinar se uma função estatisticamente significativa pode ou não ser obtida para separar dois (ou mais) grupos.

2.2.1 Estágio 1: Objetivos da Análise Discriminante

Para implementar a técnica de Análise Discriminante a um conjunto de dados, é necessário averiguar se quatro objetivos são satisfeitos:

I. Na análise, precisa-se de no mínimo dois grupos, para identificar se há diferença estatisticamente significativa entre os perfis de escore médio no conjunto das variáveis estudadas;
II. Definir as variáveis independentes que explicam o máximo de diferenças nos perfis de escore médio dos dois ou mais grupos;
III. O pesquisador deve definir os procedimentos para classificar indivíduos em grupos, tendo como base seus escores em um conjunto de variáveis independentes;
IV. Por fim, deve-se definir o número e a composição das dimensões de discriminação dos grupos formados, baseando-se no conjunto de variáveis independentes.

2.2.2 Estágio 2: Projeto de Pesquisa para Análise Discriminante

Para que a Análise Discriminante tenha um bom resultado, deve ser levada em consideração a seleção de variáveis dependentes e independentes e o tamanho da amostra para se estimar as funções discriminantes e a sua divisão.

Em muitos casos a amostra é dividida em duas sub-amostras, uma usada para a estimação da função discriminante e a outra para fins de validação. É fundamental que cada uma delas, tenha o tamanho adequado. O procedimento padrão divide a amostra total aleatoriamente em dois grupos: em amostra de análise, ou treinamento; e em amostra de teste. A primeira amostra é usada para desenvolver a função discriminante, ou treinar o classificador. A segunda é necessária para testar a adequação da função discriminante. Esse procedimento de validação da função é chamado de validação cruzada ou partição da amostra.

Um outro método de validação cruzada, bastante utilizado na prática e também usado em diversos pacotes estatísticos é conhecido como *Leave-One-Out* (deixar um de fora). Neste método, apenas um indivíduo é removido do grupo total de cada vez. Após isso, este indivíduo que foi separado é testado nas funções discriminantes para verificar se ele é classificado corretamente ou não.

Por último, é também usado o método conhecido como resubstituição, o qual utiliza todos os indivíduos para calcular os coeficientes das funções discriminantes. Após esta etapa, toda a amostra é novamente utilizada para testar o grau de acurácia, isto é, classificação correta, obtido pela Análise Discriminante.

2.2.3 Estágio 3: Suposição da Análise Discriminante

As suposições necessárias para determinar a função discriminante: normalidade multivariada das variáveis independentes; das estruturas (matrizes) de dispersão e covariância desconhecida (mas igual), para os grupos como definidos pela variável dependente. Matrizes de covariância desiguais podem afetar negativamente o processo de classificação.

Os dados devem obedecer à suposição de normalidade multivariada necessária pela Análise Discriminante, pois a falta de normalidade também pode causar problemas na estimação da função discriminante.

O pesquisador não deve esquecer que a falta de multicolinearidade entre as variáveis independentes pode ser outro fator a afetar o resultado. A multicolinearidade mostra se duas ou mais variáveis independentes estão altamente correlacionadas.

Hair et al. (2005) e Teixeira, (2006) afirmam que o teste de MANOVA, é necessário para verificar se as variáveis possuem significância. Caso elas não apresentem significância estatística, são eliminadas do estudo.

Pizzol (2003) empregou em seu trabalho a Análise Discriminante para validar o resultado da pesquisa. Seu objetivo foi apresentar e discutir um método de tipificação de sistemas de produção dividido em duas etapas. Na primeira etapa foram usados grupos focais, e na segunda empregou-se a Análise Discriminante para validar os resultados obtidos nas entrevistas em grupos.

2.2.4 Estágio 4: Estimação do Modelo Discriminante e Avaliação do Ajuste Geral

Para se definir a função discriminante, deve-se decidir o método de estimação e, conseqüentemente, determinar o número de funções. Com as funções estimadas, o ajuste geral do modelo pode ser avaliado de diversas maneiras. O escore *Z* discriminante, também chamado de escore *Z*, pode ser calculado para cada indivíduo. A comparação das médias dos grupos nos escores *Z* fornece uma medida de discriminação entre grupos.

Dois métodos computacionais são usados para determinar a função discriminante: o método simultâneo (direto) e o método seqüencial (*stepwise*). O primeiro método inclui todas as variáveis independentes na análise. O segundo envolve a inclusão das variáveis independentes na função discriminante, uma por vez, com base em seu poder discriminatório.

Muitos critérios são empregados para determinar a significância da variável. Quando é aplicado o método *stepwise*, utilizam-se as medidas D^2 de *Mahalanobis* e *V* de *Rao*. A primeira baseia-se na distância euclidiana quadrada generalizada, a qual se adapta a variâncias desiguais. O critério de significância convencional é de 0,05 ou acima.

Então, assim que as funções discriminantes são definidas, deve-se verificar o ajuste geral das funções discriminantes encontradas no estudo. Primeiramente, calcula-se o escore *Z* discriminante para cada observação. A seguir avaliam-se as diferenças de grupos nos escores *Z* discriminantes, e, por fim, avalia-se a precisão de previsão de pertinência a grupos.

Os escores *Z* para qualquer função discriminante podem ser calculados para cada indivíduo, conforme a equação (2). Depois que a função é ajustada, deve-se avaliar as diferenças entre os grupos.

A avaliação da precisão preditiva de pertinência ao grupo é necessária para verificar se cada observação foi corretamente classificada. Para isso, é preciso levar em consideração a concepção estatística e prática para desenvolver matrizes de classificação, a determinação do escore de corte, a construção das matrizes de classificação e os padrões para avaliar a precisão de classificação na análise.

As matrizes de classificação (conhecidas como matrizes de confusão) são importantes para determinar a habilidade preditiva de uma função discriminante, ou seja, para mostrar uma avaliação mais precisa do poder discriminatório da função, ou acurácia de classificação.

A determinação do escore de corte determina o grupo em que o indivíduo deve ser classificado. A construção das matrizes de classificação define se o escore de corte é ótimo ou não. Esse corte depende do tamanho dos grupos, se eles são iguais ou diferentes. O escore para dois grupos de mesmo tamanho é definido conforme a equação (3), e de tamanhos diferentes conforme a equação (4):

$$Z_{CE} = \frac{Z_A + Z_B}{2} \tag{3}$$

onde:
Z_{ce} = valor do escore crítico para grupos de mesmo tamanho;
Z_A = centróide de grupo A;
Z_B = centróide de grupo B.

$$Z_{CU} = \frac{N_A Z_B + N_B Z_A}{N_A + N_B} \tag{4}$$

onde:
Z_{CU} = valor do escore crítico para grupos com tamanhos diferentes;
N_A = número no grupo A;
N_B = número no grupo B;

Z_A = centróide para o grupo A;
Z_B = centróide para o grupo B.

As Figuras 3 e 4 mostram os escores de corte ótimo para grupos iguais e diferentes.

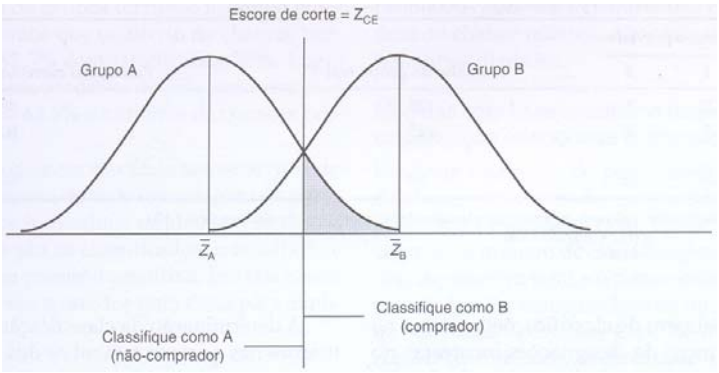

Figura 3 - Escore de corte ótimo com iguais tamanhos de amostra.

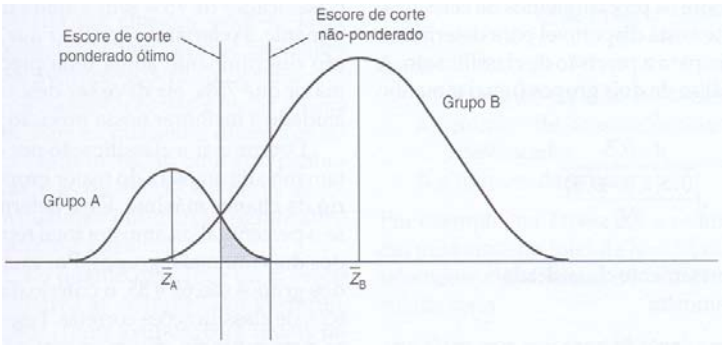

Figura 4 - Escore de corte ótimo com tamanhos diferentes de amostra.

Para validar a função discriminante, devem-se obter amostras aleatórias, criando-se dois grupos: Um grupo (amostra de análise) é utilizado para a obtenção da função discriminante; e o outro (análise de teste) para validar a função, criando a matriz de classificação. O critério de classificação de cada indivíduo no grupo é definido como:

Classifique um indivíduo no grupo A se $Z_n < Z_{ct}$
Classifique um indivíduo no grupo B se $Z_n > Z_{ct}$

Onde:
Z_n : escore Z discriminante para o n-ésimo indivíduo
Z_{ct}: valor do escore de corte crítico

O teste "t" para o procedimento de classificação é definido conforme a equação (5):

$$t = \frac{p - 0{,}5}{\sqrt{\frac{0{,}5(1{,}0 - 0{,}5)}{N}}} \qquad (5)$$

onde:
p = proporção corretamente classificada
N = tamanho da amostra

2.2.5 Estágio 5: Interpretação dos Resultados

O pesquisador deve determinar a importância relativa de cada variável independente na discriminação entre os grupos. Para isso, utiliza os pesos discriminantes padronizados, cargas discriminantes ou contribuições das variáveis preditoras, a cada função separadamente (correlação de estrutura) e valores *F* parciais. No primeiro caso, analisa-se o sinal e a magnitude do peso discriminante das funções discriminantes. No segundo, mede-se a correlação linear simples entre cada variável independente e a função discriminante. E, no terceiro, aponta-se o nível associado de significância para cada variável.

3 ESTUDO DE CASO: CONCURSO VESTIBULAR 2007

Esta pesquisa analisou o banco de dados dos escores dos candidatos ao concurso Vestibular 2007 da UFSM, visando o estudo de adequabilidade de um índice de discriminação por disciplina através da aplicação da técnica de Análise Discriminante.

3.1 Análise Descritiva por Situação

As variáveis independentes são representadas por onze provas: Biologia, Física, Química, Geografia, História, Literatura Brasileira, Língua Estrangeira, Língua Portuguesa, Matemática, Filosofia e Redação.

A variável dependente é definida como situação. Nesta pesquisa, essa variável é o grupo dos candidatos não-selecionados[1], selecionados[2] e classificados[3], representados por 0, 1 e 2, respectivamente. A Tabela 1 apresenta as variáveis estudadas e seus tipos.

Tabela 1 - Variáveis estudas na pesquisa.

Variável	Tipo da variável	
Código do curso	identificadores	
Nome do curso	identificadores	
Inscritos	identificadores	1-18020
Sexo	identificadores	M, F
Idade	identificadores	0-107
Bio	independente	0-15
Fis	independente	0-15
Qui	independente	0-15
Geo	independente	0-15
His	independente	0-15
Lit	independente	0-15
Lin Est	independente	0-15
LP	independente	0-15
Mat	independente	0-15
Fil	independente	0-15
Red	independente	0-7,5
Sit	dependente	0, 1, 2
Class	identificadores	1-nro

Fonte: Setor de Estatística e Informática da COPERVES.

[1]Candidatos não-selecionados para a correção da prova de redação.

[2]Candidatos selecionados para a correção da prova de redação, mas não classificados no Vestibular.

[3]Candidatos classificados na prova de redação e, portanto, no Vestibular.

3.2 Análise Descritiva Geral

A COPERVES possui em seu sistema 21.053 candidatos inscritos no concurso Vestibular 2007 da UFSM. Deste total de inscritos, a instituição apresentou a freqüência de 18.020 candidatos em todo o concurso.

A análise foi realizada por curso, onde foram selecionados três dos sessenta e seis cursos oferecidos pela instituição. Escolheram-se os cursos de forma a testar amostras de diferentes tamanhos, sendo uma amostra, ou curso, de tamanho grande, uma de tamanho médio e uma de tamanho pequeno.

3.3 Análise Descritiva por Curso

3.3.1 Medicina

O curso de Medicina apresenta a freqüência de 2.015 candidatos. Deste total, 1.176 são do sexo feminino e 839 do masculino. O presente curso apresenta idade média de 20,3 anos, mediana de 20,0 anos e desvio padrão de 3,0 anos; freqüência de 1.826 (90,6%) candidatos não-selecionados, 109 (5,4%) selecionados e 80 (4%) classificados.

As notas médias por prova entre os três grupos, não-selecionados, selecionados e classificados, foram comparadas inicialmente através da Análise de Variância Multivariada (MANOVA). Aplicando o teste de *lambda* de *Wilks* ao conjunto de dados do curso de Medicina, verificou-se a significância do fator situação. Constata-se que o valor de significância é menor que 0,05, e, portanto, o fator situação é significativo. Existe significância na diferença entre as notas médias por prova para pelo menos um dos três grupos utilizando a análise de variância para o fator situação.

Neste caso, foi explorado através da diferença mínima significativa (teste DMS), onde houve diferenças significativas.

Comparando-se os três grupos nas provas de Biologia, Física, Química, Geografia, História, Literatura, Língua Estrangeira, Língua Portuguesa e Matemática para o curso de Medicina, constata-se que o grupo dos não-selecionados possui notas médias inferiores ao grupo dos selecionados e dos classificados.

Assim, considerando a nota média geral dos candidatos por prova nos três grupos, conclui-se que o grupo dos não-selecionados apresenta notas médias inferiores em todas as provas, se comparados aos grupos dos selecionados e dos classificados. Entretanto, o grupo dos selecionados e dos classificados apresenta notas médias estatisticamente iguais em quase todas as provas. Apenas na prova de Filosofia o grupo dos selecionados possui nota média estatisticamente inferior ao grupo dos classificados ($p = 0,001$).

Prosseguindo a análise, foi utilizado o teste "*t*" para a comparação entre as notas médias de redação para os grupos selecionados e classificados, já que o grupo dos não-selecionados não possui a redação. Novamente, a nota média para o fator situação neste caso é significativa ($p = 0,001$), portanto, a nota média de redação do grupo dos selecionados é inferior à nota média dos classificados. A nota da prova de redação discrimina o grupo dos selecionados dos classificados, já que o grupo dos selecionados possui 9,3 em nota média a menos que o grupo dos classificados.

3.3.2 Direito (diurno)

O curso de Direito apresenta a freqüência de 483 candidatos. Deste total, 281 são do sexo feminino e 202 do masculino. O presente curso apresenta idade média de 20,7 anos, mediana de 19,0 anos e desvio padrão de 6,8 anos; freqüência de 419 (86,7%) candidatos não-selecionados, 32 (6,6%) selecionados e 32 (6,6%) classificados.

A estatística de *lambda* de *Wilks* no curso de Direito constatou-se que é significativa para o fator situação.

Houve significância estatística na análise de variância para todas as provas, rejeitando-se a hipótese nula de que a nota média por prova em cada prova dos três grupos é igual, a favor da hipótese alternativa de que pelo menos uma das notas médias dos três grupos difere entre si.

Comparando-se os três grupos para as provas de Biologia, Física, Química, Geografia, História, Literatura, Língua Portuguesa, Matemática e Filosofia quanto à nota média, nota-se que há diferença significativa entre o grupo dos não-selecionados e dos selecionados. As notas médias do grupo dos não-selecionados são inferiores às notas médias do grupo dos selecionados e dos classificados.

Comparando-se o grupo dos selecionados com o dos classificados, constata-se uma diferença significativa para as provas de Biologia (p = 0,001), Física (p = 0,000), Química (p = 0,029) e Matemática (p = 0,000), e uma diferença não significativa para as provas de literatura (p = 0,636), Língua Estrangeira, (p = 0,509), Língua Portuguesa (p = 0,483) e Filosofia (p = 0,567). Todas as análises levaram em consideração um nível de significância de 5% para a ANOVA e também para os testes "*post hoc"*, os quais também formam o teste de DMS.

Utiliza-se o teste "*t*" para comparar as notas médias das redações, através do qual se obtém nota média de 53,09 para os selecionados e 68,03 para os classificados, a qual apresenta diferença significativa. A nota da prova de redação discrimina o grupo dos selecionados dos classificados (p = 0,006), já que o grupo dos selecionados apresenta nota média 14,94 a menos que o grupo dos classificados.

3.3.3 Música (Licenciatura Plena)

O curso de Música teve 37 candidatos. Deste total, 9 são do sexo feminino e 28 do masculino. A idade média é de 22,7 anos, mediana de 21,0 anos e desvio padrão de 8,0 anos, freqüência de 13 (35,1%) candidatos não selecionados, 12 (32,4%) selecionados e 12 (32,4%) classificados.

Para o presente curso, também se constatou significância estatística na análise de variância para todas as provas. A estatística de *lambda* de *Wilks* para o curso de Música é significativa para o fator situação.

Comparando-se os três grupos para todas as provas quanto à nota média, também se constatam que há diferença significativa entre o grupo dos não- selecionados e dos selecionados. Portanto, as notas médias do grupo dos não- selecionados são inferiores às notas médias do grupo dos selecionados e dos classificados.

O teste "*t*" para grupo dos selecionados e dos classificados, apresentou diferença de 8,92, a qual não foi significativa a 5% (p = 0,097, ou seja, 9,7%). Entretanto, considerando-se um nível de significância de 10%, essa diferença é significativa. Isso caracteriza que existe tendência forte de que a nota média da redação para o grupo dos selecionados seja estatisticamente inferior à nota média dos classificados.

Comparando-se o grupo dos selecionados com o dos classificados, observa-se uma diferença significativa para as provas de Literatura (p = 0,048), Língua Estrangeira (p = 0,034), Língua Portuguesa (p = 0,024) e Filosofia (p = 0,026), e uma diferença não significativa para as provas de Biologia (p = 0,336), Física (p = 0,183), Química (p = 0,219), Geografia (p = 0,180), História (p = 0,062) e Matemática (p = 0,229). A nota da prova de redação discrimina o grupo dos selecionados dos classificados, já que o grupo dos selecionados possui nota média de 8,9 a menos que o grupo dos classificados.

4 ANÁLISE DISCRIMINANTE

Dos três métodos de classificação que podem ser usados, resubstituição, validação cruzada por *Leave-One-Out – LOO* e validação cruzada com 50% das amostras para análise e 50% para teste, neste trabalho serão aplicados os dois primeiros.

4.1 Medicina

A Análise Discriminante para os três grupos estimou duas funções discriminantes: a primeira função discriminante é responsável por 98,6% da variabilidade total dos casos, enquanto que a segunda função discriminante é responsável por apenas 1,4%.

As duas funções discriminantes canônicas para o curso de Medicina, segundo as dez provas e aplicando-se a equação (2), são definidas da seguinte forma:

$$FDC_1 = -3{,}236 - 0{,}045Bio + 0{,}111Fis + 0{,}082Qui - 0{,}009Geo + 0{,}048His + 0{,}116Lit + 0{,}007LE - 0{,}034LP + 0{,}064Mat + 0{,}110Fil$$

$$FDC_2 = -0{,}941 - 0{,}147Bio - 0{,}040Fis + 0{,}072Qui - 0{,}022Geo + 0{,}071His - 0{,}057Lit + 0{,}076LE - 0{,}030LP - 0{,}206Mat + 0{,}438Fil$$

Onde:
FDC_1: primeira função discriminante canônica
FDC_2: segunda função discriminante canônica

A função de classificação é definida pelas notas das provas de cada candidato que será aplicada na função de classificação do grupo dos não-selecionados, dos selecionados e dos classificados. Portanto, a função que mais atribuir probabilidade ao candidato será o grupo ao qual pertencerá. A primeira função refere-se ao grupo dos não-selecionados (0), a segunda ao dos selecionados (1) e a terceira ao dos classificados (2):

$$FC(0) = -10{,}440 + 0{,}950Bio - 0{,}041Fis - 0{,}016Qui + 0{,}373Geo + 0{,}069His + 0{,}059Lit + 0{,}081LE + 0{,}420LP - 0{,}559Mat + 0{,}730Fil$$

$$FC(1) = -17{,}350 + 0{,}899Bio + 0{,}166Fis + 0{,}115Qui + 0{,}362Geo + 0{,}140His + 0{,}279Lit + 0{,}079LE + 0{,}365LP - 0{,}402Mat + 0{,}837Fil$$

$$FC(2) = -19{,}415 + 0{,}819Bio + 0{,}185Fis + 0{,}174Qui + 0{,}349Geo + 0{,}187His + 0{,}292Lit + 0{,}115LE + 0{,}341LP - 0{,}473Mat + 1{,}067Fil$$

Onde:
FC(0): função de classificação do grupo 0
FC(1): função de classificação do grupo 1
FC(2): função de classificação do grupo 2

A ilustração gráfica da Figura 5 permite a visualização dos candidatos em torno do centróide para os três grupos conjuntamente.

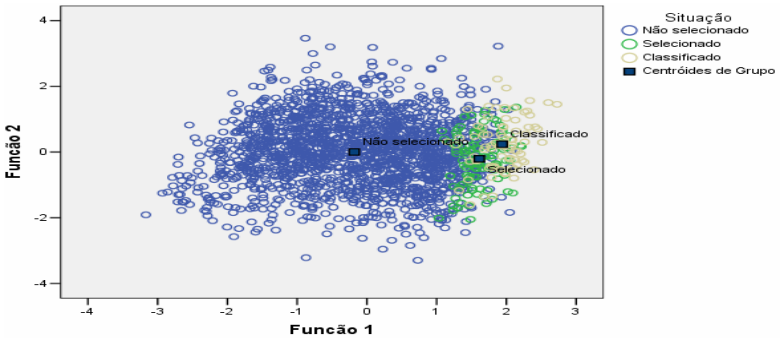

Figura 5 – Função Discriminante Canônica para os três grupos (Medicina).

O resultado de classificação para os três grupos através da matriz de classificação ou confusão, resumindo os dois métodos empregados na pesquisa: o método da resubstituição (emprego de toda a amostra) e o método *Leave-One-Out* (retirada de um elemento da amostra). O primeiro método classifica corretamente 1.384 (75,8%) dos 1.826 candidatos do grupo dos não-selecionados, 72 (66,1%) dos 109 candidatos do grupo dos selecionados e 52 (65%) dos 80 candidatos do grupo dos classificados, com média geral de 74,8% para os três grupos. No segundo método (LOO) o percentual de classificação correta é menor (74,2%). Comparando a média geral entre o método da resubstituição e o método *Leave-One-Out*, conclui-se que não há diferença significativa.

O método *Leave-One-Out* é um método de teste, apesar de não ter separado 50% da amostra, conforme o método de teste estudado na metodologia deste trabalho. A classificação correta dos grupos pode ser observada na diagonal principal da matriz de classificação de 1.384 (75,8%) para o grupo dos não-selecionados, 65 (59,6%) para o dos selecionados e 46 (57,5%) para o dos classificados.

Analisando o método da resubstituição através da matriz de classificação, observa-se que 33,9% dos candidatos selecionados poderiam ser enquadrados como classificados, e por outro lado, 35% dos candidatos classificados poderiam ser enquadrados como selecionados. Tal confusão pode ser explicada por dois motivos: (i) pelo empate na pontuação final (somatório do número de acertos nas provas de múltipla escolha com o número de acertos da prova de redação); (ii) pela mínima diferença entre os escores de cada candidato. No entanto, percebe-se que a diferença mínima influenciará somente o curso de Medicina, observada entre as posições 79, 80 e 81 (mesma pontuação final). Neste caso, a COPERVES utiliza um critério de desempate.

A Análise Discriminante para os dois grupos (candidatos selecionados e classificados) estimou uma função discriminante. O autovalor encontrado para a função é de 1,883, conforme a variância de 100%.

A função discriminante canônica para o curso de Medicina, segundo a prova de redação, é:

$$\begin{aligned} FDC1 = {} & -48{,}755 + 0{,}399Bio + 0{,}346Fis + 0{,}306Qui + 0{,}409Geo + 0{,}424His \\ & + 0{,}351Lit + 0{,}410LE + 0{,}479LP + 0{,}385Mat + 0{,}442Fil + 0{,}027Red \end{aligned}$$

As duas funções de classificação, incluindo também a prova de redação, são definidas pelas equações:

$$\begin{aligned} FC(0) = {} & -1232{,}415 + 31{,}024Bio + 18{,}393Fis + 14{,}264Qui + 21{,}141Geo + 19{,}009His \\ & + 18{,}942Lit + 20{,}716LE + 24{,}146LP + 20{,}510Mat + 16{,}880Fil + 0{,}908Red \end{aligned}$$

$$\begin{aligned} FC(1) = {} & -1367{,}700 + 32{,}127Bio + 19{,}350Fis + 15{,}110Qui + 22{,}272Geo + 20{,}181His \\ & + 19{,}912Lit + 21{,}847LE + 25{,}469LP + 21{,}575Mat + 18{,}102Fil + 0{,}982Red \end{aligned}$$

Utilizando o método da resubstituição, dos 109 candidatos do grupo dos selecionados, classificaram-se corretamente 108 (99,1%), ou seja, apenas um candidato foi mal classificado, e dos 80 candidatos do grupo dos classificados, 76 (95%) foram corretamente classificados e 4 incorretamente. Na média geral, tem-se 97,4% de classificação correta pelo método da resubstituição e 94,2% pelo método da validação cruzada ou método *Leave-One-Out* (retirada de um candidato da amostra). Neste caso, o curso de Medicina também apresenta um bom poder de classificação para os dois métodos, pois os percentuais são altos e próximos, comprovando a importância de uma amostra grande na aplicação dos métodos em estudo.

4.2 Direito (diurno)

No curso de Direito a primeira função discriminante é responsável por 97,5% da variabilidade total, enquanto que a segunda função discriminante é responsável por 2,5%. As funções discriminantes canônicas para o curso de Direito são as seguintes:

$$FDC_1 = -3{,}911 + 0{,}130Bio + 0{,}157Fis + 0{,}131Qui + 0{,}016Geo + 0{,}108His + 0{,}056Lit + 0{,}006LE - 0{,}053LP + 0{,}104Mat + 0{,}034Fil$$

$$FDC_2 = -1{,}011 - 0{,}199Bio - 0{,}121Fis + 0{,}148Qui + 0{,}165Geo - 0{,}039His + 0{,}228Lit + 0{,}027LE + 0{,}141LP - 0{,}306Mat + 0{,}014Fil$$

As funções de classificação foram definidas da seguinte forma:

$$FC\ (0) = -9{,}695 + 0{,}929Bio + 0{,}223Fis + 0{,}271Qui + 0{,}344Geo + 0{,}164His + 0{,}156Lit + 0{,}082LE - 0{,}015LP - 0{,}110Mat + 0{,}705Fil$$

$$FC\ (1) = -21{,}783 + 1{,}136Bio + 0{,}537Fis + 0{,}666Qui + 0{,}470Geo + 0{,}405His + 0{,}412Lit + 0{,}111LE - 0{,}070LP - 0{,}022Mat + 0{,}794Fil$$

$$FC\ (2) = -27{,}427 + 1{,}444Bio + 0{,}804Fis + 0{,}670Qui + 0{,}342Geo + 0{,}550His + 0{,}269Lit + 0{,}093LE - 0{,}248LP + 0{,}353Mat + 0{,}817Fil$$

A Figura 6 ilustra a representação do mapa territorial para cada candidato em torno do centróide para os três grupos conjuntamente.

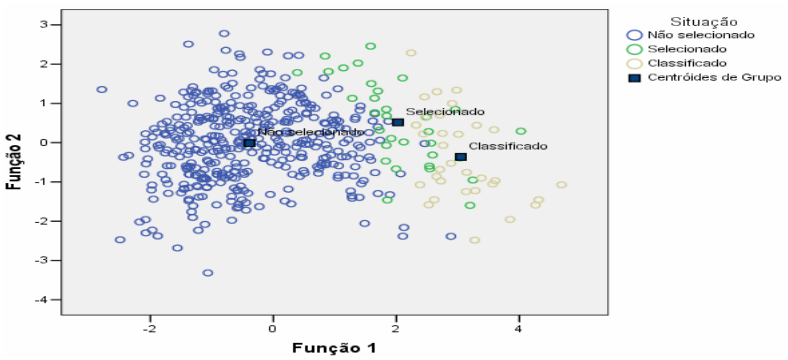

Figura 6 – Função Discriminante Canônica para os três grupos (Direito).

A classificação correta dos três grupos pode ser observada através da diagonal principal da matriz de classificação pelo método de resubstituição e pelo método *Leave-One-Out*. O primeiro método classifica corretamente 360 (85,9%) dos 419 candidatos do grupo dos não-selecionados, 21 (65,6%) dos 32 candidatos do grupo dos selecionados e 23 (71,9%) dos 32 candidatos do grupo dos classificados, com média geral de 83,6% para os três grupos. O segundo classifica corretamente 353 (84,2%) dos 419 candidatos do grupo dos não-selecionados, 20 (62,5%) dos 32 candidatos do grupo dos selecionados e 22 (68,8%) dos 32 candidatos do grupo dos classificados, com média geral de 81,8% para os três grupos.

O autovalor da função para o curso de Direito na prova de redação é de 1,702, e a função discriminante canônica é descrita a seguir:

$$FDC_1 = -20{,}018 + 0{,}339Bio + 0{,}156Fis + 0{,}072Qui + 0{,}118Geo + 0{,}317His + 0{,}145Lit + 0{,}169LE + 0{,}063LP + 0{,}185Mat + 0{,}142Fil + 0{,}039Red$$

As duas funções de classificação para a prova de redação são definidas da seguinte maneira:

$$FC\ (0) = -242{,}836 + 7{,}660Bio + 0{,}258Fis + 2{,}222Qui + 2{,}913Geo + 6{,}683His + 8{,}103Lit + 4{,}209LE + 5{,}502LP + 2{,}453Mat + 4{,}415Fil + 0{,}580Red$$

$$FC\ (1) = -294{,}250 + 8{,}531Bio + 0{,}659Fis + 2{,}408Qui + 3{,}215Geo + 7{,}499His + 8{,}475Lit + 4{,}643LE + 5{,}665LP + 2{,}927Mat + 4{,}779Fil + 0{,}679Red$$

4.3 Música (Licenciatura Plena)

No curso de Música a primeira função discriminante é responsável por 92,4% da variabilidade total e separação, enquanto que a segunda função discriminante é responsável por 7,6%. As funções discriminantes canônicas para o curso de Música são as seguintes:

$$FDC_1 = -7{,}532 + 0{,}192Bio + 0{,}391Fis - 0{,}267Qui + 0{,}066Geo + 0{,}400His + 0{,}040Lit + 0{,}514LE - 0{,}256LP + 0{,}227Mat - 0{,}064Fil$$

$$FDC_2 = -2{,}828 - 0{,}284Bio + 0{,}096Fis + 0{,}323Qui - 0{,}033Geo + 0{,}026His - 0{,}085Lit - 0{,}200LE + 0{,}346LP - 0{,}057Mat + 0{,}406Fil$$

As funções de classificação foram definidas da seguinte forma:

$$FC\ (0) = -21{,}356 + 0{,}284Bio + 1{,}881Fis - 0{,}425Qui + 0{,}633Geo + 2{,}086His - 0{,}542Lit + 2{,}051LE + 0{,}109LP + 1{,}561Mat + 0{,}829Fil$$

$$FC\ (1) = -35{,}219 + 1{,}042Bio + 2{,}760Fis - 1{,}409Qui + 0{,}829Geo + 3{,}056His - 0{,}358Lit + 3{,}527LE - 0{,}870LP + 2{,}184Mat + 0{,}269Fil$$

$$FC\ (2) = -53{,}341 + 1{,}050Bio + 3{,}546Fis - 1{,}504Qui + 0{,}905Geo + 3{,}777His - 0{,}387Lit + 4{,}188LE - 0{,}920LP + 2{,}513Mat + 0{,}619Fil$$

A Figura 7 ilustra a representação do mapa territorial para os três grupos. Os números 1, 2 e 3 e o asterisco (*) representam o grupo dos não-selecionados (0), dos selecionados (1), dos classificados (2) e o grupo de centróides, respectivamente.

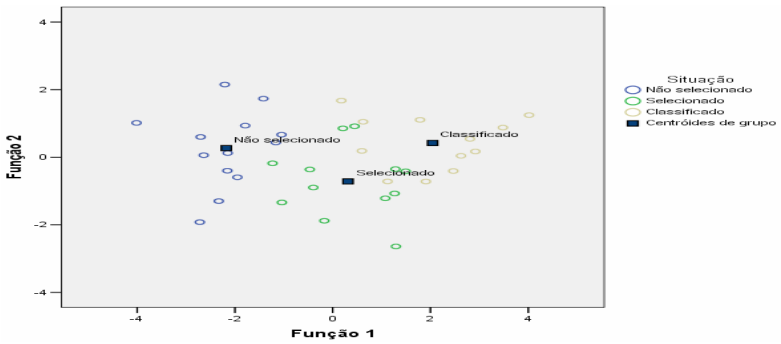

Figura 7 – Função Discriminante Canônica para os três grupos (Música).

A classificação correta dos três grupos pode ser observada através da diagonal principal da matriz de classificação pelo método de resubstituição e pelo método *Leave-One-Out*. A média geral para o primeiro método é de 89,2% e para o segundo de 59,5%. Comparando a média geral dos dois métodos, conclui-se que há diferença significativa. Isto porque como a amostra possui poucos elementos, com a retirada de um candidato ocorre a queda do poder de discriminação do método frente às amostras desconhecidas quando se usa o método *Leave-One-Out* ao invés do método de resubstituição.

O autovalor da função para o curso de Música na prova de redação é de 2,929, e a função discriminante canônica é descrita a seguir:

FDC1= - 13,630 + 0,100Bio + 1,127Fis - 0,782Qui - 0,258Geo + 0,083His
+ 0,300Lit + 0,338LE - 0,240LP + 0,326Mat + 0,504Fil + 0,109Red

As duas funções de classificação para a prova de redação são definidas da seguinte maneira:

FC (0) = - 77,403 + 1,995Bio + 13,028Fis - 10,259Qui - 2,698Geo + 1,191His
+ 3,775 Lit + 5,424LE - 4,121LP + 5,072Mat + 4,818Fil + 1,424Red

FC (1) = - 122,069 + 2,323Bio + 16,721Fis - 12,821Qui - 3,545Geo + 1,463His
+ 4,758Lit + 6,532LE - 4,907LP + 6,141Mat + 6,468Fil + 1,782Red

A média geral para o método da resubstituição é de 95,8% e para o método *Leave-One-Out* é de 58,3%. Neste caso no qual a redação faz parte da equação da função discriminante canônica e da função de classificação, nota-se que a Análise Discriminante também possui um baixo poder de classificação quando os grupos são pequenos. Esse baixo poder é observado pela diferença significativa de acurácia entre os dois métodos (37,5%).

5 CONCLUSÃO

Considerando os aspectos descritos neste artigo, foram planejados como objetivos para este trabalho: (i) avaliar um método de classificação para os inscritos no Vestibular 2007 em relação aos escores individuais por disciplina e curso pretendido; (ii) analisar as estatísticas do Vestibular 2007 da UFSM de modo a identificar o grau de discriminação entre os candidatos, em relação às provas e por curso pretendido; (iii) observar a relação entre os escores individuais por disciplina e classificação final do candidato por curso pretendido.

Conforme a proposta da pesquisa, a capacidade de discriminação entre candidatos com base na Análise Discriminante (AD) foi avaliado separadamente para cada curso. Os escores discriminantes da AD podem ser entendidos como um índice. A partir dessas notas originais, foram obtidas duas funções discriminantes e três funções de classificação, considerando conjuntamente os três grupos (não-selecionados, selecionados e classificados) e uma função discriminante e duas funções de classificação para dois grupos (selecionados e classificados).

Se o valor da função de classificação discriminante de um determinado candidato para a situação “não-selecionado” (0) for maior do que da situação “selecionado” (1) e “classificado” (2), esse candidato é atribuído ao primeiro grupo (0), se o valor da função de classificação discriminante da situação 1 for maior do que a situação 0 e 2, o candidato pertence ao segundo grupo (1) e se o valor função discriminante da situação 2 for maior que os valores das outras duas funções, então o candidato fará parte do terceiro grupo (2).

Analisando os três cursos separadamente, representando gradativamente amostras grandes, médias e pequenas, respectivamente pelos curso de Medicina, Direito e Música, observa-se que quanto maior o número de candidatos no curso, ou seja, quanto maior o tamanho da amostra, o método de classificação pela resubstituição apresenta um grau de classificação muito acurado. Quando se retira um candidato na fase do cálculo das funções e depois o testa através da aplicação do método *Leave-One-Out*, percebe-se que sempre haverá um percentual de classificação menor comparado com a resubstituição, entretanto também elevado no caso de grandes amostras.

Como a amostra para o curso de Música é muito pequena, a estimação das covariâncias entre e dentro dos grupos, as correlações, e por conseguinte a estimação dos coeficientes das funções discriminantes se tornam bastante imprecisas. Nestes casos, pode-se obter um grau relativamente aceitável de acurácia na classificação quando se usa a

resubstituição, mas ao apresentar um elemento estranho ao grupo que computou os coeficientes das funções discriminantes pelo *Leave-One-Out*, percebe-se uma grande probabilidade de cometer um erro de classificação, devido à baixa confiabilidade dos parâmetros estimados para esse novo indivíduo, pois qualquer pequena variação nos dados aparentará aos coeficientes ser completamente diferentes. Assim, observa-se, por exemplo no curso de Música, 95,8% de classificação correta, em média, quando utiliza o método de resubstituição e apenas 58,3% quando emprega o *Leave-One-Out*.

Conclui-se, que a Análise Discriminante possui um bom poder de classificação quando os grupos são compostos de amostras de proporções grandes e medianas em relação ao número de parâmetros estimados, e um baixo poder de classificação quando os grupos são formados por amostras pequenas, pois elas apresentam uma baixa confiabilidade na estimação dos parâmetros das funções discriminantes e consequentemente um maior grau de classificação incorreta.

6 REFERÊNCIAS BIBLIOGRÁFICAS

BARBOSA, E. F. et. al. **Implantação da qualidade total na educação**. Belo Horizonte: Fundação Chistiano Ottoni, 1995.

HAIR, J. F.; ANDERSON, R. E.; TATHAM, R. L.; BLACK, W. C. **Análise multivariada de dados**. 5. ed. Porto Alegre: Bookman, 2005.

PIZZOL, S. J. S. **Revista de economia e sociologia rural.** 10 jul. Disponível em: <http:www.scielo.br/scielo.php>. Acesso em: 10 jul. 2007.

SANTOS, O. J. S.; MILIONI, A. Z. **Composição de especialistas para classificação de dados.** 23 abr. Disponível em: <http: //www.scielo.oces.mctes.pt/pdf/iop/v25n1/v25n1a06.pdf>. Acesso em: 23 abr. 2007.

TEIXEIRA, L. L. **O uso de técnica de estatística multivariada no prognóstico de desistência de alunos em IES privados: um estudo de caso na cidade de Foz do Iguaçu-PR**. 06 de abr. Disponível em: <http://www. dspace.c3sl.ufpr.br/dspace/bitstream/1884/7553/1/disserta%c7%c3O+levi>. Acesso em: 06 abr. 2007.