MARCELO RIBEIRO DA LUZ
MARCOS FABIANE KUFNER

Comparação entre Análise Discriminante e Regressão Logística como método de melhor resultado na verificação de fatores que influenciam a evasão de alunos do Curso de Estatística da UFPR.

Trabalho apresentado para a disciplina de Laboratório de Estatística II do Curso de graduação em Estatística da Universidade Federal do Paraná.
Orientadora: Profa. Sonia Isoldi M. Muller

CURITIBA
2008

RESUMO

O estudo apresentado neste trabalho tem como objetivo comparar duas técnicas estatísticas, a Análise Discriminante e a Regressão Logística, como sendo a melhor técnica para verificar os fatores que influenciam na evasão de alunos do Curso de Estatística. Neste estudo foram coletadas informações de 166 alunos que ingressaram no Curso de Estatística da Universidade Federal do Paraná entre os anos de 1998 e 2000, sendo que foi tomado como base nota e freqüência das 5 disciplinas cursadas no 1° semestre do curso, e variáveis como gênero, estado civil, classificação no vestibular, escores no vestibular das matérias de matemática, física, química, português, biologia, história, língua estrangeira moderna e redação. Como a variável resposta é dicotômica, ou seja, apresenta as categorias desistência do curso ou a não desistência do curso, pode-se aplicar estas duas análises neste estudo. Conclui-se pelos resultados obtidos nas análises, que não há diferença entre os métodos utilizados e, que as variáveis nota e freqüência têm uma maior influência na evasão do aluno no Curso de Estatística da Universidade Federal do Paraná.

Palavras-chaves: Regressão Logística, Análise Discriminante, Evasão.

SUMÁRIO

1. INTRODUÇÃO

1.1 O PROBLEMA

No Brasil, evasão escolar entende-se como a interrupção do ciclo de estudos, o que é uma realidade em todas as IES (Instituição de Ensino Superior) do país. Esse abandono trás prejuízos tanto para o aluno que não terminou o curso e não terá em seu currículo o título de formação, quanto para as instituições que perdem em prestigio externo ou internamente e também a sociedade com investimentos mal aproveitados, conforme CASTRO (2005).

Por que um jovem ou uma jovem que, por meio de todos os esforços possíveis, conseguiu uma vaga universitária abandona a escola? A desistência na educação superior é relacionada a grande diversidade do sistema e à especificidade de cada instituição.

Na busca de respostas para as causas desse fenômeno deve-se analisar o que está sendo efetivamente implementado para favorecer as condições acadêmicas do aluno e, conseqüentemente, melhorar o sistema de ensino nacional. Conforme enfatiza CASTRO (2005), a evasão é sempre um processo individual, se bem que pode constituir-se em fenômeno coletivo a ser estudado como associado à eficiência do sistema.

Pode haver decepções, também, quanto às expectativas levantadas em relação à vida universitária, à estrutura e metodologia do trabalho acadêmico e ao excesso de aulas teóricas nos primeiros semestres, quando o aluno, mesmo com o pouco conhecimento específico, almeja o exercício da profissão.

Candidatos à educação superior, em decorrência de suas condições sociais e financeiras, desistem desde o início, da tentativa de ingressar em um curso mais concorrido, portanto, de mais difícil acesso, e optam por outro menos procurado, mesmo com pouco interesse em exercer a profissão correspondente. Esperam que a opção por áreas menos concorridas possibilite o ingresso a um nível educacional, cujo título poderá facilitar a ascensão social.

Neste aspecto, observa que a universidade construiu em seu interior um sistema semelhante ao resto do sistema escolar. Algumas carreiras fazem parte do sonho da maioria dos candidatos, e chegam a selecionar os mais preparados, seja qual for o critério de seleção.

Com o significativo crescimento da iniciativa privada na educação superior brasileira, fatores econômicos ligados ao trabalho e ao estudo podem ser mais decisivos que a qualidade.

Conforme CASTRO (2005), as raízes das diferentes formas de abandono são distintas e as ações preventivas para tratarem desses comportamentos também devem ser diferentes. Antes de iniciar programas de manutenção dos estudantes na universidade, é indispensável conhecer as formas de evasão. Não basta saber quem e quantos abandonam, mas o porquê da decisão e avaliar o grau de integração universitário, a fim de buscar o desenvolvimento dos sistemas.

Inúmeros estudos, teses de mestrados, doutorado tentam entender os aspectos comuns entre os estudantes que evadem os cursos superiores, na tentativa de criar uma ferramenta que possa identificar características visando auxiliar a IES a desenvolver programas que ajudem a reduzir os números da evasão.

Neste estudo analisou-se diferentes variáveis de alunos matriculados no ano de 1998/1999/2000 no Curso de Estatística da Universidade Federal do Paraná, notas e freqüências nas 5 disciplinas do 1º semestre do curso, gênero, estado civil, classificação no vestibular, escore em matemática, português, biologia, química, geografia, física, história, língua estrangeira moderna e redação. Também foram comparados dois métodos estatísticos para verificar qual deles é o melhor. Além disso, pretende-se identificar se o que desestimula o aluno é a dificuldade/ interesse nas disciplinas.

1.2 HIPÓTESES

- O procedimento estatístico de Análise Discriminante é mais eficiente do que o de Regressão Logística
- As co-variáveis nota e freqüência das disciplinas cursadas no 1º semestre do curso são mais importantes na evasão que as demais.

2. OBJETIVOS

2.1OBJETIVO GERAL

O objetivo geral deste trabalho é verificar qual das metodologias estatísticas, neste caso a Análise Discriminante e a Regressão Logística têm o melhor desempenho na verificação dos fatores que influenciam na evasão dos alunos do Curso de Estatística da UFPR.

2.2OBJETIVO ESPECÍFICO

Identificar quais as co-variáveis são mais significativas na detecção de futuros alunos que ingressam no Curso de Estatística e que venham a desistir do curso após cursarem o primeiro período.

3. JUSTIFICATIVA:

O Curso de Estatística da UFPR historicamente tem um número elevado de alunos que não concluem o curso, muitos deles já desistem logo após completarem o 1° semestre, sendo assim este estudo visa tentar detectar as possíveis razões desta evasão, e que num futuro próximo seja possível trabalhar com os aspectos que mais influenciam na evasão para que este tipo de situação ocorra o menor numero de vezes. Na Estatística existem métodos de concepção diferentes para respostas dicotômicas. Escolheu-se fazer uma comparação entre o método de Análise Discriminante e o método de Regressão Logística.

4. MATERIAL E MÉTODOS

4.1 DADOS COLETADOS

Este estudo foi realizado com 166 alunos que ingressaram no Curso de Estatística da Universidade Federal do Paraná nos anos de 1998, 1999 e 2000, tendo como objetivo investigar quais as co-variáveis apresentam-se significativas para a variável resposta: evasão (54 observações) ou não evasão do curso (112 observações).

4.2 METODOLOGIA ESTATÍSTICA

4.2.1 Regressão Logística

A escolha da técnica estatística a ser utilizada segundo GIOLO (2004) deve ser levada em conta com relação à natureza da variável resposta, neste caso a variável resposta é dicotômica, ou seja, apresenta as categorias evasão do curso ou não evasão do curso, e ainda do ponto de vista matemático, fácil de ser usada e de interpretação bem simples. Sendo assim optou-se pela utilização da Regressão Logística, na tentativa de encontrar um modelo explicativo da variável resposta em função das variáveis explicativas.

A Regressão Logística parte da função de distribuição logística que é dada por:

$$F(x) = \frac{1}{1 + e^{-x}} = \frac{e^{x}}{1 + e^{x}}, \text{ para } x = -\infty, \ldots\ldots, +\infty \qquad (1)$$

A função de distribuição logística toma valores entre zero e um, assume valor zero em uma parte do domínio das variáveis explicativas, um em outra parte do domínio e cresce suavemente na parte intermediária possuindo uma particular curva em forma de "S".

O modelo de Regressão Logística é expresso por:

$$\theta(\underline{x}) = P(Y = 1 \mid \underline{x}) = \frac{\exp\{\beta_0 + \sum_{k=1}^{p} \beta_k x_k\}}{1 + \exp\{\beta_0 + \sum_{k=1}^{p} \beta_k x_k\}} \qquad (2)$$

Para descrever a variação entre os $\theta(\underline{x})=E(Y|\underline{x})$ foi, então, foi proposto a utilização do modelo acima citado, onde $Y_i = 1$ significa a presença da resposta, $\underline{x}$ é o vetor que representa as *p* co-variáveis (fatores de risco), isto é, $\underline{x}=(x_1, x_2,..., x_p)$. O parâmetro β_0 é o intercepto e β_k (k=1,..., p) são os p parâmetros da regressão. Nota-se que este modelo retornará uma estimativa da probabilidade do indivíduo ter a resposta dado que o mesmo possui, ou não, determinados fatores de risco. Conseqüentemente,

$$1 - \theta(\underline{x}) = \frac{1}{1 + \exp\{\beta_0 + \sum_{k=1}^{p} \beta_k x_k\}} \qquad (3)$$

retornará uma estimativa de probabilidade do indivíduo não ter a resposta dado que o mesmo possui ou não determinados fatores de risco.

Observe, ainda, que fazendo-se:

$$\text{Log}\left(\frac{\theta(x)}{1 - \theta(x)}\right) = \beta_0+\sum_{k=1}^{p} \beta_k x_k \qquad (4)$$

Tem-se um modelo linear para seus parâmetros, e dependendo da variação de $\underline{x}$, pode ser contínuo e variar de $-\infty$ a $+\infty$. A estimação dos parâmetros em Regressão Logística geralmente é feita pelo método da máxima verossimilhança. Para a aplicação, deste método é necessário construir inicialmente a função de verossimilhança a qual expressa à probabilidade dos dados observados como uma função dos parâmetros desconhecidos. Os estimadores de máxima verossimilhança dos parâmetros serão os valores que maximizam esta função.

Para encontrar esses valores no modelo de Regressão Logística, considera-se a variável resposta Y codificada como 0 ou 1. Da expressão (2) pode-se obter a probabilidade condicional de que Y=1 dado $\underline{x}$, Isto é, $\theta(\underline{x}) = P(Y=1|\underline{x})$ e, que $1 - \theta(x)$ fornece a probabilidade condicional de que Y=0 dado x. Assim, $\theta(\underline{x})$ será a contribuição para a função de verossimilhança dos pares (Y_i, x_i) em que $Y_i = 1$ e $1 - \theta(x_i)$, a contribuição dos pares em que $Y_i=0$.

Assumindo então que as observações são independentes tem-se a seguinte expressão:

$$L(\underline{\beta}) = \prod (\theta(x_i))^{Y_i} (1- \theta(x_i))^{1-y_i} \qquad (5)$$

As estimativas de $\underline{\beta}$ serão os valores que maximizam a função de verossimilhança dada em (5). Algebricamente é mais fácil trabalhar com o logaritmo desta função, isto é, com:

$$l(\underline{\beta})=\log L(\underline{\beta}) = +\sum_{i=1}^{n} y_i \log(\theta(x_i)) + (1 - y_i) \log(1 - \theta(x_i)) \quad (6)$$

Para obter os valores de $\underline{\beta}$ que maximizam $l(\underline{\beta})$ basta diferenciar a respectiva função com respeito a cada parâmetro β_j (j = 0, 1,..., p) obtendo-se, assim, o sistema de (p+1) equações,

$$\sum_{i=1}^{n}(y_i - \theta(x_i)) = 0$$

$$\sum_{i=1}^{n} x_{ij}(y_i - \theta(x_i)) = 0 \qquad j= 1,..., p$$

que, quando igualadas a zero, produzem como solução as estimativas de máxima verossimilhança de β. Os valores ajustados para o modelo de regressão logístico são, portanto, obtidos substituindo-se as estimativas de β em (2).

As p + 1 equações são chamadas equações de verossimilhança e por serem não-lineares nos parâmetros β_j (j = 0, 1,..., p), requerem métodos especiais para suas soluções. Os métodos iterativos de Newton-Raphson e o escore de Fischer são algorítmos numéricos comumente utilizados com essa finalidade.

O método de estimação de variâncias-covariâncias dos coeficientes estimados seguem da teoria da estimação de máxima verossimilhança a qual estabelece que os estimadores são obtidos pela matriz de derivadas parciais de segunda ordem do logaritmo da função de verossimilhança. Essa derivada tem a seguinte forma geral:

$$\frac{\delta^2 \log L(\underline{\beta})}{\delta \beta^2_j} = - \sum^{n} x_{ij}^2 \theta(x_i)(1-\theta(x_i)) \qquad \text{para } j,l=0,1,..,p. \quad (7)$$

$$\frac{\delta^2 \log L(\underline{\beta})}{\delta \beta_j \delta \beta_l} = -\sum^{n} x_{ij} x_{il} \theta(x_i)(1-\theta(x_i)) \qquad \text{para } j,l=0,1,..,p. \quad (8)$$

A matriz contendo o negativo dos termos dados nas equações (7) e (8) será denotada por $I(\beta)$ e é chamada matriz de informação. As variâncias e co-variâncias dos coeficientes estimados serão obtidas pela inversa da matriz, será denotada por $\sum(\beta)= I^{-1}(\beta)$. O j-ésimo elemento da diagonal dessa matriz, denotado por $\sigma^2(\beta_j)$, corresponde a variância de $\underline{\beta}_j$ e, o elemento na j-ésima linha e l-ésima coluna, dessa matriz, denotado por $\sigma(\beta j, \beta l)$, corresponde a co-variância entre $\underline{\beta}_j$ e, $\underline{\beta}_l$. Os estimadores das variâncias e co-variâncias, denotados por $\sum(\underline{\beta})$, são obtidos por avaliar $\sum(\beta)$ em $\underline{\beta}$.

Em notação matricial, a matriz de informação $I(\beta)=X'VX$ em que X é uma matriz com *n* linhas e p + 1 colunas contendo um vetor de uns e as co-variáveis dos indivíduos, e V é matriz diagonal de *n* linhas e *n* colunas com elementos $\theta(x)(1 - \theta(x))$ na diagonal.

4.2.1.1 Estatística Qui-quadrado

A estatística Qui-quadrado é utilizada para, considerando um determinado p-valor, verificar quais as variáveis são significativas no modelo, sendo assim os totais marginais n_{+1} e n_{+2} são fixos e, portanto, sob a hipótese nula H_0, de não existência de significância para o estudo, a distribuição de probabilidade associada é a hipergeométrica. Assim o valor esperado de nij é:

$$E(N_{ij} \mid H_0) = \frac{(n_{i+})(n_{+j})}{n} = m_{ij}$$

e a variância:

$$V(N_{ij} \mid H_0) = \frac{(n_{1+})\,(n_{2+})\,(n_{+1})\,(n_{+2})}{n^2(n - 1)} = v_{ij}.$$

Para uma amostra suficientemente grande, n_{11} tem aproximadamente uma distribuição normal, o que implica que:

$$Q = \frac{(n_{11} - m_{11})^2}{v_{11}}$$

tem aproximadamente uma distribuição qui-quadrado com um grau de liberdade. Não importa como as linhas e colunas sejam arranjadas, Q assumirá sempre o mesmo valor, uma vez que:

$$| n_{11} - m_{11}| = | n_{ij} - m_{ij} | = \frac{| n_{11}n_{22} - n_{12}n_{21}|}{n}$$

Uma estatística relacionada a Q é a estatística de *Pearson* dada por:

$$Q_P = \sum_{i=1}^{2} \sum_{j=1}^{2} \frac{(n_{ij} - m_{ij})^2}{m_{ij}} = \frac{n}{(n - 1)} Q.$$

Se as contagens (freqüências) nas caselas forem suficientemente grandes, Q_P segue uma distribuição qui-quadrado com um grau de liberdade. Ainda, quando n cresce, Q_P e Q convergem. Uma regra útil para determinar o tamanho amostral adequado para Q e Q_P é que o valor esperado m_{ij} seja maior do que 5 para todas as caselas.

4.2.1.2 Sensibilidade e Especificidade

Estas medidas determinam a eficiência do modelo selecionado detectar a verdade. A sensibilidade é definida como a proporção de resultados positivos que o estudo apresenta, quando realizado em sujeitos conhecidos terem a doença, ou seja, é a proporção de verdadeiros positivos. A especificidade, por outro lado, é definida como a proporção de resultados negativos que o estudo apresenta, quando realizado em sujeitos conhecidos estarem livres da doença (proporção de verdadeiro negativo). O desejado de um exame (teste) é que ele tenha, simultaneamente, alta sensibilidade e especificidade. Conforme tabela abaixo:

Tabela 1 - Quadro de medidas usadas para determinar eficiência de um teste

Medida estatística	Definição	Valor esperado
Sensibilidade	Proporção de resultados positivos	Alto
Especificidade	Proporção de resultados negativos	Alto

4.2.1.3 Poder Preditivo do Modelo

O poder preditivo do modelo pode também ser obtido com a finalidade de avaliar a qualidade do modelo ajustado. Para isso, faz-se necessário estabelecer uma probabilidade, denominada “ponto de corte", a partir da qual se estabeleça que:

- a variável resposta receba o valor 1, isto é, Y = 1 para probabilidades estimadas pelo modelo que sejam maiores ou iguais a esse ponto de corte e, ainda, que
- a variável resposta receba o valor 0, isto é, Y = 0 para probabilidades estimadas pelo modelo que sejam menores do que esse ponto de corte.

4.2.1.4 *Deviance* Residual e Resíduos de *Pearson*

As estatísticas Qui-quadrado de *Pearson* (Q_p) e *deviance* (Q_L), são usadas para verificar a qualidade de ajuste do modelo de Regressão Logística, fornece um único número o qual resume a concordância entre os valores observados e os ajustados. PREGIBON (1981) estendeu os métodos de diagnóstico de regressão linear para a Regressão Logística e argumenta que, como as estatísticas Q_p e Q_L são duas medidas usadas para verificar a qualidade do modelo ajustado, faz sentido analisar os componentes individuais dessas estatísticas, uma vez que estes componentes são funções dos valores observados e preditos pelo modelo.

Assim, se em uma tabela de contingência de dimensão s x 2, tem-se para cada uma das s linhas n_{i+} sujeitos dos quais n_{i1} apresentam a resposta de interesse (sucesso) e θ_{i1} denota a probabilidade predita de sucesso para a i-ésima linha (grupo), define-se o i-ésimo resíduo por:

$$c_i = \frac{n_{i1} - ((n_{i+})\,\theta_{i1})}{\sqrt{(n_{i+})\,\theta_{i1}\,(1-\theta_{i1})}} \qquad i = 1...s.$$

Esses resíduos são conhecidos como resíduos de *Pearson*, uma vez que a soma deles ao quadrado resulta em Q_P. O exame dos valores residuais c_i auxiliam a determinar quão bem o modelo se ajusta aos grupos individuais.

Freqüentemente, resíduos excedendo o valor |2,0| (ou |2,5|) indicam falta de ajuste. Similarmente, a *deviance* residual é um componente da estatística *deviance* e é expressa por:

d_i=sinal($n_{i1} - y_{i1}$)[2 n_{i1} log(n_{i1}/y_{i1})+2($n_{i+} - n_{i1}$) log(($n_{i+} - n_{i1}$)/($n_{i+} - y_{i1}$))]$^{1/2}$, onde y_{i1}=(n_i+) $\underline{\theta}_{i1}$.

A soma das *deviances* residuais ao quadrado resulta na estatística *deviance* Q_L. A partir do exame dos resíduos *deviance* pode-se observar a presença de resíduos não usuais (demasiadamente grandes), bem como a presença de *outliers* ou, ainda, padrões sistemáticos de variação indicando, possivelmente, a escolha de um modelo não muito adequado.

As estatísticas de diagnóstico apresentadas permitem, ao analista, identificar padrões de co-variáveis que não estão se ajustando bem ao modelo. Após estes padrões serem identificados, pode-se, então, avaliar a importância que eles têm na análise.

4.2.1.5 Gráfico *Q-Qplot* com Envelope Simulado

No caso em que a variável resposta é assumida ser normalmente distribuída, é comum que afastamentos sérios da distribuição normal sejam verificados por meio do gráfico de probabilidades normal dos resíduos. No contexto de modelos lineares generalizados, em que distribuições diferentes da normal são também consideradas, gráficos similares com envelopes simulados podem ser também construídos com os resíduos gerados a partir do modelo ajustado. A inclusão do envelope simulado no *Q-Qplot* auxilia a decidir se os pontos diferem significativamente de uma linha reta, (GIOLO, 2006), para gerar tais gráficos em: regressão gama, logística, Poisson e binomial negativa, além da normal. Para que o modelo ajustado seja considerado satisfatório, faz-se necessário que as *deviances* residuais caiam dentro do envelope simulado.

4.2.2 Reconhecimento de Padrões e Classificação

De acordo com JOHNSON & WICHERN (1988), análise discriminante é uma técnica multivariada interessada com a separação de uma coleção de objetos (observações) distintos e que alocam novos objetos em grupos previamente definidos. A Análise Discriminante quando empregada como procedimento de classificação não é uma técnica exploratória, uma vez que ela conduz a regras bem distribuídas, as quais podem ser utilizadas para classificação de novos objetos.

As técnicas estatísticas de discriminação e classificação estão incorporadas num contexto mais amplo, que é o do reconhecimento de padrões. Participa junto com técnicas de programação matemática e redes neurais na formação do conjunto de procedimentos usados no reconhecimento e classificação de objetos e indivíduos.

Os objetivos imediatos da técnica quando usada para discriminação e classificação são, respectivamente, os seguintes:

1. Descrever algebricamente ou graficamente as características diferenciais dos objetos (observações) de várias populações conhecidas, no sentido de achar "discriminantes" cujos valores numéricos sejam tais que as populações possam ser separadas tanto quanto possível.
2. Grupar os objetos (observações) dentro de duas ou mais classes determinadas. Tenta-se encontrar uma regra que possa ser usada na alocação ótima de um novo objeto (observação) nas classes consideradas.

Uma função que separa, pode servir como alocadora, e da mesma forma uma regra alocadora, pode sugerir um procedimento discriminatório. Na prática, os objetivos 1 e 2, freqüentemente, sobrepõem-se e a distinção entre separação e alocação torna-se confusa.

A terminologia de “discriminar” e “classificar” foi introduzida por FISCHER (1936) no primeiro tratamento moderno dos problemas de separação.

4.2.2.1 Problema Geral de Classificação

Para ilustrar a complexidade desse tipo de sistema, considere o seguinte exemplo citado por ASCENSO E FRED (2003): Uma indústria recebe dois tipos de peixe, salmão e robalo. Os peixes são recebidos em uma esteira, e o processo de classificação é manual. A indústria gostaria de automatizar esse processo, usando para isso uma câmera. Primeiramente devemos encontrar as características que distinguem um salmão de um robalo. Altura, largura, coloração, posição da boca, e etc... Dado as diferenças entre as populações de Salmão e Robalo, podemos dizer que cada uma possui um modelo específico. Com base em duas variáveis x_1 = altura e x_2 = claridade, obtem-se dois grupos n_1 = robalos e n_2 =salmão, assim teremos a representação gráfica a seguir:

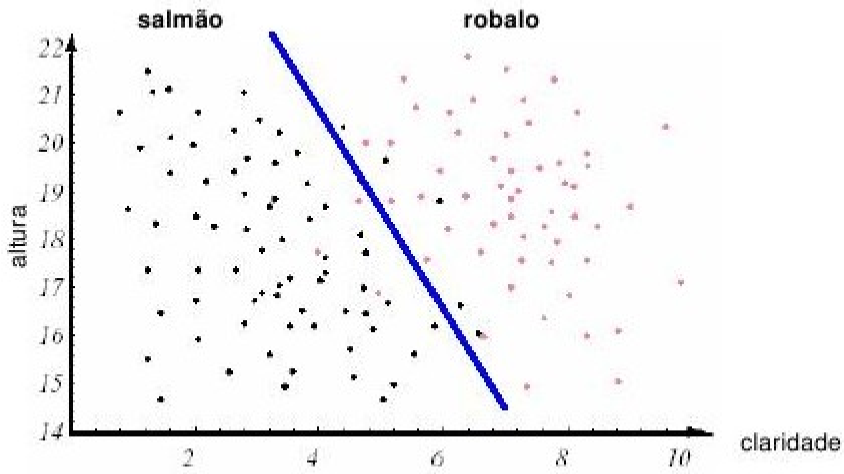

Figura 1 – Representação gráfica dos dados no espaço discriminante

Observa-se na figura 1 que:

1) robalos tendem a ser mais claros;
2) claridade parece discriminar melhor que altura
3) existem mistura entre grupos.

Dado que existe mistura e conseqüentemente classificações erradas, a idéia é criar uma regra (regiões R_1 e R_2) que minimize a chance de fazer esta mistura. Um bom procedimento resultará pouca mistura de elementos grupais. Pode ocorrer que de uma classe ou população exista maior probabilidade de ocorrência do que de outra classe. Uma regra de classificação ótima deve levar em conta as probabilidades de ocorrência a "priori". Outro aspecto da classificação é o custo. Suponha que classificar um item em Π_1 quando na verdade ele pertencente a Π_2 represente um erro mais sério do que classificar em Π_2 quando o item pertencente a Π_1. Então deve-se levar isso em conta.

Seja $f_1(\underline{x})$ e $f_2(\underline{x})$ as f.d.p.'s associadas com o vetor aleatório $\underline{X}$ de dimensão p das populações Π_1 e Π_2, respectivamente. Um objeto, com as medidas $\underline{x}$, deve ser reconhecido como de Π_1 ou de Π_2. Seja Ω o espaço amostral, isto é, o conjunto de todas as possíveis observações $\underline{x}$. Seja R_1 o conjunto de valores $\underline{x}$ para os quais nós classificamos o objeto como Π_1 e $R_2 = \Omega - R_1$ os remanescentes valores $\underline{x}$ para os quais nós classificaremos os objetos como Π_2. Os conjuntos R_1 e R_2 são mutuamente exclusivos.

Para p = 2, podemos ter a figura:

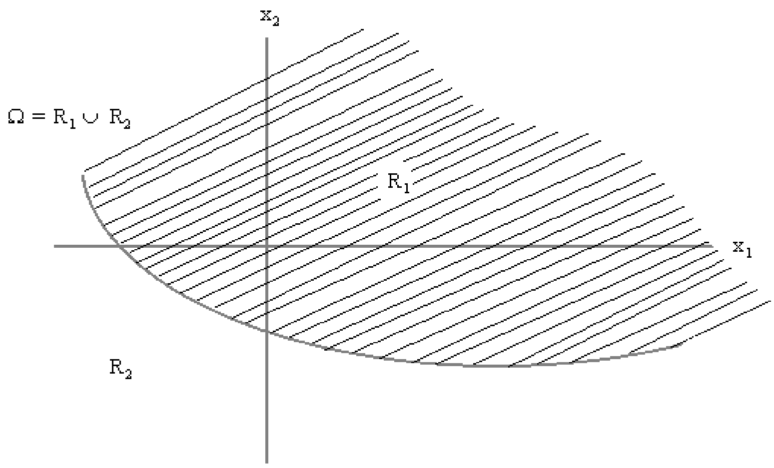

Figura 2: Regiões de classificação para duas populações

A probabilidade condicional, de reconhecer um objeto como de Π_2 quando na verdade ele é de Π_1 é:

$$P(2|1) = P(\underline{X} \in R_2 \,|\, \Pi_1) = \int_{R_2 = \Omega - R_1} f_1(\underline{x})d\underline{x}$$

Da mesma forma:

$$P(1|2) = P(\underline{X} \in R_1 \,|\, \Pi_2) = \int_{R_1} f_2(\underline{x})d\underline{x}$$

P(2|1) representa o volume formado pela f.d.p. $f_1(\underline{x})$ na região R_2.

Sendo p = 1 (caso uni-variado) tem-se:

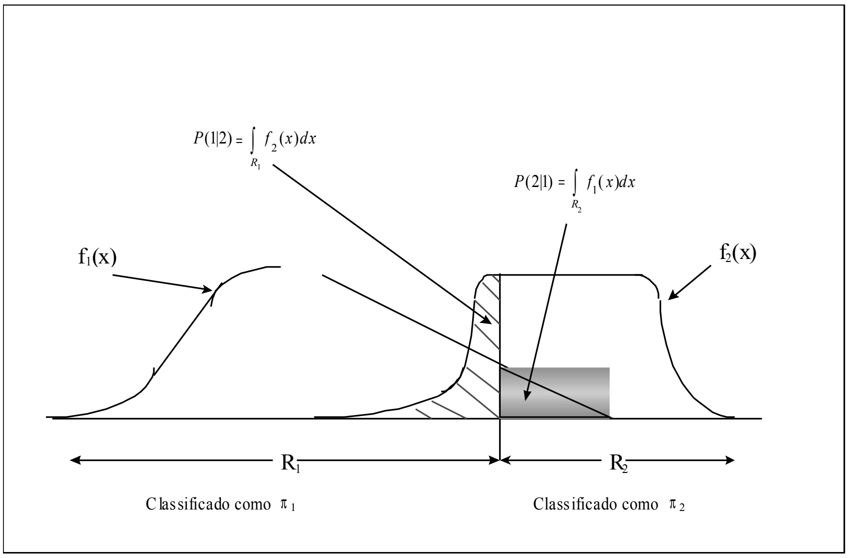

Figura 3: Classificação das Regiões para Duas Populações.

Seja p_1 a probabilidade a "priori" de Π_1 e p_2 ser a probabilidade a "priori" de Π_2, onde $p_1 + p_2 = 1$. As probabilidades de reconhecer corretamente ou incorretamente são dados por:

P(rec. correta/te. como Π_1) = P($\underline{X} \in \Pi_1$ e é rec. correta/te como Π_1) =

= P($\underline{X} \in R_1 | \Pi_1$)P($\Pi_1$) = P(1|1)$p_1$

P(rec. incorreta/te como Π_1) = P($\underline{X} \in \Pi_2$ e é rec. incorreta/te como Π_1) =

= P($\underline{X} \in R_1 | \Pi_2$)P($\Pi_2$) = P(1|2)$p_2$

P(rec. correta/te como Π_2) = P($\underline{X} \in \Pi_2$ e é rec. correta/te como Π_2) =

= P($\underline{X} \in R_2 | \Pi_2$)P($\Pi_2$) = P(2|2)$p_2$

P(rec. incorreta/te como Π_2) = P($\underline{X} \in \Pi_1$ e é rec. incorreta/te como Π_2) =

= P($\underline{X} \in R_2 | \Pi_1$)P($\Pi_1$) = P(2|1)$p_1$

Regras de reconhecimento são freqüentemente avaliados em termos de suas probabilidades de reconhecimento errado.

Tabela 2: Matriz do Custo de Reconhecimento Errado

		Reconhecimento como	
		Π_1	Π_2
População verdadeira	Π_1	0	c(2|1)
	Π_2	c(1|2)	0

Para qualquer regra, a média, ou o custo esperado de reconhecimento (classificação) errado é dado pela soma dos produtos dos elementos fora da diagonal principal pelas respectivas probabilidades:

ECM = c(2|1)P(2|1)p(1) + c(1|2)P(1|2)p(2)

Uma regra razoável de reconhecimento deve ter ECM muito baixa, tanto quanto possível.

As regiões R_1 e R_2 que minimizam o ECM são definidas pelos valores de $\underline{x}$ tal que valem as desigualdades:

$$R_1 = \frac{f_1(\underline{x})}{f_2(\underline{x})} \geq \left[\frac{c(1|2)}{c(2|1)}\right] \cdot \left[\frac{p_2}{p1}\right]$$

$$\begin{bmatrix}\text{Razão das}\\ \text{densidades}\end{bmatrix} \geq \begin{bmatrix}\text{Razão dos}\\ \text{custos}\end{bmatrix} \cdot \begin{bmatrix}\text{Razão das}\\ \text{probabilidades `a priori}\end{bmatrix}$$

$$R_2 = \frac{f_1(\underline{x})}{f_2(\underline{x})} < \left[\frac{c(1|2)}{c(2|1)}\right] \cdot \left[\frac{p_2}{p1}\right]$$

$$\begin{bmatrix}\text{Razão das}\\ \text{densidades}\end{bmatrix} < \begin{bmatrix}\text{Razão dos}\\ \text{custos}\end{bmatrix} \cdot \begin{bmatrix}\text{Razão das}\\ \text{probabilidades `a priori}\end{bmatrix}$$

4.2.2.2. Critério TPM

Outro critério, além do ECM, pode ser usado para construir procedimentos ótimos. Assim, pode-se ignorar o ECM e escolher R_1 e R_2 que minimizam a probabilidade total de erro de classificação (TPM).

TPM = P($\underline{x} \in \Pi_1$ e é classificada errada) + P($\underline{x} \in \Pi_2$ e é classificada errada)

$$TPM = p_1 \int_{R_2} f_1(\underline{x})\, d\,\underline{x} + p_2 \int_{R_1} f_2(\underline{x})\, d\,\underline{x}$$

Matematicamente, isto é equivalente a minimizar ECM quando os custos de classificação errada são iguais. Assim, podemos alocar uma nova observação $\underline{x}_0$ para a população com a maior probabilidade posteriori $P(\Pi_i|\underline{x}_0)$, onde:

$$P(\Pi_1 \mid \underline{x}_0) = \frac{P(\Pi_1 \text{ ocorre e observa-se } \underline{x}_0)}{P(\text{observa-se } \underline{x}_0)}$$

$$= \frac{P(\text{observa-se } \underline{x}_0) \mid \Pi_1) P(\Pi_1)}{P(\text{observa-se } \underline{x}_0) \mid \Pi_1) p(\Pi 1) + P(\text{observa-se } \underline{x}_0) \mid \Pi_2) p(\Pi_2)}$$

$$= \frac{p_1 f_1(\underline{x}_0)}{p_1 f_1(\underline{x}_0) + p_2 f_2(\underline{x}_0)}$$

e

$$P(\Pi_2 \mid \underline{x}_0) = 1 - P(\Pi_1 \mid \underline{x}_0) = \frac{p_2 f_2(\underline{x}_0)}{p_1 f_1(\underline{x}_0) + p_2 f_2(\underline{x}_0)}$$

e classifica-se $\underline{x}_0$ em Π_1 quando $P(\Pi_1 | \underline{x}_0) > P(\Pi_2 | \underline{x}_0)$

4.2.2.3. Classificação com Duas Populações Normais Multivariadas

Assume-se que $f_1(\underline{x})$ e $f_2(\underline{x})$ são densidades normais multivariadas, a primeira com $\underline{\mu}_1$ Σ_1 e a segunda com $\underline{\mu}_2$ Σ_2, então supondo $\Sigma_1 = \Sigma_2$, a Função Discriminante Linear (F.D.L.) de Fisher pode ser usada para classificação e corresponde a um caso particular da regra de classificação com base em ECM, conforme CHAVES (2005). Assim, Seja $\underline{X}'$ = [$X_1, X_2, ..., X_p$] para populações Π_1 e Π_2 e

$$f_i(\underline{x}): \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp[- ½ (\underline{x} - \underline{\mu}_i)' \Sigma^{-1} (\underline{x} - \underline{\mu}_i)] \quad \text{para } i = 1,2$$

Suponha que os parâmetros Π_1 e Π_2 e Σ, são conhecidos, tem-se as regiões de mínimo ECM.

$$R_1: \frac{f_1(\underline{x})}{f_2(\underline{x})} := \frac{(1/((2\pi)^{p/2}|\Sigma|^{1/2}))\exp[- ½ (\underline{x} - \underline{\mu}_1)' \Sigma^{-1} (\underline{x} - \underline{\mu}_1)]}{(1/((2\pi)^{p/2}|\Sigma|^{1/2}))\exp[- ½ (\underline{x} - \underline{\mu}_1)' \Sigma^{-1} (\underline{x} - \underline{\mu}_1)]} =$$

$$= \exp [- ½ (\underline{x} - \underline{\mu}_1)' \Sigma^{-1} (\underline{x} - \underline{\mu}_1) + ½ (\underline{x} - \underline{\mu}_2)' \Sigma^{-1} (\underline{x} - \underline{\mu}_2)] \geq \left(\frac{c\ (1\ 2)}{c\ (2\ 1)}\right)[p_2 / p_1]$$

$$R_2: \frac{f_1(\underline{x})}{f_2(\underline{x})}: = \exp [- ½ (\underline{x} - \underline{\mu}_1)' \Sigma^{-1} (\underline{x} - \underline{\mu}_1) + ½ (\underline{x} - \underline{\mu}_2)' \Sigma^{-1} (\underline{x} - \underline{\mu}_2)] < \left(\frac{c\ (1\ 2)}{c\ (2\ 1)}\right)[p_2 / p_1]$$

Sejam as populações Π_1 e Π_2 normais multivariadas. A regra de reconhecimento que minimiza ECM é dada por: reconhecer $\underline{x}_0$ como sendo de Π_1 se

$$(\underline{\mu}_1 - \underline{\mu}_2)' \Sigma^{-1} (\underline{x}_0 - ½ (\underline{\mu}_1 - \underline{\mu}_2)' \Sigma^{-1} (\underline{\mu}_1 + \underline{\mu}_2) \geq \ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right]$$

e $\underline{x}_0$ como sendo de Π_2 em caso contrário.

Em situações em que μ_1 i=1,2 são desconhecidas e Σ também, a regra deve ser modificada. Tem-se a seguinte regra do ECM mínimo para duas populações normais (regra amostral).

Alocar $\underline{x}_0$ em Π_1 se

$$(\bar{\underline{x}}_1 - \bar{\underline{x}}_2)'S_p^{-1}\,\underline{x}_0 - \tfrac{1}{2}(\bar{\underline{x}}_1 - \bar{\underline{x}}_2)'S_p^{-1}(\bar{\underline{x}}_1 + \bar{\underline{x}}_2) \geq \ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right]$$

Alocar $\underline{x}_0$ em Π_2 em caso contrário.

O primeiro termo da regra de classificação e reconhecimento, $(\bar{\underline{x}}_1 - \bar{\underline{x}}_2)'\,S_p^{-1}\,x$, é a função linear obtida por Fisher que maximiza a variabilidade univariada entre as amostras relativamente a variabilidade dentro das amostras. A expressão inteira

$$w = (\bar{\underline{x}}_1 - \bar{\underline{x}}_2)'S_p^{-1}\,\underline{x}_0 - \tfrac{1}{2}(\bar{\underline{x}}_1 - \bar{\underline{x}}_2)'S_p^{-1}(\bar{\underline{x}}_1 + \bar{\underline{x}}_2) = (\bar{\underline{x}}_1 - \bar{\underline{x}}_2)'S_p^{-1}[\underline{x} - \tfrac{1}{2}(\bar{\underline{x}}_1 + \bar{\underline{x}}_2)]$$

é conhecida como função de classificação de Anderson.

Quando $\Sigma_1 \neq \Sigma_2$, tem-se a classificação quadrática. Supondo as matrizes de co-variância Σ_1 para $\underline{x} \in \Pi_1$ e Σ_2 para $\underline{x} \in \Pi_2$ em $\Sigma_1 \neq \Sigma_2$, as regras de reconhecimento de padrões tornam-se mais complicadas.

Seja então $\underline{x} \sim N_p(\underline{\mu}_i, \Sigma_i)$ i=1,2 com $\underline{\mu}_1 \neq \underline{\mu}_2$ e $\Sigma_1 \neq \Sigma_2$. A probabilidade total de reconhecimento errada (TPM) e o custo esperado de reconhecimento errado dependem da razão de densidades:

$$\frac{f_1(\underline{x})}{f_2(\underline{x})}$$

ou, equivalentemente, do logaritmo das razões das densidades

$$\ln[f_1(\underline{x}) / f_2(\underline{x})] = \ln[f_1(\underline{x})] - \ln[f_2(\underline{x})]$$

Sejam as populações Π_1 e Π_2 descritas por densidades normais multivariadas $N_p(\underline{\mu}_1, \Sigma_1)$ e $N_p(\underline{\mu}_2, \Sigma_2)$. Então a regra de reconhecimento que minimiza o ECM é dada por:

$$R_1 = -\tfrac{1}{2}\,\underline{x}'_0(\Sigma_1^{-1} - \Sigma_2^{-1})\,\underline{x}_0 + (\underline{\mu}'_1\Sigma_1^{-1} - \underline{\mu}'_2\Sigma_2^{-1})\,\underline{x}_0 - k$$

Substituindo as expressões das densidades normais multivariadas tem-se:

$$R_1: \frac{f_1(\underline{x})}{f_2(\underline{x})} := \frac{(1/((2\pi)^{p/2}|\Sigma|^{1/2}))\exp[-\tfrac{1}{2}(\underline{x} - \underline{\mu}_1)'\ \Sigma^{-1}(\underline{x} - \underline{\mu}_1)]}{(1/((2\pi)^{p/2}|\Sigma|^{1/2}))\exp[-\tfrac{1}{2}(\underline{x} - \underline{\mu}_1)'\ \Sigma^{-1}(\underline{x} - \underline{\mu}_1)]} \geq \ln\left(\frac{c(1|2)}{c(2|1)}\right)[p_2/p_1]$$

Na prática, a regra de reconhecimento estabelecida é implementadas substituindo-se os, parâmetros $\underline{\mu}_1$, $\underline{\mu}_2$, Σ_1, e Σ_2, pelas suas estimativas $\underline{x}_1, \underline{x}_2$, S_1 e S_2 , tal que:

Alocamos $\underline{x}_0$ em Π_1 se:

$$- \tfrac{1}{2}\ \underline{x}'_0 (S_1^{-1} - S_2^{-1})\ \underline{x}_0 + (\underline{x}'_1 S_1^{-1} - \underline{x}'_2 S_2^{-1})\ \underline{x}_0 - k \geq \left(\ln \frac{c\ (1|2)}{c\ (2|1)} \right) [p_2 / p_1]$$

Alocamos $\underline{x}_0$ em Π_2, caso contrário

4.2.2.4 Discriminação e Classificação entre Duas Populações - Método de Fischer

Basicamente, o problema consiste em separar duas classes de objetos ou fixar um novo objeto em uma das duas classes. Deste modo, é interessante alguma exemplificação. A tabela 3 a seguir mostra diversas situações onde a Análise Discriminante pode ser empregada. É comum denominar as classes (populações) de Π_1 e Π_2, e os objetos separados ou classificados com base nas medidas de p variáveis aleatórias são associadas com vetores do tipo:

$$\underline{X}' = [X_1, X_2, \dots, X_P] ,$$

onde as variáveis X_i , i = 1 , 2 , ... , p, são as medidas das características investigadas nos objetos, conforme JOHNSON & WICHERN (1988).

Os valores observados de $\underline{X}$ podem diferir de uma classe para outra, sendo que a totalidade dos valores da 1ª classe é a população dos valores $\underline{x}$ para Π_1 e aqueles da 2ª classe é a população dos valores de $\underline{X}$ para Π_2. Assim, estas populações podem ser descritas pelas funções densidade de probabilidade f_1 ($\underline{x}$) e f_2 ($\underline{x}$).

Tabela 3: Situações - Exemplos

Problema	entradas	saidas
Reconhecimento de Voz	sinais de voz	Palavra, identidade do locutor
Testes não invasivos/destrutivos	Ultra-sons, emissão de ondas acústicas	Presença / ausência de anomalia
Detecção/Doagnósticosde doenças	ECG, EEG, ultra-sons	Tipos de condições cardíacas, classe de estados cerebrais, patologias
Identificação de Recursos Naturais	imagens multi-espectrais	Formas de terrenos, vegetação
Reconhecimento Aéreo	Infravermelhos, imagens de Radar	Tanques, campos de cultivo, estradas, tráfego
Robótica	Imagens de interiores e exteriores em 3D, luz estruturada, laser, imagem estéreo	Identificação de objetos, tarefas industriais

A idéia de Fischer foi transformar as observações multivariadas $\underline{X}$'s nas observações univariadas y's tal que os y's das populações Π_1 e Π_2 sejam separadas tanto quanto possível. Fischer teve a idéia de tomar combinações lineares de $\underline{X}$ para criar os y's, dado que as combinações lineares são funções de $\underline{X}$ e por outro lado são de fácil cálculo matemático.

Seja μ_{1y} a média dos y's obtidos dos $\underline{X}$'s pertencentes a Π_1 e μ_{2y} a média dos y's obtidos dos $\underline{X}$'s pertencentes a Π_2, então Fischer selecionou a combinação linear que maximiza a distância quadrática entre μ_{1y} e μ_{2y} relativamente à variabilidade dos y's. Assim, seja:

$\underline{\mu}_1 = E(\underline{X}|\Pi_1)$ = valor esperado de uma observação multivariada de Π_1.

$\underline{\mu}_2 = E(\underline{X}|\Pi_2)$ = valor esperado de uma observação multivariada de Π_2.

e supondo a matriz de co-variância

$$\Sigma = E(\underline{X} - \underline{\mu}_i)(\underline{X} - \underline{\mu}_i)' \quad i = 1, 2$$

como sendo a mesma para ambas as populações, e considerando a Combinação Linear.

$$\underset{1x1}{Y} = \underset{lxp}{\underline{c}'}\ \underset{px1}{\underline{X}}$$

tem-se

$$\mu_{1y} = E(Y|\Pi_1) = E(\underline{c}'\underline{X}|\Pi_1) = \underline{c}'E(\underline{X}|\Pi_1) = \underline{c}'\underline{\mu}_1 \quad ,$$

$$\mu_{2y} = E(Y|\Pi_2) = E(\underline{c}'\underline{X}|\Pi_2) = \underline{c}'E(\underline{X}|\Pi_2) = \underline{c}'\underline{\mu}_2 \quad ,$$

e

$$V(Y) = \sigma_y 2 = V(\underline{c}'\underline{X}) = \underline{c}'V(\underline{X})\,\underline{c} = \underline{c}'\Sigma\,\underline{c} \quad ,$$

que são a mesma para ambas as populações. Segundo Fischer, a melhor combinação linear é a derivada da razão entre o "quadrado da distância entre as médias" e a "variância de Y".

$$\frac{(\mu_{1y} - \mu_{2y})^2}{\sigma_y^2} = \frac{(\underline{c}'\underline{\mu}_1 - \underline{c}'\underline{\mu}_2)^2}{\underline{c}'\Sigma\,\underline{c}} = \frac{\underline{c}'(\underline{\mu}_1 - \underline{\mu}_2)(\underline{\mu}_1 - \underline{\mu}_2)'\underline{c}}{\underline{c}'\Sigma\,\underline{c}} = \frac{(\underline{c}'\,\underline{\delta})^2}{\underline{c}'\Sigma\underline{c}}$$

onde $\underline{\delta} = \underline{\mu}_1 - \underline{\mu}_2$.

Seja $\underline{\delta} = \underline{\mu}_1 - \underline{\mu}_2$ e $Y' = \underline{c}'\underline{X}$, então $\dfrac{(\underline{c}'\underline{\delta})^2}{\underline{c}'\Sigma\underline{c}}$ é maximizada por:

$\underline{c} = k\,\Sigma^{-1}\,\underline{\delta} = k\Sigma^{1}(\underline{\mu}_1 - \underline{\mu}_2)$ para qualquer $k \neq 0$.

Escolhendo k = 1 tem-se:

$\underline{c} = \Sigma^{-1}(\underline{\mu}_1 - \underline{\mu}_2)$ e $Y = \underline{c}'\underline{X} = (\underline{\mu}_1 - \underline{\mu}_2)'\Sigma^{-1}\,\underline{X}$,

que é conhecida como ***função discriminante linear de Fischer***.

A função discriminante linear de Fischer transforma as populações multivariadas Π_1 e Π_2 em populações univariadas, tais que as médias das populações univariadas correspondentes sejam separadas tanto quanto possível relativamente a variância populacional, considerada comum.

Assim tomando-se

$$y_0 = (\underline{\mu}_1 - \underline{\mu}_2)'\Sigma^{-1}\, \underline{x}_0$$

como o valor da função discriminante de Fischer para uma nova observação $\underline{x}_0$, e considerando o ponto médio entre as médias das duas populações univariadas,

$$m = ½ \,(\underset{y}{\underline{\mu}_1} - \underset{y}{\underline{\mu}_2})$$

como

$$m = ½ \,(\underline{c}_1' \underline{\mu}_1 + \underline{c}_2' \underline{\mu}_2)$$

$$m = ½ \,[(\underline{\mu}_1 - \underline{\mu}_2)' \,\Sigma^{-1} \underline{\mu}_1 + (\underline{\mu}_1 - \underline{\mu}_2)' \,\Sigma^{-1} \underline{\mu}_2]$$

$$m = ½ \,[\,(\underline{\mu}_1 - \underline{\mu}_2)' \,\Sigma^{-1} (\underline{\mu}_1 + \underline{\mu}_2)\,],$$

tem-se que:

$$E(Y_0|\Pi_1) - m \geq 0$$

$$E(Y_0|\Pi_2) - m < 0\,,$$

ou seja, se $\underline{X}_0$ pertence a Π_1 , se espera que Y_0 seja igual ou maior do que o ponto médio. Por outro lado se $\underline{X}_0$ pertence a Π_2, o valor esperado de Y_0 será menor que o ponto médio. Desta forma a regra de classificação é :

- alocar $\underline{x}_0$ em Π_1 se $y_0 - m \geq 0$
- alocar $\underline{x}_0$ em Π_2 se $y_0 - m < 0$

Geralmente, os parâmetros $\underline{\mu}_1$, $\underline{\mu}_2$ e Σ são desconhecidos, então supondo que se tenha n_1 observações da v.a. multivariada X_1 da população Π_1 e n_2 observações da v.a. multivariada X_2 da população Π_2, então os resultados amostrais para aquelas quantidades são:

$$\bar{\underline{x}}_1 = \frac{1}{n_1} \sum_{i=1}^{n1} \underline{x}_{i1} \;;\; S_1 = \frac{1}{n_1 - 1} \sum_{i=1}^{n1} (\underline{x}_{i1} - \bar{\underline{x}}_1)(\underline{x}_{i1} - \bar{\underline{x}}_1)'$$

$$\bar{\underline{x}}_2 = \frac{1}{n_2} \sum_{i=1}^{n2} \underline{x}_{i2} \;;\; S_2 = \frac{1}{n_2 - 1} \sum_{i=1}^{n2} (\underline{x}_{i2} - \bar{\underline{x}}_2)(\underline{x}_{i2} - \bar{\underline{x}}_2)'$$

mas uma vez que se assuma que as populações sejam assemelhadas é natural considerar a variância como a mesma, daí estima-se a matriz de covariância comum Σ por:

$$S_p = \frac{(n_1 - 1)\, S_1 + (n_2 - 1)\, S_2}{(n_1 + n_2 - 2)}$$

que é um estimador não-viciado daquele parâmetro.

conseqüentemente, a função discriminante linear de Fischer amostral é dada por:

$$y = \underline{c}'\, x = (\bar{x}_1 - \bar{x}_2)'\, S_p^{-1}\, \underline{x}$$

a estimativa do ponto médio entre as duas médias amostrais univariadas

$$\bar{y}_1 = \underline{c}'\, \bar{x}_1 \quad \text{e} \quad \bar{y}_2 = \underline{c}'\, \bar{x}_2$$

é dada por:

$$m = ½\, (\bar{y}_1 + \bar{y}_2) = ½\, [(\bar{\underline{x}}_1 - \bar{\underline{x}}_2)'\, S_p^{-1} \bar{\underline{x}}_1 + (\bar{\underline{x}}_1 - \bar{\underline{x}}_2)'\, S_p^{-1} \bar{\underline{x}}_2]$$

$$m = ½\, (\bar{\underline{x}}_1 - \bar{\underline{x}}_2)'\, S_p^{-1}\, (\bar{\underline{x}}_1 + \bar{\underline{x}}_2)$$

e finalmente a regra de classificação é a seguinte:

- alocar $\underline{x}_0$ em Π_1 se $y_0 = (\bar{\underline{x}}_1 - \bar{\underline{x}}_2)'\, S_p^{-1}\, \underline{x}_0 \geq m$
- alocar $\underline{x}_0$ em Π_2 se $y_0 = (\bar{\underline{x}}_1 - \bar{\underline{x}}_2)'\, S_p^{-1}\, \underline{x}_0 < m$

ou melhor se:

$y_0 - m \geq 0$ $\quad$ $\underline{x}_0$ é alocado em Π_1

$y_0 - m < 0$ $\quad$ $\underline{x}_0$ é alocado em Π_2

A combinação linear particular $=y = \underline{c}'\, x = (\bar{x}_1 - \bar{x}_2)'\, S_p^{-1}\, \underline{x}$ maximiza a razão:

$$\frac{(\bar{y}_1 - \bar{y}_2)^2}{S_y^2} = \frac{(\hat{\underset{\sim}{c}}_1 \bar{\underset{\sim}{x}}_1 - \hat{\underset{\sim}{c}}_2 \bar{\underset{\sim}{x}}_2)^2}{\hat{\underset{\sim}{c}}'\, S_p\, \hat{\underset{\sim}{c}}} = \frac{(\hat{\underset{\sim}{c}}'\, \underset{\sim}{d})^2}{\hat{\underset{\sim}{c}}'\, S_p\, \hat{\underset{\sim}{c}}}$$

onde

$$\underline{d} = \bar{\underline{x}}_1 - \bar{\underline{x}}_2$$

e

$$S_y^2 = \frac{\sum_{i=1}^{n_1} (y_{i1} - \bar{y}_1)^2 + \sum^{n_2} (y_{i2} - \bar{y}_2)^2}{n_1 - n_2 - 2}$$

4.3 ESTATÍSTICA *KAPPA*

Esta estatística é utilizada para comparar performances dos métodos utilizados em um mesmo estudo, é tomado como base nos cálculos as tabelas de contingências obtidas através da classificação de cada método (MULLER , 1997). A estatística *Kappa* (K) é calculada utilizando as seguintes proporções:

P_o = Proporção de concordância (Obtida na diagonal)

P_e = Proporção esperada de concordância (obtida no total das marginais)

Então *Kappa* é dado por:

$$K = \frac{P_o - P_e}{1 - P_e}$$

onde:

$$P_o = \frac{\sum_{i=1}^{m} x_{ii}}{N}$$

e

$$P_o = \frac{\sum_{i=1}^{m} x_{i+} * x_{+i}}{N^2}$$

onde:

m = número de linhas na tabela de contingência.

x_{ii} = número de observações na linha *i* e coluna *i*.

x_{i+} = total de observações na linha *i.*

x_{+i} = total de observações na coluna *i*.

N = número total de observações.

A estatística *Kappa* incorpora os elementos que não estão na diagonal principal da matriz com um produto de linhas e colunas, enquanto que outras estatísticas só utilizam os dados classificados corretamente.

Uma das principais vantagens da estatística *Kappa* é a habilidade para usar este valor como uma base para determinar a significância estatística de alguma tabela dada ou a diferença entre tabelas. O teste é baseado em uma variância estimada de *Kappa* e usando um Teste Z para determinar se há diferença significativa entre as tabelas.

Então uma aproximação da variância do coeficiente verificada assintóticamente quando N é grande (≥ 30), é dada por :

$$\sigma_k^2 = \frac{P_o\,(1 - P_o)}{N(1 - P_e)^2}$$

Sabendo que a diferença entre duas estatísticas *Kappa* quando N é suficientemente grande, tem uma distribuição aproximadamente normal, o teste de significância da diferença entre dois coeficientes k independentes pode ser feito por:

$$Z = \frac{k_i - k_j}{\sqrt{(\sigma_{ki}^2 + \sigma_{kj}^2)}} \qquad \text{com } i \neq j \text{ e } i \text{ e } j = 1,2.$$

onde:

k_i = coeficiente *Kappa* para o primeiro método comparado.

k_j = coeficiente *Kappa* para o segundo método comparado.

Σ_{ki}^2 = variância amostral do coeficiente *Kappa* para o primeiro método comparado.

Σ_{kj}^2 = variância amostral do coeficiente *Kappa* para o primeiro método comparado.

Sendo que a hipótese do teste é:

H_o = Não há diferença entre as performances dos dois métodos

H_1 = Há diferença entre as performances dos dois métodos.

4.4 RECURSOS COMPUTACIONAIS

A análise dos dados foi feita através dos softwares R e Statgraphics Centurion, sendo que foram utilizadas: Regressão Logística, estatística qui-quadrado, *deviance* residual, resíduos de *Pearson*, *Q-Qplot* com envelope simulado, sensibilidade, especificidade, valor preditivo positivo e negativo, falsos negativos e falsos positivos, poder preditivo do modelo, Estatística *Kappa* e Análise Discriminante.

5. RESULTADOS E DISCUSSÕES

5.1 – ANÁLISE DE REGRESSÃO LOGÍSTICA

Inicialmente foi realizada uma análise descritiva dos dados, dos 166 alunos acompanhados no estudo 54 alunos desistiram do curso e os outros 112 concluíram o curso. Para analisar quais variáveis são mais relevantes neste estudo realizou-se um teste Qui-quadrado de *Pearson*. Considerando um p-valor limite de 0,20, verificam-se as variáveis significativas, ou seja, rejeita-se a hipótese nula.

Tabela 4: Teste Qui-quadrado entre as variáveis explicativas e a variável resposta

		Ocorrência de evasão escolar					
		Sim		Não		Q	Qp
Variável	Categoria	Absoluto	Relativo	Absoluto	Relativo	* p	
	Masculino	37	34,58%	70	65,42%		
Gênero	Feminino	17	28,81%	42	71,19%	0,4492	0,4479
	Solteiro	30	26,09%	85	73,91%		
Estado Civil	Casado	16	45,71%	19	54,29%		
	Outros	8	50,00%	8	50,00%	0,0283	0,0272
Classificação	<=33	28	34,57%	53	65,43%		
Vestibular	>33	26	30,59%	59	69,41%	0,5843	0,5855
Escore	<=320	14	27,45%	37	72,55%		
	320< <540	34	37,36%	57	62,64%		
Português	>= 540	6	25,00%	18	75,00%	0,3371	0,3349
Escore	<=320	27	38,03%	44	61,97%		
	320< <540	13	27,08%	35	72,92%		
Matemática	>= 540	14	29,79%	33	70,21%	0,4112	0,4091
Escore	<=320	13	48,15%	14	51,85%		
	320< <540	27	30,00%	63	70,00%		
Biologia	>= 540	14	28,57%	35	71,43%	0,1660	0,1642
Escore	<=320	28	33,73%	55	66,27%		
	320< <540	19	28,36%	48	71,64%		
Química	>= 540	7	43,75%	9	56,25%	0,4736	0,4714
Escore	<=320	10	34,48%	19	65,52%		
	320< <540	25	28,74%	62	71,26%		
Geografia	>= 540	19	38,00%	31	62,00%	0,5234	0,5213
Escore	<=300	27	31,40%	59	68,60%		
	300< <480	22	37,93%	36	62,07%		
Física	>= 480	5	22,73%	17	77,27%	0,4119	0,4097

* p: nível descritivo de associação obtido pelo teste Qui-quadrado.
Fonte: Dados analisados no software R

Tabela 4: Teste Qui-quadrado entre as variáveis explicativas e a variável resposta

Variável	Categoria	Ocorrência de evasão escolar – Sim – Absoluto	Sim – Relativo	Não – Absoluto	Não – Relativo	Q * p	Qp
Escore	<=320	18	26,87%	49	73,13%		
	320< <540	29	34,94%	54	65,06%		
História	>= 540	7	43,75%	9	56,25%	0,3492	0,3470
Escore	<=320	17	33,33%	34	66,67%		
Língua Estrang.	320< <540	25	37,31%	42	62,69%		
Moderna	>= 540	12	25,00%	36	75,00%	0,3788	0,3765
Escore	<=270	8	23,53%	26	76,47%		
	270< <430	39	34,82%	73	65,18%		
Redação	>= 430	7	35,00%	13	65,00%	0,4564	0,4542
Freqüência	<45	26	76,47%	8	23,53%		
Estatística Geral I	>=45	28	21,21%	104	78,79%	0,0004	0,0004
Nota	<50	41	61,19%	26	38,81%		
Estatística Geral I	>=50	13	13,13%	86	86,87%	0,0000	0,0001
Freqüência	<45	26	76,47%	8	23,53%		
Probabilidade I	>=45	28	21,21%	104	78,79%	0,0004	0,0004
Nota	<50	47	52,81%	42	47,19%		
Probabilidade I	>=50	7	9,09%	70	90,91%	0,0003	0,0002
Freqüência	<45	35	66,04%	18	33,96%		
Cálculo I	>=45	19	16,81%	94	83,19%	0,0001	0,0001
Nota	<50	9	11,39%	70	88,61%		
Cálculo I	>=50	45	51,72%	42	48,28%	0,0011	0,0010
Freqüência	<45	42	45,65%	50	54,35%		
Lógica	>=45	12	16,22%	62	83,78%	0,0406	0,0386
Nota	<50	49	37,12%	83	62,88%		
Lógica	>=50	5	14,71%	29	85,29%	0,0131	0,0129
Freqüência	<45	30	78,95%	8	21,05%		
Lab. Informática	>=45	24	18,75%	104	81,25%	0,0000	0,0000
Nota	<50	33	70,21%	14	29,79%		
Lab. Informática	>=50	21	17,65%	98	82,35%	0,0001	0,0001
Total		54	32,53%	112	67,47%		

* p: nível descritivo de associação obtido pelo teste Qui-quadrado.
Fonte: Dados analisados no software R

A partir da tabela 4, nota-se que as estatísticas Q e Qp são significativas para as co-variáveis: Estado Civil, Escore Biologia, Freqüência e nota em Estatística Geral I, Freqüência e nota em Probabilidade I, Freqüência e nota em Calculo I, Freqüência e nota em Lógica, Freqüência e nota em Laboratório de Informática, logo foram selecionadas para dar início à modelagem.

5.1.2 – Ajuste do Modelo de Regressão Logístico

Utilizando-se o método stepwise foram gerados vários modelos, dos quais 3 são apresentados conforme tabela a seguir:

Tabela 5 - modelos ajustados

Modelo	AIC
Evasão ~ Nota Prob. I + Freq. Calculo I + Escore Biologia + Nota Lab. Inf. + Freq. Lab. Inf.+ Freq. Estat. Geral I	4,27
Evasão ~ Nota Prob. I + Freq. Calculo I + Escore Biologia + Nota Lab. Inf. + Freq. Lab. Inf.	6,58
Evasão ~ Nota Prob. I + Freq. Calculo I + Escore Biologia + Nota Lab. Inf.	6,74

Fonte: Dados analisados no software R

Para a escolha do melhor modelo foi utilizado entre outros o critério do menor AIC, sendo assim o modelo selecionado foi o que possui as seguintes co-variáveis: Escore Biologia, Freqüência em Estatística Geral I, Nota em Probabilidade I, Freqüência em Cálculo I, Freqüência e nota em Laboratório de Informática. As estimativas dos parâmetros estão descritos na tabela a seguir:

Tabela 6 - Estimativa dos parâmetros

Parâmetros	Coeficientes	p-valor
Intercepto	-0,16372	0,00000
Nota em Probabilidade I	0,00436	0,00306
Freqüência em Cálculo I	0,00456	0,06433
Escore em Biologia	0,00050	0,00762
Nota em Laboratório de informática	0,00458	0,00310
Freqüência em Laboratório de Informática	0,00886	0,01845
Freqüência em Estatística Geral I	0,00727	0,03770

Fonte: Dados analisados no software R

Na tabela 6 é testado se o ajuste da regressão foi satisfatório ou não. A hipótese nula refere-se a uma não satisfação do modelo de regressão logístico ajustado, ou seja, H_0: $\beta_k = 0$. Na análise de *deviance* da Tabela 6, percebe-se que a regressão para o modelo escolhido foi significativa com um p-valor muito baixo, portanto rejeita-se a hipótese nula e fica válido o modelo de regressão logístico e fica ajustado como abaixo:

$$P(Y=1/X=x) = \theta(x) = \frac{e^y}{1+e^y} \quad (6)$$

Onde :

$y=-0,164+0,0044*probnt+0,005*calcfreq+0,0005*Biob+0,005*infnt+0,009*inffreq+0,007*estgerfreq$

5.1.2.1. Qualidade do Modelo Ajustado

Para verificar a qualidade do modelo ajustado, foi realizada a análise de resíduos, onde foi utilizada a análise de resíduos de *Pearson*, *deviance* residual e o gráfico *Q-Qplot* dos resíduos com envelope simulado, como mostra as figuras 4, 5 e 6 respectivamente.

Na figura 4 e 5, espera-se que os pontos fiquem dentro do intervalo -2,5 até 2,5, para que se possa concluir que o modelo é satisfatório. Como constatado nestes dois gráficos são poucos os pontos que estão fora deste intervalo, portanto apresenta um bom ajuste pela análise de resíduos de *Pearson* e pela *deviance* residual.

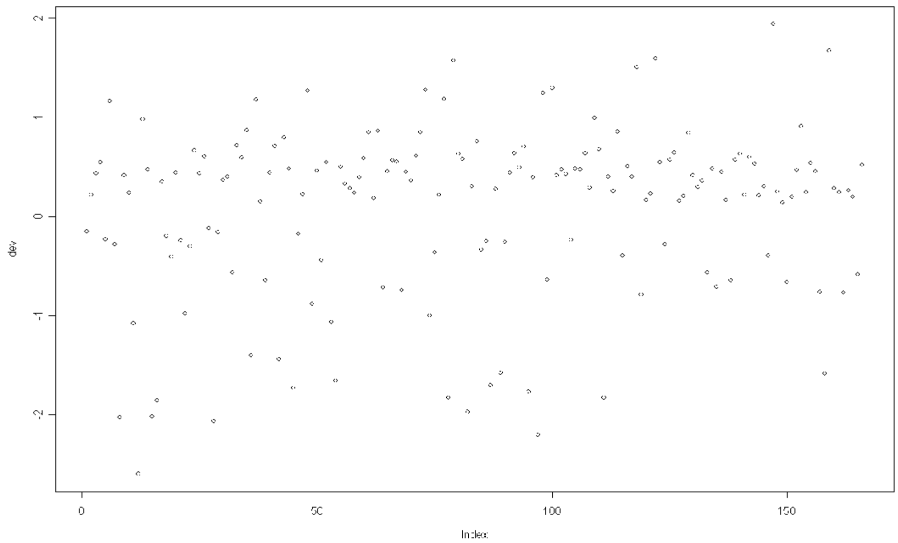

Figura 4: Gráfico da *Deviance* Residual

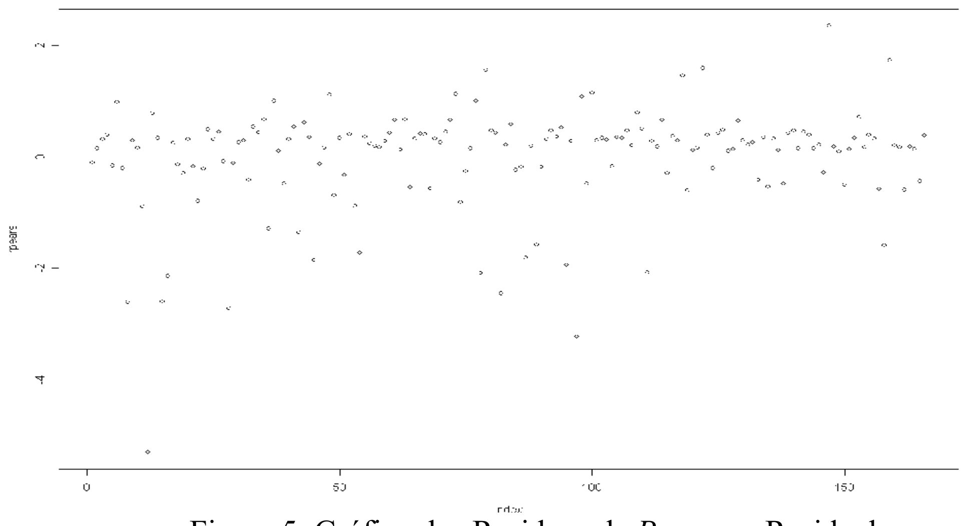

Figura 5: Gráfico dos Residuos de *Pearsone* Residual

Com relação à figura 6, os pontos devem apresentar-se dentro do envelope, o que se verifica na análise, portanto apresenta-se satisfatório.

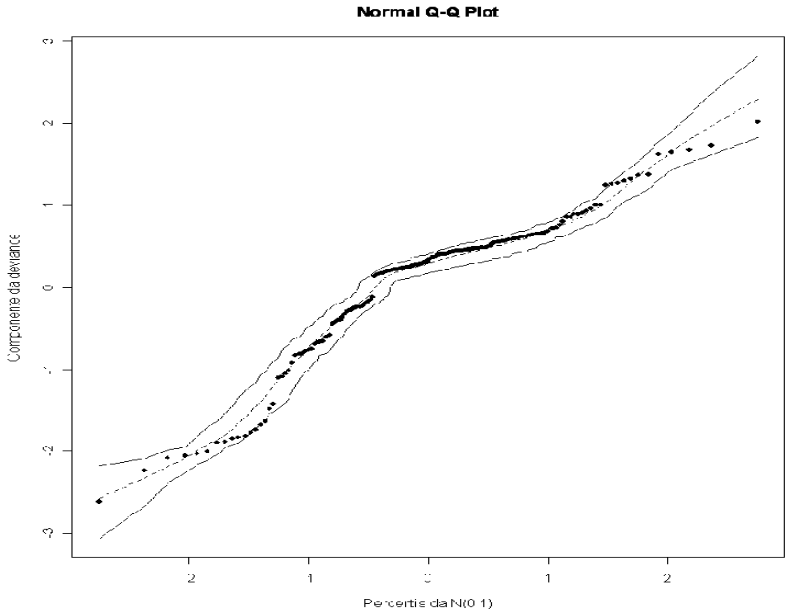

Figura 6: Gráfico *Q-Qplot* com Envelope Simulado

Assim, pelos três gráficos apresentados, conclui-se que o modelo ajustado é satisfatório.

5.1.2.2. Poder Preditivo do Modelo

A tabela 7 apresenta a classificação apresentada pelo modelo de Regressão Logística, onde 0 (zero) representa a ocorrência de evasão e 1 (um) a não ocorrência de evasão, com um ponto de corte de 0,5.

Tabela 7 - Classificação pelo modelo

Observado	Classificado pelo modelo 0 (-)	Classificado pelo modelo 1 (+)	Total
0 (-)	37	17	54
1 (+)	10	102	112
Total	47	119	166

Fonte: Dados analisados no software R

Na tabela 8 apresentam-se os cálculos de sensibilidade, especificidade, valor preditivo do modelo, taxa de falsos negativos e falsos positivos e valores preditivos positivo e negativo para o modelo ajustado.

Tabela 8 - Poder preditivo do modelo

Especificidade	68,52%
Sensibilidade	91,07%
Preditivo negativo	78,72%
Preditivo positivo	85,71%
Falso positivo	8,93%
Falso negativo	31,48%
Preditivo do modelo	83,73%

Fonte: Dados analisados no software R

O valor preditivo do modelo foi de 83,73% de chance de classificar corretamente um individuo corretamente, levando em conta a ocorrência de ocorrer ou não a evasão do aluno, o que quer dizer um bom resultado para o modelo ajustado. Com relação à sensibilidade e especificidade os valores são satisfatórios, ou seja, o modelo está classificando bem os alunos com relação a variável resposta de ocorrer a evasão do aluno do Curso de Estatística.

5.2 ANÁLISE DISCRIMINANTE

5.2.1 Analise Discritiva

A Análise Discriminante é um procedimento utilizado para ajudar a distinguir entre dois ou mais grupos de dados, com base em um conjunto p de variáveis quantitativas observadas. Isso é feito construindo funções discriminantes que são combinações lineares das variáveis. O objetivo de tal análise geralmente é um ou ambos os seguintes:

- Descrever matematicamente casos observados de forma que os separe em grupos, tão bem quanto possível.
- Classificar novas observações como pertencentes a um ou outro grupo

Foi utilizado o procedimento de Análise Discriminante para descrever a relação entre a resposta e as variáveis explicativas. A variável resposta utilizada foi situação do aluno no curso, definida como 0 ou 1. A probabilidade de pertencer a cada um dos grupos (evasão ou não evasão foi considerada a priori igual a 50%).

As variáveis utilizadas nessa parte da análise foram àquelas significativas pelo modelo de Regressão Logística considerado anteriormente, ou seja, Escore em biologia, freqüência em Estatística Geral 1, freqüência em Cálculo 1, freqüência em Laboratório de Informática e notas em Probabilidades 1 e Laboratório de Informática.

5.2.2 Poder Preditivo do Modelo

A tabela 9 mostra a quantidade de itens que foram classificados pelo modelo de Análise Discriminante, onde 0 (zero) representa a ocorrência de evasão e 1 (um) a não ocorrência de evasão:

Tabela 9 - Classificação pelo modelo

Observado	Classificado pelo modelo 0 (-)	Classificado pelo modelo 1 (+)	Total
0 (-)	37	17	54
1 (+)	12	100	112
Total	49	117	166

Fonte: Dados analisados no software Statgraphics

Na tabela a seguir são apresentados os cálculos de sensibilidade, especificidade, valor preditivo do modelo, taxa de falsos negativos e falsos positivos e valores preditivos positivo e negativo para a análise em questão.

Tabela 10 - Poder preditivo do modelo

Especifidade	68,52%
Sensibilidade	89,29%
Preditivo negativo	75,51%
Preditivo positivo	85,47%
Falso positivo	10,71%
Falso negativo	31,48%
Preditivo do modelo	82,53%

Fonte: Dados analisados no software Statgraphics

Pelos resultados acima percebemos que o modelo de Análise Discriminante é satisfatório, pois se tem valores altos para sensibilidade (89,29%), especificidade (68,52%) e valor preditivo do modelo (82,53%), sendo assim um modelo confiável para a questão em estudo.

A tabela de classificação completa onde são apresentados os resultados obtidos para os indivíduos do estudo encontra-se no apêndice 1, tal tabela mostra os dois grupos que receberam os maiores escores para os casos selecionados. Ela mostra:

- *Primeiro e segundo grupo mais prováveis* – os dois grupos com maiores escores
- *Valores* – os valores dos escores calculados para os dois grupos
- *Distância quadrada* – a distância quadrada de Mahalanobis das observações dos centróides dos grupos, no espaço das funções discriminantes. Quanto mais afastada a observação está do centróide do grupo, menos provável que pertença a esse grupo
- *Probabilidade* – a probabilidade estimada de que o caso pertença a um grupo particu-

lar. A probabilidade é baseada na razão da altura da função de densidade normal e a distância das observações do centróide de cada grupo e as probabilidades a priori.

Para o aluno 1, por exemplo, o valor mais provável foi o de ele pertencer ao grupo 0, com probabilidade de 0,9996 e de pertencer ao grupo 1 com probabilidade de 0,0004. O valor verdadeiro da resposta situação foi 0, o que indica que foi uma classificação correta. Para os outros casos as interpretações são similares.

O gráfico a seguir é uma representação útil para identificar quão bem a função discriminante separa os grupos.

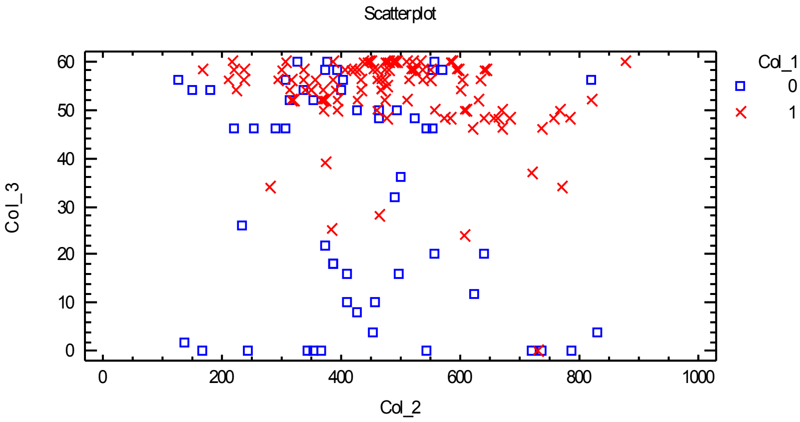

Figura 7: Representação gráfica dos dados no espaço discriminante

O diagrama de dispersão 3D para três das variáveis é apresentado a seguir, que dá uma idéia de como as variáveis servem para classificação.

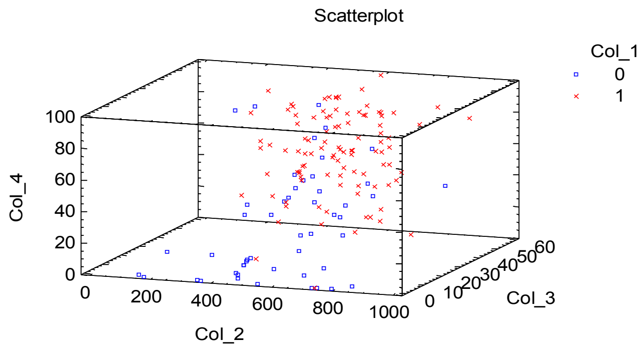

Figura 8: Representação gráfica 3D dos dados no espaço discriminante

A função discriminante é apresentada na tabela a seguir:

Tabela 12 – Função discriminante

Função discriminante	Autovalor	% Relativa	Correlação canônica
1	0,85779	100	0,67950

Fonte: Dados analisados no software Statgraphics

Como só temos uma função discriminante, a porcentagem relativa da variação contada por pela função é 100%. A correlação canônica, que representa a habilidade relativa de distinguir entre grupos é de 0,6795.

Tabela 13 – Valores para Lambda de Wilks, qui-quadrado, G.l. e p-valor para a função

Funções	Wilks Lambda	Qui-quadrado	gl	p-valor
1	0,538274	99,72	6	0,00000

Fonte: Dados analisados no software Statgraphics

O p-valor mostrado na tabela 13 foi estatisticamente significativo, indicando que o método pode ser utilizado com essas variáveis para classificar corretamente a evasão dos alunos.

A função de coeficientes de classificação para a variável situação é:

Tabela 14 – Função de coeficientes de classificação

	Evasão	*Não-evasão*
Biob	0,02420	0,02840
EstgerIfreq	0,11622	0,17699
ProbInota	-0,03293	0,00357
Calcfreq	0,00984	0,04800
Labinffreq	0,15592	0,08182
Labinfnota	-0,03306	0,00523
Constante	-9,46176	-15,79850

Fonte: Dados analisados no software Statgraphics

Assim, a função utilizada para evasão é:

-9,462+0,0242*Biob+0,116*Estgerfreq-0,033*Probnota+0,010*Calcfreq+0,156*inffreq - 0,033*infnt

Se for considerado que os dados vêm de uma distribuição normal multivariada, então os escores são relacionados com as probabilidades de que uma observação pertença a um grupo particular.

A primeira função discriminante padronizada é dada por:

0,325*Biob+0,459*Estgerfreq+0,418*Probnt+0,330*Calcfreq-0,633*Inffreq+0,596*infnt

Pela magnitude relativa dos coeficientes na equação acima, pode-se determinar como as variáveis independentes serão usadas para discriminar entre os grupos.

A seguir é apresentada, tabela com medidas estatísticas das variáveis incluídas na equação:

Tabela 15 – Medidas estatísticas das variáveis

SITUAÇÃO	**Evasão**		**Não-evasão**	
n	54		112	
	média	desvio padrao	média	desvio padrao
Biob	428,481	167,678	484,125	144,006
Estgerfreq	32,444	22,869	53,143	8,704
Probnota	18,741	22,491	52,366	22,520
Calcfreq	26,204	23,110	51,348	13,105
Labinffreq	31,019	22,928	49,205	12,863
Labinfnota	32,667	39,667	72,652	25,081

Fonte: Dados analisados no software Statgraphics

A matriz de co-variância entre as variáveis é apresentada na tabela a seguir:

Tabela 16 – Matriz de covariância entre as variáveis

	Biob	EstgerIfreq	ProbInota	Calcfreq	Labinffreq	Labinfnota
Biob	23122,2	-464,26	-229,86	-444,84	-759,39	-1152,14
EstgerIfreq	-464,26	220,29	173,17	181,74	201,7	286,08
ProbInota	-229,86	173,17	506,73	203,47	184,59	300,05
Calcfreq	-444,84	181,74	203,47	288,83	191,07	305,06
Labinffreq	-759,39	201,7	184,59	191,07	281,86	417,3
Labinfnota	-1152,14	286,08	300,05	305,06	417,3	934,25

Fonte: Dados analisados no software Statgraphics

Na tabela 17 é apresentada a matriz de correlação entre as variáveis:

Tabela 17 – Matriz de correlação entre as variáveis

	Biob	EstgerIfreq	ProbInota	Calcfreq	Labinffreq	Labinfnota
Biob	1	-0,21	-0,07	-0,17	-0,3	-0,25
EstgerIfreq	-0,21	1	0,52	0,72	0,81	0,63
ProbInota	-0,07	0,52	1	0,53	0,49	0,44
Calcfreq	-0,17	0,72	0,53	1	0,67	0,59
Labinffreq	-0,3	0,81	0,49	0,67	1	0,81
Labinfnota	-0,25	0,63	0,44	0,59	0,81	1

Fonte: Dados analisados no software Statgraphics

Com base nas informações da tabela anterior pode-se verificar que as variáveis Calcfreq e labinffreq tem valores mais altos de correlação com as demais variáveis.

5.3 COMPARATIVO ENTRE OS METODOS ESTUDADOS

A seguir são apresentadas as tabelas 18 e 19 com resumo dos resultados encontrados durante a análise deste estudo, resumo este que tem por finalidade ajudar a auxiliar na escolha do método mais adequado para futuros estudos nesta área, cabe aqui salientar que as 6 variáveis finais do modelo são: Escore no vestibular em biologia, freqüência em Estatística Geral 1, freqüência em Cálculo 1, freqüência em Laboratório de Informática e notas em Probabilidades 1 e Laboratório de Informática.

Tabela 18 – Quadro comparativo do percentual de classif. Correta entre os modelos estudados

	Regressão Logistica	Análise discriminante
Com todas as Variáveis	84,93%	84,94%
Com as 6 variaveis escolhidas	83,73%	82,53%

Fonte: Dados analisados nos softwares R e Statgraphics

Tabela 19 - Quadro comparativo de medidas estatísticas calculadas

	Regressão Logística	Análise Discriminante
Especifidade	68,52%	68,52%
Sensibilidade	91,07%	89,29%
Preditivo negativo	78,72%	75,51%
Preditivo positivo	85,71%	85,47%
Falso positivo	8,93%	10,71%
Falso negativo	31,48%	31,48%
Preditivo do modelo	83,73%	82,53%

Fonte: Dados analisados nos softwares R e Statgrphics

Com base na tabela 18, verifica-se que os valores são bem próximos, tanto se levar em conta a Regressão Logística e a Análise Discriminante, e o modelo com todas as variáveis e o modelo só com as 6 variáveis. Na tabela 19 verifica-se igualdade nos valores nos métodos para as medidas de especificidade e falsos negativos, com relação as demais medidas o método de Regressão Logística tem valores percentuais maiores.

Utilizando os resultados obtidos nas tabelas 7 e 9, calcula-se estatística *Kappa* para verificar se existe diferença significativa nos métodos de Regressão Logística e Análise Discriminante.

A fórmula utilizada é a seguinte:

$$Z = \frac{k_1 - k_2}{\sqrt{(\boldsymbol{\sigma_{k1}}^2 + \boldsymbol{\sigma_{k2}}^2)}}$$

k_1 = 0,8373 = coeficiente *Kappa* para o método de Regressão Logística.

K_2 = 0,8253 = coeficiente *Kappa* para o método de Análise Discriminante.

$\Sigma_{k1}{}^2$ =0,0046=variância amostral do coef. *Kappa* para método de Regressão Logística.

$\Sigma_{k2}{}^2$ =0,0043=variância amostral do coef. *Kappa* para método de Análise Discriminante.

Logo: Z = 0,1272.

sendo as hipóteses estatísticas:

H_o = Não há diferença entre as performances dos dois métodos

H_1 = Há diferença entre as performances dos dois métodos.

e observando o valor de Z padronizado ao nível de significância (probabilidade de rejeitar H_0 dado que é verdadeira) de 5 % que corresponde ao intervalo (-1,96,+ 1,96) pode-se ver que o valor de Z= 0,1272, correspondente a estatística Z calculada pelo teste e que compara os dois métodos, encontra-se dentro do intervalo proposto (dentro da região de aceitação de H_0), o que indica uma aceitação da hipótese H_0 e conclui-se que os dois métodos tem desempenho iguais ao classificar a evasão do Curso de Estatística.

6. CONCLUSÕES

De acordo com os resultados apresentados nos testes, pode-se concluir que o modelo ajustado apresentou-se satisfatório, com um poder preditivo de 83,73%, utilizando se a Regressão Logística, no método de classificação de Análise Discriminante obteve-se o valor de 82,53 %, o que pode ser considerado um bom resultado também. Sendo assim este estudo pode ser tomado como base para eventuais análises de evasão.

Com base no resultado da estatística *Kappa*, conclui-se que não há diferença entre as performances dos 2 métodos utilizados, ou seja, fica a critério do responsável pelo estudo qual método utilizar, pois deve se levar em conta o método disponível no momento bem como o conhecimento do responsável pela análise em questão.

Com relação à quantidade de variáveis utilizadas para o estudo fica provado pelos resultados anteriores que as variáveis Escore no vestibular em biologia, freqüência em Estatística Geral 1, freqüência em Cálculo 1, freqüência em Laboratório de Informática e notas em Probabilidades 1 e Laboratório de Informática, explicam grande parte do modelo, visto que na análise em Regressão Logística com todas as variáveis obteve-se 84,93 % como valor preditivo, já no método de analise discriminante obteve-se o valor de 84,94 %, sendo assim desnecessário o uso das demais variáveis.

Na análise das 6 variáveis do modelo é interessante verificar que 3 delas são referentes a freqüência do aluno, 2 referentes a notas e uma referente ao escore no vestibular, logo podemos destacar a importância da freqüência do aluno nas aulas como um fator relevante na continuidade do curso.

Podemos também verificar que as variáveis notas e freqüências nas disciplinas são mais importantes, pois apenas escore em biologia foi significativa no modelo. Além disso, pode-se detectar o aluno com probabilidade grande de evasão logo no primeiro semestre cursado. Podendo este fato ser um ponto de partida para futuros estudos e para se evitar a evasão dos alunos do Curso de Estatística da Universidade Federal do Paraná.

7. REFERÊNCIAS BIBLIOGRÁFICAS

ASCENSO, João; FRED, Ana. **Reconhecimento de Padrões,** 2003. Disponível em: http://ltodi.est.ips.pt/jascenso/padroes/teoricas/Aula%202%20-%20Introdu%C3%A7%C3%A3o%20ao%20RP.pdf . Acessado em 16/04/08 as 16:00 min

CASTRO**, O fenômeno da evasão escolar na educação superior no Brasil,** 2005. Disponível em http://www.iesalc.unesco.org.ve/programas/Deserci%C3%B3n/Informe%20Deserci%C3%B3n %20 Brasil%20-%20D%C3%A9bora%20Niquini.pdf acessado 06/03/08 as 16:10 min

FISHER, R.A., The Statistical Utilizacion of Multiple Measurements, **Annals os Eugenics**. 8 (1938), 376-386.

GIOLO, Suely Ruiz. Apostila de Análise de Regressão, 2003.

GIOLO, Suely Ruiz. Apostila de Análise de Dados Discretos, 2004.

GIOLO, Suely Ruiz. Introdução a Análise de Dados Categóricos, 2006.

JOHNSON, R. A., WICHERN, D.W. **Applied Multivariate Statistical Analysis.** Prentice Hall International, Inc. New Jersey, 1988.

MULLER, Sonia I. M. Gama. Comparação entre os Métodos de Máxima Veros-similhança, e o Método de Fisher para Reconhecimento de Padrões em Imagens Coloridas, 1997.

MULLER, Sonia I. M. Gama. Sistema Integrado de Avaliação com Aplicação na Engenharia, 2007.

CHAVES, Anselmo Neto. Apostila de Análise Multivariada II, 2005.

PREGIBON, D. Logistic Regression Diagnostics, Annals of Statistics, v.9, 1981.

APÊNDICES

APÊNDICE 1

Tabela 11- Classificação para os indivíduos do estudo

Aluno	*Grupo Real*	*Grupo mais provável*	*Maior Escore*	*Distância Quadrada*	Probabilidade	*2º grupo mais provável*	*2º maior escore*	*Distância Quadrada*	Probabilidade
1	0	0	2,49	3,66	1	1	-3,18	15,01	0
2	0	0	4,34	2,13	0,99	1	-0,44	11,69	0,01
3	0	*1	13,27	0,01	0,86	0	11,49	3,57	0,14
4	0	0	10,99	0,77	0,55	1	10,79	1,17	0,45
5	0	*1	12,7	0,65	0,97	0	9,19	7,67	0,03
6	0	*1	11,98	0,02	0,84	0	10,3	3,36	0,16
7	0	*1	13,2	0,02	0,84	0	11,58	3,28	0,16
8	0	0	0,01	3,18	1	1	-5,4	14,02	0
9	0	0	8,5	0,53	0,97	1	5,15	7,24	0,03
10	0	0	4,49	2,61	0,99	1	-0,6	12,79	0,01
11	0	0	5,65	0,41	0,66	1	4,99	1,73	0,34
12	0	0	5,08	1,46	0,99	1	0,79	10,05	0,01
13	0	0	-3,61	5,2	1	1	-10	17,97	0
14	0	0	2,86	4,05	1	1	-3	15,78	0
15	0	0	13,85	0,31	0,95	1	10,83	6,35	0,05
16	0	*1	11,4	0,46	0,64	0	10,81	1,64	0,36
17	0	0	7,07	0	0,88	1	5,08	3,99	0,12
18	0	*1	13,12	0,4	0,66	0	12,44	1,75	0,34
19	0	*1	7,23	0,11	0,78	0	5,96	2,65	0,22
20	0	0	-5,75	3,24	1	1	-11,2	14,14	0
21	0	0	5,15	0,97	0,98	1	1,3	8,67	0,02
22	0	0	11,19	0,56	0,61	1	10,73	1,47	0,39
23	0	*1	12,6	0,21	0,74	0	11,58	2,26	0,26
24	0	0	12,81	0,06	0,92	1	10,4	4,88	0,08
25	0	0	8,83	0,02	0,84	1	7,16	3,36	0,16
26	0	0	8,35	0,08	0,8	1	6,97	2,84	0,2
27	0	0	10,69	0,95	0,98	1	6,86	8,6	0,02
28	0	*1	5,52	0,01	0,85	0	3,8	3,45	0,15
29	0	*1	17,13	0,03	0,83	0	15,57	3,17	0,17
30	0	0	2,33	1,57	0,99	1	-2,05	10,31	0,01
31	0	0	5,23	2,1	0,99	1	0,46	11,63	0,01
32	0	*1	9,66	0,34	0,68	0	8,88	1,89	0,32
33	0	*1	12,71	0,83	0,53	0	12,58	1,1	0,47
34	0	0	3,32	2,4	0,99	1	-1,64	12,33	0,01
35	0	*1	16,16	0,36	0,68	0	15,41	1,86	0,32
36	0	*1	15,87	0,01	0,85	0	14,16	3,42	0,15
37	0	0	7,2	0,28	0,95	1	4,24	6,21	0,05
38	0	*1	11,99	0,11	0,78	0	10,72	2,65	0,22
39	0	0	3,62	1,18	0,98	1	-0,43	9,28	0,02
40	0	0	8,14	0,13	0,93	1	5,5	5,41	0,07
41	0	0	-4,06	1,83	0,99	1	-8,63	10,97	0,01
42	0	0	7,81	0,56	0,97	1	4,42	7,35	0,03

Fonte: Dados analisados no Software Statgrhapics

Tabela 11- Classificação para os indivíduos do estudo. (Cont.)

Aluno	*Grupo Real*	*Grupo mais provável*	*Maior Escore*	*Distância Quadrada*	Probabilidade	*2º grupo mais provável*	*2º maior escore*	*Distância Quadrada*	Probabilidade
43	0	0	15,65	0,02	0,84	1	14,03	3,27	0,16
44	0	0	9,37	0,39	0,96	1	6,23	6,69	0,04
45	0	0	1,56	1,15	0,98	1	-2,46	9,19	0,02
46	0	0	10,98	0	0,86	1	9,13	3,69	0,14
47	0	0	2,84	0	0,87	1	0,95	3,78	0,13
48	0	*1	22,03	0,94	0,5	0	22,01	0,98	0,5
49	0	0	10,84	0,19	0,94	1	8,06	5,75	0,06
50	0	0	8,19	0,52	0,97	1	4,86	7,19	0,03
51	1	1	17,91	0,99	0,98	0	14,03	8,74	0,02
52	1	1	17,67	0,04	0,91	0	15,37	4,64	0,09
53	1	1	11,22	0	0,88	0	9,25	3,95	0,12
54	1	1	8,68	0,15	0,94	0	5,99	5,53	0,06
55	1	1	8,47	0,44	0,65	0	7,85	1,69	0,35
56	1	1	16,12	0,01	0,9	0	13,96	4,33	0,1
57	1	1	14,4	0,27	0,95	0	11,47	6,14	0,05
58	1	1	11,15	0,09	0,92	0	8,64	5,1	0,08
59	1	1	12,21	0,04	0,82	0	10,67	3,12	0,18
60	1	1	7,19	0	0,87	0	5,3	3,78	0,13
61	1	1	10,79	0,25	0,95	0	7,89	6,05	0,05
62	1	1	12,92	0,13	0,93	0	10,28	5,41	0,07
63	1	1	11,39	0,08	0,8	0	10,02	2,82	0,2
64	1	1	12,83	0	0,86	0	11,02	3,61	0,14
65	1	1	7,96	0,24	0,72	0	7,01	2,15	0,28
66	1	*0	6,97	0,31	0,69	1	6,15	1,96	0,31
67	1	1	19,3	1,95	0,99	0	14,64	11,26	0,01
68	1	1	10,75	0,1	0,93	0	8,22	5,17	0,07
69	1	1	8,43	0,04	0,82	0	6,9	3,09	0,18
70	1	1	10,53	0,16	0,76	0	9,4	2,42	0,24
71	1	1	12,36	0,04	0,91	0	10,04	4,69	0,09
72	1	1	19,26	0,89	0,98	0	15,49	8,43	0,02
73	1	*0	9,29	0,85	0,53	1	9,18	1,08	0,47
74	1	1	11,62	0,06	0,92	0	9,21	4,88	0,08
75	1	1	13,85	0	0,88	0	11,9	3,9	0,12
76	1	1	15,59	0	0,86	0	13,73	3,71	0,14
77	1	1	16,02	0,31	0,95	0	13,01	6,32	0,05
78	1	1	17,58	0,48	0,96	0	14,31	7,02	0,04
79	1	1	18,53	0,67	0,97	0	15	7,73	0,03
80	1	1	13,89	0,06	0,92	0	11,49	4,85	0,08
81	1	1	11,23	0,37	0,67	0	10,5	1,82	0,33
82	1	1	18,7	1,16	0,98	0	14,67	9,21	0,02
83	1	1	16,96	0,56	0,61	0	16,51	1,46	0,39
84	1	1	14,47	0,03	0,91	0	12,2	4,57	0,09

Fonte: Dados analisados no Software Statgrhapics

Tabela 11- Classificação para os indivíduos do estudo. (Cont.)

Aluno	*Grupo Real*	*Grupo mais provável*	*Maior Escore*	*Distância Quadrada*	Probabilidade	*2º grupo mais provável*	*2º maior escore*	*Distância Quadrada*	Probabilidade
85	1	1	13,99	0,02	0,84	0	12,36	3,29	0,16
86	1	1	11,24	0	0,88	0	9,29	3,91	0,12
87	1	1	13,56	0,03	0,9	0	11,32	4,52	0,1
88	1	1	15,26	0,02	0,84	0	13,63	3,28	0,16
89	1	1	12,02	0,69	0,57	0	11,73	1,27	0,43
90	1	*0	8,98	0,32	0,69	1	8,18	1,93	0,31
91	1	1	17,41	1,01	0,98	0	13,51	8,8	0,02
92	1	*0	6,29	0,45	0,65	1	5,68	1,66	0,35
93	1	*0	5,01	0,06	0,81	1	3,58	2,92	0,19
94	1	1	16,69	0,05	0,82	0	15,2	3,03	0,18
95	1	1	11,93	0,25	0,72	0	10,99	2,14	0,28
96	1	1	17,65	0,38	0,96	0	14,52	6,65	0,04
97	1	1	11,41	0,23	0,73	0	10,44	2,18	0,27
98	1	1	18,83	0,48	0,96	0	15,55	7,03	0,04
99	1	1	12,07	0,02	0,84	0	10,42	3,33	0,16
100	1	1	15,03	0	0,88	0	13,01	4,04	0,12
101	1	1	16,6	0,14	0,77	0	15,41	2,51	0,23
102	1	1	13,28	0,1	0,93	0	10,73	5,19	0,07
103	1	*0	9,12	0,51	0,63	1	8,6	1,55	0,37
104	1	1	15,12	0,08	0,92	0	12,64	5,05	0,08
105	1	1	18,59	0	0,87	0	16,7	3,79	0,13
106	1	1	20,47	0,03	0,91	0	18,19	4,6	0,09
107	1	1	14,35	0,03	0,9	0	12,11	4,49	0,1
108	1	1	14,64	0	0,86	0	12,85	3,59	0,14
109	1	1	14,44	0,54	0,97	0	11,08	7,27	0,03
110	1	1	10,64	0,73	0,56	0	10,4	1,22	0,44
111	1	1	9,97	0,15	0,76	0	8,79	2,49	0,24
112	1	1	18,22	0,02	0,9	0	15,99	4,48	0,1
113	1	1	18,79	0,53	0,97	0	15,44	7,22	0,03
114	1	1	12,46	0,38	0,67	0	11,75	1,8	0,33
115	1	1	17,96	0	0,87	0	16,07	3,78	0,13
116	1	1	13,8	0,1	0,93	0	11,27	5,16	0,07
117	1	*0	3,19	0,07	0,8	1	1,78	2,88	0,2
118	1	*0	7,45	0	0,88	1	5,45	3,99	0,12
119	1	1	14,84	0,08	0,8	0	13,46	2,83	0,2
120	1	1	18,73	0,92	0,51	0	18,69	1	0,49
121	1	1	13,1	0,06	0,81	0	11,64	2,97	0,19
122	1	1	21,61	1,44	0,99	0	17,34	9,99	0,01
123	1	1	25,3	0,59	0,97	0	21,87	7,46	0,03
124	1	1	18,17	0,49	0,64	0	17,62	1,6	0,36
125	1	1	16,44	3,63E-007	0,87	0	14,52	3,84	0,13
126	1	1	17,22	0,28	0,95	0	14,27	6,18	0,05

Fonte: Dados analisados no Software Statgrhapics

Tabela 11- Classificação para os indivíduos do estudo. (Cont.)

Aluno	*Grupo Real*	*Grupo mais provável*	*Maior Escore*	*Distância Quadrada*	Probabilidade	*2º grupo mais provável*	*2º maior escore*	*Distância Quadrada*	Probabilidade
127	1	1	17,83	0,09	0,92	0	15,33	5,1	0,08
128	1	1	12,43	0,07	0,92	0	9,99	4,95	0,08
129	1	1	14,3	1,76	0,99	0	9,78	10,8	0,01
130	1	1	13,43	0,1	0,78	0	12,14	2,69	0,22
131	1	1	13,02	0,15	0,76	0	11,86	2,48	0,24
132	1	1	17,86	0,78	0,97	0	14,2	8,08	0,03
133	1	1	12,55	0,01	0,85	0	10,85	3,42	0,15
134	1	1	11,95	0,38	0,67	0	11,25	1,8	0,33
135	1	1	14,61	1,07	0,98	0	10,65	8,98	0,02
136	1	1	16,31	0,32	0,95	0	13,28	6,38	0,05
137	1	*0	8,05	0,54	0,97	1	4,69	7,25	0,03
138	1	1	14,82	0,67	0,97	0	11,3	7,73	0,03
139	1	1	19,47	2,14	0,99	0	14,69	11,72	0,01
140	1	1	12,06	0,97	0,98	0	8,22	8,66	0,02
141	1	1	13,07	0	0,88	0	11,1	3,94	0,12
142	1	*0	19,32	0,92	0,51	1	19,29	1	0,49
143	1	1	20,04	0,66	0,97	0	16,53	7,68	0,03
144	1	1	13,94	0,01	0,85	0	12,21	3,47	0,15
145	1	1	14,72	0,01	0,9	0	12,57	4,31	0,1
146	1	*0	15,45	0,01	0,89	1	13,39	4,13	0,11
147	1	1	18	0,34	0,96	0	14,93	6,48	0,04
148	1	1	12,85	0,45	0,96	0	9,61	6,93	0,04
149	1	1	14,97	0,56	0,97	0	11,58	7,35	0,03
150	1	1	13,3	1,35	0,99	0	9,1	9,75	0,01
151	1	1	16,2	0,01	0,85	0	14,48	3,45	0,15
152	0	0	8,19	0,52	0,97	1	4,86	7,19	0,03
153	0	0	8,86	2,31	0,99	1	3,96	12,12	0,01
154	0	1	13,26	0,01	0,9	0	11,11	4,32	0,1
155	0	0	6,2	0,28	0,71	1	5,32	2,05	0,29
156	1	0	0,04	2,49	0,99	1	-4,97	12,51	0,01
157	1	0	13,67	0,23	0,73	1	12,7	2,18	0,27
158	1	1	16,34	0,82	0,98	0	12,65	8,2	0,02
159	1	1	15,22	0,06	0,92	0	12,83	4,84	0,08
160	1	1	15,03	0,06	0,81	0	13,6	2,91	0,19
161	1	1	16,48	0,2	0,94	0	13,69	5,78	0,06
162	1	1	13,68	0,03	0,9	0	11,44	4,49	0,1
163	1	1	8,97	0,08	0,8	0	7,61	2,81	0,2
164	1	1	15,77	1,51	0,99	0	11,44	10,17	0,01
165	1	1	21,13	0,58	0,97	0	17,72	7,4	0,03
166	1	1	17,89	0,01	0,89	0	15,8	4,19	0,11

Fonte: Dados analisados no Software Statgrhapics

APÊNDICE 2

Variáveis Utilizadas na Análise de Evasão de Alunos do Curso de Estatística da **UFPR**

1- Gênero: 1-Masculino; 2- Feminino

2- Estado Civil: 1 - Solteiro(a); 2 - Casado(a) e 3 - Outro

3 - Classificação no vestibular

4 - Escore no vestibular em Português

5 - Escore no vestibular em Matemática

6 - Escore no vestibular em Biologia

7 - Escore no vestibular em Química

8 - Escore no vestibular em Geografia

9 - Escore no vestibular em Física

10 - Escore no vestibular em História

11 - Escore no vestibular em Língua Estrangeira Moderna

12 - Escore no vestibular em Redação

13 - Freqüência na disciplina de Estatística Geral I: de 0 a 60 horas

14 - Nota na disciplina de Estatística Geral I: de 0 a 100

15 - Freqüência na disciplina de Cálculo de Probabilidade I: de 0 a 60 horas

16 - Nota na disciplina de Cálculo de Probabilidade I: de 0 a 100

17 - Freqüência na disciplina de Cálculo com Geometria. Analítica I: de 0 a 60horas

18 - Nota na disciplina de Cálculo com Geometria Analítica I: de 0 a 100

19 - Freqüência na disciplina de Lógica: de 0 a 60 horas

20 - Nota na disciplina de Lógica: de 0 a 100

21 - Freqüência na disciplina de Laboratório de Informática: de 0 a 60 horas

22 - Nota na disciplina de Laboratório de Informática: de 0 a 100

APÊNDICE 3

Comandos utilizados no Software R

```
dados<-matrix(c(a,b,c,d),nc=2)
 dados
 Qp<-chisq.test(dados,correct=F)
 Qp
 n<-sum(dados)
 Q<-((n-1)/n)*Qp$statistic
 Q
 p<-1-pchisq(Q,1)
 p
----------------------------------------------------------------
dados<-read.table("C:/ex221.txt",header=T)
attach(dados)
ajust<-glm(as.matrix(dados[,c(1,2)])~Nota ProbI, FreqCalcI,EscBiol, NotaLabInf,
FreqLab Inf,FreqEsT GerI,family=binomial, data=dados)
ajust<-glm(as.matrix(dados[,c(1,2)])~Nota ProbI, FreqCalcI,EscBiol, NotaLabInf,
FreqLab Inf,FreqEsT GerI,family=binomial(link="logit"),data=dados)
 ajust
 anova(ajust)
 anova(ajust,test="Chisq")
 summary(ajust)
 ajust$fitted.values
 ajust$y
 ajust$residuals
 dev<-residuals(ajust,type='deviance')
 dev
 QL<-sum(dev^2)
 QL
 p1<-1-pchisq(QL,6)
 p1
 rpears<-residuals(ajust,type='pearson')
 rpears
 QP<-sum(rpears^2)
 QP
 p2<-1-pchisq(QP,6)
 p2
----------------------------------------------------------------
ajust1<-glm(evento~ NotaProbI+ FreqCalcI+EscBiol+NotaLabInf+FreqLabInf+
FreqEsTGerI, family=binomial(link="logit"))
ajust1
summary(ajust1)
anova(ajust1,test="Chisq")
ajust2<-glm(evento~ NotaProbI+ FreqCalcI+EscBiol+NotaLabInf+FreqLabInf+
FreqEsTGerI,family=binomial(link="logit"))
ajust2
summary(ajust2)
anova(ajust2, test="Chisq")
cbind(dc,sexo,ecg,idade,ajust2$fitted.values)
```

```
dev<-residuals(ajust2,type='deviance')
dev
plot(dev)
rpears<-residuals(ajust2,type='pearson')
rpears
plot(rpears)
--------------------------------------------------------
fit.model<-ajust2
  par(mfrow=c(1,1))
  X <- model.matrix(fit.model)
  n <- nrow(X)
  p <- ncol(X)
  w <- fit.model$weights
  W <- diag(w)
  H <- solve(t(X)%*%W%*%X)
  H <- sqrt(W)%*%X%*%H%*%t(X)%*%sqrt(W)
  h <- diag(H)
  td <- resid(fit.model,type="deviance")/sqrt(1-h)
  e <- matrix(0,n,100)
 for(i in 1:100){
   dif <- runif(n) - fitted(fit.model)
   dif[dif >= 0 ] <- 0
   dif[dif<0] <- 1
   nresp <- dif
   fit <- glm(nresp ~ X, family=binomial)
   w <- fit$weights
   W <- diag(w)
   H <- solve(t(X)%*%W%*%X)
   H <- sqrt(W)%*%X%*%H%*%t(X)%*%sqrt(W)
   h <- diag(H)
   e[,i] <- sort(resid(fit,type="deviance")/sqrt(1-h))}
   e1 <- numeric(n)
   e2 <- numeric(n)
 for(i in 1:n){
   eo <- sort(e[i,])
   e1[i] <- eo[5]
   e2[i] <- eo[95]}
 med <- apply(e,1,mean)
 faixa <- range(td,e1,e2)
 par(pty="s")
 qqnorm(td,xlab="Percentis da N(0,1)",
 ylab="Componente da deviance", ylim=faixa, pch=16)
 par(new=T)
 qqnorm(e1,axes=F,xlab="",ylab="",type="l",ylim=faixa,lty=1)
 par(new=T)
 qqnorm(e2,axes=F,xlab="",ylab="", type="l",ylim=faixa,lty=1)
 par(new=T)
 qqnorm(med,axes=F,xlab="", ylab="", type="l",ylim=faixa,lty=2)
```