# 5
# Aplicação e comparação dos métodos

Foram feitas comparações do STLR-Tree com os seguintes métodos de classificação: *Generalized Additive Models (GAM), Regressão Logística, Redes Neurais, Análise Discriminante, k-Nearest Neighbor (k-NN)* e *Classification and Regression Trees (CART)*. O STLR-Tree e as Redes Neurais foram programados no Matlab 6.1 e os demais métodos foram rodados no programa R 2.6.2, utilizando, além das funções existentes no próprio, bibliotecas tais como: tree (CART), VGAM (GAM) e kknn (k-NN), disponíveis na própria página do programa na internet:

http://cran.r-project.org.

No Apêndice B encontram-se alguns comandos do R utilizados para o CART, GAM, k-NN e Regressão Logística.

## 5.1
## Bases de dados

A aplicação do STLR-Tree e a comparação com outros métodos de classificação foram feitas para três bases. Duas encontradas em exemplos do livro (14). Uma delas utilizada para fazer a distinção entre as mensagens que são realmente *e-mail* e as consideradas *spam* e outra para a ocorrência ou não de um possível infarto do miocárdio. Ambas disponibilizadas na página do próprio livro na internet:

http://www-stat.stanford.edu/ElemStatLearn.

A terceira provém de um estudo feito pra uma empresa do setor de energia elétrica no Estado do Rio de Janeiro, onde deseja-se classificar eventuais fraudes ou irregularidades de seus usuários. Este é o único exemplo no qual Redes Neurais foram utilizadas em comparação com os demais métodos de Classificação.

Uma breve descrição dos bancos segue a seguir:

– E-mail/Spam: contém dados de 4601 mensagens de e-mail, em que a variável dependente é igual a 0 se a mensagem foi considerada um e-mail

de fato ou 1 caso tinha sido caracterizada como um *spam* (mensagem que não é verdadeiramente um e-mail particular). Originalmente, a base possui 57 preditores dos quais foram selecionados 16 com base nos resultados de significância das mesmas disponíveis no livro onde aparecem como exemplo. Das 4601 mensagens, foram selecionadas 2000 aleatoriamente para compor a base *in sample* e 1000 para a base *out-of-sample*. As variáveis explicativas são as seguintes: percentual de palavras no e-mail que correspondam a: our, over, remove, internet, free, business, hp, hpl, george, 1999, re e edu; percentual de caracteres no e-mail que correspondam a: ! (char_!) e $ (char_$); comprimento da mais longa seqüência ininterrupta de letras maiúsculas (CAPMAX) e soma do comprimento das seqüências ininterruptas de letras maiúsculas (CAPTOT).

– Doenças Cardíacas na África do Sul (DCAS): possui informações de 462 indivíduos homens com idades entre 15 e 64 anos, para as variáveis: pressão arterial (sbp); consumo de tabaco em *kg* (tobacco); colesterol ldl (ldl); índice de obesidade (obesity); consumo de álcool (alcohol); idade em anos (age) e a variável resposta binária corresponde a ocorrência ou não de infarto do miocárdio até a data da coleta dos dados. Dessa 462 observações, 362 foram selecionadas para a base *in sample* e 100 para *out-of-sample*.

– Fraude/Irregularidade no Consumo de Energia Elétrica: a empresa possui cerca de 452 mil clientes inspecionados em baixa tensão com perfis de consumo de energia diferentes, distribuídos em 2 regiões de estudo (Leste e Oeste). Essas regiões estão subdivididas e foi uma dessas subdivisões que selecionamos nesse exemplo. Ela possui 2430 clientes (*in sample*) e 2941(*out-of-sample*) que são classificados através de uma variável binária (indic_irregul_cod) com o valor 0 para os clientes normais e 1 para os supostos clientes irregulares. As demais variáveis são: consumo no mês (consumo), consumo no ano anterior (comsumo_ano_ant), consumo no ano base (consumo_ano_base), média 3 meses (media_3), média 6 meses (media_6), média dos meses 1 ao 12 (media_12), média dos meses 13 ao 24 (media_12_24), indicador trimestral 1_2 (indic_trimestral_1), indicador trimestral 2_3 (indic_trimestral_2), indicador anual (indic_anual), indicador ajuste (indic_ajuste), indicador tendência (indic_tendencia), temperatura mínima (temperatura_min), temperatura máxima (temperatura_max ), carga.

Para cada um dos métodos analisados e comparados tentou-se ajustar o melhor modelo para cada um deles, confrontando as melhores classificações

que cada um resultou.

Vale ressaltar que em todas as comparações com o GAM, ajustamos o mesmo usando um *Suavizador Spline Cúbico* com 4 graus de liberdade para cada preditor e o método k-NN foi ajustado para um $k = 10$.

No Apêndice C as tabelas C.1, C.2, e C.3 apresentam os valores de algumas Estatísticas Descritivas das variáveis de cada uma das bases descritas anteriormente.

### 5.1.1 Aplicação: E-mail/Spam

Todas as 16 covariáveis selecionadas previamente foram colocadas como candidatas à variável de transição, fazendo parte do conjunto $\mathbf{x}$ e o conjunto de variáveis $\mathbf{z}$ foi composto por: our, over, remove, internet, free, business, hp, hpl, george, re, edu, char_! e char_$, CAPMAX e CAPTOT.

O ajuste a esses dados gerou uma árvore com 2 nós terminais e profundidade 1, contra um CART com 13 nós terminais e profundidades igual a 7. A figura 5.1 ilustra a estrutura do STLR-Tree para o ajuste final onde as pertinências de cada regime estão sendo mostradas.

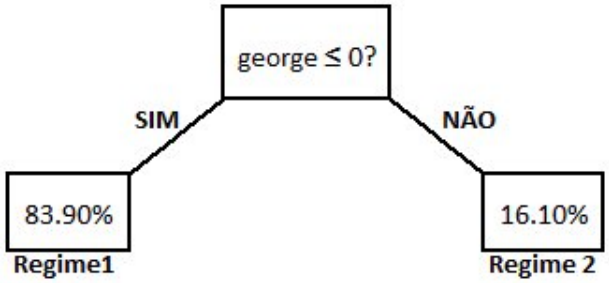


Figura 5.1: Estrutura do modelo - Spam

As equações referentes a cada um dos regimes encontrado são dadas por

$$\begin{aligned} Regime\ \ 1 = \ & (-0.39 - 3.15our - 10.77over + 9.87remove - 5.67internet + \\ + \ & 3.11free + 0.04business + 2.4hp + 1.8hpl - 2.88george - \\ - \ & 7.04remove - 14.61edu + 3.45char\_! + 19.25char\_\$ + \\ + \ & 1.53CAPMAX + 0.02CAPTOT) * G(george; 0, 8, 0, 1) \end{aligned}$$

$$\begin{aligned} Regime\ \ 2 = \ & (+150.51 + 784.57our - 658.04over + 7299.91remove + \\ + \ & 396.22internet + 48.4free - 21.48business - \\ - \ & 722.10hp + 148.85hpl + 379.96re - 38.98edu - \end{aligned}$$

$$- \quad 265.25char\_! - 955.31char\_\$ - 1.59CAPMAX +$$
$$+ \quad 2.3CAPTOT) * [1 - G(george; 0.8, 0.1)]$$

A seguir apresenta-se a tabela de classificação da análise *in sample*, tabela 5.1, e na seqüência a consolidação dos valores de sensitividade, especificidade e total de acertos, tabela 5.2, além da tabela com os métodos ordenados pelas taxas de acerto, tabela 5.3.

O STLR-Tree apresenta o segundo melhor desempenho para a taxa total de acertos 91.20%, ficando apenas atrás do GAM que obteve 95.20%.

Analisando a tabela 5.2 ressaltamos ainda as taxas de erro total (100% - taxa de acerto total) de classificação: GAM (4.80%), STLR-Tree (8.80%), Regressão Logística (8.95%), CART (10.05%), k-NN (28.85%) e Análise Discriminante (12.25%).

Tabela 5.1: Tabela de Classificação (*in sample*) - Spam

| Observado (y) | Predito (ŷ) | | |
|---|---|---|---|
| | 0 | 1 | |
| 0 | 1142 | 70 | 1212 |
| 1 | 105 | 683 | 788 |
| | 1247 | 753 | 2000 |

Tabela 5.2: Comparação das Taxas de Acerto (*in sample*) - Spam

| | Tx. de acertos Total | Tx. de acertos para 1 (Sensitividade) | Tx. de acertos para 0 (Especificidade) |
|---|---|---|---|
| GAM | 95.20% | 92.26% | 97.11% |
| STLR-Tree | 91.25% | 86.68% | 94.22% |
| Reg. Logística | 91.05% | 86.80% | 93.81% |
| CART | 89.95% | 80.46% | 96.12% |
| k-NN | 89.21% | 81.32% | 94.15% |
| Análise Discrim. | 87.75% | 78.17% | 93.98% |

Tabela 5.3: Métodos de Classificação Ordenados por Taxas de Acerto (*in sample*) - Spam

| Tx. de acertos Total | Tx. de acertos para 1 (Sensitividade) | Tx. de acertos para 0 (Especificidade) |
|---|---|---|
| GAM (0.952) | GAM (0.922) | GAM (0.971) |
| STLR-Tree (0.912) | Reg. Logística (0.86) | CART (0.961) |
| Reg. Logística (0.910) | STLR-Tree (0.866) | STLR-Tree (0.942) |
| CART (0.899) | k-NN (0.813) | k-NN (0.941) |
| k-NN (0.892) | CART (0.804) | Análise Discrim. (0.939) |
| Análise Discrim. (0.877) | Análise Discrim. (0.781) | Reg. Logística (0.938) |

Nas tabelas seguintes temos as mesmas informações, porém para a análise *out-of-sample*, 5.4 e 5.5. Como esperado o valor das taxas diminui, mantendo a mesma ordem encontrada na análise (*in sample*) para a taxa total de acertos.

Tabela 5.4: Comparação das Taxas de Acerto (*out-of-sample*) - Spam

| | Tx. de acertos Total | Tx. de acertos para 1 (Sensitividade) | Tx. de acertos para 0 (Especificidade) |
|---|---|---|---|
| GAM | 84,40% | 83,65% | 91,58% |
| STLR-Tree | 85,70% | 85,08% | 91,58% |
| Reg. Logística | 85,60% | 84,86% | 92,63% |
| K-NN | 87,30% | 87,29% | 87,37% |
| Análise Discrim. | 78,20% | 76,91% | 90,53% |

Tabela 5.5: Métodos de Classificação Ordenados por Taxas de Acerto (*out-of-sample*) - Spam

| Tx. de acertos Total | Tx. de acertos para 1 (Sensitividade) | Tx. de acertos para 0 (Especificidade) |
|---|---|---|
| K-NN (0,873) | K-NN (0,873) | Reg. Logística (0,926) |
| STLR-Tree (0,857) | STLR-Tree (0,851) | STLR-Tree (0,916) |
| Reg. Logística (0,856) | Reg. Logística (0,849) | GAM (0,916) |
| GAM (0,844) | GAM (0,836) | Análise Discrim. (0,905) |
| Análise Discrim. (0,782) | Análise Discrim. (0,769) | K-NN (0,874) |

A quantidade de parâmetros do STLR-Tree foi de 16 para cada nó terminal, que somados aos outros 2 parâmetros não-lineares, resultam em um total de $r$=34 parâmetros. Gam estimou 16 parâmetros lineares e a parte não-linear possui um parâmetro para cada observação. Regressão Logística e Análise Discriminante 15 cada.

A tabela D.1 do Apêndice D apresenta os parâmetros lineares de cada um dos métodos e a tabela D.5 os coeficientes dos parâmetros não-lineares do STLR-Tree.

### 5.1.2 Aplicação: Doenças Cardíacas na África do Sul (DCAS)

No modelo selecionado o conjunto de variáveis **z** é composto por: sbp, tobacco, ldl, alcohol e age. Seus resultados são mostrados a seguir.

O ajuste a esses dados gerou um modelo com 2 nós terminais e uma profundidade, contra um CART com 13 nós terminais e profundidades igual a 7, figura 5.2.

Os modelos encontrados foram

$$Regime \quad 1 = \qquad (-288.658 + 2.477sbp + 2.293tobacco -$$

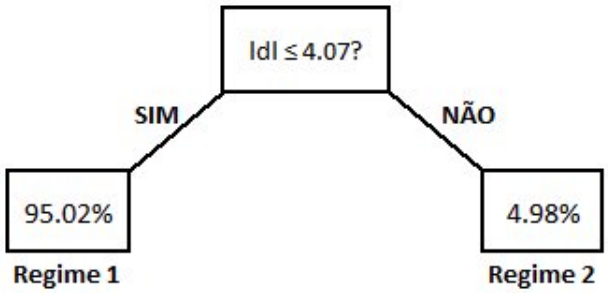


Figura 5.2: Estrutura do modelo - DCAS

$$\begin{aligned} & - \quad 6.156obesity - 1.255alcohol + \\ & + \quad 2.937age) * G(ldl; 4.07, 50) \end{aligned}$$

$$\begin{aligned} Regime \quad 2 = \quad & (-2.743 - 0.006sbp + 0.083tobacco + \\ & + \quad 0.008obesity + 0.005alcohol + \\ & + \quad 0.05age) * [1 - G(ldl; 4.07, 50)] \end{aligned}$$

Na tabela 5.6 é mostrada a tabela de classificação para os dados desta base (*in sample*).

Tabela 5.6: Tabela de Classificação (*in sample*) - DCAS

| Observado (y) | Predito (ŷ) | | |
|---|---|---|---|
| | 0 | 1 | |
| 0 | 268 | 34 | 302 |
| 1 | 85 | 75 | 160 |
| | 353 | 109 | 462 |

As duas tabelas seguintes, 5.7 e 5.8, pode ser verificado que, para esta base, as taxas de classificação de todos os métodos não foram tão eficientes quanto aquelas apresentadas para a base anterior. Estas tabelas se referem à análise *in sample*.

O STLR-Tree apresenta a terceira melhor taxa de acerto total com 72.65%. Os dois métodos que melhora classificaram foram GAM (82.60%) e CART (75.14%).

Com a tabela 5.7 podemos concluir que as taxas de erro total foram razoavelmente altas para todos os métodos: GAM (17.40%), STLR-Tree (27.35%), CART (24.86%), Regressão Logística (28.73%), Análise Discriminante (32.32%) e k-NN (29.87%).

As mesmas comparações anteriores foram feitas para a base *out-of-sample*, 5.9 e 5.10.

Tabela 5.7: Comparação das Taxas de Acerto (*in sample*) - DCAS

| | Tx. de acertos Total | Tx. de acertos para 1 (Sensitividade) | Tx. de acertos para 0 (Especificidade) |
|---|---|---|---|
| GAM | 78.35% | 60.63% | 87.75% |
| STLR-Tree | 74.24% | 46.88% | 88.74% |
| CART | 72.94% | 50.63% | 84.77% |
| Reg. Logística | 70.78% | 47.50% | 83.11% |
| Análise Discrim. | 69.05% | 71.88% | 67.55% |
| K-NN | 66.23% | 56.52% | 70.37% |

Tabela 5.8: Métodos de Classificação Ordenados por Taxas de Acerto (*in sample*) - DCAS

| Tx. de acertos Total | Tx. de acertos para 1 (Sensitividade) | Tx. de acertos para 0 (Especificidade) |
|---|---|---|
| GAM (0.783) | Análise Discrim. (0.718) | STLR-Tree (0.887) |
| STLR-Tree (0.742) | GAM (0.606) | GAM (0.877) |
| CART (0.729) | K-NN (0.565) | CART (0.847) |
| Reg. Logística (0.707) | CART (0.506) | Reg. Logística (0.831) |
| Análise Discrim. (0.690) | Reg. Logística (0.47) | K-NN (0.703) |
| K-NN (0.662) | STLR-Tree (0.468) | Análise Discrim. (0.675) |

Como na aplicação anterior, a ordem de classificação que diz respeito a taxa total de acertos, se manteve, porém com diminuição dos valores percentuais.

Tabela 5.9: Comparação das Taxas de Acerto (*out-of-sample*) - DCAS

| | Tx. de acertos Total | Tx. de acertos para 1 (Sensitividade) | Tx. de acertos para 0 (Especificidade) |
|---|---|---|---|
| GAM | 77,00% | 51,72% | 87,32% |
| STLR-Tree | 71,00% | 55,17% | 77,46% |
| Reg. Logística | 74,00% | 41,38% | 87,32% |
| Análise Discrim. | 69,00% | 72,41% | 67,61% |
| K-NN | 70,00% | 55,17% | 76,06% |

Tabela 5.10: Métodos de Classificação Ordenados por Taxas de Acerto (*out-of-sample*) - DCAS

| Tx. de acertos Total | Tx. de acertos para 1 (Sensitividade) | Tx. de acertos para 0 (Especificidade) |
|---|---|---|
| GAM (0,77) | Análise Discrim. (0,724) | GAM (0,873) |
| Reg. Logística (0,74) | STLR-Tree (0,552) | Reg. Logística (0,873) |
| STLR-Tree (0,71) | K-NN (0,552) | STLR-Tree (0,775) |
| K-NN (0,7) | GAM (0,517) | K-NN (0,761) |
| Análise Discrim. (0,69) | Reg. Logística (0,414) | Análise Discrim. (0,676) |

A quantidade de parâmetros em cada um dos dois nós terminais foi 6, assim o STLR-Tree tem 12 parâmetros lineares além de mais 2 não-lineares, em um total de $r$=14 parâmetros. O número de parâmetros lineares estimados pelo

GAM foi 6 e, como dito anteriormente, o número de parâmetros não lineares é igual ao número de observações de cada variável. Em Regressão Logística o número encontrado foi igual ao encontrado na parte linear do GAM, 6. Já Análise Discriminante estimou um total de 5. Na tabela D.2 do Apêndice D podemos verificar seus valores. Os coeficientes não-lineares do STLR-Tree são apresentados na tabela D.5.

### 5.1.3 Aplicação: Fraude/Irregularidade no Consumo de Energia Elétrica

Como este foi o único exemplo em que os métodos comparados incluíram Redes Neurais cabe fazer algumas observações quanto à sua programação.

No estudo original sobre as fraudes e irregularidades do setor Elétrico, encontrado em (25), o autor, visando um melhor treinamento das Redes Neurais, dividiu em cinco bases de treinamento/validação para o aprendizado (ou ajuste) do modelo e um arquivo de teste. Os dados relativos ao treinamento/validação foram coletados de dois períodos distintos: de janeiro de 2001 a dezembro de 2005 e os meses de março a junho de 2006, a fim de avaliar períodos distintos (verão e inverno de 2006).

Através de um comitê, onde foram gerados cinco modelos a partir das cinco bases de treinamento, onde cada um deles foi testado com a base de teste. Essas cinco redes treinadas possuíam uma camada de 8 neurônios escondidos.

Dentre elas, a que obteve a melhor classificação foi a utilizada para fazer as comparações deste trabalho.

O modelo para os dados de Fraude/Irregularidade tem seus resultados mostrados a seguir, com a mesma seqüência de figura e tabelas mostrada nos modelos anteriores. Nele o conjunto de variáveis **z** contém todas as variáveis de **x** menos a indic_trimestral_1 e a indic_trimestral_2.

Sua estrutura foi a maior dentre todas as estruturas dos demais exemplos, tendo 5 nós terminais e profundidade igual a 4, figura 5.3. A estrutura do CART tem 4 nós terminais e profundidade 3.

Para cada regime abaixo está descrito a equação dos modelos

$$\begin{aligned} Regime \quad 1 = \quad & (-2.38 + 6.4consumo - 1.48comsumo\_ano\_ant + \\ + \quad & 35.2consumo\_ano\_base + 0.28media\_3 + 0.14media\_6 - \\ - \quad & 0.78media\_12 + 0.03media\_12\_24 - 1.51indic\_anual - \\ - \quad & 0.55indic\_ajuste + 2.79indic\_tendencia - \\ - \quad & 1.62temperatura\_min + 0.95temperatura\_max + \\ + \quad & 0.59carga) * [1 - G(temperatura\_min; 3.54, 10.21)] \end{aligned}$$

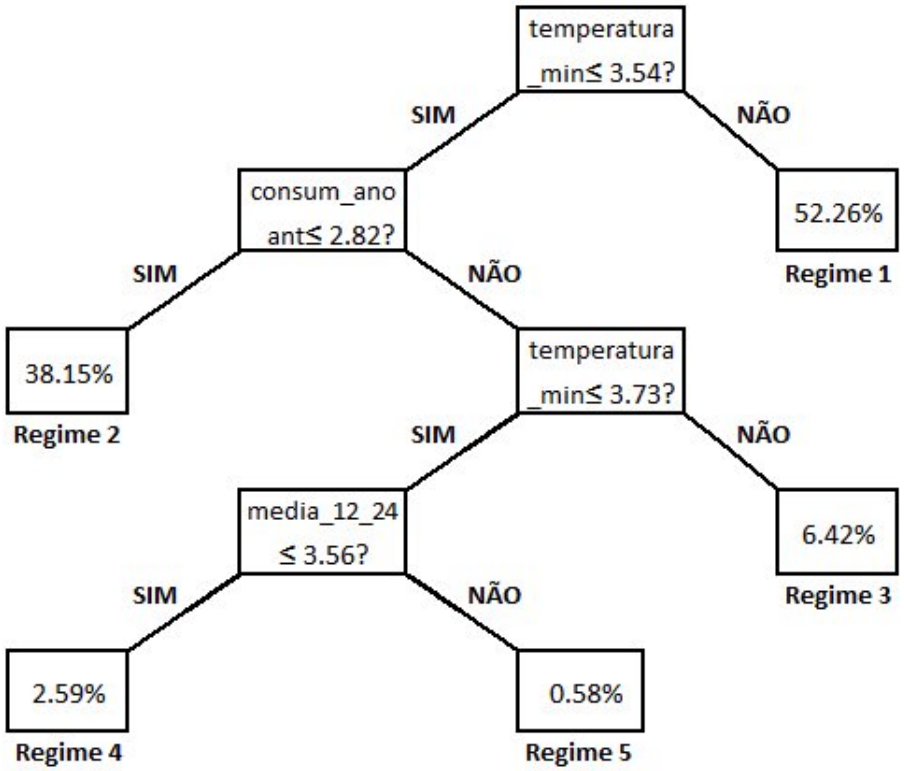


Figura 5.3: Estrutura do modelo - Consumo de Energia

$$
\begin{aligned}
Regime\ \ 2 = \quad & (-0.54 + 33.76 consumo - 23.65 comsumo\_ano\_ant - \\
- \quad & 6.64 consumo\_ano\_base + 0.46 media\_3 - \\
- \quad & 7.7 media\_6 - 0.12 media\_12 - 1.54 media\_12\_24 + \\
+ \quad & 0.15 indic\_anual + 3.56 indic\_ajuste - \\
- \quad & 4.65 indic\_tendencia + 3.62 temperatura\_min - \\
- \quad & 0.59 temperatura\_max + 0.74 carga) * \\
* \quad & G(comsumo\_ano\_ant; 2.82, 50) * \\
* \quad & G(temperatura\_min; 3.54, 10)
\end{aligned}
$$

$$
\begin{aligned}
Regime\ \ 3 = \quad & (-3.86 + 7.82 consumo - 1.65 comsumo\_ano\_ant + \\
+ \quad & 1.03 consumo\_ano\_base - 0.11 media\_3 - 0.03 media\_6 - \\
- \quad & 3.32 media\_12 + 6.5 media\_12\_24 - 5.74 indic\_anual + \\
+ \quad & 1.96 indic\_ajuste - 0.92 indic\_tendencia + \\
+ \quad & 0.6 temperatura\_min - 0.36 temperatura\_max - 4.7 carga) * \\
* \quad & [1 - G(temperatura\_min; 3.73, 50)] *
\end{aligned}
$$

$$
\begin{aligned}
& * \quad [1 - G(comsumo\_ano\_ant; 2.82, 50)] * \\
& * \quad G(temperatura\_min; 3.54, 10.21)
\end{aligned}
$$

$$
\begin{aligned}
Regime \quad 4 = \quad & (33.85 - 35.36consumo - 32.45comsumo\_ano\_ant + \\
& + \quad 15.22consumo\_ano\_base + 0.98media\_3 - 19.19media\_6 + \\
& + \quad 6.09media\_12 + 40.83media\_12\_24 - 75.25indic\_anual + \\
& + \quad 35.18indic\_ajuste - 25.79indic\_tendencia - \\
& - \quad 18temperatura\_min - 29.1temperatura\_max + 56.38carga) * \\
& * \quad G(media\_12\_24; 3.56, 100) * G(temperatura\_min; 3.73, 50) * \\
& * \quad [1 - G(comsumo\_ano\_ant; 2.82, 50)] * G(temperatura\_min; 3.54, 10)
\end{aligned}
$$

$$
\begin{aligned}
Regime \quad 5 = \quad & (-16.84 - 0.69consumo - 0.23comsumo\_ano\_ant - \\
& - \quad 0.01consumo\_ano\_base + 0.98media\_3 - 2.61media\_6 + \\
& + \quad 1.51media\_12 - 0.06media\_12\_24 - 0.72indic\_anual - \\
& - \quad 0.24indic\_ajuste - 0.15indic\_tendencia + \\
& + \quad 1.9temperatura\_min + 0.34temperatura\_max - 1.23carga) * \\
& * \quad [1 - G(media\_12\_24; 3.56, 100)] * G(temperatura\_min; 3.73, 50) * \\
& * \quad [1 - G(comsumo\_ano\_ant; 2.82, 50)] * G(temperatura\_min; 3.54, 10)
\end{aligned}
$$

A seguir está a tabela de classificação, tabela 5.11.

Tabela 5.11: Tabela de Classificação (*in sample*) - Consumo de Energia

| Observado (y) | Predito (ŷ) 0 | Predito (ŷ) 1 | |
|---|---|---|---|
| 0 | 775 | 440 | 1215 |
| 1 | 397 | 818 | 1215 |
| | 1172 | 1258 | 2430 |

Nas tabelas 5.12 e 5.13 (análise *in sample*) podemos verificar o bom desempenho do modelo na taxa de acertos total, ficando com um percentual de 68.48%, estando abaixo apenas de Redes Neurais (71.77%). Aqui GAM, que estava sempre com um melhor desempenho, detém o terceiro melhor resultado para a referida taxa, 67.74%.

Analisando ainda a tabela 5.12, nota-se que as taxas de erro total de classificação de todos os métodos foram: Redes Neurais (28.23%),

GAM (32.26%), STLR-Tree (31.52%), Análise Discriminante (39.84%), CART (40.08%), Regressão Logística (40.21%), e k-NN (48.85%).

Tabela 5.12: Comparação das Taxas de Acerto (*in sample*) - Fraude no Consumo de Energia Elétrica

| | Tx. de acertos Total | Tx. de acertos para 1 (Sensitividade) | Tx. de acertos para 0 (Especificidade) |
|---|---|---|---|
| Redes Neurais | 71.77% | 69.38% | 74.16% |
| GAM | 66.58% | 64.20% | 68.97% |
| STLR-Tree | 65.56% | 67.33% | 63.79% |
| Análise Discrim. | 60.16% | 58.85% | 61.48% |
| CART | 59.92% | 29.71% | 90.12% |
| Reg. Logística | 59.79% | 57.20% | 62.39% |
| K-NN | 54.94% | 56.90% | 53.41% |

Tabela 5.13: Métodos de Classificação Ordenados por Taxas de Acerto (*in sample*) - Fraude no Consumo de Energia

| Tx. de acertos Total | Tx. de acertos para 1 (Sensitividade) | Tx. de acertos para 0 (Especificidade) |
|---|---|---|
| Redes Neurais (0.717) | Redes Neurais (0.693) | CART (0.901) |
| GAM (0.665) | STLR-Tree (0.673) | Redes Neurais (0.741) |
| STLR-Tree (0.655) | GAM (0.642) | GAM (0.689) |
| Análise Discrim. (0.601) | Análise Discrim. (0.588) | STLR-Tree (0.637) |
| CART (0.599) | Reg. Logística (0.57) | Reg. Logística (0.623) |
| Reg. Logística (0.597) | K-NN (0.56) | Análise Discrim. (0.614) |
| K-NN (0.549) | CART (0.297) | K-NN (0.534) |

A análise *out-of-sample* aplicada aos dados de consumo de energia, consta nas tabelas a seguir, 5.14 e 5.15. O STLR-Tree tem um desempenho não tão satisfatório quanto o apresentado para a base *in sample*, tendo sua taxa de acertos total caindo de 68.48% para 56.81%.

Tabela 5.14: Comparação das Taxas de Acerto (*out-of-sample*) - Fraude no Consumo de Energia Elétrica

| | Tx. de acertos Total | Tx. de acertos para 1 (Sensitividade) | Tx. de acertos para 0 (Especificidade) |
|---|---|---|---|
| GAM | 55,19% | 55,25% | 55,15% |
| STLR-Tree | 56,81% | 40,63% | 68,70% |
| Redes Neurais | 59,54% | 46,47% | 66,77% |
| Reg. Logística | 52,98% | 40,36% | 59,96% |
| Análise Discrim. | 57,43% | 55,92% | 58,27% |
| K-NN | 50,80% | 65,46% | 42,68% |

Em cada nó terminal foram estimados 14 parâmetros totalizando 70 coeficientes calculados através do STLR-Tree para sua parte linear. Com

Tabela 5.15: Métodos de Classificação Ordenados por Taxas de Acerto (*out-of-sample*) - Fraude no Consumo de Energia Elétrica

| Tx. de acertos Total | Tx. de acertos para 1 (Sensitividade) | Tx. de acertos para 0 (Especificidade) |
|---|---|---|
| Redes Neurais (0,595) | K-NN (0,655) | STLR-Tree (0,687) |
| Análise Discrim. (0,574) | Análise Discrim. (0,559) | Redes Neurais (0,668) |
| STLR-Tree (0,568) | GAM (0,552) | Reg. Logística (0,6) |
| GAM (0,552) | Redes Neurais (0,465) | Análise Discrim. (0,583) |
| Reg. Logística (0,53) | STLR-Tree (0,406) | GAM (0,552) |
| K-NN (0,508) | Reg. Logística (0,404) | K-NN (0,427) |

esses tivemos os outros 8 parâmetros não lineares, que, juntos, somam 78 parâmetros no total. GAM tem 13 na parte linear, Regressão Logística 9 e Análise Discriminante 15, como pode ser visto na tabela D.3 do Apêndice D. No mesmo apêndice, tabela D.5, encontram-se os valores dos coeficientes não-lineares do STLR-Tree.

Os pesos das variáveis em cada um dos 8 neurônios da camada oculta calculados pela Rede Neural em um total de 128 valores estão na tabela D.4 do mesmo apêndice.

Para ilustrar o que foi mostrado anteriormente, as figuras 5.4, 5.5 e 5.6 a seguir representam os valores das taxas de erro total de cada um dos métodos comparados, que foram plotadas para cada um dos exemplos onde se pode ver claramente a posição do modelo STLR-Tree em relação aos demais.

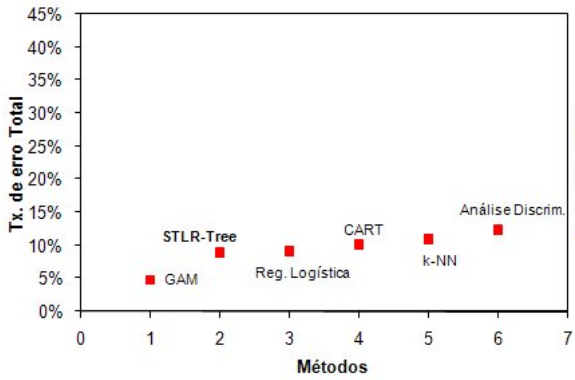


Figura 5.4: Gráfico das taxas de erro total - E-mail/Spam

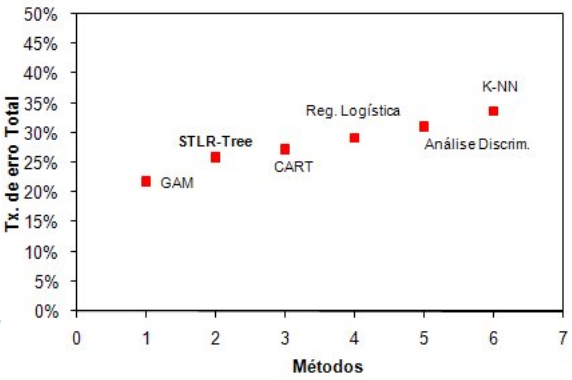


Figura 5.5: Gráfico das taxas de erro total - DCAS

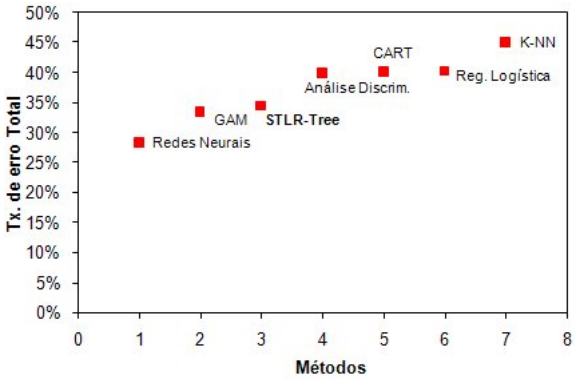


Figura 5.6: Gráfico das taxas de erro total - Fraude no Consumo de Energia Elétrica