

**Universidade do Porto
Faculdade de Economia**

**Experimentação e avaliação de modelos para um problema de
atribuição de Crédito.**

Tese de mestrado em análise de dados e sistemas de apoio à decisão

Fevereiro de 2005

**Trabalho realizado por António Sarmento
Orientação: Prof. Drº Pavel Brazdil**

Agradecimentos

Ao Professor Doutor Pavel Brazdil pela forma paciente e sabedora com que conduziu esta presente tese.

Ao Professor Doutor Alípio Jorge pelo empenho e amizade que sempre demonstrou nas horas mais difíceis.

Ao Zé Luís pelo entusiasmo e ajuda que sempre soube incutir e pelo testemunho através da partilha de experiências que permitiram relativizar as ocasiões mais difíceis.

Ao Drº Ulisses Carneiro que desde o primeiro momento se disponibilizou em fornecer os elementos imprescindíveis para a elaboração desta tese.

Pelo grupo de CVX, pelo extraordinário apoio espiritual e amigo.

E finalmente à minha Mulher Inês e às minhas três filhas, Francisca, Leonor e Isabel que tanto as privei da minha presença. Em especial a Inês que sempre soube manter o equilíbrio, sendo uma das verdadeiras obreira desta tese.

Resumo

A performance dos diversos algoritmos de “Data Mining” não depende exclusivamente destes, mas também da utilização de um conjunto de técnicas, denominadas de pré-processamento, que precedem a aplicação dos referidos algoritmos.

Assim sendo, nesta dissertação procurou-se sistematizar e experimentar técnicas que permitissem discretizar variáveis contínuas e agrupar valores de variáveis nominais oriundas de um conjunto de dados referentes a um caso de atribuição de crédito. De forma a melhorar aquelas técnicas, reformulamos algumas e propusemos alterações. Para avaliar a pertinência nas nossas propostas, efectuamos o pré-processamento dos referidos dados usando, por um lado as técnicas convencionais e por outro as nossas propostas. Em ambas situações utilizamos o Discriminante Linear de Fisher e a Regressão logística e avaliamos a performance dos modelos de forma a avaliar as nossas propostas face às técnicas convencionais de pré-processamento. A performance foi vista segundo dois modos diferentes: O primeiro pela taxa de erro de uma regra de classificação; o segundo pela discriminação dos modelos descrito pelas curvas ROC e medido pelo índice de Gini.

Abstract

The performance of the “Data Mining” algorithms does not depend exclusively on themselves, but also on the use of a group of technics called “Data preprocessing operations” which precede the application of those algorithms.

So, this thesis tries to explore and try out technics that permit discretization of continuous attributes and creation of group values on nominal attributes. To carry out those tasks we used a credit data set provided by a financial institution specialized on credit granting. In order to improve those technics, we reformulate some and propose some modifications. To see how pertinent our proposals are, we applied the preprocessing procedures on the data set, using conventional technics on the one hand, and our proposals on the other. In both situations we, posteriorly used, Fisher’s Discriminant and Logistic Regression and evaluate the performance of the models in order to judge if our proposals are more interesting. The performance was valued in two different ways: the first by measuring the error rate of a classification rule; the second by discriminatory power of the models, described by the ROC curves and measured by the Gini index.

1- Introdução.	1
2- O problema da concessão de crédito.	6
2.1 A decisão de crédito	6
2.2 Breve resenha histórica do “credit scoring”	10
2.3 Resumo de algumas metodologias tipicamente utilizadas na construção de um “credit scoring”	11
3- Técnicas convencionais utilizadas no processo de descoberta de conhecimento (“KDD process”)	17
3.1 “Data Mining” e Descoberta de Conhecimento em Bases de Dados (“Knowledge Discovery in Databases”)	17
3.2 Formulação do problema e selecção dos dados	19
3.3 Pré-processamento de dados e respectivas técnicas básicas	20
3.4 “Data mining” e respectivos algoritmos	21
3.5 Avaliação de um algoritmo e respectivas técnicas básicas	29
4- Um caso de concessão de crédito e os primeiros resultados	34
4.1 Um problema de concessão de crédito	34
4.2 Sobre o caso de crédito	34
4.3 Os Primeiros Resultados	37
5- Técnicas de Pré-processamento de dados	39
5.1 Selecção de atributos	39
5.2 Métodos de Selecção de atributos do tipo “wrapper”	43
5.3 Estratégia de Selecção de atributos Multi-variada e/ou uni-variada	43
5.4 Binarização de atributos nominais ou de atributos discretizados	44
5.5 Tratamento de valores omissos (“missing Values.”)	45
5.6 Discretização de atributos	46
5.7 Agrupamento de factores (“Grouping attributes levels”)	47
6- Proposta de melhoria de algumas técnicas de pré-processamento e experimentação das mesmas na modelização do caso prático	52
6.1 Proposta de melhoria do agrupamento de factores	52
6.2 Proposta de melhoria do “entropy-based discretization”	63
6.3 Modelização do caso prático utilizando as propostas de melhoria	67
6.4 Resultados e avaliação de performance	74
7. Avaliação de Modelos com Curvas “ROC”.	78
7.1 Deficiências na medida – Taxa de erro	78
7.2 “ROC analysis”	79

7.3 Medidas de avaliação do poder discriminatório	82
7.4 “Roc analysis” do caso prático	85
8. Conclusões	89
ANEXOS	93
Anexo 1	93
Anexo 2	95
Anexo 3	97
Referências Bibliográficas	98

1- Introdução.

É um lugar comum afirmar que “Informação é poder”. A evolução tecnológica a que temos vindo a assistir nas últimas décadas, em particular a evolução computacional, propicia o cumprimento desta máxima. Esta evolução tornou possível o registo electrónico massivo daquilo que somos e o que vamos fazendo no quotidiano, permitindo que o legítimo proprietário da informação identifique necessidades, oportunidades, interesses, enfim, conhecimento.

Se, por um lado, a tecnologia actual permite o armazenamento electrónico avultado de dados, ao ponto de no nosso planeta duplicar cada vinte meses, por outro não foram desenvolvidos mecanismos de extracção de conhecimento dos mesmos. Acontece que os dados “per si” carecem de valor, pois só terão utilidade (ou poder) se forem percebidos. Outro problema que se levanta é o de o processo de extracção de conhecimento a partir dos dados ter que ser automático, pois caso contrário, dado o volume dos dados armazenados, não seria humanamente possível fazê-lo.

O processo de extracção de conhecimento a partir de bases de dados - “KDD process”- “Knowledge Discovery Process”- proposto por Fayyad [16] e unanimemente aceite pela comunidade científica sistematiza o conjunto de etapas a percorrer com vista à obtenção do conhecimento. Tal como proposto por Fayyad, “Data Mining” é uma das etapas do processo, que sendo a mais mediática, normalmente é confundida com o “KDD process”. “Data Mining”[4] é a actividade que, pelo uso de técnicas de “Machine Learning”¹, representa os dados num modelo ou estrutura, permitindo reconhecer, automatizadamente, padrões ou regularidades, ajudando assim a produzir conhecimento.

Na área do Risco de crédito os Modelos de “Credit scoring” procuram extrair, automatizadamente, conhecimento das suas bases de dados de forma a melhor conhecer o cliente e a prever se este será um “Bom” ou “Mau” pagador. Portanto o “Credit scoring” procura responder aos mesmos desafios que o “Data Mining” e de uma forma geral do “KDD process”. Se assim é, podemos-nos questionar sobre a diferença entre um (“Credit scoring”) e os outros (“Data Mining” e “KDD process”). Uma resposta poderá ser, ambos têm objectivos comuns, mas diferenciam-se no âmbito do escopo e no conjunto de técnicas utilizadas.

Os Modelos de “Credit scoring” alicerçaram-se, desde sempre, nas técnicas estatísticas convencionais, nomeadamente na Regressão Logística. Por outro lado, o “Data Mining” procurou para além daquelas, outras linguagens de representação que aperfeiçoassem a representatividade dos modelos. As árvores de decisão, Redes Neurais, Regras de associação, e muitos outros compõem a “paleta” de modelos ou algoritmos de “Data Mining” tendo em conta o tipo de problema a resolver. Só muito recentemente os técnicos de “credit scoring”, se iniciaram na experimentação dos modelos de “Data Mining”, constatando-se que a comunhão entre “Credit scoring” e “Data Mining”

¹ Machine Learning [4] é a actividade que agrega o conjunto de técnicas destinadas a serem usadas em “Data Mining” e de uma forma geral no “KDD process”.

constitui já uma realidade. Este sinal é-nos dado por algumas publicações, que finalmente começam a aparecer, versando a temática de “Credit scoring” [1], [2].

Para além da modelização dos dados, existem outras tarefas previstas no âmbito do “Credit scoring” e do “KDD process”, nomeadamente as tarefas de pré-processamento. A grande quantidade de publicações abordando, modificando e reinventando os algoritmos de “Data Mining” tem remetido para segundo plano a importância que as tarefas de pré-processamento detêm no “KDD process” e o impacto considerável que aqueles detêm no acerto ou discriminação dos modelos. No “Credit scoring” este tipo de tarefas é considerado importante havendo uma prática unanimemente instalada. No entanto a maior parte dos procedimentos adoptados pelos técnicos de “Credit scoring”, como veremos, carecem de alguma objectividade.

Actualmente, a cada vez maior diversidade dos dados provoca verdadeiros entraves à modelização. Tanto os atributos nominais, cada vez mais frequentes, como os numéricos, necessitam de ser pré-processados, quando possuem grande quantidade de valores diferentes. A necessidade de pré-processamento surge por duas razões: em primeiro lugar, atributos com muitos valores diferentes, aumenta o tempo de processamento do algoritmo; em segundo lugar, em algoritmos de “Data Mining” que resultem da combinação linear de atributos (como por exemplo a Regressão Logística e o Discriminante Linear), nos atributos nominais com muitos valores, aumenta, pela binarização, o número de variáveis do modelo podendo provocar sobre-ajustamento do mesmo. Para além do que acaba de ser evocado existem outras razões que fundamentam a discretização de variáveis numéricas. O cumprimento da assumpção de linearidade das variáveis, face ao valor estimado da variável dependente obriga, quando necessário, à discretização das variáveis numéricas [3]. Por outro lado a existência de valores omissos, por vezes também obriga a que se discretize a variável.

As práticas de pré-processamento existentes em “Machine Learning” para resolver aqueles problemas enunciados variam consoante se trate de atributos numéricos ou de nominais. Para os atributos numéricos a constituição de intervalos, denominada de discretização é relativamente frequente sendo o “entropy-based discretizer” [4] o algoritmo de discretização mais conhecido. Para os atributos nominais a constituição de grupos de valores ou factores, denominada por Quinlan de “Grouping attribute values”² [10], constitui das poucas abordagens, disponíveis na literatura, sobre este problema. Tanto o “Entropy-based discretizer” como “Grouping attribute values” de Quinlan, são algoritmos de procura heurística do tipo “subida da colina” (Hill Climbing). Por sua vez no “Credit scoring” o “coarse Classifying” [1] é a técnica mais usual. Esta técnica usa a semelhança de “odds” para unir pequenos intervalos de valores numéricos ou agrupar factores, de modo a obter intervalos e grupos de factores homogéneos. No agrupamento de factores a semelhança conceptual entre factores é argumento de agrupamento, razão pela qual Lyn C. Thomas afirma que o “coarse classifying” é “tanto de arte como de ciência”.

Sobre as técnicas acima brevemente descritas colocam-se algumas dúvidas. Os quatro parágrafos seguintes sistematizam as questões e apresentam as respectivas soluções.

² Quinlan [10] propõe este método especificamente para árvores de decisão como o C4.5.

O algoritmo subjacente ao “Grouping attribute values” de Quinlan [10], das experiências efectuadas, no âmbito do caso prático apresentado neste trabalho, agrega os factores invariavelmente em dois grupos. Para além disso a agregação de factores é feita independentemente do número de instâncias por factor permitindo que factores não significativos, fragilizados na sua representatividade, sejam agrupados, podendo enviesar o agrupamento. Por outro lado o “coarse classifying” [1], para além de constituir um método intuitivo de agregação, não define um critério objectivo quanto ao número de grupos. Portanto, tanto o “Grouping attribute values” como o “coarse classifying” não acautelam o grau de abstracção do agrupamento. É que o mote do agrupamento é em prole da robustez e da generalização e portanto importa saber a partir de que número de grupos é um determinado agrupamento generalizável ou não. Por outro lado a generalização é facilmente obténivel, mas sempre à custa da perda de informação por via da agregação. No limite podemos transformar as variáveis nominais em binárias garantindo a generalização, no entanto nunca se sabe se com três ou quatro grupos se obtém o mesmo grau de generalização, mas com menos perda de informação.

Neste trabalho propomos duas melhorias ao método de agrupamento de factores proposto por Quinlan. Em primeiro lugar, incorporar na procura heurística um conjunto de validação independente, com o objectivo de testar em cada iteração o ganho de informação do grupo obtido no treino com o ganho de informação obtido no conjunto de validação. Assim, através de uma determinada estatística de teste poderemos aferir a generalização do grupo obtido no treino, segundo a métrica de avaliação – ganho de informação. O algoritmo pára quando um determinado P-value (previamente definido pelo decisor) resultante da estatística de teste for atingido. Em segundo lugar, de forma a combater o enviesamento introduzido pelos factores não significativos preconizamos a criação de um factor “BOLO” que agrega os factores menos frequentes. Ensaia-se vários factores “BOLO”, ordenando a frequência relativa dos factores por ordem crescente e estabelecendo vários cortes no sentido ascendente. Para cada corte efectua-se normalmente o agrupamento de factores. O corte vencedor será aquele cujo agrupamento detiver maior ganho de informação.

Foram efectuadas experiências utilizando o “Entropy-based discretizor” tendo-se constatado que a eleição do ponto de corte (processo idêntico à determinação de um teste num nó de uma árvore de decisão C4.5), para a constituição do intervalo, é um processo demorado por vezes impraticável. Para além disso, o critério de paragem inerente ao algoritmo (MDL - Minimum description length) frequentemente pára na primeira iteração, resultando na não discretização do atributo. Alguns autores [4] e [9] alertam que nem sempre o MDL resulta, enunciando as situações em que tal acontece.

Assim, são também propostas duas melhorias ao “Entropy-based discretizor”. Em primeiro lugar, de forma a diminuir o tempo de processamento do algoritmo, procede-se à pré-discretização da variável em dez percentis e procede-se de seguida à junção de intervalos adjacentes ou “vizinhos”, utilizando o mesmo princípio do agrupamento de factores. Atingindo o P-value, previamente definido o algoritmo pára. Em segundo lugar o critério de paragem, tal como atrás foi dito, passa a ser o mesmo que o agrupamento de factores, visto que o MDL nem sempre discretiza.

O objectivo do presente trabalho consiste em avaliar as duas propostas referentes ao agrupamento de factores e à discretização de variáveis numéricas, por forma a verificar se beneficiam a performance dos modelos de Regressão Logística e do Discriminante Linear, comparativamente com as restantes técnicas de pré-processamento em uso.

Assim, no segundo Capítulo abordamos questões relacionadas com o crédito ao consumo em Portugal, nomeadamente a evolução da decisão de crédito e a sua relação com o “Credit scoring”. Referimos também as práticas tipicamente aplicadas em “Credit scoring” relacionadas, entre outras, com a selecção da população de desenvolvimento, a maturidade da população e a definição da variável a explicar.

No terceiro Capítulo fazemos referência ao “KDD process” e às suas fases mais importantes. São referidos técnicas simples de pré-processamento, e apresentados os algoritmos de “Data Mining” utilizados neste trabalho. A Regressão Logística e o Discriminante Linear são os algoritmos de “Data Mining” a experimentar no âmbito do presente trabalho.

No quarto Capítulo descrevemos o caso prático e apresentamos os primeiros resultados produzidos pela aplicação das técnicas referidas no terceiro Capítulo. Relativamente ao caso prático inventariamos e descrevemos os atributos, explicamos como se procedeu relativamente à selecção da população de desenvolvimento, e explicamos o método pelo qual se definiu a variável a explicar. Apresentamos os primeiros resultados das taxas de erro respeitantes à Regressão Logística e ao Discriminante Linear, cujo pré-processamento consistiu apenas na binarização das variáveis nominais. As respectivas taxas de erro foram estimadas a partir de um “holdout group test” que representa 30% do conjunto de dados.

Dada a importância do pré-processamento no presente trabalho, dedicamos o quinto Capítulo a descrever as técnicas de pré-processamento mais comuns, tais como: Selecção de atributos; Binarização de atributos nominais ou discretizados; Tratamento de valores omissos; Discretização de atributos e agrupamento de factores.

No sexto Capítulo apresentamos as nossas propostas relativas ao agrupamento de factores e à discretização e procedemos à modelização do caso prático utilizando as seguintes técnicas de pré-processamento: Agrupamento de factores (segundo nossa proposta); Discretização de atributos numéricos (segundo a nossa proposta); Binarização das variáveis nominais ou discretizadas; selecção de atributos. No final procede-se ao treino e teste dos modelos de Regressão Logística, Discriminante Linear. No sentido de avaliar o impacto das nossas propostas procedemos à modelização do caso prático utilizando duas estratégias usuais de pré-processamento: A primeira contendo apenas a binarização de variáveis nominais (já modelizada no Capítulo quarto); A segunda contendo a binarização de variáveis nominais e a Selecção de atributos utilizando o “Stepwise Foward”. Procedeu-se à avaliação das três estratégias de pré-processamento a partir da taxa de erro dos respectivos classificadores obtidos a partir da Regressão Logística e do Discriminante Linear, tendo sido estimada num “holdout group test” equivalente a 30% do conjunto de dados.

No sétimo Capítulo explicamos um modo alternativo para avaliação de classificadores, denominado de “ROC analysis”. Descrevemos, também as métricas que avaliam o poder discriminante dos modelos. À luz destes conceitos, avaliamos as três estratégias de pré-processamento. Assim sendo, iremos avaliar os classificadores segundo duas métricas, o erro e a discriminação.

Por último gostaríamos de referir que todo trabalho de pré-processamento, binarização de variáveis, os algoritmos de discretização e de agrupamento de factores, selecção univariada de variáveis e estimação dos parâmetros para as Regressões Logísticas e Discriminantes Lineares foi efectuado pela criação de funções no software “R”. Pontualmente foi utilizado SPSS e quando assim é vem referenciado.

2- O problema da concessão de crédito.

Pretendemos que este capítulo informe o leitor sobre:

- O processo de decisão de crédito e a sua evolução;
- História do “credit scoring”;
- Algumas metodologias tipicamente aplicadas na construção de um “credit scoring”.

2.1 A decisão de crédito

Em Portugal a concessão de crédito, até 1989, era maioritariamente efectuada pela Banca de retalho essencialmente vocacionada para a concessão de crédito a Empresas. No que toca ao crédito ao consumo³ este era normalmente concedido pelos próprios comerciantes, que com o propósito de aumentar as suas vendas, facultavam “facilidades de pagamento” através de cheques pré-datados ou letras. Para este efeito o comerciante usava regras de bom-senso para a atribuição de “fiado”. Se o cliente candidato a crédito fosse assíduo ou conhecido ou residente nas proximidades do estabelecimento então o manancial de informação disponível era maior pelo que o teor daquela determinava a atribuição, ou não de crédito. Por outro lado, se o cliente candidato fosse desconhecido ou residente em local geograficamente distante, na falta de informação qualitativa do candidato o comerciante socorria-se junto de outros comerciantes no sentido de classificar a experiência de crédito do cliente. Também por vezes, quebrando o sigilo bancário, contactava o banco do cliente para obter informações (Saldo médio da conta bancária, existência ou não de incidentes de crédito ou cheques sem provisão).

Se por um lado o comerciante incentivava as suas vendas através da concessão de “facilidades de pagamento”, por outro, as suas necessidades de fundo de maneo aumentavam, tendo que descontar, junto da banca, os títulos aceites pelo cliente. Esta para além de se fazer cobrar de uma elevada comissão de desconto, ficava isenta de qualquer risco de incumprimento do cliente, ficando esse do lado do comerciante.

Inequivocamente o comerciante foi o 1º decisor de crédito ao consumo. A história e a actualidade estão repletas de exemplos da constituição de Sociedades financeiras no seio de grupos da grande distribuição. Tal facto comprova que a actividade creditícia e a actividade mercantil são inseparáveis sendo válido quer se trate de um comerciante de “bairro”, quer se trate de um hipermercado.

Em finais de 1989, o então ministro das finanças Miguel Cadilhe, sentindo a necessidade de aumentar a procura interna, promove o crédito ao consumo, através da criação das “SFAC” – Sociedades financeiras de aquisição a crédito.

³ Entende-se por Crédito ao consumo todo aquele que é concedido a particulares e que não se destine ao suporte de qualquer actividade empresarial ou geradora de valor. O crédito hipotecário à habitação não se considera Crédito ao consumo.

Sendo instituições financeiras, com capacidade de concessão de crédito e com exigências substancialmente inferiores de capital social mínimo obrigatório, comparativamente com a banca, estas proliferaram a partir dessa data, indo substituir o comerciante no seu papel de fornecedor creditício, aliviando-o do risco e das necessidades de fundo de Maneio. De 1989 a 1995, pese embora as restrições do decreto-lei de vendas a prestações, as SFAC's, incrementaram consideravelmente o crédito ao consumo concedido, essencialmente no mercado das viaturas novas e usadas.

A criação das SFAC's inaugurou um novo processo de relacionamento, desta feita tripartido: Cliente, Comerciante e SFAC. Esse relacionamento encontra-se ilustrado na Figura 1

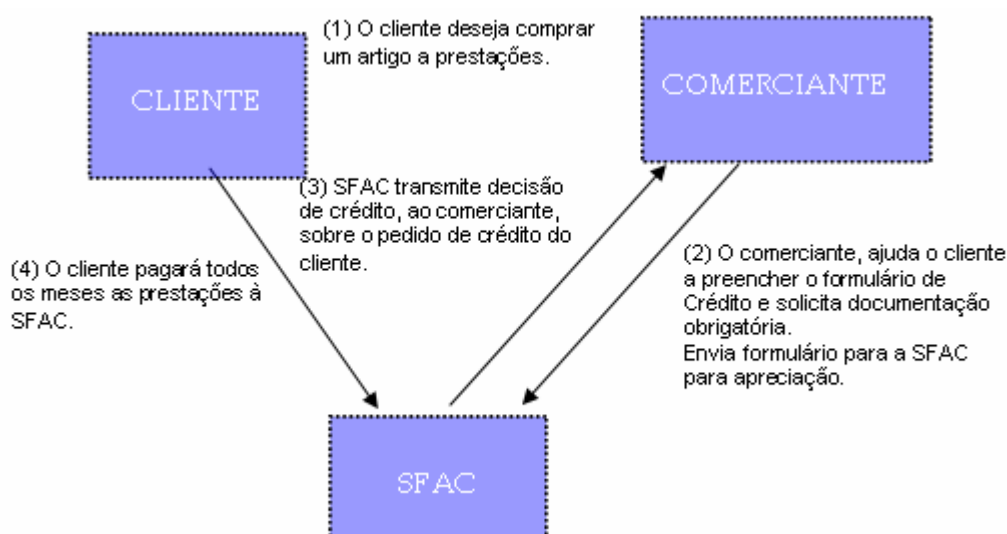


Figura 1 – Novo Circuito na venda a prestações

Até, aproximadamente, 1995 o processo decisório era essencialmente intuitivo estruturando-se no “feeling” e na experiência dos analistas de crédito. Habitualmente eram analisadas variáveis, tais como a taxa de endividamento (ratio entre os custos mensais do agregado familiar e o respectivo ordenado líquido); Bens móveis e imóveis do agregado; Profissão; tipo de contrato de trabalho; estabilidade no emprego averiguável pela antiguidade na entidade patronal; Nível dos saldos médios nas contas bancárias do cliente; Entrada inicial face ao Valor de preço de venda ao público e Idades dos proponentes. Tendo em conta estes parâmetros, o analista, fazia o seu julgamento ponderando os prós e os contras, colocando-os numa “balança mental”.

Após 6-7 anos de actividade das SFAC's, constata-se:

- Os primeiros incumprimentos começam-se a avolumar, e o número de contratos a entrar em contencioso cresce continuamente.
- Face ao número cada vez crescente de SFAC's a competitividade, entre elas, aumenta, sendo o tempo de decisão, bem como a taxa de juro,

factores distintivos. Com a baixa das taxas de juro a gestão dos custos operacionais torna-se imprescindível e quanto maior o volume de negócios, mantendo-se o esquema decisório acima referido, igual incremento sofrerá a equipa de analistas de crédito.

- O “feeling” do analista, varia conforme o analista, não havendo um padrão consistente de critérios de aceitação/análise. Por sua vez, em muitas SFAC’S o analista acumula a função de comercial, e a pressão dos comerciantes para a aprovação dos negócios é cada vez maior.
- Enfim, o gestor sente-se confuso, pois pensava que tinha a melhor equipa de analistas, e no entanto não consegue medir, tão simplesmente a probabilidade de um cliente, com determinado perfil, entrar em atraso ao fim de 12 meses, após celebrar o contrato de crédito. Mesmo que queira diminuir a concessão de crédito em x% para a prazo diminuir o risco, não o conseguirá fazer adequadamente, pois os parâmetros de análise estão na “cabeça do analista”, que sempre decidiu de uma determinada forma e dificilmente decidirá doutra.

Perante este cenário e sentindo a necessidade da profissionalização da gestão do risco, os gestores começaram a adoptar, algumas técnicas cujo sucesso além fronteiras se encontrava devidamente comprovado. Tanto o Reino Unido, E.U.A, França como inclusivamente a própria Espanha, já concebiam modelos estatísticos com o intuito de otimizar a decisão de crédito no sentido de aceitar a maior quantidade de clientes com o menor risco possível. Apelidados de “credit scoring”, estes modelos constroem-se com base em informação histórica onde a entidade cliente se encontra multi-variadamente caracterizada, por variáveis sócio demográficas (e.g. Idade; Profissão, estado Civil, salário, e outras) e classificados em “Bom” ou “Mau” pagador em conformidade com a experiência de crédito. Muito simplesmente o “credit scoring”, ou o modelo estatístico que lhe está associado, transforma a realidade multi-variadamente caracterizada – “entidade cliente” –, numa escala uni variada de valores, ou eixo, em que a entidade cliente é sintetizada num “score” (pontuação) daquela mesma escala. Geometricamente falando, é como se estivessemos a projectar um ponto representado num espaço multi-dimensional, com tantos eixos quantas as variáveis, num único eixo ou escala. Dada uma instância, quanto maior o valor do seu “score” nessa escala, maior a probabilidade de o cliente ser “Bom” pagador. Esta nova escala ou eixo, é o resultado da combinação linear dos atributos ou das variáveis sócio/demográficas e é constituída por tantos pontos, quantos os “scores” de todas as instâncias. A combinação linear representa-se da forma:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \dots + \beta_n X_n \quad (1)$$

Cujos pesos (ou betas) a atribuir a cada variável, resulta de um processo de estimação que tenta maximizar a separação entre Bons e Maus pagadores.

Para melhor compreensão do leitor, considere o seguinte exemplo:

Após reunião de toda informação histórica sobre os atributos que caracterizam os clientes, e após a categorização de cada um em “Bom” ou “Mau”, segundo uma regra definida pelo gestor, estimam-se os seguintes parâmetros para as variáveis: X_1 Número de dependentes; X_2 Valor da entrada inicial; X_3 Prazo da operação (meses) e X_4 Antiguidade da conta bancária (anos).

Quadro 1-Parâmetros ou pesos da Regressão

Pesos
$\beta_0 = -1,628$
$\beta_1 = -0,178$
$\beta_2 = -0,00311$
$\beta_3 = 0,099$
$\beta_4 = 0,063$

Consideremos 2 clientes, caracterizados da seguinte forma multi-variada e sobre os quais desconhecemos a sua situação face ao incumprimento:

Quadro 2 – Dois Clientes caracterizados pelas variáveis

Cliente \ variável	Nº Dependentes X_1	Valor da Entrada inicial X_2	Prazo da operação X_3	Antiguidade da conta bancária X_4
Cliente1	0	500,00 €	4	20
Cliente2	2	100,00 €	6	10

O valor total dos Y ou o score de cada cliente, resultante da soma dos pesos descritos no Quadro1 multiplicados pelos respectivos valores das variáveis constantes no Quadro 2, traduz-se no quadro seguinte:

Quadro 3 – Os “scores” dos dois Clientes

Clientes	“Score”
Cliente 1	0,78823
Cliente 2	-1,64071

A cada score, corresponde uma determinada probabilidade de ser “Bom”, pressupondo que os dados foram extraídos de uma população que segue uma determinada lei de probabilidade multi-variada. Assim sendo, dada as características do cliente 1 a probabilidade⁴ de este ser “Bom” é de 68,75%, contra 16,24% do cliente 2 ser “Bom”. Assim sendo o cliente 1 tem mais probabilidade de ser Bom pagador que o cliente 2.

Uma vez definida a escala de valores, definem-se intervalos de “scores” (também denominadas de “tranches”), em que em cada intervalo perfaz um percentil de 10% e associa-se a cada um deles a taxa de sinistralidade dos clientes que pertencem a esse

⁴ As probabilidades de 68,75% e 16,24%,determinam-se exponenciando os respectivos “scores” na equação da regressão logística expressa em (2.2) e explicada na secção 3.4.2

intervalo. Necessariamente à medida que aumenta o “score” menor a taxa de sinistralidade. Assim sendo o Gestor poderá fixar a sua estratégia de aceitação ou de análise de crédito para futuras propostas de crédito, fixando muito simplesmente um “cut-off”, que significa, o valor na escala acima do qual aceito a proposta de crédito, caso contrário recuso. A fixação da estratégia de aceitação, através da definição do “cut-off”, é feita seguindo o pressuposto que a população que foi alvo de modelação é estatisticamente semelhante às populações vindouras.

2.2 Breve resenha histórica do “credit scoring”

Nos anos 30, algumas empresas de venda por correspondência (“Mail order Companies”) de Chicago, introduziram “scorings” numéricos [1], com o objectivo de resolver disparidades ou inconsistências dos analistas de crédito.

Em 1936 o método de discriminante linear de Fisher [6], constitui uma das primeiras abordagens, no domínio da estatística, na predição de grupos, dado um conjunto de características. Fisher conseguiu prever a qual das três espécies da planta “íris” pertencerá, uma dada planta com determinadas características ou variáveis. Para tal predição, Fisher, constituiu a função discriminante como combinação linear das variáveis, às quais está associado um vector de pesos, cuja estimação resulta da maximização do quociente entre dispersão entre grupos e a dispersão intra grupos. Quanto maior for este quociente, mais a função discriminante separa os grupos.

Logo imediatamente a seguir ao final da segunda Guerra Mundial, o método discriminante de Fisher [6], passou a ser a técnica de eleição para o desenvolvimento de “credit scoring”. A primeira empresa de consultadoria a se especializar na concepção de “credit scorings”, foi formada em São Francisco por Bill Fair e Earl Isaac.

Em 1960, com a banalização dos cartões de crédito (especialmente nos E.U.A.), fez com que os bancos, bem como outros emissores de cartões de crédito, identificassem utilidade do “credit scoring”, pois dada a volumetria de pedidos de adesão o esquema clássico de análise de crédito tornava-se insustentável, pois os recursos teriam que aumentar significativamente.

O “credit scoring”, não se esgota na função de substituir o analista de crédito, automatizando a decisão. A verdade é que com o aumento de pedidos de adesão de cartões de crédito e consequente aplicação do “credit scoring”, concluiu-se que este é melhor classificador que o julgamento humano, que às vezes cai em preconceitos que não tem qualquer cabimento estatístico. Assim sendo constatou-se que a taxa de maus pagadores [1] caiu para valores superiores a 50%.

Em 1980, face ao sucesso do “credit scoring” no domínio dos cartões de crédito, os bancos expandiram o âmbito da sua aplicação para empréstimos pessoais. Foi também em 1980 que a regressão logística destronou o método discriminante linear de Fisher, entretanto a linguagem de representação mais usada [1].

Em 1990, com a revolução do Marketing Directo, os princípios do “credit scoring” passaram a ser aplicados para melhorar a taxa de resposta das campanhas de “Direct Mail”. Neste caso concreto o conceito alvo muda de Bom/Mau pagador para Responde/Não responde ao “mail”. Aliás, esta técnica já teria sido utilizada pela Sears, em 1950, com o objectivo de determinar para quem enviar o catálogo.

Embora o “credit scoring” seja eminentemente aplicado ao crédito a particulares, não poderemos deixar de referenciar o trabalho efectuado por Altman (1968), no sentido de experimentar e desenvolver “credit scoring” para prever o risco de falência das empresas. Embora este trabalho tenha produzido resultados interessantes quanto à associação de determinados rácios com o risco de falência, no entanto a capacidade preditiva dos modelos era substancialmente inferior ao “credit scoring” para particulares.

Actualmente os princípios do “credit scoring” são aplicáveis em todas as fases do ciclo de vida do cliente de crédito, a seguir enunciadas:

- Recrutamento do cliente – “scoring” de abertura ou de aceitação;
- Aumentos de “plafond” ou atribuição de mais crédito – “Behaviourial scoring” ou score de comportamento, em que, tradicionalmente as variáveis raramente são de cariz sócio/demográfico, mas sim de comportamento do cliente face ao crédito;
- Marketing directo: “Response score”, já comentado acima;
- Recuperação de crédito ou cobrança de atrasos: “Collections scoring”, no sentido da eficácia, visa criar prioridades no trabalho de cobrança, segmentando a carteira de atrasos por gravidade e adequar a afectação de recursos em função da probabilidade de incobrabilidade de um determinado cliente em mora.

2.3 Resumo de algumas metodologias tipicamente utilizadas na construção de um “credit scoring”

Não se pretende, nesta Secção, descrever exhaustivamente todo o processo de construção de “credit scoring”, mas apenas explicar as partes ou fases que o tornam singular comparativamente às modelizações convencionais.

2.3.1 Selecção da população de desenvolvimento

Quanto maior a dimensão dos dados e das características maiores garantias teremos de obter um modelo eficaz. Tal como veremos a seguir, a dimensão dos dados determinará o modo como se selecciona a população de desenvolvimento, bem como a definição do conceito alvo de “Bom” e “Mau” pagador. Entende-se por população de desenvolvimento, o conjunto de dados onde o algoritmo efectuará a sua aprendizagem ou treino. Tanto quanto pudemos apurar, os “scoring makers”, no tocante à selecção da população de desenvolvimento, utilizam a metodologia descrita nas duas Secções seguintes:

- **Apuramento da maturidade de uma população.**

Para o que a seguir se explica, importa reter a noção nuclear de exposição. Entende-se por exposição ou tempo de exposição o tempo decorrido desde o início do contrato até à presente data. Utiliza-se a expressão exposição, para definir o tempo durante o qual o cliente revela o seu comportamento face ao (in) cumprimento das prestações mensais.

O interesse em determinar a exposição de um cliente ou de um conjunto, prende-se com a necessidade de apurar, dentro do período de exposição, qual o momento a partir do qual a taxa de incumprimento, de um conjunto de clientes ou o incumprimento de um cliente, não evolui mais. A este momento denominamos por maturidade de um cliente ou de um conjunto de clientes. Conhecendo assim a maturidade da população, estamos em condições de seleccionar a população de desenvolvimento constituída pelos clientes cuja maturidade é igual ou superior à maturidade global e que portanto seguramente os poderemos classificar em “Bons” ou “Maus” pagadores. Se o tempo de maturidade de uma população é reduzido, então a probabilidade de se obter uma população de desenvolvimento numerosa é alta, sendo o contrário obviamente mais difícil.

No entanto a complexidade aumenta se considerarmos que dentro de um conjunto de clientes podemos ter vários subconjuntos de clientes com maturidades substancialmente diferentes, normalmente associados a meses de abertura diferentes. Posto isto, deveremos aferir a semelhança das maturidades por meses de abertura, pois caso não sejam homogéneas incorremos no risco de integrar clientes na população de desenvolvimento cuja a classificação de “Bons” ou “Maus” poderá não ser correcta, perigando a eficácia do modelo obtido. Assim sendo é normalmente construído, à boa maneira da estatística descritiva, um gráfico – “Delinquency dynamic report”, que tenta relacionar o Mês de abertura do cliente, o tempo de exposição e a taxa de incumprimento.

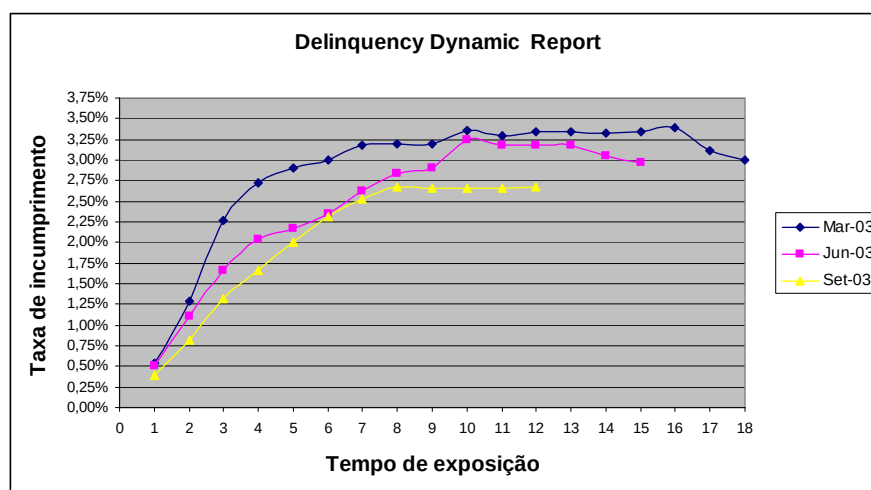


Figura 2 – DDR-“Delinquency Dynamic Report”

Na Figura 2 vemos representados, no eixo das ordenadas a taxa de incumprimento, que se define pelo quociente entre o nº de clientes, abertos⁵ num determinado mês, com mais de 2 mensalidades em atraso sobre a totalidade de clientes abertos nesse mês. A taxa de incumprimento é medida tendo em conta todos os tempos de exposição possíveis de um cliente. Cada linha na Figura 2 representa o “Dynamic delinquency” de cada mês de abertura. Pela tendência das linhas observamos que são semelhantes à curva do ciclo de vida do produto: (1) Arranque; (2) Expansão e (3) Estabilização. A estabilização corresponde à maturidade, que na Figura 2 se observa a partir do décimo mês de exposição. Daqui se conclui que a população, para este caso concreto, a ser seleccionada, terá que ter um período de exposição aproximadamente de 10 meses, ou em média de 10 meses. Alguns autores, tais como Murray Bailey [2] e Lyn C.Thomas [1], sugerem maturidades standard de 12 a 18 meses.

- Estratégias de selecção da população tendo em conta a maturidade.

Quanto à selecção da população de desenvolvimento, tendo em conta a maturidade da carteira de crédito, normalmente praticam-se 2 estratégias:

- Tempo de exposição fixa

Selecciona-se a população obedecendo ao princípio que todos os meses de abertura têm uma exposição igual à maturidade da carteira.

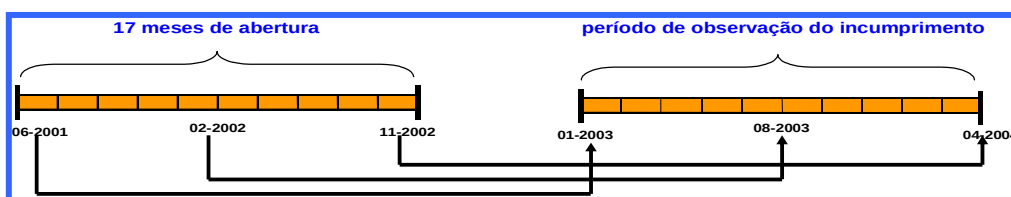


Figura 3 – Tempo de exposição fixa

Vantagens: Garante-se para todos os meses, uma exposição semelhante proporcionando equidade na observação do incumprimento para cada mês de abertura. A exposição é igual à maturidade da carteira

Desvantagens: Obriga a um universo de clientes maior e mais antigo.

- Tempo de exposição móvel

Selecciona-se a população que em média terá uma exposição aproximadamente igual à maturidade.

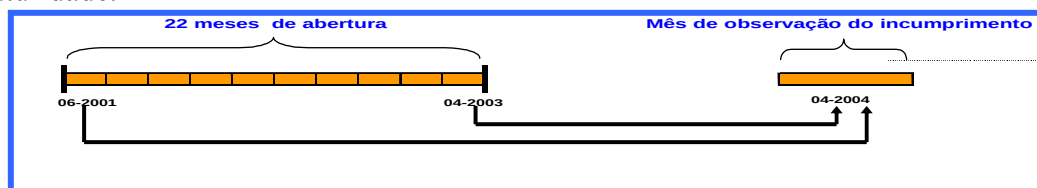


Figura 4 – Tempo de exposição móvel

⁵ Este termo é utilizado para referir os clientes contratados num determinado periodo de tempo.

Vantagens: Garantimos uma população mais numerosa e mais recente. Permitindo que o ciclo de vida do modelo seja maior, pois tem em conta população mais recente

Desvantagem: Existência de uma parte da população com exposição inferior à maturidade.

2.3.2 Definição do conceito alvo (ou definição variável dependente)

É unânime quer nos autores (Lyn C. Thomas [1] e Murray Bailey [2]) quer nos “scoring makers”, classificar os clientes, quanto ao incumprimento, da seguinte forma:

- É considerado “Mau” aquele que, num determinado período de observação, detém ou deteve mais de duas mensalidades em atraso;
- É considerado “Bom” aquele que, num determinado período de observação, não tem ou não teve mensalidades em atraso;
- É considerado “Indeterminado” aquele que, num determinado período de observação, detém ou deteve uma ou duas mensalidades em atraso;

A dimensão do período de observação é um parâmetro importante a considerar. Se o (i) período de exposição disponível não é suficientemente alargado para se poder apurar, com segurança, a maturidade, ou se a (ii) população disponível é escassa, é aconselhável observar o risco de cada cliente em mais do que um mês (consecutivo). Se (i) e (ii) não se verificarem, os “scoring makers” tendencialmente optam pela redução, a um mês, do período de observação do risco de cada cliente.

Importa sublinhar que os Indeterminados, são retirados da população de desenvolvimento.

2.3.3 Inferência dos recusados (“Reject inference”)

Um dos maiores problemas no processo de desenvolvimento de modelos de “scoring” para a gestão de atribuição de crédito, consiste na evidência que somente as propostas de crédito aprovadas e concretizadas, poderão ser classificados como: “Bons”, “Maus”, “indeterminados”. Para as propostas recusadas, no passado, apenas detemos as variáveis dos proponentes, mas obviamente não possuímos a informação de “Bons” ou “Maus”. Se estes clientes, recusados, forem ignorados e retirados da população de desenvolvimento, provocará um “bias”, quando o modelo, construído sobre os “Bons” e os “Maus”, classificar um proponente recusado anteriormente. Àquele “bias” denomina-se de “viés dos recusados” (“reject bias”). Assim sendo, e no sentido de ultrapassar o “reject bias”, [1] e [2] sistematizam o seguinte conjunto de técnicas:

- **Extrapolação (“Extrapolation”)**

Este método foi sugerido por Hand e Henley (1993) [1], aparentemente simples e eficaz, não nos parece que esteja a ser aplicado com frequência pelos “scoring makers”.

Para se poder aplicar esta técnica é condição obrigatória que o conjunto de variáveis pertencentes ao antigo modelo de “scoring” (X_{Old}) esteja contido no conjunto de variáveis a modelar (X_{New}).

Utilizando somente as variáveis antigas (X_{Old}), dividimos a população de desenvolvimento em 2 partes: Na primeira, os recusados e sobre os quais não temos informação se são “Bons” ou “Maus”; Na segunda, os aceites sobre os quais temos informação de “Bons” e “Maus”.

De seguida e utilizando as variáveis antigas, modelamos os bons e os maus e usamos o resultado deste modelo para aferir a probabilidade de um recusado ser “bom”.

Assim sendo a partir do momento que podemos extrapolar a Probabilidade de um recusado ser “Bom”, podemos facilmente integrá-lo na população de desenvolvimento e ajustar o Modelo, então elaborado somente para os bons e para os maus.

- **“Augmentation”**

Este método um pouco mais complexo e trabalhoso, normalmente é usado quando a premissa X_{Old} não está contido em X_{New} .

Passamos a descrever os passos:

- Construir um modelo sobre os bons e Maus, somente para a população aceite, no sentido de estimar a seguinte probabilidade condicionada:

$P(\text{Bom} | x \text{ dado que é aceite})$, isto é, dada a instância x e dado que foi aceite, qual a probabilidade de o respectivo cliente ser “Bom” pagador

- Construir um modelo, cuja variável dependente passa a ser “Aceite” ou “Recusado”, com o objectivo se obter a seguinte probabilidade condicionada:

$P(\text{Aceite} | \text{“score” } s \text{ da instância } x)$, isto é, dado um determinado “score” s (do modelo aceite / recusado) da instância x qual a probabilidade de ser aceite.

- Hsia (1978), segundo Lyn C.Thomas [1], estabelece, como premissa, a seguinte igualdade de probabilidades condicionadas:

$P(\text{Bom} | \text{“score” } s \text{ de } x \text{ dado que é recusado}) = P(\text{Bom} | \text{“score” } s \text{ de } x \text{ dado que é aceite})$, isto é, para o mesmo “score” s a probabilidade de um recusado ser Bom é igual à probabilidade, de um aceite ser Bom.

Assim sendo aos recusados com “score” s (do modelo aceite / recusado), será atribuída a $P(\text{Bom} | \text{score} = s)$ dado que é aceite).

- **Outras Técnicas**

Existem outras abordagens possíveis:

- Dada a impossibilidade de aferir, precisamente, a $P(\text{Bom} | \text{score})$, isto é, dado um determinado “score”, que implica a recusa, qual a probabilidade de ser “Bom”, a solução passa pelo seu efectivo apuramento, aceitando clientes que em circunstâncias normais seriam recusados. O problema está em determinar se o ganho de acerto do novo Modelo compensará os riscos envolvidos na aceitação de clientes de Risco?
- Outra hipótese será, sobre os clientes recusados numa determinada instituição financeira e se aprovados noutra, obter informação sobre o seu comportamento no pagamento.

Finalmente, é considerar todos os que foram recusados como Maus. Esta prática nada aconselhável, ainda colhe adeptos, actualmente. No entanto estas abordagens dado o seu custo ou exequibilidade, não são bem encaradas pelos gestores.

3- Técnicas convencionais utilizadas no processo de descoberta de conhecimento (“KDD process”)

Pretendemos que este capítulo informe o leitor sobre:

- Breve abordagem ao conceito de “Data Mining” e à sua relação com *Descoberta de Conhecimento em Bases de Dados* (“Knowledge Discovery in Databases”);
- Breve descrição das fases mais importantes do “KDD process” bem como das respectivas técnicas utilizadas.

3.1 “Data Mining” e *Descoberta de Conhecimento em Bases de Dados* (“Knowledge Discovery in Databases”)

Nos nossos dias, graças às tecnologias computacionais, somos capazes de guardar dados em grande quantidade. Alguns dos nossos comportamentos quotidianos estão guardados e muitos deles nem imaginamos que estão a ser alvo de arquivo electrónico ou de tratamento informático (independentemente da legitimidade do uso dos dados). Vejamos alguns exemplos:

- Quando passamos com a nossa viatura na “Via verde”, a Brisa tem possibilidade de apurar a regularidade com que passamos numa determinada portagem, as horas de ida e de regresso, a velocidade média, etc.
- Quando utilizamos o nosso cartão de crédito ou cartão de débito, é possível apurar a regularidade de compra de um indivíduo, em que loja, a que horas e no limite associar o código da compra ao número de transacção de forma associar o (s) artigo (s) que comprou, etc.
- Quando efectuamos um telefonema, na rede fixa, e em alguns casos na rede móvel, é possível apurar os hábitos de uso do telefone, a que horas do dia, com que frequência, a duração das chamadas, para onde falamos e para que nº de telefone falamos, etc.

Estima-se que actualmente a quantidade de dados electronicamente armazenados no nosso planeta dobra de 20 em 20 meses [4]. No entanto os dados “per si” carecem de utilidade, pois só terão valor ou utilidade se forem percebidos. O grande volume de dados disponível electronicamente, tem sido possível graças à evolução das tecnologias computacionais, mas contrariamente aquela tendência, estes dados não estão a ser transformados em conhecimento ou informação. Para o serem, o processo de transformação de dados em conhecimento, terá que ser igualmente automático, pois caso contrário não seria possível tratar todos os dados.

Genericamente o processo de transformação de dados brutos ou de “inputs”, em informação ou conhecimento, comporta uma série sequencial de tarefas (descritas nas secções seguintes) a começar na selecção de um conjunto de dados que caracterize o

fenómeno a estudar e a acabar na procura ou reconhecimento de padrões ou similitudes nos e entre os dados seleccionados.

Este processo de transformação, provavelmente fazemo-lo todos os dias de forma intuitiva, quando tentamos identificar padrões no comportamento dos clientes, de forma a reorientar as abordagens de marketing, ou quando o vitivinicultor tenta identificar padrões nas variadas formas de podar a vinha com o grau obtido na colheita seguinte. Desta forma, os padrões reconhecidos irão nos ajudar, perante novos dados ou “inputs”, a predizer ou no mínimo especular sobre o que vai acontecer. Assim o processo automático de reconhecimento de padrões, uma vez terminado, e averiguada a sua capacidade de abstracção, culminará necessariamente na predição de novos dados ou “inputs”.

Assim “data mining”, numa primeira abordagem, poder-se-á considerar uma das actividades que participa no processo de captura de conhecimento dos dados, procurando, automatizadamente, o reconhecimento de padrões nos dados armazenados electronicamente.

Para que o “data mining” promova, eficazmente, a compreensão dos dados e/ou a predição de novos, é importante que os padrões reconhecidos, bem como a forma como são expressos, reúnam duas condições: Em primeiro lugar os padrões reconhecidos deverão descobrir informação útil, evitando o reconhecimento desnecessário de padrões triviais. Por padrão trivial entende-se aquele que constata informação inútil, tal como a seguir se exemplifica: “quem não compra pentes são os carecas”; “quem não vai ao obstetra são os homens”. Em segundo lugar os padrões deverão ser expressos de forma inteligível, ou seja, representados numa estrutura que possa ser examinada, reflectida e usada em futuras decisões. Ian H. Witten e Eibe Frank [4] qualificam esta faculdade da expressão dos padrões como *estruturantes* e associam o modo de expressão de padrões à opacidade da caixa onde são gerados.

Importa reiterar que tudo o que atrás referimos acerca de “data mining” não esgota o conceito de *descoberta de conhecimento em dados armazenados electronicamente*, pois tal como a seguir se descreve, a actividade de “Data mining” é uma das componentes, e não a única, que participa no processo de descoberta de conhecimento em bases de dados. Fayad et al [16] foram os primeiros, em 1996, a sistematizar a *Descoberta de Conhecimento em Bases de Dados*, num processo exploratório, no qual se identificam etapas e os respectivos conjuntos de técnicas. Este processo exploratório, tal como foi definido [16] por Fayad denomina-se de *Processo de descoberta do conhecimento* (“Knowledge Discovery Process” ou “KDD process”) e é hoje unanimemente aceite na comunidade científica.

Cada uma das seguintes secções corresponde às quatro fases mais importantes no “KDD process” e onde serão, respectivamente e sucintamente descritas os conjuntos de técnicas convencionais utilizadas em cada fase.

3.2 Formulação do problema e selecção dos dados

Sendo a primeira etapa no “KDD process”, é aqui que cabe, em primeiro lugar, a definição inequívoca do problema a estudar, em segundo lugar, a selecção dos dados que serão alvo de tratamento nas fases seguintes do processo. Esta fase é essencialmente interrogativa, pois o *descobridor*, deverá questionar-se para garantir que o projecto de *Descoberta do conhecimento* se encontre devidamente estruturado nos seus conceitos fundamentais.

3.2.1 Formulação do problema

A formulação do problema faz-se pela resposta a duas perguntas essenciais:

- Qual o fenómeno a estudar?
- Pretendemos compreender ou prever?

Por outras palavras deveremos identificar, cabalmente, qual o assunto que nos motiva a descobrir conhecimento. Tal como identificaremos no Capítulo 4, o presente trabalho, pretende estudar o fenómeno do incumprimento no crédito ao consumo, concedido por uma determinada instituição financeira.

É respondendo à segunda questão, que se define o verdadeiro objectivo do processo de descoberta do conhecimento. Dado o fenómeno pretendemos apenas compreendê-lo interpretando os padrões reconhecidos? Ou prever o valor de determinada variável a partir de novos casos? A definição deste objectivo é importante, pois condiciona opções futuras em fases posteriores do *processo de descoberta de conhecimento*, nomeadamente no caso em que se opte compreender os padrões. Neste caso a forma de expressar o reconhecimento de padrões deverá ser inteligível (Ver Secção anterior), pelo que a escolha dos algoritmos de “data mining” deverá ser efectuada em função desta opção.

Tal como identificamos no Capítulo 4, o presente trabalho tem como objectivo a predição de uma variável dicotómica.

3.2.2 Selecção de dados

Consiste em seleccionar um conjunto de dados que, em primeiro lugar, contenha um número suficientemente grande de observações de forma a garantir a inferência do conhecimento obtido. Em segundo lugar, que contenha um conjunto de atributos, suficientemente caracterizador da realidade que pretendemos conhecer ou prever. Enfim o conjunto de dados, seleccionado, e que irá ser alvo de tratamento nas fases seguintes, deverá retratar fielmente o fenómeno que pretendemos compreender ou prever. Nas Secções 2.3.1 e 2.3.2, encontram-se explicadas algumas das técnicas, aplicadas ao “credit scoring”, que se enquadram neste âmbito.

Finalmente, importa perceber, em prole da garantia da qualidade dos dados, como é que os dados são capturados e por quem? Quais são os processos de validação automáticos inerentes à sua inserção nos sistemas de informação? Quais são as rotinas e procedimentos de modificação e actualização dos dados capturados? Quais são as rotinas de auditoria e controle, caso existam, à fiabilidade dos dados e respectivos relatórios? Quais os dados omissos e porque é que são omissos? Enfim todas estas questões visam perceber a qualidade dos dados de forma a definir as acções a imprimir no sentido de uma melhor selecção dos dados, garantindo o resultado do *Processo de Descoberta de Conhecimento*.

3.3 Pré-processamento de dados e respectivas técnicas básicas

Considera-se por técnicas básicas de pré-processamento todas as acções realizadas imediatamente antes da aplicação dos algoritmos de “data mining” e que imprimem transformações nos dados, indispensáveis à aplicação dos referidos algoritmos. Estas acções a que denominamos técnicas básicas ou convencionais de pré-processamento resumem-se a duas, que a seguir descrevemos.

3.3.1 Binarização de variáveis nominais

Cada vez mais as variáveis nominais proliferam nos conjuntos de dados. Se por um lado existem algoritmos de “data mining” que tratam as variáveis qualquer que seja a sua natureza (numérica ou nominal) como é o exemplo das árvores de decisão C4.5 [10], por outro existem algoritmos que não são indiferentes à natureza das variáveis, sendo o exemplo disso os algoritmos, de cariz estatístico, associados às tarefas de classificação ou de regressão. Nestes modelos a variável dependente estimada resulta de uma soma pesada das variáveis, pelo que se torna impossível de fazer participar uma variável não numérica naquela soma, sem proceder a uma transformação da mesma. Para contornar este problema opta-se por uma das seguintes alternativas: A primeira alternativa será radicalmente excluir as variáveis nominais; A segunda consiste em transformar os k valores possíveis da variável em $k-1$ variáveis numéricas. Esta última denominada de *binarização de variáveis nominais* será alvo de detalhe no Capítulo 5.

3.3.2. Tratamento de valores omissos (“missing values”)

A existência de valores omissos nos dados é tão natural quanto a existência dos próprios dados. É intrínseco. Torna-se necessário o seu tratamento, pois os algoritmos de “data mining”, por si só não lidam com valores omissos. A Secção 5.5 fará referência aos vários tipos de valores omissos, bem como as várias estratégias de os tratar.

Resta-nos concluir, dizendo que para além das técnicas básicas de pré-processamento, atrás referenciadas como indispensáveis à aplicação dos algoritmos de “data mining”, existem outras técnicas a que denominamos de técnicas avançadas de pré-processamento, que sendo dispensáveis à aplicação dos referidos algoritmos, podem melhorar os resultados finais. Tudo isto se encontra descrito no Capítulo 5.

3.4 “Data mining” e respectivos algoritmos

Iremos enumerar e explicar, não toda a algoritmia frequentemente utilizada em “Data Mining”, mas aquela que irá ser alvo de estudo.

3.4.1 Discriminante Linear

A análise discriminante foi originalmente desenvolvida na botânica e a sua aplicação teve como objectivo fazer a distinção de grupos de plantas com base no tamanho e tipo de folhas, para que posteriormente, fosse possível compreender e prever novas espécies encontradas (Fisher 1936) [6]. O interesse pela análise discriminante cedo se generalizou a outros campos do conhecimento, sempre que foi possível encontrar grupos de indivíduos e conhecer quais as características que os distinguem uns dos outros.

- **O método de Fisher para dois grupos⁶**

Consideremos uma população G constituída por dois grupos G_1 e G_2 (tal que se conhece os indivíduos n_1 que pertencem a G_1 e n_2 que pertence a G_2 , em que $n_1+n_2=N$) e caracterizada por o conjunto X de p variáveis. Com vista a efectuar um estudo descritivo de um fenómeno, de forma a compreender, a importância das variáveis na diferenciação dos grupos ou classes, Fisher propôs que se encontrasse uma combinação linear de X de tal modo que se maximizasse o coeficiente **(1.2)** entre a diferença de médias das combinações lineares para os dois grupos relativamente à variância.

$$Z = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \dots \beta_n X_n \quad (1.1)$$

Por outras palavras, o novo eixo constituído como combinação linear **(1.1)** dos pontos de X deverá ter uma inclinação tal que, uma vez todos os exemplos projectados no novo eixo separem G_1 e G_2 de forma a maximizar o quociente entre a dispersão entre grupos (representado tradicionalmente por B) e a dispersão intra grupos (representado tradicionalmente por W). Quanto maior for este quociente, mais a função discriminante separa os grupos, pois interessa que o numerador seja o maior possível (significa grupos mais afastados um do outro) e o denominador o menor possível (significa maior homogeneidade dentro de cada grupo). No entanto chamamos a atenção que a função discriminante é calculada sob dois pressupostos. O primeiro é que as médias de G_j correspondam às verdadeiras médias μ_j (Para $j=1,2$). O segundo é a equicovariância, isto é, a dispersão intra grupo para G_1 e G_2 deverão ser estatisticamente semelhantes, traduzindo-se pela igualdade das matrizes de variância e covariância.

⁶ Segundo Elisabete Reis [6]

Maximizando (1.2) e igualando-a a zero obtemos (1.3), onde se apura os parâmetros da função discriminante da forma expressa em (1.1).

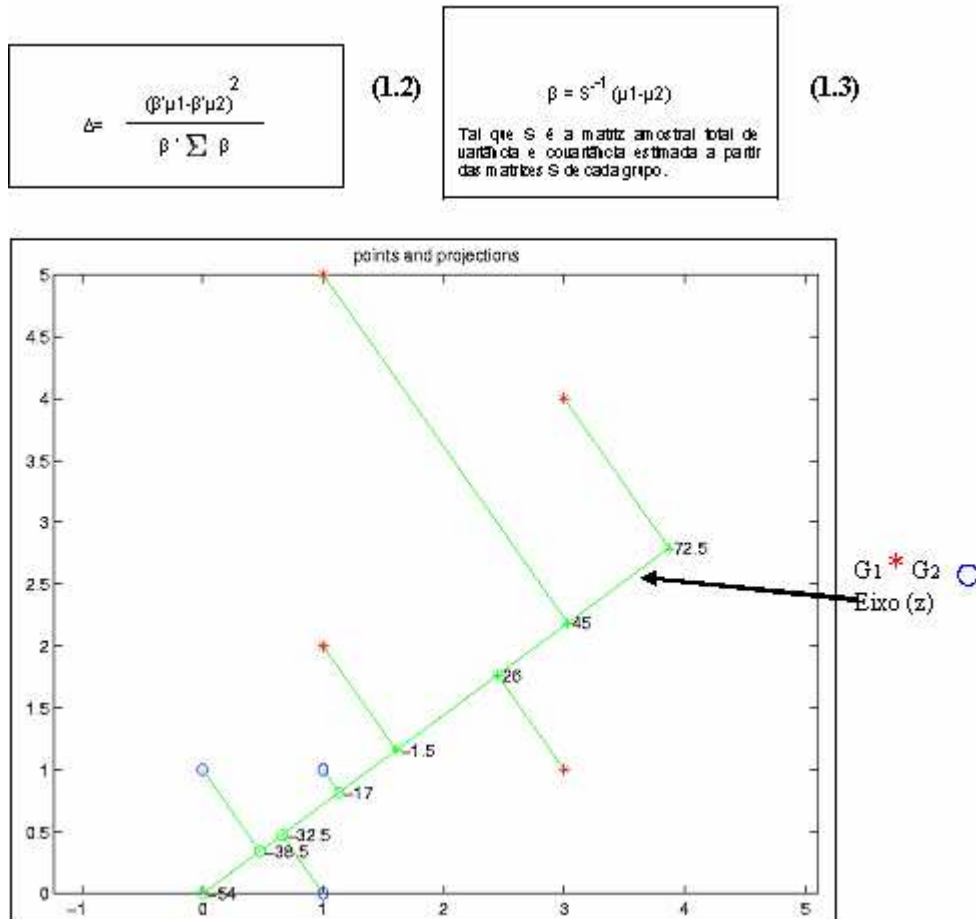


Figura 5 – A projecção das instâncias no novo eixo, calculado pelo critério de Fisher.

A Figura 5 ilustra elementos de G1 e G2 que projectados sobre o eixo (z), cuja inclinação resulta do valor apurado do vector de coeficientes β , conforme (1.3), separam correctamente os elementos de G1 e G2. O valor resultante da projecção de um exemplo no eixo, é denominado de “score” tal que $Z = \beta'X$. Podemos observar na Figura 5 que os “scores” estão compreendidos entre -54 e 72.5, sendo que para valores inferiores ou iguais a -17, encontrámos somente elementos do grupo G2, para valores superiores a -17, encontrámos somente elementos do Grupo G1. Neste sentido o “score” -17 poderá suportar uma regra de afectação ou classificação de novas instâncias, aos dois grupos. Explicaremos na alínea seguinte dois modos de classificação de novas instâncias.

- **Classificação – Como afectar uma nova observação a um grupo?**

1) Método de “Cut-off”

Para a função discriminante de Fisher, a regra para classificar um indivíduo i , com características dadas por p atributos, num grupo de entre dois, deverá ser classificado no G_1 se estiver mais próximo da média deste grupo do que da média do G_2 , isto é, se a distância entre o seu “score” e a média dos “scores” do G_1 for menor ou igual que a mesma distância relativamente ao G_2 . Se considerarmos Y_c , ponto crítico ou ponto médio entre as médias dos “scores” dos dois grupos, tal como se encontra definido em (1.6),

$$Y_c = \frac{\bar{Y}_1 + \bar{Y}_2}{2} = \frac{\beta \bar{X}_1 + \beta \bar{X}_2}{2} = \frac{1}{2}(\bar{X}_1 - \bar{X}_2)^T \sum_{i=1}^n (\bar{X}_1 + \bar{X}_2) \quad (1.6)$$

Para:
 $\beta = \sum_{i=1}^n (\mu_1 - \mu_2)$
 Tal que $\sum_{i=1}^n \text{pool}_{i,i}$ é a média ponderada das matrizes Variância e covariância de cada Grupo

a regra de classificação consistirá em :

Classificar no G_1 se $Y_i > Y_c$, ou classificar no G_2 se $Y_i < Y_c$. Para que esta classificação seja óptima existem dois pressupostos fundamentais: A primeira consiste na aderência da população, de onde o problema foi extraído, à lei de probabilidade multi-normal; a segunda consiste na igualdade da dispersão intra grupo, já anteriormente referida.

Assim a regra de “cut-off” consiste em subtrair à função discriminante a expressão Y_c (1.6), pelo que se maior que zero será afecta a um grupo, caso contrário ao outro. O que acabamos de formular expressa-se em (1.7)

$$x^T \Sigma^{-1}(\mu_1 - \mu_2) - \frac{1}{2}(\mu_1 + \mu_2)^T \Sigma^{-1}(\mu_1 - \mu_2) > 0 \quad (1.7)$$

2) “Likelihood ratio” utilizando o teorema de “Bayes”.

Consideremos as seguintes probabilidades condicionadas, $P(\text{classe}=1 \mid X_{i\text{-ésima instância}})$ cujo significado indica a probabilidade de ser Bom pagador (classe=1), dada uma determinada instância e $P(\text{classe}=0 \mid X_{i\text{-ésima instância}})$ cujo significado indica a probabilidade de ser Mau pagador (classe=0), dada uma determinada instância. Sabendo que $P(\text{classe}=0 \mid X_{i\text{-ésima instância}}) = 1 - P(\text{classe}=1 \mid X_{i\text{-ésima instância}})$, então se o quociente (ratio) formulado entre $P(\text{classe}=1 \mid X_{i\text{-ésima instância}})$ e $P(\text{classe}=0 \mid X_{i\text{-ésima instância}})$ for superior a 1, significa que a probabilidade de uma determinada instância pertencer a uma classe é superior à probabilidade de pertencer à outra. Este é o princípio do “Likelihood ratio” (doravante denominado por “LR”) e representa-se como se expressa em (1.8).

$$LR = \frac{P(\text{classe}=1 \mid x\text{-ésima instância})}{P(\text{classe}=0 \mid x\text{-ésima instância})} > 1 \quad (1.8)$$

Para determinar ambas probabilidades condicionadas constantes em “LR”, teremos que aplicar o teorema de “Bayes”, tal como se expressa em (1.9) e se explica em (a), (b) e (c).

$$P(\text{classe}=1 \mid x\text{-ésima instância}) = \frac{P(x\text{-ésima instância} \mid \text{classe}=1) \times P(\text{classe}=1)}{P(x\text{-ésima instância})} \quad (1.9)$$

Tal que:

- (a) $P(\text{classe}=1 \mid x\text{-ésima instância})$ denomina-se probabilidade à posteriori;
- (b) $P(\text{classe}=1)$ denomina-se probabilidade à priori;
- (c) $P(x\text{-ésima instância} \mid \text{classe}=1)$ denomina-se probabilidade condicionada, reflectindo, qual a probabilidade de observarmos $x\text{-ésima instância}$, dada a classe 1. Se assumirmos que os atributos têm distribuições contínuas, então aquela probabilidade condicionada reflecte a densidade de probabilidade das observações ou instâncias oriundas da classe=1.

Aplicando o teorema de “bayes” em (1.8), chegamos à expressão (1.10), que após as transformações a seguir indicadas, culmina na regra de classificação expressa em (1.13)

$$LR = \frac{P(x\text{-ésima instância} \mid \text{classe}=1) \times P(\text{classe}=1)}{P(x\text{-ésima instância} \mid \text{classe}=0) \times P(\text{classe}=0)} > 1 \quad (1.10)$$

Substituindo, as probabilidades condicionadas, pelas respectivas funções de densidade de probabilidade, obtém-se a expressão (1.11).

$$LR = \frac{f_1(x)}{f_0(x)} > \frac{P(\text{classe}=0)}{P(\text{classe}=1)} \quad (1.11)$$

Pressupondo que $f_1(x)$ e $f_0(x)$, seguem a lei de probabilidade multi-normal e que as probabilidades à priori, $P(\text{classe}=0)$ e $P(\text{classe}=1)$ são iguais, simplificamos (1.11) em (1.12)

$$LR = \frac{\frac{1}{\sqrt{|2\pi\Sigma|}} \exp\left(-\frac{1}{2}(\mathbf{x} - \mu_1)^T \Sigma^{-1}(\mathbf{x} - \mu_1)\right)}{\frac{1}{\sqrt{|2\pi\Sigma|}} \exp\left(-\frac{1}{2}(\mathbf{x} - \mu_2)^T \Sigma^{-1}(\mathbf{x} - \mu_2)\right)} > 1 \quad (1.12)$$

Transformando a expressão (1.12) através do logaritmo neperiano, chegamos à expressão matemática da regra.

$$LR = \ln \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} > \ln(1) \Leftrightarrow \mathbf{x}^T \Sigma^{-1}(\mu_1 - \mu_2) - \frac{1}{2}(\mu_1 + \mu_2)^T \Sigma^{-1}(\mu_1 - \mu_2) > 0 \quad (1.13)$$

Comparando as regras de classificação obtidas em 1) e 2) chegamos à mesma expressão. Ambos os métodos são equivalentes se pressupor-mos que as probabilidades a priori são iguais e que as funções de densidade de probabilidade seguem a lei multi-normal.

3.4.2 Regressão Logística

Motivação

Como “ferramenta” de classificação a regressão logística não impõe restrições quanto à lei Multi-variada de probabilidade de uma população. Contrariamente ao discriminante linear de Fisher, que para a classificação ser ótima a população terá que seguir uma distribuição multi-normal. Não obriga ao cumprimento do pressuposto da equicovariância, pois critério de preferência, não é a maximização do ratio dispersão entre grupos / intra grupos.

Enquanto pertencente à família dos GLM's (“General linear models”), contrariamente a estes, cuja variável de resposta (ou dependente) é contínua (compreendida entre $-\infty$ e $+\infty$), a regressão logística modeliza as relações entre uma variável de resposta (dicotômica ou binária) e as variáveis independentes.

A regressão logística tem sido usada, na análise estatística multi-variada, essencialmente a partir dos trabalhos executados por Truett, Cornfield e Kannel em 1967⁷, no campo da medicina, em particular no estudo das doenças coronárias.

⁷ Tal como se encontra citado em Hosmer e Lemeshow [3]

- **A regressão Logística**

Pretende-se determinar a probabilidade condicionada $P(Y=1|x)$, que reflecte a probabilidade de $Y=1$, (tal que Y tem dois valores possíveis $\{1,0\}$) dado um determinado vector de atributos (ou exemplo). Para determinar tal probabilidade, usando a distribuição logística como lei de probabilidade, utiliza-se a equação da regressão logística que se encontra representada em (2.2) e (2.3):

Para $P(Y=1|x)$

$$P(Y=1|x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}} \quad (2.2)$$

Para $P(Y=0|x)$

$$P(Y=0|x) = 1 - P(Y=1|x) = 1 - \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}} = \frac{1}{1 + e^{\beta_0 + \beta_1 x}} \quad (2.3)$$

A função logística pode ser transformada conduzindo a uma nova interpretação. Definindo-se “odds” (probabilidade) Como:

$$\text{"odds"} = \frac{P(Y=1|x)}{1 - P(Y=1|x)} \quad (2.4)$$

Substituindo na expressão (2.4), as expressões em (2.3) e (2.2), fica

$$\text{"odds"} = \frac{P(Y=1|x)}{1 - P(Y=1|x)} = e^{\beta_0 + \beta_1 x} \quad (2.5)$$

Tomando o logaritmo da expressão (2.5), chegamos á conclusão que o logaritmo neperiano do quociente das probabilidades é uma equação linear:

$$g(x) = \ln(e^{\beta_0 + \beta_1 x}) = \beta_0 + \beta_1 x$$

(2.6)

À transformação operada em (2.6) denomina-se de **Logit**.

Assim concluímos, após as transformações operadas em (2.4), (2.5) e (2.6), que o ratio entre as probabilidades condicionadas de cada classe ou grupo, se exprime como uma combinação linear das variáveis, faltando estimar os respectivos parâmetros.

- **Processo de estimação de parâmetros**

Obtendo amostras, independentes, retiradas das 2 classes, a função verosimilhança para os parâmetros β_0, β_1 terá a seguinte representação:

$$l(\beta_0, \beta_1) = \prod_{\{ \text{amostra } Y=1 \}} P(Y=1|x) \prod_{\{ \text{amostra } Y=-1 \}} P(Y=0|x)$$

(2.7)

Fazendo substituir, em (2.7) as expressões (2.2) e (2.3), procede-se à estimação dos parâmetros β_0, β_1 que maximizam aquela expressão. Basicamente este método, determina (ou estima) os valores desconhecidos dos parâmetros, que maximizam a probabilidade de obter o correspondente aos valores observados.

Entretanto a expressão (2.7) carece ser transformada, para melhor manuseamento matemático, pelo que efectuar-se-á o logaritmo neperiano de (2.7), ficando:

$$\ln l(\beta_0, \beta_1) = \sum_{i=1}^n \{ y_i \ln [P(Y=1|x_i)] + (1-y_i) \ln [1-P(Y=1|x_i)] \}$$

Em y_i, x_i referem-se ao valor da variável dicotómica e ao valor da variável independente, respectivamente, da i -ésima observação em n .

(2.8)

Contrariamente à função de verosimilhança no método dos mínimos quadrados, para regressões simples, a função verosimilhança para β_0, β_1 , na regressão logística é não linear. Tal significa que β_0, β_1 têm que ser encontrados através de métodos iterativos, conforme proposto por Cox (1966) e Day & Kerridge (1967), segundo Hosmer e Lemeshow [3].

- **Diferenças entre o modelo de regressão linear e a regressão logística.**

Para melhor esclarecer, sistematizamos de seguida as diferenças entre a regressão logística e a regressão linear.

a) Enquanto numa regressão linear a variável dependente é contínua, variando entre $-\infty$ e $+\infty$, numa regressão logística a variável dependente varia entre 0 e 1.

b) Outra diferença, diz respeito à natureza da relação entre a variável dependente e as independentes. Em qualquer problema de regressão linear, o objectivo é quantificar o valor esperado da variável dependente, dado um determinado exemplo (vector de atributos ou variáveis independentes.). Tal quantidade, denomina-se de esperança condicionada e terá a seguinte notação: “ $E(Y|x)$ ”, onde Y constitui a variável dependente e x o valor da variável independente.

Numa regressão linear, assume-se que a quantidade $E(Y|x)$ é expressa como sendo o resultado de uma equação linear da forma:

$$E(Y|x) = \beta_0 + \beta_1 x, \text{ tal que } E(Y|x) \text{ está compreendido entre } -\infty \text{ e } +\infty \quad (2.1)$$

Numa regressão logística, onde a variável de resposta é dicotómica, $E(Y|x)$, tal que $Y \in \{1, 0\}$, variará entre 0 e 1, pois a $E(Y|x)$ revela a proporção esperada de $Y=1$, dado o exemplo x. Neste caso $E(Y|x)$, traduz-se numa Probabilidade de acontecer $Y=1$, dado X. Como veremos mais tarde a função de distribuição associada será a distribuição logística.

c) Estimativa dos parâmetros.

Em qualquer regressão os parâmetros ($\beta_0, \beta_1, \beta_2, \dots, \beta_n$) têm que ser estimados, de acordo com um qualquer critério de preferência.

Numa regressão linear o método mais usado é o dos “Mínimos quadrados”. Neste método são escolhidos os β 's que minimizam a soma do quadrado da diferença entre valores de Y observados e preditos pelo modelo. Na regressão logística o método dos mínimos quadrados demonstrou-se desajustado pelo que inviabiliza a sua aplicação neste tipo de regressão. O método dos mínimos quadrados, pertence a uma família mais vasta de estimativas que é o **método da máxima Verosimilhança**. Basicamente este método, determina (ou estima) os valores desconhecidos dos parâmetros, que maximizam a probabilidade de obter o correspondente aos valores observados. Para aplicar este método, deveremos em primeiro lugar construir a **função de verosimilhança**, em que esta função expressa a probabilidade de observar o verdadeiro valor de Y, em função de $\beta_0, \beta_1, \beta_2, \dots, \beta_n$. Deste modo os parâmetros encontrados, serão os que melhor “conduzirão” o ajustamento do modelo aos valores observados. Para estimar os parâmetros de uma regressão logística, utilizaremos, também o **método da**

máxima verosimilhança, tal como se encontra explicado na alínea – *processo de estimação de parâmetros*.

- **Regra de classificação**

Definido o modelo, torna-se necessário definir uma regra a partir da qual classificamos uma determinada instância na classe 1 (Bom pagador) ou na classe 0 (mau pagador). A fixação dessa regra é determinada, pelo “LR” (“Likelihood ratio”), tal como foi preconizado na Secção 3.4.1. alínea 2. Assim o quociente das probabilidades condicionadas terá que ser superior a 1, para classificar uma instância na classe=1, caso contrário será classificada como classe=0. Como a regressão logística representa directamente a $P(\text{classe}=1 \mid x_i\text{-ésima instância})$, através do cálculo da expressão (2.2), dados os parâmetros, não necessitaremos do teorema de “Bayes”. Assim sendo, basta imprimir a regra de $P(\text{classe}=1 \mid x_i\text{-ésima instância}) > 50\%$ para classificar a $x_i\text{-ésima instância}$ na classe=1. Se as probabilidades à priori de ambas as classes são diferentes, então devera-se adicionar à expressão (2.6), o logaritmo neperiano de $(P(\text{Classe}=1) / P(\text{Classe}=0))$, estimar os parâmetros e proceder ao cálculo da expressão (2.2) para obter a probabilidade condicionada $P(\text{classe}=1 \mid x_i\text{-ésima instância})$ [5].

3.5 Avaliação de um algoritmo e respectivas técnicas básicas

São vários os autores que sistematizam as variadíssimas abordagens sobre esta problemática. Ian H. Witten e Eibe Frank [4] e D.Michie [5], ambos reservam um capítulo para inventariar as técnicas de medição da performance dos algoritmos.

Se queremos prever, é fundamental medir a capacidade preditiva do modelo, dada uma determinada linguagem de representação.

É normal em “Data Mining” fazer correr vários algoritmos sobre o mesmo problema (Por exemplo a Árvore de Decisão: C4.5; discriminante linear; K-NN, etc) e medir a performance de cada um de forma a apurar qual o melhor. Ora medir a performance de vários algoritmos, e escolher o melhor, é um modo de inferir que a linguagem de representação do algoritmo eleito se aproxima da verdadeira função que rege o problema. Assim a predição será correcta.

Genericamente, existem 4 pontos importantes a considerar quando abordamos a temática da performance dos algoritmos:

3.5.1 A definição de Taxa de erro

Existem duas noções; A primeira, *Taxa de erro empírico* é a taxa de instâncias mal classificadas no conjunto de treino, de um algoritmo treinado no conjunto de treino. Esta taxa também se denomina de erro de re-substituição. A segunda, *Taxa de erro de generalização* é a taxa de instâncias mal classificadas num conjunto independente de dados, respeitantes a um algoritmo treinado conjunto de treino.

Por razões óbvias a taxa de erro de generalização é o indicador mais correcto e portanto mais usado para medir a performance do modelo. A partir de agora onde se lê taxa de erro devera-se ler taxa de erro de generalização.

3.5.2 A estimativa da taxa de erro

Dependendo do número de instâncias disponíveis, de seguida apresentamos as várias estratégias praticadas:

- *“Holdout test”*: Consiste em dividir, aleatoriamente, o conjunto de dados em 2 grupos distintos. Um deles para conjunto de treino; O outro como conjunto de teste para estimação do erro de generalização. O *“Holdout test”*, é uma estratégia normalmente usada quando existe abundância de dados. É importante referir que os dois grupos deverão manter a proporção de classes existentes no conjunto de dados inicial.
- *“K-Cross-Validation” ou “leave one out”*: Este método consiste em dividir o conjunto de dados em k partições disjuntas. Das k partições, k-1 constituem o conjunto de treino e o restante o conjunto de teste. Itera-se este procedimento K vezes (normalmente K=10), tal que a partição de teste usada em cada iteração é sempre uma partição diferente. (exemplo ilustrado na Figura 6). Assim sendo a taxa de erro será a média das taxas de erro obtidas nas diferentes partições de teste. Este método de amostragem é usado quando existe escassez de dados.

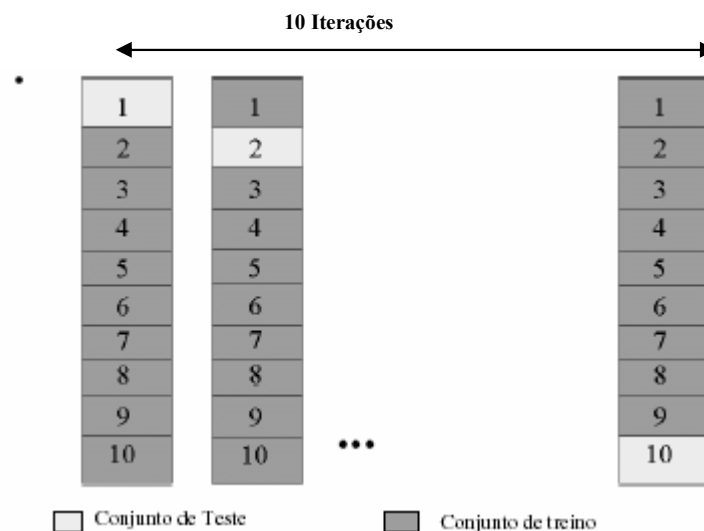


Figura 6

- “Bootstrap”: Este método consiste em dividir o conjunto de dados em n partições. Das n partições, constituem-se $n-1$ partições de treino. Até aqui o método é igual ao “Cross-Validation”. A diferença consiste em cada partição de treino das $n-1$, são obtidas com reposição, podendo haver exemplos repetidos nas $n-1$ partições de treino. O conjunto de teste é constituído pelos exemplos do conjunto de dados que não pertencem ao conjunto de treino. Este procedimento itera-se k vezes, em que K , normalmente = 10. A taxa de erro é a média das taxas de erro apuradas nas k iterações diferentes.

3.5.3 Inferir a taxa de erro determinada segundo umas das estratégias da Secção 3.5.2

Tal como é preconizado por Ian e Eibe [4], a taxa de erro é inferida recorrendo á inferência intervalar, sabendo que a variável aleatória “Número de erros de um algoritmo” segue uma distribuição binomial, onde a taxa de erro é a proporção binomial. Segundo o que se pode interpretar em R. Guimarães e J. Sarsfield Cabral [8].

Sabendo que ϵ é a proporção estimada (binomial) de erro tal que ϵ resulta do quociente entre o número de erros e o número de elementos da amostra (N), tal que N é uma amostra de grande dimensão. Se considerarmos que o processo de amostragem é aleatório simples, em que: (i) A averiguação do sucesso em cada instância, só pode ter dois resultados possíveis - Certo ou errado. (ii) A probabilidade de ocorrência de cada resultado se mantêm inalterada de instância para instância (no grupo teste). (iii) Os resultados associados a cada instância são independentes.

Cumpridas que estão aquelas três propriedades estamos na presença das “*Experiências de Bernoulli*”, sendo que a variável aleatória “Número de erros de um algoritmo” segue uma distribuição Binomial $B(N, \epsilon)$. Se N é grande então a v.a. “número de erros” aproximar-se-á à distribuição normal, cujos parâmetros serão os descritos em (3.1)

$$\begin{aligned} \mu &= N \cdot \epsilon \\ \sigma^2 &= \frac{\epsilon \times (1 - \epsilon)}{N} \end{aligned} \quad (3.1)$$

Assim sendo os limites do intervalo de confiança serão: $\epsilon \pm \alpha \cdot \sigma$

3.5.4 Comparar Taxas de erro estimadas, para o mesmo conjunto de dados, de diversos algoritmos

Se compararmos duas ou mais estimativas de erro de diferentes algoritmos, é conveniente avaliar se as taxas de erro estimadas são estatisticamente diferentes. Para esse efeito passamos a explicar dois testes:

- **Teste-t (“t-test”)**

Como maioria dos testes paramétricos de localização, o “t-test”, irá apurar se as médias das 2 populações, estão significativamente distantes. Se o estiverem, então diz-se que as 2 populações são diferentes, no que toca à localização.

Mas esta conclusão parte do cumprimento de um pressuposto fundamental: Normalidade de ambas populações, quando estas não são consideradas grandes ($n < 30$). Caso contrário se $n > 30$, pelo teorema do limite central⁸, a média amostral segue a distribuição normal, bastando aplicar o teste z.

Consideremos duas amostras independentes, oriundas de duas populações, tal que S_A^2 e S_B^2 representam as variâncias amostrais; N_A e N_B as dimensões amostrais e $\bar{X}_A$ e $\bar{X}_B$ representam as respectivas médias amostrais.

$$ET = \frac{(\bar{X}_A - \bar{X}_B) - a}{(S_A^2/N_A + S_B^2/N_B)^{1/2}} \quad (3.2)$$

A expressão (3.2) traduz a estatística de teste aplicada a duas amostras de pequenas dimensão extraídas de duas populações Normais, em que as variâncias amostrais são diferentes. As hipóteses do teste são:

$$H_0: \mu_a = \mu_b \text{ ou } \mu_a - \mu_b = a$$

$$H_1: \mu_a - \mu_b \neq a \text{ ou } \mu_a - \mu_b > a, \mu_a - \mu_b < a$$

Testar as médias da mesma variável aleatória de duas populações independentes, através das respectivas amostras, designa-se de “paired t-test”. Por outro lado, designar-se-á de “one-sample t-test”, se pretendermos testar uma média representada por um valor fixo com a média amostral extraída de uma determinada população [21]. A ET (3.2) ou estatística de teste deverá registar valores baixos, se H_0 verdadeira, caso

⁸ Conforme se pode apurar segundo R.Guimarães J.A.Cabral [8]

contrário deverá registar valores elevados, quer positivos quer negativos. Quando a H_0 é verdadeira, a estatística de teste segue uma distribuição tGL (o cálculo de GL da distribuição t, para efeitos desta ET, baseia-se numa expressão que não reproduzirei aqui).

- **Para variáveis não aderentes á lei normal ou de lei de probabilidade desconhecida (para amostras consideradas pequenas): Wicoxon-Man-Whitney test.**

O objectivo deste teste é o de avaliar se as medianas de duas populações A e B, com a mesma forma, se localizam no mesmo ponto. Na hipótese nula, admite-se que as medianas de ambas as populações coincidem, e na hipótese alternativa postula-se que a diferença entre medianas é não nula, é positiva ou é negativa.

Sejam N_A e N_B 2 amostras, independentes, retiradas das populações A e B, tal que a dimensão de N_A é menor ou igual N_B . O cálculo da estatística de teste compreende os seguintes passos:

- 1- Ordena-se de forma crescente e em conjunto as N observações ($N = N_A + N_B$);
- 2- Atribui-se um nº de ordem a cada observação, começando por 1, na que se encontra em primeiro lugar no vector ordenado, e terminando em N , na que se encontra em último lugar;
- 3- A estatística de teste, W , corresponde à soma de números de ordem relativos à amostra de menor dimensão:

$$ET = W \text{ (soma dos números de ordem da menor amostra).}$$

Se a estatística W tomar um valor w muito baixo (ou muito alto), tal significa que as observações referentes à menor amostra são predominantemente inferiores (ou superiores) às observações da outra amostra. Tal facto contraria a hipótese nula, isto é, contraria a hipótese de que ambas as populações com a mesma forma têm a mesma localização.

Uma vez acabado o presente Capítulo e terminada a explicação das técnicas convencionais aplicadas no “KDD process”, resta-nos experimentá-las e publicar os respectivos resultados. Todo este trabalho será objecto de tratamento no Capítulo seguinte.

4- Um caso de concessão de crédito e os primeiros resultados

Pretendemos que este capítulo informe o leitor sobre:

- A apresentação de um problema de concessão de crédito;
- Breve descrição dos dados, bem como o modo como foram seleccionados;
- Os primeiros resultados, produzidos a partir da aplicação das técnicas convencionais enunciadas no Capítulo anterior.

4.1 Um problema de concessão de crédito

Pretendemos estudar o fenómeno do incumprimento, no âmbito do crédito ao consumo, com o objectivo de predizer, dadas as características de um contraente e de uma compra a crédito, o comportamento daquele face ao regular pagamento das prestações mensais. Como veremos a seguir, a presente formulação obriga a que se crie uma variável alvo (variável a explicar) que defina ou tipifique “o comportamento regular face ao pagamento das prestações mensais”, em duas categorias, como por exemplo “Bom” ou “Mau” pagador.

4.2 Sobre o caso de crédito

Os dados foram gentilmente cedidos por uma instituição bancária, que se dedica à concessão de crédito ao consumo há mais de uma década. Os dados foram extraídos a 31 de Outubro de 2003, tendo sido aleatoriamente seleccionados os contratos de crédito a termo, cujo início contratual se verificou entre 1 de Abril de 2002 e 31 de Dezembro de 2002. A natureza dos dados extraídos pode-se estruturar em três tipos distintos:

- Caracterização do cliente;
- Caracterização da operação de crédito;
- Comportamento do cliente.

A primeira tem a ver com a informação que caracteriza o cliente na sua esfera sócio demográfica, sendo exemplo disso a idade, profissão, estado civil, etc; a segunda tem a ver com a caracterização da operação de crédito, isto é, o bem, o preço de venda ao público, o valor a financiar, o prazo da operação, etc; e finalmente a terceira, prende-se com as informações respeitantes ao comportamento dos clientes, apurando mensalmente o número de prestações em atraso, durante o período de exposição⁹ que termina a 31 de Outubro de 2003. Tanto a primeira como a segunda, são recolhidas no momento em que o cliente propõe o crédito e constituirão as variáveis independentes dos algoritmos de “data mining” que iremos utilizar. Por sua vez a terceira servirá, à luz dos critérios expostos anteriormente na Secção 2.3.2, para definir a variável dependente ou seja, para

⁹ A noção explicada na Secção 2.3.1

cada instância caracterizada pelas variáveis independentes, a partir daquela, classificamos em “Bom” ou “Mau” pagador.

Importa referir duas restrições importantes ao presente estudo: Em primeiro lugar a instituição bancária tem interesse que o enfoque do presente estudo incidisse sobre todos os contratos celebrados a partir de 1 de Abril de 2002 e oriundos de um determinado mercado; Em segundo lugar a instituição disponibilizou para o estudo 1199 instâncias de cada classe (“Bons” e “Maus”), com o intuito de não tornar do conhecimento público a verdadeira proporção populacional dos “Bons” e “Maus” pagadores.

4.2.1 Selecção dos dados

À luz do que se explicou nas Secções 2.3.1 e 2.3.2, justificaremos as seguintes decisões tomadas no âmbito da selecção da população, pese embora as duas restrições legitimamente impostas pela Instituição.

Embora não caiba aqui divulgar, observou-se o “Delinquency dynamic report” respeitante aos meses de abertura anteriores a Abril de 2002 e constatou-se uma maturidade equivalente a 12 meses. Assim sendo e tal como já foi explicado, estamos em condições de optar, ou por uma estratégia de selecção a tempo de exposição fixo ou móvel.

Consideremos o quadro 4 a seguir apresentado.

Quadro 4 – Exposição Média dos clientes, vistos em Outubro 2003

Mês	Distrib(%)	Exposição(Mês)	Exp. Média ponderada carteira
Abr-02	1,49%	18	0,27
Mai-02	2,09%	17	0,36
Jun-02	1,66%	16	0,27
Jul-02	2,48%	15	0,37
Ago-02	5,60%	14	0,78
Set-02	8,38%	13	1,09
Out-02	15,51%	12	1,86
Nov-02	27,39%	11	3,01
Dez-02	35,40%	10	3,54
Grand Total	100,00%		11,55

Observando o Quadro 4, conseguimos perceber através da coluna Exposição (Mês) a exposição de cada mês de abertura, tendo Outubro de 2003 como mês de observação. Desde já concluímos que todos os meses de abertura, exceptuando Novembro e Dezembro de 2002, detêm exposições iguais ou superiores à maturidade (12 meses). Assim sendo, se optarmos por uma estratégia de selecção a tempo de exposição fixo, garantindo que todos os meses de abertura respeitam a maturidade, seleccionamos os meses de Abril a Outubro. No entanto estes meses, tal como se constata pela leitura da coluna Distrib (%), perfazem apenas 37.21% dos contratos, pelo que a população assim seleccionada seria manifestamente escassa. Perante esta dificuldade, se considerarmos uma estratégia de selecção com tempo de exposição móvel, certamente iríamos obter mais população, importando no entanto assegurar que o tempo de exposição médio não

deverá divergir da maturidade. Pela leitura da coluna “Exposição Média ponderada Carteira”, conclui-se que os meses de Novembro e Dezembro, fazem baixar a exposição média, pelo que o valor médio global de 11.55 é muito próximo da maturidade. Assim sendo, optamos por uma estratégia de selecção a tempo de exposição móvel, para os meses de abertura que figuram no Quadro 4. Importa referir, que embora não estejam referidos, no Quadro 4, os meses de abertura de Janeiro de 2003 em diante, constatou-se que se incluirmos o mês de Janeiro de 2003 na população seleccionada, a exposição média global desceria para 11 meses, sendo a importância relativa daquele mês de 17.51% e uma exposição de 9 meses, pelo que achamos prudente não incluir este mês.

Definida que está a população e o respectivo mês de observação, tal como já foi referido, daquela, foram seleccionados aleatoriamente, 2.398 instâncias sendo metade “Bons” e a outra metade “Maus”. Dasquelas foram excluídas 18 instâncias, por se considerarem inválidas, resultando num total de 2.380 observações em que 1.182 são “Bons” e 1.198 são “Maus”.

Finalmente, o conceito de “Bom” e “Mau” foi definido de acordo com o que consta na Secção 2.3.2, com o período de observação igual a um mês (31 de Outubro de 2003).

4.2.2 Variáveis e sua descrição

Quadro 5 – Nomes das variáveis e sua descrição

Nome da variável	Descrição	Tipo	Número de factores (“levels”)	Valores omissos (S/N)	Natureza dos dados extraídos
Ndepend	Nº de dependentes	Numérica	-----	N	Característica Cliente
País	Naturalidade	Nominal	12	N	Característica Cliente
Bem	Tipo bem a Financiar	Nominal	8	N	Característica Operação de crédito
Canal	Canal de envio da proposta	Nominal	2	N	Característica Operação de crédito
Vprest	Valor da prestação mensal a pagar (euros)	Numérica	-----	N	Característica Operação de crédito
Vpvp	Valor de venda ao público do bem (euros)	Numérica	-----	N	Característica Operação de crédito
Vei	Valor entrada inicial (vpvp-vf) (euros)	Numérica	-----	N	Característica Operação de crédito
Vf	Valor a financiar (euros)	Numérica	-----	N	Característica Operação de crédito
Prazo	Prazo de pagamento escolhido (meses)	Numérica	-----	N	Característica Operação de crédito
Idade1	Idade do proponente em (anos)	Numérica	-----	N	Característica Cliente
Estciv1	Estado civil do proponente	Nominal	7	N	Característica Cliente
Prof1	Profissão do proponente	Nominal	93	N	Característica Cliente
Origtelpe	Tipo de telefone de contacto	Nominal	6	N	Característica Cliente
Estatuto1	Tipo de contrato de trabalho	Nominal	6	N	Característica Cliente
Vrend1	Valor do rendimento mensal (euros)	Numérica	-----	S	Característica Cliente
Tipohab1	Tipo habitação	Nominal	8	S	Característica Cliente
Antemp1	Antiguidade no trabalho (euros)	Numérica	-----	S	Característica Cliente
Antbanc1	Antiguidade no banco (euros)	Numérica	-----	N	Característica Cliente
Anthab1	Antiguidade na habitação	Numérica	-----	N	Característica Cliente
Sex1	Sexo do proponente	Nominal	2	N	Característica Cliente
Distrito	Distrito de Residência	Nominal	19	N	Característica Cliente
Classe	Variável – classe “Bom” ou “Mau”	Nominal	2	N	Característica Comportamento Cliente

O Quadro 5 apresenta detalhadamente os nomes das variáveis bem como a descrição de cada uma delas. Para além de indicarmos a natureza das mesmas, indicamos para as variáveis nominais o número de valores possíveis ou factores (“Levels”). Esta informação é importante para compreendermos a importância do tratamento destas variáveis, especialmente aquelas, com um número elevado de factores. Teremos oportunidade de abordar em profundidade este problema nos Capítulos 5 e 6.

4.3 Os Primeiros Resultados

Nesta Secção apresentamos os resultados obtidos com a aplicação da Regressão logística e do Discriminante Linear. Para além disso descrevemos alguns passos de pré-processamento indispensáveis à aplicação dos algoritmos. Passemos, sucintamente, a enunciar as tarefas efectuadas:

4.3.1 Pré-processamento

Em conformidade com o explicado na Secção 3.3 e dado que a algoritmia de “data mining” a utilizar estrutura-se em combinações lineares de variáveis, procedeu-se à binarização de todas as variáveis nominais. Tal processo aumentou o número de variáveis nominais, inicialmente de dez para 153 variáveis binárias. A binarização foi efectuada em R, usando uma função criada para o efeito.

Foram tratados os valores omissos respeitantes às variáveis *vrendl* e *antempl*. Os valores omissos, não se devem a falta de informação, mas apenas à impossibilidade de estas variáveis possuírem valores em determinadas circunstâncias. A variável *antempl* que significa antiguidade na entidade patronal, não é aplicável aos reformados; Para o *Vrendl* que significa salário do proponente não é aplicável aos empresários ou profissionais liberais. De notar que o número de valores omissos em *Vrendl* é de 37 instâncias e em *antempl* é de 43. Assim sendo foi assumido o valor zero.

4.3.2 Algoritmos de “data mining”: Regressão Logística e o Discriminante linear

Com vista a comparar resultados, ensaiamos os dois algoritmos e aplicámos as respectivas regras de classificação, descritas nas Secções 3.4.1 e 3.4.2. Para aquele efeito pressupomos para ambos os algoritmos, igualdade de probabilidades à priori e a Multi-normalidade da população para o Discriminante Linear.

4.3.3 Resultados e avaliação de Performance

Dado que o presente caso prático é constituído por um conjunto de dados em número suficiente e dado igualmente, que as melhorias que pretendemos testar e que se encontram descritas no Capítulo 6, dificilmente se aplicaria o “*K-Cross-Validation*”, optamos pela constituição de um “*holdout group test*” para estimar a taxa de erro dos modelos obtidos.

O “*holdout group test*” é constituído por 714 instâncias (30% o conjunto inicial dos dados), tendo sido aleatoriamente extraído do conjunto inicial, preservando a igualdade de proporção das classes. Assim sendo o Conjunto de treino totaliza 1.666 observações.

O quadro 6 ilustra as taxas de erro estimadas no “*holdout group teste*”, bem como as respectivas estimativas intervalares com o fito de aferir a amplitude da variação das taxas de erro. Os intervalos de confiança foram constituídos a partir de um nível de confiança de 95% e tendo por base o que foi explicado na Secção 3.5.3.

Quadro 6- Taxa de erro dos dois classificadores

Classificadores	Taxa de erro	Intervalo de Confiança
Regressão Logística	0,269	0,269 ± 0,0325
Discriminante	0,276	0,276 ± 0,0328

A taxa de erro da Regressão Logística é inferior ao Discriminante, mas a diferença não é significativa. Se justapormos os respectivos intervalos de confiança, concluimos que ambos intervalos se interceptam quase na sua totalidade, explicando assim que, na população, aquelas taxas variam dentro dos mesmos limites.

Assim sendo conclui-se que, face aos resultados dos testes, os dois algoritmos aqui apresentados não se diferenciam no que diz respeito ao acerto na classificação, no presente caso prático.

5- Técnicas de Pré-processamento de dados

Pretendemos que este capítulo informe o leitor sobre:

Uma breve descrição das técnicas de pré-processamento básicas e avançadas.

No processo de modelação, desde o primeiro passo – extracção dos dados e constituição de um conjunto de dados, até ao último – Modelação e Validação, ocorrem uma série de tarefas, genericamente denominadas por Pré-processamento de dados. No âmbito do presente trabalho distinguimos técnicas básicas de técnicas avançadas de pré-processamento de dados. Consideramos a primeira, tal como o interpretamos no Capítulo 4, como um processo essencial na modelização, tal como a binarização das variáveis nominais e substituição de valores omissos, ao passo que as tarefas avançadas de pré-processamento, sendo importantes são dispensáveis. Estas tarefas, visam tratar os dados com o objectivo de diminuir a complexidade/dimensão dos dados sem prejudicar a capacidade preditiva do modelo e de, também, diminuir o tempo de aprendizagem.

5.1 Selecção de atributos

Quando se seleccionam dados no âmbito de um problema de classificação, a tendência é acrescentar o maior número de variáveis possíveis, de forma a bem caracterizar o problema. Acontece, normalmente, que muitas das variáveis pouco ou nada estão associadas ao conceito-alvo, havendo nestes casos dois tipos de variáveis: As variáveis completamente irrelevantes, ou seja em nada distinguem o conceito-alvo e as variáveis redundantes, ou seja, em nada acrescentam á discriminação do conceito-alvo.

O propósito da selecção de atributos consiste em, a partir de um conjunto inicial de F atributos, seleccionar um subconjunto G , tal que $G < F$, tendo sido G apurado segundo um determinado critério que permita identificar as variáveis relevantes para o problema em análise. A eliminação de atributos inúteis, permite reduzir a dimensão dos dados de treino e a sua complexidade e portanto, reduzir o tempo de processamento dos métodos a aplicar nas fases seguintes. Além disso, Hosmer e Lemeshow [3], apelam para importância da selecção de variáveis, pois tendencialmente, com um menor número de variáveis o modelo será mais generalizável e robusto. Este assunto, muito querido dos estatísticos, é válido, pois quanto maior o número de graus de liberdade subjacentes ao modelo, maior será a dependência do modelo ao conjunto de treino e portanto maior a sua variabilidade.

A selecção de atributos, deverá eleger o subconjunto de atributos, com maior relevância para o conceito-alvo, não perdendo de vista duas condições: A primeira é o de a capacidade preditiva do modelo não diminuir significativamente. A segunda é que as probabilidades condicionadas $P(x | \text{Bom})$ e $P(x | \text{Mau})$, que representam as funções de densidade de probabilidade para cada um dos grupos, “Bom” e “Mau” se mantenham semelhantes, para todos os elementos de ambos os grupos, antes e depois da selecção de atributos.

Idealmente, o subconjunto de atributos a eleger, deveria resultar de uma procura completa a todo o espaço de hipóteses (todas as combinações possíveis de atributos) e

que segundo determinada métrica de avaliação seleccionaria a melhor combinação. Ora tal, dependendo da dimensão do conjunto de dados, torna-se computacionalmente inviável, pelo que vários autores, dado este constrangimento, propõe estratégias variadas, mas todas com uma mecânica comum e que poderemos sintetizar na seguinte sequência de tarefas [9]:

- 1) Procedimento de geração de subconjuntos possíveis de atributos candidatos;
- 2) Avaliação dos candidatos gerados em 1);
- 3) Critério de paragem (“Stopping criteria”) pára a geração de candidatos;
- 4) Após paragem, validação do Subconjunto eleito.

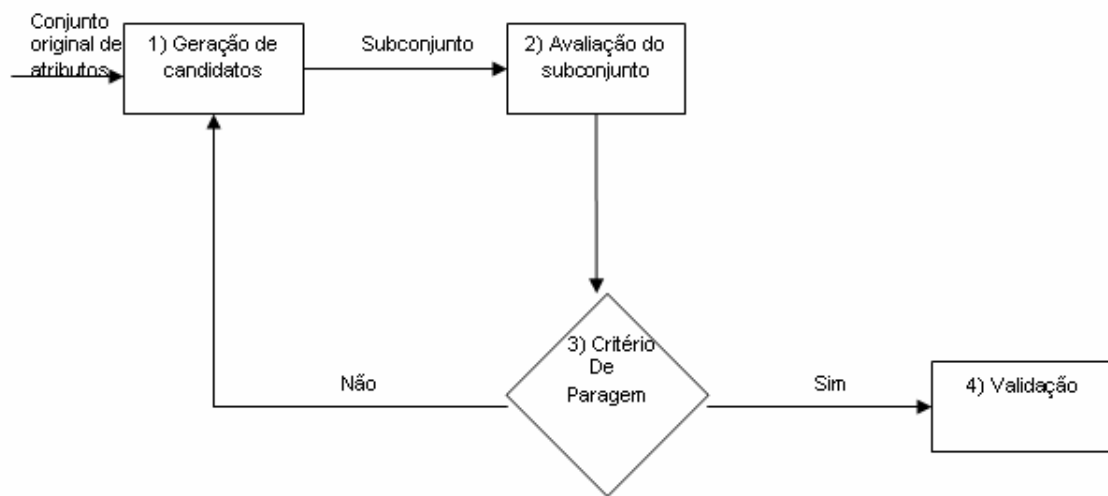


Figura 7 – Processo de procura do melhor conjunto de atributos

Os passos ilustrados na Figura 7 passamos a detalhar a seguir:

Em 1) o procedimento de geração de candidatos, funciona como um “motor” que gera candidatos segundo três modos possíveis:

- O sistema arranca sem candidatos e em cada iteração acrescenta um novo;
- O sistema arranca com todos os candidatos e em cada iteração elimina um;
- O sistema, sem qualquer sequência gera candidatos aleatoriamente.

Em 2) cada subconjunto candidato gerado é avaliado segundo uma determinada função de avaliação, cujo valor é comparado com a melhor avaliação obtida em candidatos anteriores. Dos dois selecciona-se o melhor, substituindo o melhor subconjunto anterior.

Em 3) define-se, em que condições o sistema deverá parar. A necessidade de definir um critério de paragem reside no facto de a selecção do melhor subconjunto de variáveis ser computacionalmente complexo e também no facto de a partir de determinado momento não se identificar interesse em gerar mais subconjuntos candidatos. Assim o critério de paragem pode ser estabelecido numa das três formas seguintes: a primeira, se atingido um determinado número de iterações ou se atingido uma determinada dimensão do

subconjunto de variáveis, o ciclo pára; a segunda, se numa nova iteração e segundo a métrica de avaliação não se obtém ganho face à anterior, o ciclo pára; a terceira, se atingido um determinado valor de avaliação previamente definido o ciclo pára.

Resumindo, cada iteração compreende o processo que gera um candidato, por incremento ou eliminação, avalia-o segundo a métrica escolhida, selecciona o melhor subconjunto comparativamente com o melhor anterior e finalmente verifica o critério de paragem. Se aquele se verifica pára, caso contrário continua em nova iteração.

Em 4) após “disparo” do critério de paragem, procede-se à validação do subconjunto seleccionado. Esta fase não é considerada intrinsecamente, como fazendo parte do processo de selecção de atributos e muitas vezes é obviada, pelo que não será objecto de tratamento e desenvolvimento.

5.1.1 Estratégias de procura no espaço de atributos

Se o espaço original de atributos for igual a N , então o número total de subconjuntos de atributos a serem gerados será 2^N o que inviabiliza uma geração total dos candidatos. Para resolver este problema existem, essencialmente, três abordagens possíveis:

Procura

- Completa;
- Heurística;
- Aleatória.

A primeira abordagem de geração de candidatos executa uma procura completa do subconjunto óptimo de candidatos de acordo com a função de avaliação escolhida. Procura completa, não significa exaustiva, sendo que são usadas algumas heurísticas, que reduzindo substancialmente a exaustividade da procura, não prejudicam a possibilidade de encontrar o subconjunto de atributos óptimo.

A segunda abordagem mencionada acima é a procura heurística. A diferença é que neste tipo de procura, em cada iteração, todos os atributos não seleccionados em iterações anteriores continuarão a ser sempre candidatos, excepto quando seleccionados. Na sua versão mais simples e mais usual, e tal como vimos em 1) a geração de subconjuntos de atributos é incremental, tal significa que numa dada iteração o subconjunto de atributos a avaliar é composto pelo subconjunto avaliado na iteração anterior, acrescido/diminuído de um atributo. Contrariamente ao anterior este procedimento de geração é incremental, daí resultando um espaço de procura de N^2 .

Finalmente a procura aleatória é bastante diferente dos anteriores, pois apesar de possuir um espaço de procura de 2^N , não o percorre todo, pois um número de iterações limite é fixado, percorrendo-o de forma aleatória.

5.1.2 Função avaliação

A função avaliação permite medir a capacidade que um atributo ou um conjunto de atributos tem em distinguir o conceito-alvo. Genericamente consideram-se 4 tipologias de métricas, actualmente em uso, que poderão formar uma função avaliação e que se encontram descritas a seguir.

Distâncias

Esta métrica é conhecida como medida de separação ou discriminação. Se tomarmos como exemplo um conjunto de dados com 2 atributos X e Y e uma variável-classe binária, X será preferido relativamente a Y, se aquele conseguir separar melhor as probabilidades condicionadas: $P(x | C1)$ e $P(x | C2)$. A distância euclidiana entre as instâncias de dois grupos $\{C1, C2\}$ é a medida mais frequentemente usada.

Entropia

A entropia mede a quantidade média de informação necessária (em bits) para identificar a classe de uma determinada instância. Ora para um atributo ou um conjunto de atributos, medirá a quantidade média de informação necessária, para aqueles atributos identificarem a classe de uma determinada instância. Quanto maior a quantidade média de informação, mais entropia existe e menos os atributos distinguem o conceito-alvo.

A entropia é um conceito oriundo da teoria da informação [10] que foi utilizado, pela primeira vez por Quinlan, no seu famoso algoritmo, antecessor do C4.5, ID3, para a selecção dos nós e dos respectivos testes da árvore de decisão.

A entropia da variável i representa-se da seguinte forma:

$$\text{Entropia}_i = - p(\text{Bom} | x_i) \times \log_2(\text{Bom} | x_i) - p(\text{Mau} | x_i) \times \log_2(\text{Mau} | x_i) \quad (5.1)$$

A entropia da variável-alvo ou classe representa-se da seguinte forma:

$$\text{Entropia}_c = - p(\text{Bom}) \times \log_2(\text{Bom}) - p(\text{Mau}) \times \log_2(\text{Mau}). \quad (5.2)$$

O “information gain” representa-se: $\text{Entropia}_c - \text{Entropia}_i$, quanto maior a diferença entre a entropia de i face a c, maior a capacidade do atributo i em explicar c.

Medidas de associação

Estamos a falar das medidas convencionais de correlação ou de associação entre variáveis. Exemplo dessas medidas: R de Pearson, Qui-Quadrado. O princípio é o seguinte: se a associação entre o atributo A e a classe C for maior que a associação entre o atributo B e a classe C, então selecciona-se A.

“Minimum description length” (MDL)

Dado um conjunto de dados e dois modelos possíveis, o princípio do MDL reitera que a melhor generalização é aquela que minimiza o número de bits ou o tamanho de informação, se desejarmos comunicar a teoria ou a generalização mais o conjunto de dados, dada a teoria, através de um canal sem a existência de ruído. Aplicando à selecção de atributos [9], segundo o MDL interessa seleccionar o subconjunto de atributos que minimize o número de bits necessários para transmitir as classes das respectivas instâncias. A função avaliação expressa-se numa equação que quantifica o número de bits da informação veiculada por um determinado subconjunto de atributos.

5.2 Métodos de Selecção de atributos do tipo “wrapper”

Nestes, a função avaliação a utilizar mede a capacidade do atributo ou dos atributos, em distinguir a classe, usando o método de aprendizagem, que irá ser empregue, posteriormente na modelização. Isto é, seja um discriminante linear ou uma regressão logística, todos estes modelos, estimam os seus parâmetros com base no erro cometido no conjunto de treino, sendo que os processos de estimação são diferentes entre eles. A ideia é aplicar este princípio na selecção de atributos, seleccionando os atributos com menor erro, segundo o tipo de modelo que irá ser utilizado na fase de modelação.

Aos processos de selecção de atributos que utilizam esta métrica de avaliação denominam-se de “wrapper methods”, os descritos anteriormente costumam ser denominados de “Filter methods”.

5.3 Estratégia de Selecção de atributos Multi-variada e/ou uni-variada

Diz-se que uma selecção de atributos é uni-variada, quando independentemente dos outros, se avalia a capacidade de um atributo em distinguir a variável-classe. Assim sendo e segundo esta estratégia apura-se o melhor subconjunto de atributos, pela avaliação individual de cada um deles. Este tipo de selecção parte do princípio (ingénuo ou “naive”) que os atributos são independentes uns dos outros. Por oposição a esta, a selecção multi-variada avalia a capacidade de um subconjunto de atributos em distinguir a variável-classe. Esta estratégia, que aliás está subjacente à presente Secção, parte do princípio que as variáveis não são completamente independentes e que portanto um determinado atributo isoladamente terá um determinado peso explicativo, hipoteticamente diferente, se co-avaliado com outros atributos.

Apesar daquelas duas estratégias serem aparentemente antagónicas, há autores [3] que sustentam que elas são complementares e que por tal preconizam duas fases sequenciais na selecção de atributos:

- Selecção “wrapper” uni-variada ou teste de significância (uni-variado) aos coeficientes.

Para cada variável e partindo de uma determinada linguagem de representação mede-se no conjunto de treino a capacidade do atributo isoladamente maximizar a probabilidade de uma dada instância ser “Bom”, quando é efectivamente observado como “Bom”. A medida de avaliação, no caso da regressão logística, é o LRT – Likelihood **ratio Test**, no caso do Discriminante Linear – **Lambda de Wilks**. Estes serão brevemente explicados no capítulo seguinte. Esta primeira selecção serve para despistar variáveis insignificantes do ponto de vista explicativo, de forma a tornar menos trabalhosas as fases subsequentes. Tal como será explicado tanto o LRT como o Lambda de Wilks, seguem uma distribuição de Qui-quadrado, pelo que a selecção uni-variada se fará em função de um P-value previamente fixado. Hosmer e Lemeshow [3], preconizam, apoiando-se para o efeito no estudo efectuado por “Mickey e Greenland”, que se deverá eliminar variáveis cujo P-value seja superior a 0,25.

- Selecção “wrapper” Multivariada de atributos.
Entre outros Hosmer e Lemeshow [3] propõem o método “passo a passo” (“stepwise”), que tal com anteriormente explicamos corresponde a uma selecção de atributos com uma estratégia de geração de candidatos: “heurística”, com uma métrica de avaliação: “Erro do classificador ou do algoritmo” e com um stopping criteria: eliminação ou entrada de uma nova variável não melhora o LRT.

5.4 Binarização de atributos nominais ou de atributos discretizados

Como é sabido as variáveis qualitativas não poderão ser, directamente, modeladas através dum discriminante ou duma regressão linear. A estas variáveis, Hosmer e Lemeshow [3] denominam-nas de “Design variables” e deverão ser transformadas, ou codificadas em valores numéricos.

Consideremos que a variável V tem os seguintes valores possíveis (factores):

$$\{Val_1, Val_2, Val_3, Val_4, Val_5, \dots, Val_N\}.$$

Pretende-se criar N-1 variáveis binárias de forma a representarem os factores de V. Pela leitura do Quadro 7, ilustramos os valores de cada variável binária criada, para cada factor ou valor da variável original.

Quadro 7 – Binarização de variáveis

V	V1	V2	V3	V4	V5	...	VN-1
Val1	0	0	0	0	0	...	0
Val2	1	0	0	0	0	...	0
Val3	0	1	0	0	0	...	0
Val4	0	0	1	0	0	...	0
Val5	0	0	0	1	0	...	0
...	...	...	...	...	...	...	...
ValN	0	0	0	0	0	...	1

Variável Original

Variáveis binárias criadas

A título de exemplo o factor Val₃ corresponderá ao seguinte vector de valores das variáveis binárias {0,1,0,0,0, ..., 0}.

5.5 Tratamento de valores omissos (“missing Values.”)

Difícilmente se encontram conjuntos de dados onde não existam valores omissos ou “missing values”, pelo que é imperativo saber qual a melhor forma de lidar com estes para que não afectem a eficácia do modelo. Michie et al. [5] sintetiza quais os tipos de valores omissos com que normalmente nos deparamos, quais as técnicas para os ultrapassar de forma a diminuir o impacto negativo no acerto do modelo.

Existem dois tipos de “missings”: (i) Os verdadeiramente omissos, isto é, a informação existe mas desconhecemos o seu respectivo valor ou conteúdo; (ii) os que efectivamente não existem. Exemplo disso é, considerando dois atributos “profissão de um individuo” e “antiguidade na entidade patronal”, em que este último não é determinável se considerarmos um individuo com a profissão de reformado. A técnica a utilizar para tratar os “missings” dependerá da sua tipologia e também do tipo de atributo, isto é, se é numérico ou nominal. Se o “missing” é considerado como em (i), e se numérico é prática corrente substituí-lo pela média ou mediana do atributo; se nominal é normalmente assumido como omissos com a descrição de “desconhecido”, a alternativa é substituí-lo pelo valor mais frequente, a moda. Por sua vez se o valor omissos é considerado como em (ii) e se numérico este é considerado em geral com valor N/A (não aplicável).

É importante referir que os missings não existem somente nas variáveis independentes, pelo que se existirem instâncias com a variável dependente omissa, Michie [5] aconselha a sua remoção.

5.6 Discretização de atributos

São várias as razões que conduzem à necessidade de discretizar variáveis contínuas. Witten e Frank [4], justificam a necessidade de discretizar em função da algoritmia a utilizar:

- Naive Bayes, é um classificador intrinsecamente elaborado para o tratamento de atributos nominais. Embora preveja o tratamento de variáveis numéricas, será sempre no pressuposto, nem sempre seguro, que estas seguem uma distribuição normal. Ora para evitar tal insegurança, é aconselhável proceder-se á discretização das variáveis.
- Árvores de Decisão, este classificador trata variáveis numéricas, no entanto existem dois problemas quando estas são frequentes no conjunto de dados: Em primeiro lugar, para se proceder ao ponto de corte (“cut point”) no nó, de um atributo numérico, por vezes o tempo de processamento torna-se demorado chegando a ser inviável; em segundo lugar, o mesmo atributo numérico pode ser seleccionado várias vezes á medida que caminhamos em profundidade na árvore, correspondendo a vários nós com vários cortes diferentes. Ora tal complexifica a árvore, tornando-a de difícil interpretação. Este problema também se resolve com a discretização.
- Hosmer and Lemeshow [3] no seu livro “Applied logistic regression”, propõem a discretização de variáveis contínuas, num problema de regressão logística aplicada ao “Low birth weight data set”, quando para uma determinada variável contínua se percebe que o logaritmo do “odds ratio”, se desvia acentuadamente da linha da regressão. Neste caso a variável deverá ser discretizada, pois a regressão logística assenta no pressuposto que deverá haver uma relação linear entre a variável e o seu “logit”.
- Do nosso ponto de vista existem outras razões, pelas quais a discretização se torna, por vezes, indispensável: Em primeiro lugar, se um atributo numérico, possuir valores omissos, uma das formas será discretizar o atributo, para que se possa tratar o missing como “NÃO RESPONDEU” e assim estaremos na presença de um atributo nominal “transformado”. Em segundo lugar, nos problemas em que as regressões lineares são utilizadas, a discretização é um meio importante para fornecer robustez ao modelo resultante, tornando-o mais generalizável. A discretização é também um meio para combater os valores extremos e os “outliers” que tanto perturba a estimação dos parâmetros. No entanto a discretização, não é gratuita, faz-se à custa de perda de informação do atributo, mas em nome da abstracção. O problema está em como discretizar optimizando o binário: perda de informação versus abstracção. Veremos mais adiante como tratar este problema

5.6.1 Estratégias de Discretização

Existem 2 estratégias de discretização:

- **Discretização não supervisionada**

Os seguintes tipos de discretização, não têm em conta a classe no processo de discretização: (i) “Equal – interval binning” – A discretização é efectuada através de intervalos de igual amplitude (exemplo: 5-10;10-15;15-20,). Normalmente este tipo de discretização, provoca um desequilíbrio no número de elementos por intervalo, desvirtuando a capacidade preditiva do atributo. (ii) “Equal - frequency binning” – No sentido de eliminar as desvantagens identificadas no anterior, este permite que os intervalos se formem com amplitudes variadas, mas com o mesmo número de elementos. No entanto, ambos discretizam tendo em conta um critério que em nada se relaciona com o conceito-alvo, podendo desvirtuar a capacidade preditiva do atributo.

- **Discretização supervisionada**

Por discretização supervisionada entende-se, o conjunto de técnicas que discretizam uma variável contínua de modo a que a discretização obtida discrimine ou separe os grupos que constituem a variável-classe. De entre essas técnicas destaca-se o “Entropy-based discretization”, como sendo o mais popular [4]. Esta técnica consiste basicamente em aplicar o princípio de tratamento de atributos numéricos nas árvores de decisão, nomeadamente no C4.5 [10], quando se determina um ponto de corte (que maximize o “information gain”) para o teste do nó. A única particularidade desta técnica consiste em recursivamente encontrar pontos de corte, para a formação de intervalos, até se verificar um determinado critério de paragem. Descrevemos no ANEXO 1, a sequência lógica dos passos do “Entropy-based discretization”.

5.7 Agrupamento de factores (“Grouping attributes levels”)

Assim como é necessário “agrupar” valores numéricos, definindo intervalos, também é necessário agrupar factores de uma variável nominal. Para variáveis nominais com elevado número de factores, os problemas que surgem na construção de árvores de decisão serão idênticos aos abordados para os atributos numéricos, quer ao nível do tempo de processamento quer ao nível da complexidade (tamanho da árvore). Para os problemas de regressão que envolvem modelos de regressão linear ou para os problemas de classificação que envolvam discriminantes, os atributos nominais com k factores são transformados em k-1 variáveis binárias. Ora se um atributo nominal tem 93 factores (Como é o que consta no caso prático), 92 variáveis binárias serão constituídas aumentando a complexidade do modelo, o perigo de “overfitting” e a dificuldade de compreensão da realidade modelada. Para além do que se acaba de referir, o elevado número de factores [1] pode propiciar escassez de instâncias em alguns factores podendo portanto produzir um modelo pouco generalizável e pouco robusto. A presença cada vez mais frequente de variáveis nominais nos problemas a tratar, obriga a que se redobre a atenção à problemática do agrupamento de factores.

5.7.1 Estratégias de agrupamento de factores

Com alguma dificuldade se encontram trabalhos publicados que versem a problemática de tratamento dos factores em variáveis nominais. Tanto quanto pudemos apurar, existem pelo menos, duas referências que expõem o problema de agrupamento de factores e adiantam soluções. A primeira referência é o “Grouping attribute values” de Quinlan [10] e a Segunda é o “Coarse classifying” segundo Lyn C. Thomas [1]. Ambas se encontram descritas a seguir.

Agrupamento de factores segundo Quinlan¹⁰

No âmbito do “Machine learning” e mais propriamente no domínio das árvores de decisão, Quinlan⁸ [10] é o único que refere esta problemática e que adianta uma solução. Com objectivo de diminuir a complexidade das árvores, Quinlan serviu-se do “soybean data” de Michalski, para ensaiar e exemplificar um algoritmo que apura-se grupos de valores de atributos nominais. O atributo usado foi o “Stem Canker”, com os seguintes factores: {None}, {Below soil}, {above soil} e {above 2nd node}. O agrupamento de factores juntamente com o “Gain ratio” é uma das particularidades do C4.5 de Quinlan [10].

O esquema agregador de factores, usado por Quinlan, é semelhante ao “entropy-based discretization”, no que diz respeito à estratégia de geração de candidatos. No entanto regista uma medida de avaliação e um critério de paragem diferentes: Em primeiro lugar, a medida usada é o “Gain ratio” e não a entropia ou mesmo o ganho de informação (“information gain”). O “Gain ratio”, segundo Quinlan, constitui uma melhoria face ao ganho de informação do ID3, versão antecessora do C4.5, e expressa-se em (5.3) pelo quociente entre o ganho de informação do atributo A , dado um determinado agrupamento e o “Split info” do atributo, dado o agrupamento. Quinlan diagnosticou que o ganho de informação de um atributo com muitos valores, sejam eles factores ou números, enviesam aquela medida sobre avaliando-a. A explicação encontra-se, considerando por absurdo um atributo detentor de tantos valores diferentes quantas as instâncias, sendo que para cada valor existe uma classe. Assim sendo se considerarmos que para cada valor corresponde um subconjunto da variável, podemos afirmar que a o ganho de informação da variável A em estabelecer partições da classe é máximo. Este viés é rectificado ponderando o ganho de informação pela quantidade média de informação medida em bits, que é necessária gerar para identificar a que subconjunto da variável pertence uma determinada instância por ele caracterizada – *Split info*. Quanto maior for o número de valores distintos de um atributo, maior será a quantidade de informação a transmitir para identificar que uma determinada instância detém um determinado valor no atributo A .

$$\text{Gain ratio} = \frac{\text{Info (Classe)} - \text{info}_c(A)}{\text{Split Info}_c(A)} \quad (5.3)$$

¹⁰ Capítulo 7, página 63 a 69

Em segundo lugar, o critério de paragem não obedece ao princípio do MDL tal como no “entropy-based discretization”, apenas, tão simplesmente é despoletado quando na iteração n o “gain ratio” é menor ou igual ao “gain ratio” na iteração $n-1$.
 Descrevemos de seguida a sequência lógica dos passos do “grouping attributes values” presente no C4.5:

-
1. Ordenar alfabeticamente os valores possíveis do atributo nominal.
 2. Se o atributo detém k factores, procede-se à criação de todos as combinações, dois a dois, entre factores. O número de grupos resultante será equivalente a C_k^2 .
 3. Cálculo do “Gain ratio” para cada grupo candidato ou combinação obtida. Considerando agora, c como a combinação ou grupo de factores a ser avaliada.
 - 3.1 Cálculo do “information gain” de A para a combinação c :
 $\text{Gain}_c(A) = \text{Info}(\text{Classe}) - \text{Info}_c(A)$. Ver a expressão (A4) do ANEXO 1.
 - 3.2. Cálculo do split $\text{info}_c(A) = -P(k_1|n) \times \log_2(P(k_1|n)) - P(k_2|n) \times \log_2(P(k_2|n)) \dots P(k_i|n) \times \log_2(P(k_i|n))$, Em que k é o número de factores correspondentes à combinação c num máximo de i combinações.
 - 3.3 Cálculo de “gain ratio” segundo a expressão (5.3)
 4. Seleccionar a combinação que Max (“Gain ratio”). Remover, individualmente, os factores que constituem a combinação e substituí-los pela própria combinação eleita.
 5. Se Max (“Gain ratio”) ITERAÇÃO $n < \text{Max}(\text{“Gain ratio”})$ ITERAÇÃO $n-1$, pára, caso contrário repete 2, 3 e 4.
-

Agrupamento de factores (“Coarse classifying”) segundo Lyn C. Thomas¹¹

“Coarse classifying” é o termo utilizado na área de “scoring” para denominar o processo de discretização de variáveis contínuas e de agrupamento de factores das variáveis nominais. Lyn C. Thomas [1], entende que o agrupamento de factores tem “tanto de arte como de ciência” (“...is as much art as it is sciense”) e preconiza que o agrupamento de factores se deva efectuar de acordo com dois critérios: Dois factores se devem agrupar se os respectivos “odds¹²” forem os mais semelhantes¹³; Por outro lado, a semelhança conceptual entre factores (exemplo agrupar “ Casa alugada sem Móvel” e “Casa alugada com móvel”), constitui critério de agrupamento.

¹¹ Página 132

¹² Entende-se por “odds” o quociente entre o número de elementos da classe C_1 e a classe C_2 .

¹³ Entende-se por “odds” mais semelhantes de todos os pares possíveis de factores, o par cuja diferença dos respectivos “odds” é mínima.

Para melhor compreensão do leitor considere a variável *Tipohab1* e os respectivos factores tal como se encontra ilustrado no Quadro 8.

Quadro 8-Distribuição das instâncias de treino pertencentes a cada factor de *Tipohab1* por cada uma das classes.

Factores variável <i>Tipohab1</i>	Maus	Bons	Frequência	odds
{Fami}	367	233	600	0,634
{Arra}	130	93	223	0,715
{Out}	4	3	7	0,750
{Arcc}	93	70	163	0,752
{Empr}	7	9	16	1,285
{Prop}	160	275	435	1,718
{Aqui}	80	141	221	1,762
{Desc}	0	1	1	Inf

Dado que se identificou que é necessário agrupar os factores em dois grupos finais, segundo Lyn C. Thomas, efectuam-se os seguintes passos:

1. Identificar os factores com a frequência relativa baixa¹⁴. Destes, agrupar os que têm “odds” mais semelhantes. Pela observação do Quadro 8, identificamos os factores {Out}, {Empr} e {Desc} com frequência baixa. Destes agruparemos {Empr} com {Desc}, uma vez que {Out}, dada a semelhança de “odds” com {Arra} e {Arcc}, facilmente agrupará com um destes.
2. Identificar os factores com significado semelhante. Identificamos que *Arra* – *Arrendatário de apartamento* e *Arcc* – *Arrendatário de casa*, são aproximados no conceito e não dispõem de “odds” radicalmente diferentes.
3. Tendo em conta o número de grupos a atingir, deveremos agrupar iterativamente os restantes factores em grupos com factores que contenham “odds” semelhantes. Assim sendo, os primeiros a serem agrupados são {Arcc} e {Out}, pois detêm os “odds” mais semelhantes, resultando deste agrupamento um “odds” de 0,7528. Os segundos a serem agrupados são {Arcc, Out} com {Arra}, pois os respectivos “odds” (0,7528 e 0,715) são os mais semelhantes. Os terceiros a serem agrupados são {Prop} com {Aqui} resultando deste agrupamento um “odds” de 1,73. Em quarto lugar procede-se ao agrupamento de {Fami} com {Arra,Out,Arcc}. Por fim em quinto lugar procede-se ao agrupamento de {Prop,Aqui} com {Empr,Desc}. O resultado final encontra-se ilustrado no Quadro 9.

¹⁴ Lyn C.Thomas não define o que se entende por frequência relativa baixa. Para o exemplo que acima descrevemos entendemos por frequência relativa baixa aquela que é igual ou inferior a 1%.

Quadro 9- Resultado do agrupamento de factores segundo Lyn C.Thomas

Grupos de factores	Maus	Bons	odds
{Fami, Arra, Out, Arrc}	594	399	0,672
{Empr, Prop, Aqui, Desc}	247	426	1,725

Esta prática, utilizada na área de “scoring” e sustentada por alguns autores [1] e [2], é eminentemente intuitiva justificando a máxima utilizada – “tanto de arte como de ciência”. Lyn C.Thomas[1], assume que o julgamento sensível subjacente a esta técnica é preponderante e necessário, no entanto reclama uma parte à ciência, afirmando que algumas decisões tomadas no passo número 3 poderão ser auxiliadas pela Estatística, através de testes do Qui-quadrado.

Resta-nos afirmar que as duas técnicas de agrupamento de factores apresentadas (a de Quinlan e a de Lyn C.Thomas), são técnicas supervisionadas. Isto é, á semelhança do “Entropy-based discretizor” estas procuram, quando efectuam o agrupamento de factores, separar as classes da variável-alvo, tentando afectar o maior número de instâncias pertencentes a uma classe em detrimento da outra, no respectivo grupo obtido.

6- Proposta de melhoria de algumas técnicas de pré-processamento e experimentação das mesmas na modelização do caso prático

Pretendemos que este capítulo informe o leitor sobre:

- As propostas de melhorias do agrupamento de factores e do “entropy-based discretization”;
- Exemplos práticos sobre a aplicação das técnicas propostas;
- As tarefas inerentes à modelização do caso prático, utilizando algumas técnicas descritas no Capítulo anterior, bem como as técnicas propostas;
- Apuramento e comparação dos resultados produzidos por três estratégias de pré-processamento, contendo uma delas as propostas de melhoria apresentadas no presente Capítulo.

6.1 Proposta de melhoria do agrupamento de factores

Das duas estratégias de agrupamento de factores referidas na Secção 5.7, consideramos discutíveis os seguintes aspectos:

- Um agrupamento de factores criterioso e objectivo, não se compadece com um agrupamento de factores a partir da respectiva observação sensível e subjectiva dos “odds”. Este tipo de agrupamento de factores é preconizado por Lyn C. Thomas [1].
- Lyn C. Thomas, preocupa-se e bem, com os factores pouco significativos, no entanto contradiz-se quando afirma que estes devem ser agrupados por semelhança. Ora se são pouco significativos, incorremos num grave erro de os agrupar por aquele mesmo critério.
- Ambas estratégias, não acautelam o grau de abstracção do agrupamento obtido, medido num conjunto de validação através de uma determinada estatística de teste. É que o mote do agrupamento é em prole da robustez e da generalização e portanto importa saber a partir de que número de grupos é que um determinado agrupamento é generalizável ou não.

Por outro lado, a generalização é facilmente obtenível, mas sempre à custa da perda de informação por via da agregação. No limite podemos transformar as variáveis nominais em binárias, teremos sempre um bom P-value no teste obtido para medir abstracção, no entanto, provavelmente, com o mesmo P-value posso ter mais informação no mesmo atributo com 3 factores em vez de 2.

- Se utilizarmos o método de Quinlan [10], no nosso caso prático, o número de grupos é invariavelmente dois.

Consideramos que estas quatro alíneas constituem motivo suficiente para que se procure aperfeiçoar as técnicas existentes de pré-processamento. Tal como a seguir facilmente se pode constatar, o conteúdo da nossa proposta intercepta com a estratégia de Quinlan referida na Secção 5.7, pelo que consideramos que a nossa proposta não é uma novidade, mas sim um complemento àquela. De forma a nortear a nossa proposta e a estabelecer parâmetros de sucesso, definimos que se entende por melhor agrupamento, aquele que cumpra três requisitos: Em primeiro lugar o agrupamento obtido deverá possuir da maior capacidade de abstracção possível, medida através do P-value de uma determinada estatística de teste; Em segundo lugar minimize a perda de ganho de informação comparativamente ao atributo sem agrupamento e por último que a predictibilidade do modelo obtido, a partir de variáveis com factores agrupados segundo o presente método, não seja prejudicada significativamente.

Existem quatro passos fundamentais na nossa proposta, pelo que passaremos a descrevê-los, nas quatro Secções seguintes. Posteriormente, fazendo uso da variável nominal *Tipohab1* – *Tipo habitação*, concretizaremos o método proposto de forma a aumentar a sua compreensão face ao leitor. Os passos são os seguintes:

1. Dividir o conjunto de dados em dois subconjuntos: O Conjunto de treino e o Conjunto de validação.
2. Geração dos grupos-candidatos e a procura do melhor no Conjunto de treino.
3. Escolha da função avaliação ou da métrica de avaliação.
4. Definição do critério de paragem.

6.1.1 Dividir o conjunto de dados em 2 subconjuntos: O Conjunto de treino e o Conjunto de validação

Convencionou-se que o conjunto de treino será de 60% do conjunto de dados e deverá ser extraído de forma aleatória *estratificada*¹⁵ (Quadro 10), isto é, a proporção de cada factor do atributo (*Tipohab1*) que pretendemos tratar é preservada no conjunto de treino. A razão desta opção prende-se com o facto de, caso optasse-mos por uma amostragem aleatória simples, incorreríamos no risco de não conter a totalidade dos factores presentes no conjunto de treino, conduzindo-nos a um modelo necessariamente omissivo e portanto pouco representativo. Como a seguir se explica e se demonstra, o conjunto de treino servirá para ensaiar agrupamentos de factores com o objectivo de procurar o melhor agrupamento, agrupamento esse avaliado por uma determinada métrica (passo 3 - Secção 6.1.3). Como também veremos, a procura iterativa do melhor agrupamento será efectuada por uma heurística (passo 2 – Secção 6.1.2) uma vez que o espaço das hipóteses é elevado. Ao processo de ensaiar agrupamentos de factores, também denominamos de geração de grupos-candidatos e ao melhor grupo candidato numa determinada iteração denominamos de grupo eleito.

¹⁵ A constituição do conjunto de treino de forma aleatória e estratificada foi efectuada nos termos de [19]

Quadro 10

Conjunto de treino						
	Levels	Proportion Levels	sample size	Freq by level	Levels	Description
1	{Desc}	0.0006	1	1	1	{Desc} DESCONHECIDO
2	{Out}	0.0042	4	7	2	{Out} OUTROS
3	{Empr}	0.0096	10	16	3	{Empr} ALOJADO POR ENTIDADE
4	{Arcc}	0.0978	98	163	4	{Arcc} ARRENDATARIO DE CASA
5	{Aqui}	0.1326	133	221	5	{Aqui} EM AQUISIÇÃO
6	{Arca}	0.1338	134	223	6	{Arca} ARRENDATARIO DE APAR
7	{Prop}	0.2611	261	435	7	{Prop} PROPRIETARIO
8	{Fami}	0.3501	350	500	8	{Fami} ALOJADO PELA FAMILIA
	Total		1001	1666		

Elaborado em R

O Quadro 10 indica por cada “level” ou factor, na coluna “Proportion levels” a proporção de cada factor no conjunto total de dados composto por 1666 instâncias, sendo a coluna “Freq. by level” a correspondente quantidade de instâncias por factor. O “sample size”, indica a quantidade de instâncias por factor que constituem o grupo de treino para a geração de grupos-candidatos. As instâncias seleccionadas foram aleatoriamente obtidas por factor, obedecendo à proporcionalidade da população. Por exemplo no factor, {Prop} cuja frequência relativa na população é de 26,11%, foram aleatoriamente obtidas 261 instâncias.

Por sua vez o conjunto de validação servirá para avaliar a generalização de cada grupo eleito em cada iteração, servindo igualmente para verificar se o critério de paragem (passo 4 – Secção 6.1.4), previamente definido, é atingido. Por generalização de um grupo eleito entende-se a “capacidade” de aquele preservar, num conjunto independente, o valor da função de avaliação demonstrada no treino. Tal significa que se pretende, para um determinado grupo eleito, averiguar se a métrica é estatisticamente idêntica no conjunto de treino e no conjunto independente de validação. Tal averiguação é aferida através de um teste estatístico, cujo P-value determinará o resultado.

No conjunto de validação constituído pelos restantes 40% (no Quadro 10 resulta da diferença entre as colunas “Freq by level” e “sample size”), optou-se por o fraccionar em 10 subconjuntos, de igual dimensão, aleatoriamente constituídos. Esta opção fundamenta-se em dois aspectos: Em primeiro lugar consideramos que dez serão o maior número possível de subconjuntos, tendo em conta a dimensão dos dados. Em segundo lugar, somos da opinião que dez subconjuntos produzirão valores da métrica e dispersão da mesma (a variância), de modo a que o teste estatístico a imprimir sobre o grupo eleito seja rigoroso.

6.1.2 Geração dos grupos-candidatos e a procura do melhor no conjunto de treino

Depois da constituição dos conjuntos de treino e de validação, procedemos à geração de candidatos e procurar o melhor, segundo uma métrica (definida na Secção seguinte). À semelhança do que foi explicado anteriormente, as estratégias possíveis de geração de subconjuntos de atributos – candidatos, são aplicáveis na geração de grupos-candidatos. Passemos de seguida a aprofundar as duas principais estratégias de geração de grupos-candidatos e respectiva procura do melhor, a *Procura Completa* e a *Procura Heurística*.

Procura Completa

Consideramos um atributo w com 4 factores: A, B, C e D.

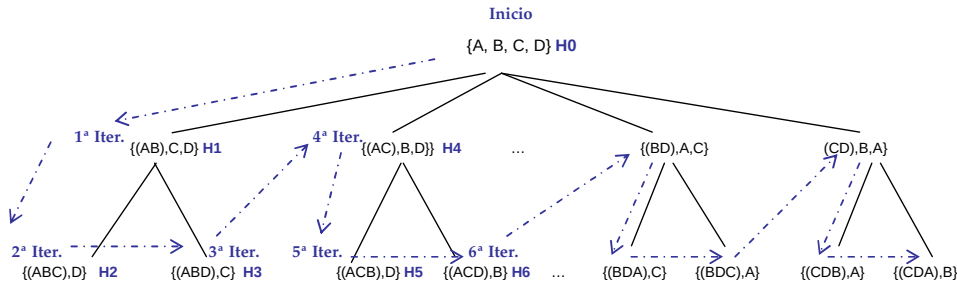


Figura 8 – Exemplo de um algoritmo de procura completo “top-down” e em profundidade primeiro

A geração *completa* de grupos-candidatos e a procura do melhor, faz-se utilizando um algoritmo de procura completa representado na Figura 8. Tal como é ilustrado, partindo da *hipótese maximamente geral* (H_0), expande-se a procura para a *hipótese imediatamente mais específica* (H_1), entendendo-se por imediatamente mais específica, um dos grupos-descendentes ($\{(AB),C,D\}$, $\{(AC),B,D\}$, $\{(AD),B,C\}$, $\{(BC),A,D\}$, $\{(BD),A,C\}$ e $\{(CD),B,A\}$) constituídos a partir de todas as combinações possíveis dos elementos do grupo-ascendente ($\{A, B, C, D\}$). Avaliado H_1 verifica-se se atinge o objectivo ou critério de paragem predefinida, e se sim o algoritmo termina a procura, caso contrário expande-se, para a hipótese imediatamente mais específica que será H_2 . Se H_2 não satisfazer, procura H_3 e assim sucessivamente. Ao algoritmo representado na Figura 8, quanto à geração de grupos-candidatos e procura do melhor, denomina-se de *cima para baixo* (“top-down”), pois a procura é efectuada a partir do topo da árvore e denomina-se de *em profundidade primeiro* [20], pois tal como se pode observar pela orientação das setas na figura 10, a procura parte do topo directamente á máxima profundidade atingida pelo grupo-candidato $\{(ABC),D\}$. Uma outra estratégia que poderia ser também adoptada seria a de *largura primeiro*.

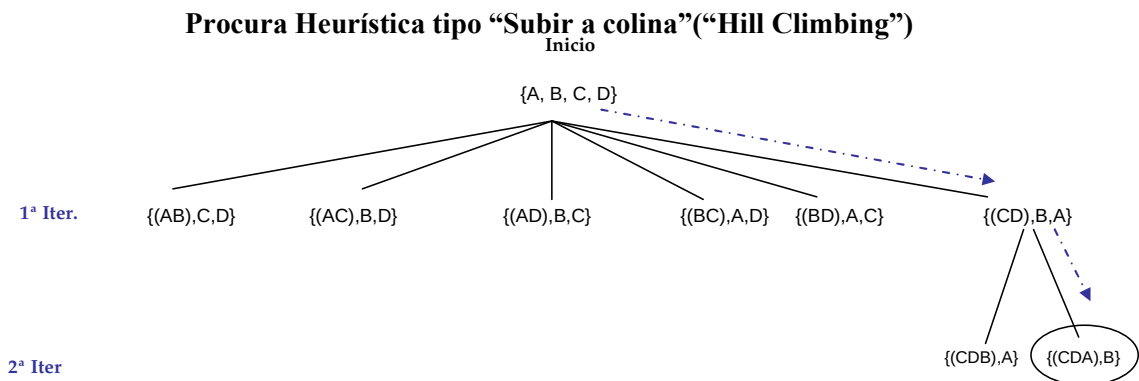


Figura 9 – exemplo de um algoritmo “Subir a colina”

Enquanto os algoritmos de procura completa, como o próprio nome indica percorrem, completa e sistematicamente o espaço das hipóteses, entendendo-se por espaço de hipóteses todas as combinações¹⁶ possíveis entre os factores, o algoritmo de “Subir a colina” (“Hill Climbing”) orienta a procura, em consonância com a métrica de avaliação escolhida, evitando a pesquisa completa no espaço das hipóteses [20]. Tal como se observa na Figura 9, a primeira iteração avalia o conjunto de grupos-candidatos imediatamente específicos de H_0 , elegendo entre eles a hipótese ou o grupo-candidato que maximiza medida escolhida para avaliar grupos-candidatos ($\{(CD),B,A\}$). Encontrado o melhor e se este não verificar o critério de paragem, a iteração seguinte faz expandir para os respectivos imediatamente mais específicos, tornando a eleger do conjunto de grupos-candidatos resultante, o melhor ($\{(CDA),B\}$). Como o grupo eleito é constituído por dois elementos, o algoritmo pára.

De referir duas características importantes das heurísticas: Em primeiro lugar, a heurística que aqui apresentamos é irreversível (“greedy search”), ou seja depois de seleccionar um candidato não retrocede. De qualquer modo existem categorias de heurísticas avançadas onde a retroacção é possível; Em segundo lugar, a heurística que aqui representamos é do tipo “Para a frente” (“Foward”), tal significa que a heurística é iniciada com conjunto vazio e em cada iteração acrescenta um novo factor.

Destas duas abordagens possíveis, a nossa opção recai sobre a *Procura heurística*, porque se trata de uma procura que do ponto de vista computacional consome menos tempo e recursos e também pelo facto daquela, na selecção de atributos (“Stepwise”) se revelar eficaz na procura obtida.

¹⁶ Formalmente é constituída pela soma das combinações: $C_3^4 + C_2^4$, para o exemplo descrito.

6.1.3 Escolha da função avaliação ou da métrica de avaliação

A escolha de um candidato é efectuada segundo uma métrica de avaliação ou função de avaliação. Das várias métricas de avaliação em uso, descritas na Secção 5.1.2, iremos optar pelo ganho de informação (“information gain”), pois têm sido registados bons resultados no âmbito da discretização de variáveis contínuas [4]. O “Gain ratio” de Quinlan, também poderia ser utilizado, não indo porém, divergir dos resultados obtidos pelo ganho de informação em cada iteração para cada grupo-candidato. Tal se justifica pelo denominador da expressão (5.3), o *split info*, manter-se estável em cada grupo-candidato no que toca à respectiva entropia.

6.1.4 Definição do critério de paragem

O critério de paragem irá parar a heurística, quando o resultado do teste estatístico entre o valor da métrica apurada no treino e na validação, para um determinado grupo eleito, atingir o P-value desejado. Desta forma garante-se que o número de grupos atingido salvguarde, minimamente, a generalização do agrupamento impedindo que o ganho de informação continue a descer pela agregação desnecessária de factores. Assim para este efeito é necessário definir dois aspectos: Em primeiro lugar, qual a estatística de teste a utilizar para comparar o valor da métrica obtida no conjunto de treino com o valor da métrica obtida no conjunto de validação, com intuito de aferir a generalização do grupo-eleito; Em segundo lugar, qual o P-value a partir do qual se considera que o grupo-eleito é generalizável, isto é, a partir de que valor do P-value o valor da métrica do agrupamento, demonstrada e medida no treino, tem “capacidade” de se manter em novos casos. Para efectuar esse teste teremos que eleger o ganho de informação como variável aleatória e determinar se esta se ajusta a alguma lei de probabilidade. Tal como se demonstra em ANEXO 2, o ganho de informação de quase a totalidade dos atributos nominais constante no caso prático, se ajusta a uma distribuição normal, sendo de prever que genericamente o ganho de informação se distribui daquela forma. Se assim se confirmar a normalidade, estamos em condições de testar o valor obtido no conjunto de treino com as médias obtidas no conjunto de validação e testar a sua igualdade através de “one-sample t-test” [21] tal como se encontra descrito na Secção 3.5.4. Caso não se confirme a normalidade do ganho de informação nos atributos nominais deverá optar-se por um teste não paramétrico, como o “Wilcoxon-Man-Whitney test” (descrito também na Secção 3.5.4). Finalmente deveremos definir o P-value. A definição deste é subjectiva e depende dos objectivos do decisor. Consideremos o agrupamento da variável *Tipohabl* e consideremos por agora que não se estabeleceu nenhum critério de paragem. Como é visível na Figura 10 à medida que as iterações vão aumentando, mais os factores serão agregados, implicando que o ganho de informação vá diminuindo, no entanto por outro lado, o P-value aumenta com as iterações. Este “trade-off” – ganho de informação versus P-value complica a fixação de uma regra de decisão, pois como é óbvio pretende-se seleccionar o agrupamento com o maior ganho de informação e o maior P-value.

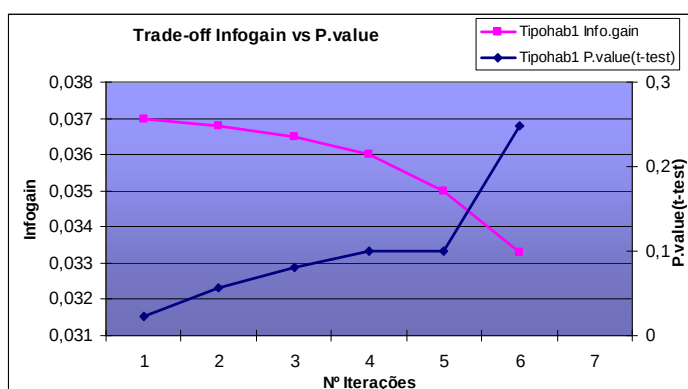


Figura 10 - “trade-off” entre o ganho de informação e o P-value

Onde se deve parar? O decisor fixará o P-value que mais lhe convier, que pode ir dos tradicionais 5% de significância até valores aproximadamente iguais a 25%.

Acabada que está a descrição dos passos inerentes à nossa proposta de agrupamento de factores, passaremos de seguida a explicar o agrupamento de factores respeitante à variável *Tipohab1*.

6.1.5 Agrupamento da variável *Tipohab1*, nos termos da nossa proposta

O Quadro 11 sintetiza o detalhe das iterações e respectivos resultados respeitantes ao agrupamento dos factores da variável *Tipohab1*. Concentremo-nos na primeira iteração – “iter 1” do Quadro 11 e consideremos a seguinte sequência de passos:

- A heurística gera $C_2^{8\ 17}$ no conjunto de treino (60% do conjunto de dados), que perfaz vinte e oito candidatos distintos. Estes encontram-se na coluna “Grupos de candidatos”.
- Para cada um deles apurou-se o ganho de informação observável na coluna “Infogain treino”.
- A heurística aplica a função avaliação: $\text{Max}(\text{infogain } i\text{-enésima iteração})$ e selecciona o candidato $\{\text{Fami}\}, \{\text{Arrc, Empr}\}, \{\text{Arra}\}, \{\text{Desc}\}, \{\text{Aqui}\}, \{\text{Out}\}, \{\text{Prop}\}$ com um ganho de informação de treino de 0,037. O candidato eleito está assinalado com X na coluna “Máximo do Infogain”.

¹⁷ Designa-se por combinações de n elementos distintos tomados k a k , os agrupamentos constituídos por k daqueles elementos que diferem uns dos outros apenas pelos elementos que neles figuram, independentemente da ordem pela qual se dispõem. A seguinte expressão formula o que se acaba de afirmar.

$$C_{n,k} = \frac{n!}{(n-k)! \times k!}$$

- Utiliza-se o agrupamento, atrás eleito, nos 10 conjuntos de validação e apura-se o ganho de informação de validação para cada um dos conjuntos. O ganho de informação de validação não consta no quadro 11, mas para melhor esclarecimento do leitor apresentamos de seguida:

Itera.1
0.09152962; 0.06470302; 0.04302353; 0.12996715; 0.01488538; 0.04967261; 0.23749805; 0.09173782; 0.09827170; 0.07724340

- Efectua-se um “one-sample t-test” aos ganhos de informação obtidos nos subconjuntos de validação, com o valor fixo do ganho de informação obtido no treino. Se P-value superior a 0,18 então pára, caso contrário prossegue. Tal significa que o critério de paragem considerado é o P-value superior ou igual a 0,18. Para as restantes iterações, são repetidos os passos anteriores, até disparo do critério de paragem.

Quadro 11 – Demonstração do agrupamento de factores, segundo nossa proposta, para a variável *Tipohabl*

Iter	Grupos Candidatos	Infogain Treino	Maximo do Infogain	One-sample t-test	Wilcoxon test
0	{Fami}, {Empr}, {Arra}, {Arrc}, {Desc}, {Aqui}, {Out}, {Prop}	0,0370	-----	-----	-----
1	{Fami,Empr}, {Arra}, {Arrc}, {Desc}, {Aqui}, {Out}, {Prop}	0,0366			--
1	{Fami,Arra}, {Empr}, {Arrc}, {Desc}, {Aqui}, {Out}, {Prop}	0,0362			
1	{Fami,Arrc}, {Arra}, {Empr}, {Desc}, {Aqui}, {Out}, {Prop}	0,0355			
1	{Fami,Desc}, {Arra}, {Arrc}, {Empr}, {Aqui}, {Out}, {Prop}	0,0356			
1	{Fami,Aqui}, {Empr}, {Arra}, {Desc}, {Arrc}, {Out}, {Prop}	0,0158			
1	{Fami,Out}, {Empr}, {Arra}, {Arrc}, {Desc}, {Aqui}, {Prop}	0,0367			
1	{Fami,Prop}, {Empr}, {Arra}, {Arrc}, {Desc}, {Aqui}, {Out}	0,0140			

1	{Fami}, {Arrc,Empr}, {Arra}, {Desc}, {Aqui}, {Out}, {Prop}	0,0370	X	0.023	0.009

1	{Fami}, {Empr}, {Arra}, {Arrc}, {Desc}, {Aqui}, {Out,Prop}	0,0354			
Candidato seleccionado na 1ª iter {Fami}, {Arrc,Empr}, {Arra}, {Desc}, {Aqui}, {Out}, {Prop}					
2	{Fami,Arrc,Empr}, {Arra}, {Desc}, {Aqui}, {Out}, {Prop}	0,0352			
2	{Fami,Arra}, {Arrc,Empr}, {Desc}, {Aqui}, {Out}, {Prop}	0,0362			
2	{Fami, Desc}, {Arrc,Empr}, {Arra}, {Aqui}, {Out}, {Prop}	0,0356			
2	{Fami, Aqui}, {Arrc,Empr}, {Arra}, {Desc}, {Out}, {Prop}	0,0157			
2	{Fami,Out}, {Arrc,Empr}, {Arra}, {Desc}, {Aqui}, {Prop}	0,0367			
2	{Fami, Prop}, {Arrc,Empr}, {Arra}, {Desc}, {Aqui}, {Out}	0,0140			
2	{Fami}, {Arrc,Empr,Arra}, {Desc}, {Aqui}, {Out}, {Prop}	0,0368	X	0.056	0.037 1

2	{Fami}, {Arrc,Empr}, {Arra}, {Desc}, {Aqui}, {Out,Prop}	0,0354			
Candidato seleccionado na 2ª iter {Fami}, {Arrc,Empr,Arra}, {Desc}, {Aqui}, {Out}, {Prop}					
3	{Fami,Arrc,Empr,Arra}, {Desc}, {Aqui}, {Out}, {Prop}	0,0349			
3	{Fami,Desc}, {Arrc,Empr,Arra}, {Aqui}, {Out}, {Prop}	0,0354			
3	-----				
3	{Fami, Out}, {Arrc,Empr,Arra}, {Desc}, {Aqui}, {Prop}	0,0365	X	0.081	0.037 1
3	-----				
Candidato seleccionado na 3ª iter {Fami, Out}, {Arrc,Empr,Arra}, {Desc}, {Aqui}, {Prop}					
4	{Fami, Out,Arrc,Empr,Arra}, {Desc}, {Aqui}, {Prop}	0,0346			
4	-----				
4	{Fami, Out}, {Arrc,Empr,Arra}, {Desc}, {Aqui,Prop}	0,0360	X	0.107	0.083
Candidato seleccionado na 4ª iter {Fami, Out}, {Arrc,Empr,Arra}, {Desc}, {Aqui,Prop}					
5	-----				
5	{Fami, Out}, {Arrc,Empr,Arra}, {Desc,Aqui,Prop}	0,0353	X	0.102	0.083
Candidato seleccionado na 5ª iter {Fami, Out}, {Arrc,Empr,Arra}, {Desc,Aqui,Prop}					
6	-----				
6	{Fami, Out,Arrc,Empr,Arra}, {Desc,Aqui,Prop}	0,0333	X	0.249	0.43
Candidato seleccionado na 6ª iter {Fami, Out,Arrc,Empr,Arra}, {Desc,Aqui,Prop} → Stop Criteria					

6.1.6 O problema dos factores não significativos

Continuando a considerar a variável nominal *Tipohab1* e tendo em conta o agrupamento eleito segundo a nossa proposta:

[1º Grupo] {Fami, Out, Arrc, Empr, Arra}

[2º Grupo] {Desc, Aqui, Prop}

Dois factos se constatarem, a partir da observação dos Quadros 12 e 13: O primeiro, é que os factores {Desc}, {Out} e {Empr}, correspondem a 1,5% das 999 instâncias de treino; O segundo, é que os factores {Prop} e {Aqui} são os factores com os “odds” mais elevados, e por esse facto foram agrupados. Estes são frequentes, correspondendo respectivamente a 26% e 13% das instâncias. Associados àqueles, o factor {Desc} que compõe o terceiro elemento do grupo, detém um “odds” superior aos restantes dois factores do grupo, no entanto não é significativo na frequência, correspondendo a 0,01% das instâncias. Isto significa, que não pondo em causa a abstracção do agrupamento na sua globalidade, no que toca ao ganho de informação, na obstante, não podemos deixar de colocar as seguintes questões: Fará sentido proceder ao agrupamento de factores não significativos, quanto à sua frequência relativa, afinal estes são similares aos “outliers”? É válido que se integre o factor {Desc}, num determinado grupo, sabendo que os “odds” que lhe estão associados, dada a escassez da frequência, podem não corresponder na realidade? Sabendo que o tipo de agrupamento imprimido é supervisionado, avaliando os candidatos pelo ganho de informação, tende-se a criar grupos a partir de factores com “odds” semelhantes, no entanto fá-lo independentemente da frequência relativa do factor. Este problema urge de ser tratado pois pode perigar a validade dos agrupamentos criados.

Quadro12 - Dados do conjunto de treino (60%)

Matriz de distribuição ascendente					
	Maus	Bons	freq	Dist%	Acum
{Desc}	0	1	1	0.001	0.001
{Out}	3	1	4	0.004	0.005
{Empr}	5	5	10	0.010	0.015
{Arrc}	52	46	98	0.098	0.113
{Aqui}	45	88	133	0.133	0.246
{Arra}	75	59	134	0.134	0.380
{Prop}	100	161	261	0.261	0.641
{Fami}	219	139	358	0.358	1.000

Quadro 13 - Dados do conjunto de treino (60%)

Matriz de "odds"				
	Maus	Bons	odds	freq
{Out}	3	1	0.333	4
{Fami}	219	139	0.634	358
{Arra}	75	59	0.786	134
{Arrc}	52	46	0.884	98
{empr}	5	5	1.000	10
{Prop}	100	161	1.610	261
{Aqui}	45	88	1.955	133
{Desc}	0	1	Inf	1

Posta a problemática impõe-se uma solução.

- **Solução para os factores não significativos**

A solução proposta assenta nos seguintes princípios:

- Se existem factores não significativos no conjunto de dados de treino, certamente existirão na população. Portanto eles devem ser tratados como tal. Assim sendo os não significativos deverão ser agrupados num grupo, que vai ser denominado de “BOLO”.
- Ordenando a frequência relativa dos factores por ordem crescente, o grupo “BOLO” será determinado, ensaiando, neste caso, três cortes no sentido ascendente. O Quadro 14 ilustra os três cortes.
- Para cada cenário de corte, efectua-se agrupamento de factores preconizado segundo a nossa proposta, onde o factor “BOLO” é um factor como outro qualquer.
- Por cada cenário de corte, com base na regra previamente decidida, sairá um agrupamento vencedor. Dos três agrupamentos vencedores, será eleito aquele que tiver maior ganho de informação. No Quadro 15 poderá observar o detalhe dos três vencedores, respeitantes aos três cortes e o vencedor final.

Quadro 14 – Três cenários de corte para o factor “BOLO”

Levels	Frequência Acum	1º cenário- Estratégia corte a 2%		2º cenário - Estratégia corte a 12%		3º cenário - Estratégia corte a 25%	
Levels	Frequência Acum	Levels	Frequência Acum	Levels	Frequência Acum	Levels	Frequência Acum
{Desc}	0,10%	BOLO	1,50%	BOLO	11,30%	BOLO	24,60%
{Out}	0,05%						
{Empr}	1,50%	{Arrc}	11,30%	{Aqui}	24,60%	{Arra}	38,00%
{Arrc}	11,30%	{Aqui}	24,60%	{Arra}	38,00%	{Prop}	64,00%
{Aqui}	24,60%	{Arra}	38,00%	{Prop}	64,00%	{Fami}	100,00%
{Arra}	38,00%	{Prop}	64,00%	{Fami}	100,00%		
{Prop}	64,00%						
{Fami}	100,00%						

Quadro 15 – O corte vencedor

Tipohab1						Grupos		
corte	Nº Levels	Infogain-Treino	Pvalue-t	Pvalue-w				
24,6%	2	0,03239	0,618	0,55	{Arra,Fami}	{Prop}	{Bolo}	
11,3%	2	0,0401	0,87	0,375	{Prop,Aqui}	{Bolo,Arra,Fami}		
1,5%	4	0,0387	0,22	0,43	{Arrc,Arra}	{Aqui,Prop}	{Fami}	{Bolo}

↓
O agrupamento vencedor é: {Prop, Aqui} e {Bolo, Arra, Fami} correspondente ao 2º cenário.

6.2 Proposta de melhoria do “entropy-based discretization”

Pela experimentação do “entropy-based discretization” (tal como o descrevemos na Secção 5.6 e respectivo ANEXO 1), diagnosticamos dois problemas cuja resolução poderá aumentar a eficácia desta técnica. Passamos de seguida a descrever brevemente os problemas e as respectivas soluções:

- 1- O “entropy-based discretization”, como vimos, elege recursivamente pontos de corte, seleccionando aquele que imprimir entropia mínima á variável em causa. Acontece que para proceder à eleição de pontos de corte é necessário gerar todos os candidatos e avaliar o ganho de informação de cada um. Este processo é computacionalmente demorado, tornando-se em alguns casos impraticável.

Temperature												
6		5			4		3	2	1			
64	65	68	69	70	71	72	75	80	81	83	85	
Yes	No	Yes	yes	yes	No	Yes	Yes	No	Yes	Yes	No	

Golf Data set (Fonte: Ian Witten e Eibe Frank)

Figura 11

Como solução do problema propomos que os atributos numéricos deverão ser alvo de uma pré-discretização, utilizando o “Equal – frequence binning” (exemplo de 10 percentis), e depois proceder à junção de intervalos adjacente ou “vizinhos”.

- 2- O critério de paragem proposto por Witten e Frank [4] é o MDL (Minimum description length). Dado que o “entropy-based discretization”, elege um número elevado de pontos de corte, o MDL enquanto critério de paragem, seria o mais adequado pois tentaria reduzir o número de intervalos de forma a diminuir a quantidade de informação (em bits) necessários para transmitir as classes das respectivas instâncias. Pela expressão (A5), constante no ANEXO 1, constata-se que o MDL actua, estabelecendo um ganho de informação mínimo que um determinado corte eleito deverá garantir para a constituição de um novo intervalo. Alguns autores [9] [4], alertam que nem sempre o MDL resulta, enunciando as situações em que tal acontece. Efectivamente, ensaiando o MDL no caso retratado na Figura 11, nenhum dos cortes seria eleito pelo critério do MDL, não havendo portanto lugar a discretização para aquela variável. No exemplo que a seguir apresentamos igual situação ocorre, facto que demonstramos no ANEXO 3. Posto isto propomos o mesmo critério de paragem utilizado no agrupamento de factores, por nós preconizado.

6.2.1 “Entropy-based discretizor” aplicado à variável Antemp1, nos termos da nossa proposta

Passamos de seguida a descrever o processo de “entropy-based discretizor” por nós proposto, mostrando a sequência de os passos a efectuar:

- Sabendo que os valores omissos desta variável foram identificados com -9999, para proceder à discretização desta variável iremos ignorá-los, numa primeira fase, excluindo-os.
- Procedeu-se á pré-discretização da variável, gerando 9 intervalos, com igual frequência relativa de observações. Os intervalos resultantes são os seguintes.

“(-1,0]” “(0,1]” “(1,2]” “(2,3]” “(3,4]” “(4,7]” “(7,10]” “(10,15]” “(15,36]”

- Por cada iteração efectuem-se todas as junções, possíveis, de dois intervalos adjacentes ou “vizinhos”. Elege-se a junção vencedora pelo critério de máximo ganho de informação. Faz-se aplicar a junção vencedora nos dez grupos de validação e afere-se a igualdade do ganho de informação obtido no treino com a média dos ganhos de informação obtidos nos conjuntos de teste. Há lugar a paragem quando o P-value atingir um valor superior a 0,18. Exemplificamos no Quadro 16 os passos relatados, à semelhança do que fizemos no agrupamento de factores. Desta vez só identificamos os infogain de treino no candidato vencedor.
- Depois da discretização dos valores numéricos, torna-se necessário integrar os valores omissos excluídos e numericamente identificados por -9999. Optamos por introduzir o intervalo] -99999, -9999] e agrupá-lo com um dos intervalos apurados no procedimento de discretização. Para tal utilizamos o agrupamento de factores por nós proposto. Como o Quadro 17 ilustra, o intervalo] -99999,9999] agrupa com o intervalo] 7,15].

Quadro 16 – Demonstração da discretização da variável *Antemp1*, segundo nossa proposta.

Iter	Grupos Candidatos	Infogain Treino	Maximo do Infogain	One –sample t-test	Wilcoxon test
0	]-1,0] ,]0,1] ,]1,2] ,]2,3] ,]3,4] ,]4,7] ,]7,10] ,]10,15] ,]15,36]	0,059	-----	-----	-----
1	]-1,1] ,]1,2] ,]2,3] ,]3,4] ,]4,7] ,]7,10] ,]10,15] ,]15,36]				
1	]-1,0] ,]0,2] ,]2,3] ,]3,4] ,]4,7] ,]7,10] ,]10,15] ,]15,36]				
1	]-1,0] ,]0,1] ,]1,3] ,]3,4] ,]4,7] ,]7,10] ,]10,15] ,]15,36]				
1	]-1,0] ,]0,1] ,]1,2] ,]2,4] ,]4,7] ,]7,10] ,]10,15] ,]15,36]				
1	]-1,0] ,]0,1] ,]1,2] ,]2,3] ,]3,7] ,]7,10] ,]10,15] ,]15,36]				
1	]-1,0] ,]0,1] ,]1,2] ,]2,3] ,]3,4] ,]4,10] ,]10,15] ,]15,36]				
1	]-1,0] ,]0,1] ,]1,2] ,]2,3] ,]3,4] ,]4,7] ,]7,15] ,]15,36]	0,062	x	0.0113	0.005
Candidato seleccionado na 1ª iter]-1,0] ,]0,1] ,]1,2] ,]2,3] ,]3,4] ,]4,7] ,]7,15] ,]15,36]					
2	]-1,1] ,]1,2] ,]2,3] ,]3,4] ,]4,7] ,]7,15] ,]15,36]				
2	]-1,0] ,]0,2] ,]2,3] ,]3,4] ,]4,7] ,]7,15] ,]15,36]				
2	]-1,0] ,]0,1] ,]1,3] ,]3,4] ,]4,7] ,]7,15] ,]15,36]	0,0617	X	0.012	0.027
2	]-1,0] ,]0,1] ,]1,2] ,]2,4] ,]4,7] ,]7,15] ,]15,36]				
2	]-1,0] ,]0,1] ,]1,2] ,]2,3] ,]3,7] ,]7,15] ,]15,36]				
2	]-1,0] ,]0,1] ,]1,2] ,]2,3] ,]3,4] ,]4,7] ,]7,36]				
Candidato seleccionado na 2ª iter]-1,0] ,]0,1] ,]1,3] ,]3,4] ,]4,7] ,]7,15] ,]15,36]					
3	]-1,1] ,]1,3] ,]3,4] ,]4,7] ,]7,15] ,]15,36]				
3	]-1,0] ,]0,3] ,]3,4] ,]4,7] ,]7,15] ,]15,36]				
3	]-1,0] ,]0,1] ,]1,4] ,]4,7] ,]7,15] ,]15,36]	0,0615	x	0.032	0.048
3	]-1,0] ,]0,1] ,]1,3] ,]3,7] ,]7,15] ,]15,36]				
3	]-1,0] ,]0,1] ,]1,3] ,]3,4] ,]4,15] ,]15,36]				
3	]-1,0] ,]0,1] ,]1,3] ,]3,4] ,]4,7] ,]7,36]				
Candidato seleccionado na 3ª iter]-1,0] ,]0,1] ,]1,4] ,]4,7] ,]7,15] ,]15,36]					
4	]-1,1] ,]1,4] ,]4,7] ,]7,15] ,]15,36]	0,060	x	0.071	0.16
4	]-1,0] ,]0,4] ,]4,7] ,]7,15] ,]15,36]				
4	]-1,0] ,]0,1] ,]1,7] ,]7,15] ,]15,36]				
4	]-1,0] ,]0,1] ,]1,4] ,]4,15] ,]15,36]				
4	]-1,0] ,]0,1] ,]1,4] ,]4,7] ,]7,36]				
Candidato seleccionado na 4ª iter]-1,1] ,]1,4] ,]4,7] ,]7,15] ,]15,36]					
5	]-1,4] ,]4,7] ,]7,15] ,]15,36]	0,059	x	0.32	0.43
5	]-1,1] ,]1,7] ,]7,15] ,]15,36]				
5	]-1,1] ,]1,4] ,]4,15] ,]15,36]				
5	]-1,1] ,]1,4] ,]4,7] ,]7,36]				
Candidato seleccionado na 5ª iter]-1,4] ,]4,7] ,]7,15] ,]15,36] → Stopping Criteria					

Quadro 17 – Agrupamento do factor -9999 (omisso), num dos intervalos descritos no Quadro anterior

Iter	Grupos Candidatos	Infogain Treino	Máximo do Infogain	t-paired test	Wilcoxon test
0	]-1,4] ,]4,7] ,]7,15] ,]15,36] ,]-99999,-9999]	0,058			
1	{]-1,4] ,]-99999,-9999] } ,]4,7] ,]7,15] ,]15,36]				
1	]-1,4] , {]4,7] ,]-99999,-9999] } ,]7,15] ,]15,36]				
1	]-1,4] ,]4,7] , {]7,15] ,]-99999,-9999] } ,]15,36]	0,0585	X	0.1832	0.322
1	]-1,4] ,]4,7] ,]7,15] , {]15,36] ,]-99999,-9999] }				
Candidato seleccionado na 1ª]-1,4] ,]4,7] , {]7,15] ,]-99999,-9999] } ,]15,36] → Stopping Criteria					

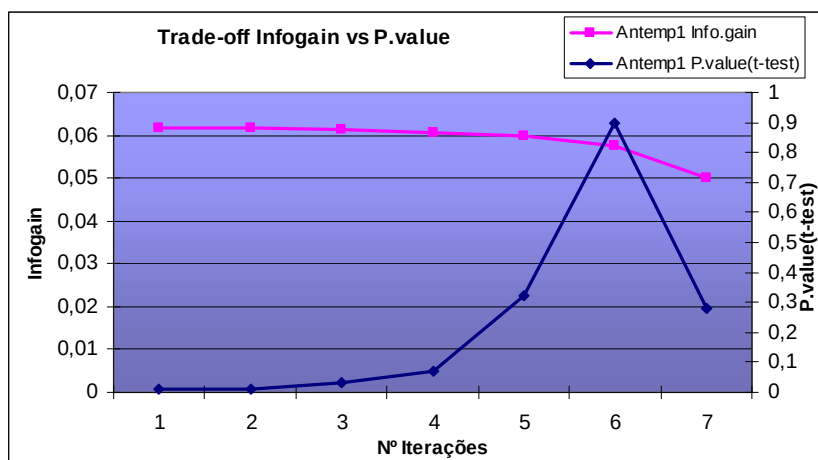


Figura 12 – “trade-off” entre o ganho de informação e o P-value

Na Figura 12 constatamos o mesmo facto, observado no agrupamento de factores na Figura 10, à medida que se vai efectuando a união de intervalos adjacentes, o *infogain* diminui, e por outro lado aumenta a probabilidade do valor daquela se situar dentro de um intervalo de valores na população. É com este sentido que utilizamos a palavra abstracção de um determinado agrupamento de valores ou junção de intervalos adjacentes. No entanto nem sempre se verifica que à medida que o *infogain* diminui, aumenta o P-value. O que apresentamos, na figura 12 é exemplo disso. Caso não houvesse critério de paragem, na 7ª iteração, aumenta a entropia, diminuindo o ganho, e diminui o P-value.

Uma vez que estão descritas as nossas preconizações de melhoria ao nível do agrupamento de factores e do “entropy-based discretization”, deveremos apurar se aquelas são vantajosas face aos resultados obtidos pela utilização das técnicas de pré-processamento usuais.

As Secções seguintes apresentarão os resultados produzidos pelas técnicas propostas que por sua vez serão comparadas de forma a concluir a oportunidade das nossas propostas.

6.3 Modelização do caso prático utilizando as propostas de melhoria

O método de pré-processamento proposto nesta Secção consiste nas seguintes fases sequenciais: Divisão do conjunto de dados em dois subconjuntos: O conjunto de treino composto por 1666 instâncias e o conjunto de teste composto por 714 instâncias (à semelhança do que foi efectuado na Secção 4.3.3); agrupamento de factores das variáveis nominais; discretização das variáveis numéricas; binarização das variáveis nominais ou discretizadas; selecção dos atributos. No final destas fases procede-se ao treino e teste dos modelos de Regressão logística e Discriminante Linear. Este método é comparado (na Secção 6.4) com a aplicação apenas da fase de binarização (sem selecção de atributos) e com a aplicação de binarização com selecção de atributos através do método de “stepwise”.

Para este estudo foi utilizado o conjunto de dados descrito na Secção 4.2. Este conjunto de dados tem 2.380 registos e 22 atributos (incluindo a classe), tendo-se destacado para treino 1.666 registos. Cada registo representa um crédito concedido, sendo caracterizado por atributos de natureza demográfica, de operação de crédito e a classe. A amostra foi recolhida de forma aleatória e apresenta igual proporção de exemplos das classes “Bom” e “Mau”.

Nas secções seguintes descreve-se a aplicação das várias fases do método proposto ao conjunto de dados em análise. A Secção 6.4 apresenta os resultados do treino e teste dos modelos com os três tipos de pré-processamento considerados.

6.3.1 Agrupamento de factores das variáveis nominais

Sobre todas as variáveis nominais, efectuou-se o agrupamento dos factores conforme a técnica proposta na Secção 6.1. A técnica referida tem como finalidade reduzir o número de factores de uma variável nominal, reduzindo a sua complexidade sem reduzir significativamente o seu valor informativo. A técnica usada é baseada no critério de ganho de informação.

Quadro 18: Resultado do agrupamento de factores ao conjunto dados em análise

Variável	Nº de factores antes	Nº de factores depois	Agrupamento de Quinlan
<i>Estcivil1</i>	7	2	2
<i>Tipohab1</i>	8	2	2
<i>Origtelpes</i>	6	3	2
<i>Bem</i>	8	2	2
<i>Estatuto1</i>	6	2	2
<i>Prof1</i>	93	5	2
<i>Distrito</i>	19	2	2
<i>País</i>	11	2	2
<i>Canal</i>	2	2	2
<i>Sex1</i>	2	2	2

O Quadro 18 apresenta os resultados obtidos pela aplicação do método proposto. Pode-se observar que das dez variáveis nominais, duas são originalmente binárias. Das restantes oito variáveis “agrupáveis”, a técnica de agrupamento proposta permitiu binarizar cinco variáveis, tendo-se obtido um número de factores superior a dois em apenas duas variáveis.

Por outro lado, a aplicação ao nosso conjunto de dados da técnica preconizada por Quinlan, ver Secção 5.7.1, resulta sempre na obtenção de variáveis binárias. Este facto parece-nos redutor, pois poderá resultar na perda excessiva do valor informativo do atributo.

Chamamos a atenção do seguinte:

Dado que se constituiu um “holdout group test” equivalente a 30% das 2.380 instâncias, restam para treino 1.666 instâncias. Tal como é preconizado pela nossa proposta o treino do agrupamento será efectuado em somente 60% das 1.666 disponíveis, sendo os restantes 40% o conjunto de validação.

6.3.2 Discretização de variáveis numéricas

No tratamento de variáveis numéricas optou-se por discretizar as variáveis com valores omissos e aquelas cujo incumprimento do pressuposto da linearidade, subjacente ao Discriminante linear e à Regressão logística, poderia pôr em perigo a validade dos modelos. Nas alíneas seguintes passamos a descrever aplicação do método aos dados em análise.

- **Tratamento das variáveis com valores omissos**

As variáveis com valores omissos nos dados em análise são *Antempl* (antiguidade no emprego) e *Vrendl* (rendimento mensal). Estas duas variáveis contêm valores omissos, não por deficiência na recolha da informação, mas pelo simples facto de não existirem efectivamente. Exemplo disso é a impossibilidade de atribuir antiguidade do emprego – *Antempl*, a um cliente reformado e salário – *Vrendl* a um trabalhador por conta própria. Assim sendo optamos por proceder à discretização da variável tal como se encontra descrita e exemplificada na Secção 6.2.1, tendo resultado os intervalos indicados no Quadro 19.

Quadro 19-Intervalos resultantes da discretização, segundo nossa proposta, para as variáveis *Antempl* e *Vrendl*

Variável	Intervalos
<i>Antempl</i> ¹⁸	] 0, 4];] 4, 7]; {] 7, 15],] -99999, -9999]};] 15, +∞]
<i>Vrendl</i> ¹⁹	{] 0, 500],] -99999, -9999]};] 500, +∞]

Na discretização das variáveis referidas, optou-se por considerar os valores omissos como um factor da própria variável. De facto, no caso destas variáveis não faria sentido substituir os valores omissos por uma média ou mediana, uma vez que não existe qualquer valor admissível. Note-se o exemplo de um reformado, classe profissional

¹⁸ O intervalo] -99999, -9999] representa os valores omissos nesta variável e que está associado à profissão de reformado.

¹⁹ O intervalo] -99999, -9999] representa os valores omissos nesta variável e que está associado ao estatuto de trabalhador por conta própria.

bastante valorizada, que poderia ver o seu score prejudicado, por lhe ser atribuído uma antiguidade no emprego não correspondente. Atribuir uma antiguidade no emprego a um reformado iria enviesar modelo.

- **Desvios ao pressuposto da linearidade dos modelos.**

O problema de desvio de linearidade dos modelos foi referido na Secção 5.6, como sendo um dos motivos para a discretização de uma variável contínua. De facto, Hosmer and Lemeshow [3] enfatizam este problema e adiantam uma técnica simples para o detectar. A técnica descrita consiste em discretizar a variável em intervalos com o mesmo número de elementos (“Equal - frequency binning”) e estimar para cada intervalo o modelo. De seguida, calcula-se a média do valor estimado pelo modelo para as observações que se situam no intervalo, e faz-se corresponder ao ponto médio do intervalo o valor médio das estimativas.

A Figura 13, representa o comportamento da variável dependente estimada quando um modelo de regressão logística é aplicado globalmente à variável *Prazo*. A Figura 14 representa o resultado da aplicação de modelos de regressão logística locais aplicados a intervalos da variável *Prazo*. (intervalos determinados pelo método “Equal - frequency binning”). A comparação das duas figuras indica que a assumpção do modelo linear global é questionável.

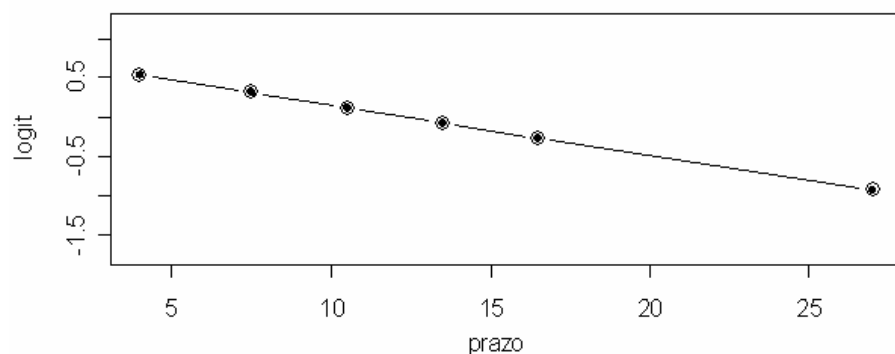


Figura 13- Representa a recta da regressão logística simples aplicada à variável *Prazo*

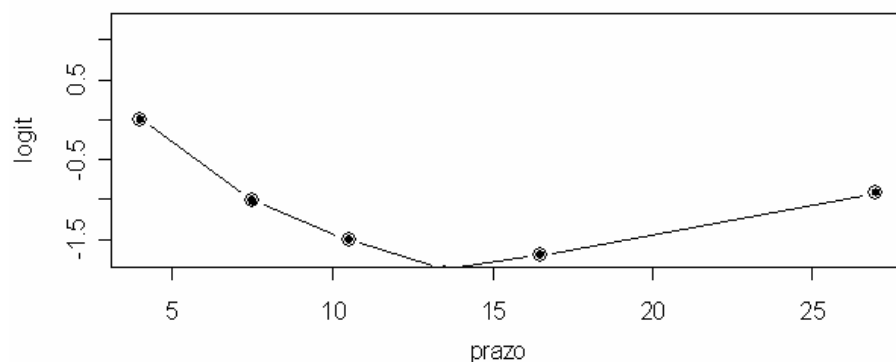


Figura 14 – Representa a linha resultante da união dos pontos se considerarmos em cada intervalo uma regressão logística distinta

Pela observação da Figura 14, se constata que a forma da curva perfaz um U (“U-shaped”). Tal significa que a variável não é crescente nem descendente no seu “logit”, percebendo-se assim a razão pela qual a recta da regressão representada na Figura 13 detém uma inclinação reduzida. O registo deste facto justifica, mais uma vez, a discretização das variáveis nestas circunstâncias, caso contrário o coeficiente obtido, pela aplicação do modelo na variável, será reduzido ou não significativo. A variável *Vpvp* tem um comportamento idêntico ao do *Prazo*.

A análise de desvio ao pressuposto da linearidade foi executada para todas as variáveis contínuas, aplicando tanto a regressão logística como o método do discriminante linear. Os resultados obtidos foram idênticos para ambos os algoritmos (Regressão Logística e Discriminante linear) tendo-se concluído que as variáveis *Vpvp* e *Prazo* deverão ser discretizadas por indicarem fortes desvios à linearidade. O Quadro 20 apresenta os intervalos apurados resultantes do processo de discretização das variáveis contínuas *Vpvp* e *Prazo*.

Quadro 20-Intervalos resultantes da discretização, segundo nossa proposta, para as variáveis *Prazo* e *Vpvp*

Variável	Intervalos
<i>Prazo</i>	] 2, 6];] 6, 9];] 9, 18];] 18, +∞]
<i>Vpvp</i>	] 298, 486];] 486, 798];] 798, 1090];] 1090, +∞]

Chamamos a atenção do seguinte:

Dado que se constituiu um “holdout group test” equivalente a 30% das 2.380 instâncias, restam para treino 1.666 instâncias. Tal como é preconizado pela nossa proposta o treino da discretização será efectuado em somente 60% das 1.666 disponíveis, sendo os restantes 40% o conjunto de validação.

6.3.3 Tratamento de dados – binarização das variáveis nominais ou discretizadas

Para que estas variáveis nominais e variáveis agora discretizadas possam ser utilizadas na Regressão logística e no discriminante, terão que ser convertidos em variáveis numéricas binárias. Este procedimento foi efectuado conforme descrito no capítulo anterior na Secção 5.4.

6.3.4 Selecção de atributos

A estratégia a utilizar foi explicada na Secção 5.3. O facto desta estratégia contemplar uma selecção preliminar uni-variada de atributos, possibilitando assim que a selecção multi-variada se efectue com menor esforço de processamento, fundamentou a nossa escolha nesta estratégia.

- **Seleccção uni-variada – Regressão logística**

Quadro 21- Matriz de “odds” da variável *Profl*, agrupada em cinco factores dos 93 originais

Matriz de "odds" <i>profl</i>				
	Maus	Bons	odds	freq
P5	216	85	0.393	301
P1	339	301	0.887	640
P4	157	189	1.203	346
P3	125	197	1.576	322
P2	4	53	13.250	57

Tomemos a variável *Profl* (após agrupados os respectivos factores) para exemplificar a selecção uni-variada. Esta consiste em estimar os parâmetros relativos às quatro variáveis²⁰ {P1,P2,P3,P4}, constantes no Quadro 21, resultantes da binarização efectuada aos cinco valores possíveis da variável *Profl*. De seguida de forma a ajuizar a pertinência explicativa da variável *Profl* formula-se um teste de significância ao vector dos coeficientes associados às quatro variáveis binárias.

Para avaliar a significância do vector de parâmetros, o principio base é comparar os valores observados da variável-classe, com os valores estimados por um modelo com a variável que pretendemos testar. Na regressão logística, a comparação entre valores observados e valores estimados, está implícita na própria função de verosimilhança. Se na expressão (2.8) substituirmos os valores dos parâmetros pelos estimados (pela maximização da função de verosimilhança em ordem aos parâmetros) obtemos um valor denominado de *logaritmo da verosimilhança* (“log likelihood”). Ora se compararmos os “log likelihood” obtidos dos dois modelos (um com a variável outro sem variável), provavelmente conseguiremos aferir se a variável, “per si”, acrescenta significativamente, valor para a predição. A métrica para aferimos aquela comparação, disponível na maior parte dos softwares, e recomendada pela generalidade dos autores (Lemeshow e Hosmer) é o *Ratio de máxima verosimilhança* (“*Likelihood ratio Test*”), doravante abreviado por *LRT*.

$$LRT = -2 \ln \frac{[\text{"Log likelihood" sem a variável}]}{[\text{"Log likelihood" com a variável}]} \quad (6.1)$$

²⁰ Tal como foi referido na Secção 5.4, se uma variável nominal detém cinco factores, então será transformada em quatro variáveis binárias.

No caso uni-variado, para apurarmos o “log likelihood” sem a variável, é o mesmo que apurarmos o “log likelihood” estimado do “intercept” ou β_0 . Assim transformando o LRT acima definido (6.1) numa expressão, fica a seguinte:

$$LRT = 2 \left\{ \sum_{i=1}^n y_i \ln [P(Y=1 | x_i)] + (1-y_i) \ln [1-P(Y=1 | x_i)] - [n_1 \cdot \ln(n_1) + n_0 \cdot \ln(n_0) - n \ln(n)] \right\} \quad (6.2)$$

Log likelihood da variável a testar
(expressão já definida em 2.8)

Log likelihood de β_0

Legenda:

Em y_i , x_i referem-se ao valor da variável dependente dicotómica e ao valor da variável independente, respectivamente, da i -ésima observação em n . A n_1 e n_0 representam o número de instâncias, respectivamente pertencentes à classe 1 e 0. n é a soma de n_1 e de n_0

Se considerarmos que a H_0 corresponde o coeficiente igual a 0 (zero), a estatística de teste LRT, seguirá uma distribuição χ^2 com um grau de liberdade. Neste caso concreto, a variável “*Profl*”, é composta por 4 variáveis binárias, o que significa que a estatística de teste LRT segue uma distribuição χ^2 com quatro graus de liberdade. Assim sendo os resultados são:

```

profl
Coefficients:
(Intercept)  P2      P3      P4      P1  log-likelihood  log-likelihood "intercept"      LRT      p-value(LRT)
-0.9326      3.5135  1.3875  1.1181  0.8137    -1089.546    -1154.706    130.3214    3.32408e-27

```

Assim se conclui que $LRT = 2x (-1089,546 - (-1154,706)) = 130,32$, e que $P(\chi^2_{(df=4)} > 130,32) = 3.32408e-27$, logo rejeita-se a hipótese nula²¹ que pelo menos um dos coeficientes é diferentes de zero. Portanto a variável *Profl* é uma variável que satisfaz plenamente os requisitos da fase uni-variada e portanto não é eliminada.

Para qualquer outra variável simples o princípio é rigorosamente igual tal como apresentamos de seguida para a variável *Vf*:

```

vf
Coefficients:
(Intercept)      vf  log-likelihood  log-likelihood "intercept"      LRT      p-value(LRT)
0.0867887    -0.0001666    -1154.096    -1154.706      1.22    0.2692128

```

Conclui-se que: $LRT = 2x (-1154,096 - (-1154,706)) = 1,22$, indicando a estatística de teste um P-value de 0.26. Tal significa que não se rejeita a hipótese²², pelo que eliminamos, de forma uni-variada, este atributo.

Tendo-se efectuado este procedimento em todas as variáveis, conclui-se que foram eliminadas as seguintes variáveis: *Canal*, *Sex1* e *Vf*.

²¹ Tal como foi descrito na Secção 5.3., devera-se eliminar variáveis (na selecção uni-variada de atributos) quando o P-value é igual ou inferior a 0.25, caso contrário a variável prossegue para a selecção multi-variada [3].

²² Aplica-se o mesmo princípio do que descrito em 20.

- **Seleccção uni-variada – Discriminante Linear**

A selecção uni-variada, usando como linguagem de representação o discriminante linear, consistirá em eliminar as variáveis, que seguramente, não contribuem significativamente para o rácio entre-grupos/intra-grupos. É a maximização deste ratio que orienta a estimação dos parâmetros, pelo que quanto menor o ratio, menor é a separação entre os grupos e maior é dispersão dentro dos grupos. Ora o Lambda de Wilks irá ser utilizado no sentido de medir o poder que uma função discriminante detém em separar os grupos, função essa composta pela variável que pretendemos testar. O lambda de Wilks formula-se como a seguir se indica:

$$\lambda \text{ Wilks} : \lambda = | \mathbf{W} | / | \mathbf{T} |, \text{ tal que } 0 < \lambda < 1 \quad (6.3)$$

tal que $| \mathbf{W} |$, determinante de $\mathbf{W}$ mede a dispersão dentro dos grupos e $| \mathbf{T} |$, determinante de $\mathbf{T}$, mede a dispersão total, ou seja, intra e entre grupos ($\mathbf{T} = \mathbf{B} + \mathbf{W}$).

A estatística $\lambda \text{ Wilks}$, preconiza na hipótese nula a igualdade das médias da variável nos grupos e pressupõe que as matrizes de variância / covariância nos K grupos sejam iguais. Sob H_0 segue uma distribuição de $\lambda \text{ Wilks}$, que normalmente não se encontra tabelada, mas pode ser aproximada quer à distribuição de F de Fisher, quer à χ^2 .

A aproximação à F de Fisher só é possível se as variáveis seguirem uma distribuição multi-normal. Neste caso a estatística de teste formalmente se reproduz a seguir:

$$F = \left(\frac{1 - \lambda}{\lambda} \right) \left(\frac{(n_1 + n_2 - K)p + 1}{p} \right) \rightarrow F(p, n_1 + n_2 - p - 1)$$

(6.4)

A aproximação ao χ^2 , formula-se como se segue:

$$\chi^2 = - [(n_1 + n_2 - 1) - \frac{1}{2}(p + K)] \ln(\lambda) \rightarrow \chi^2(p \times (K - 1))$$

(6.5)

Legenda:

- n_1 Número de instâncias da classe 1
- n_2 Número de instâncias da classe 2
- p Número de atributos
- K número de classes que neste caso é igual a dois
- λ de Wilks

Aproximando a distribuição de Wilks à χ^2 , realizamos o teste para a função discriminante contendo as 4 variáveis binárias correspondentes ao processo resultante da binarização da variável *Prof1*. Registou-se um P-value aproximadamente igual a zero significando que a hipótese nula, ou seja que os vectores de médias dos grupos são idênticos, é rejeitada. Este teste foi realizado em SPSS.

Tendo-se efectuado este procedimento em todas as variáveis, conclui-se que foram eliminadas as seguintes variáveis: *Canal*, *Sex1* e *Vf*.

- **Seleccção Multi-variada – Regressão logística**

O modo de selecção multi-variado escolhido foi o “stepwise Foward” com LRT. Se atentarmos ao explicado na Secção 5.1, aquilo que denominamos por “stepwise foward” com LRT, corresponde à selecção de atributos com as seguintes características: Em primeiro lugar, a estratégia de geração de atributos candidatos e a procura do melhor é efectuada utilizando uma heurística tal como se descreve na Secção 6.1.2; Em segundo lugar, o subconjunto de atributos a eleger é iniciado num conjunto vazio e que iterativamente acrescenta um candidato (equivalente ao termo “foward”); Em terceiro lugar, a função avaliação é do tipo “erro do classificador” ou “wrapper”, pois usando o LRT, tal significa que numa determinada iteração a variável eleita aumenta o LRT, isto é, propicia o aumento da probabilidade de uma determinada instância pertencer à classe “Bom”, quando é de facto “Bom”; E em quarto lugar, o critério de paragem é disparado quando o LRT na iteração n não aumenta face à iteração $n-1$. O “stepwise forward” com LRT foi efectuado em SPSS. O “stepwise foward” incidiu sobre as variáveis já discretizadas e agrupadas.

Como resultado desta estratégia foram seleccionadas as seguintes variáveis: *Antempl*, *bem*, *Distrito*, *Estcivill*, *Origtelpes*, *Profl*, *Tipohabl*, *Ndepend*, *Vei*, *Antbanc* e *Prazo*.

Uma vez que a variável *Antempl* foi discretizada e consequentemente binarizada em três variáveis, e que as variáveis nominais *Profl* e *Origtelpes*, foram binarizadas respectivamente em 4 e 2 variáveis, finalmente a variável *Prazo* foi discretizada e binarizada em 3, tendo sendo eliminada uma das variáveis binárias. Totalizando, o modelo conta com 18 variáveis independentes, não contando com o “intercept”. As variáveis não seleccionadas, nesta fase, foram: *Pais*, *Estatuto1*, *Idade1*, *Vpvp*, *Prazo3* e *Anthabl*.

- **Seleccção Multi-variada – Discriminante Linear**

O modo de selecção multi-variada utilizando o discriminante linear foi em tudo idêntico ao efectuado na regressão logística, exceptuando a métrica de avaliação que em vez do LRT foi o Lambda de Wilks. O “stepwise foward” com Lambda de Wilks foi efectuado em SPSS.

Os resultados deste método foram idênticos à da regressão logística acima enunciados.

6.4 Resultados e avaliação de performance

Pretende-se nesta Secção testar três estratégias de pré-processamento, usando como modelos a Regressão Logística e o Discriminante Linear. Salientamos que uma das estratégias a testar, por comparação com as restantes, fundamenta-se nas propostas explicadas nas Secções 6.1 e 6.2. O método de estimativa usado é o mesmo que foi utilizado na Secção 4.3.3., traduzindo-se num “*holdout group test*” composto por 714 instâncias das 2.380 que perfazem o conjunto dos dados. A regra de classificação utilizada para afectar uma determinada instância em “Bom” ou “Mau” é a mesma que foi utilizada na Secção 4.3.3.

Simplificando, consideremos as seguintes três abordagens:

- Classificadores obtidos usando a técnica básica de pré-processamento – binarização de variáveis nominais, de acordo com o explicado na Secção 5.4., serão identificados respectivamente por *Discrima0* e *Logis0*, quando se usa o Discriminante Linear e a Regressão Logística.
- Classificadores obtidos usando a técnica anterior acrescida da técnica de pré-processamento – selecção de atributos, de acordo com o explicado nas Secções 5.1, 5.2 e 5.3 serão identificados respectivamente por *Discrima1* e *Logis1*, quando se usa o Discriminante Linear e a Regressão Logística.
- Classificadores obtidos usando todas as técnicas de pré-processamento enunciadas anteriormente, acrescida das técnicas por nós propostas nas Secções 6.1 e 6.2, serão identificados respectivamente por *Discrima2* e *Logis2*, quando se usa o Discriminante Linear e a Regressão Logística.

No Quadro 22 apresentamos as taxas de erro apuradas pelos classificadores *Discrima0* e *Logis0* (já apuradas na Secção 4.3.3), *Discrima2* e *Logis2*. Apresentamos também as respectivas estimativas intervalares obtidas, a partir de um nível de confiança de 95% e com base com o que foi explicado na Secção 3.5.3. No âmbito do presente trabalho, cumpre-nos comparar as primeiras (taxas de erro de *Discrima0* e *Logis0*) com as segundas (taxas de erro de *Discrima2* e *Logis2*), pois desta forma pudemos aferir/medir o impacto das técnicas de pré-processamento utilizadas no presente Capítulo, no erro de generalização. Poderemos, também concluir se os objectivos a cumprir pelas técnicas de pré-processamento, enunciadas no início do Capítulo 5²³, foram ou não alcançados.

Quadro 22-Classificadores com e sem pré-processamento utilizado no presente Capítulo

Classificadores	Taxa de erro	Intervalo de Confiança
Logis0	0,269	0,269 ± 0,0325
Logis2	0,239	0,239 ± 0,0313
Discrima0	0,276	0,276 ± 0,0328
Discima2	0,245	0,245 ± 0,0315

Pela leitura do quadro 22 se conclui o seguinte: Em primeiro lugar, os classificadores obtidos com as técnicas de pré-processamento (*Discrima2* e *Logis2*) utilizadas no presente Capítulo, apresentam taxas de erro significativamente inferiores que os

²³ No primeiro parágrafo, pode-se ler “ Estas tarefas (de pré-processamento), visam tratar os dados com o objectivo de diminuir a complexidade / dimensão dos dados sem prejudicar a capacidade preditiva do modelo...”.

restantes. Consideramos diferenças significativas, não só pela diferença da taxa de erro, como pela quase não coincidência das amplitudes dos respectivos intervalos de confiança. Em segundo lugar, há semelhança do que já havia sido constatado parcialmente na Secção 4.3.3, os modelos Regressão Logística e Discriminante não se diferenciam no erro, independentemente de a utilização diferenciada das técnicas de pré-processamento. Em terceiro lugar podemos afirmar que os objectivos das técnicas de pré-processamento, utilizadas no presente capítulo foram cumpridos, pois o número de variáveis foi radicalmente reduzido (diminuindo a complexidade do modelo), reduzindo também o perigo de sobre-ajustamento (“overfitting”); A diminuição radical da complexidade do modelo, não provocou aumento na taxa de erro, pelo contrário diminuiu significativamente; A interpretabilidade do modelo aumentou, através da simplificação. Em quarto lugar e último o Classificador Logis2 é o que regista a menor taxa de erro, não se diferenciando significativamente do Discrima2.

Importa agora medir o efeito produzido pelas melhorias introduzidas no agrupamento de factores das variáveis nominais e na discretização de variáveis numéricas. Para tal propomos que seja comparada com o Discrima2 e o Logis2, os classificadores Discrima1 e Logis1 obtidos a partir das técnicas básicas de pré-processamento descritas na Secção 5.4, mais a técnica de selecção de atributos – “stepwise forward”. Pensamos assim isolar o efeito da eficácia das melhorias apresentadas.

Pela observação do quadro 23 concluímos o seguinte: Em primeiro lugar, as taxas de erro são significativamente inferiores em Discrima2 e Logis2, comparativamente com os restantes. Em segundo lugar os classificadores Discrima1 e Logis1 (ver Quadro 22), não apresentam melhorias significativas na taxa erro, comparativamente com Discrima0 e Logis0. Em terceiro lugar a complexidade do Discrima2 e do Logis2 é substancialmente inferior à do Discrima1 e do Logis1, sendo que os primeiros participam uma totalidade de 18 variáveis contra 26 no Discrima1 e Logis1. Em quarto lugar, a interpretabilidade de Discrima2 e Logis2 é superior que os restantes, não somente pelo número menor de variáveis independentes, mas também pelo facto de que no Discrima1 e no Logis1 as variáveis nominais com K factores se transformam em k-1 variáveis, em que devido à selecção de atributos, a maior parte não participa no modelo. Tal não acontece no Discrima2 e no Logis2 (exceptuando a variável *Prazo*), pois a agregação, implicitamente, não exclui nenhuma k-1 variáveis, sendo que todos os factores originais intervêm no modelo, sendo-lhe atribuído o respectivo peso, donde portanto poderemos tirar ilações.

Em suma o Discrima2 e o Logis2 obtêm um acerto significativamente superior que o Discrima1 e o Logis1, com a vantagem de ser substancialmente menos complexo, contendo menos 35% das variáveis, e deter maior interpretabilidade que o Discrima1 e o Logis1, pois a todos os factores originais, está associado um coeficiente que revela a importância do mesmo no modelo. Assim sendo consideramos que as propostas por nós apresentadas no tocante à agregação de factores e discretização de variáveis numéricas, neste caso prático, produzem resultados acrescidos, face à utilização dos métodos convencionais.

Quadro 23-Classificadores com pré-processamento utilizado no presente Capítulo e Classificadores apenas com binarização e selecção de atributos.

Classificadores	Taxa de erro	Intervalo de Confiança
Logis1	0,269	$0,269 \pm 0,0325$
Logis2	0,239	$0,239 \pm 0,0313$
Discrima1	0,275	$0,275 \pm 0,0327$
Discima2	0,245	$0,245 \pm 0,0315$

7. Avaliação de Modelos com Curvas “ROC”.

Pretendemos que este capítulo informe o leitor sobre:

- Deficiências na métrica – Taxa de acerto, como medida de avaliação de modelos.
- Outras métricas, para além da taxa de acerto, para aferir a performance de um modelo.
- Avaliação dos modelos gerados no presente caso prático, segundo a nova técnica de avaliação “Roc Analysis”.

7.1 Deficiências na medida – Taxa de erro

Como sabemos a taxa de erro de um classificador (definida e explicada na Secção 3.5), medida num conjunto independente de teste, é a métrica mais utilizada, em Aprendizagem automática e “Data Mining” para avaliar ou comparar o acerto nas predições entre classificadores. Pese embora a sua prática generalizada, alguns autores apresentam deficiências que podem por em causa a eficácia daquela medida. Peter Flach [11], colocou a seguinte questão em relação a dois classificadores e duas classes representadas aqui por (+) e (-): *Classif1* - classifica correctamente 40 em 50 (+) e 40 em 50 (-) e *Classif2* - classifica correctamente 30 em 50 (+) e 50 em 50 (-). Será que ambos são equivalentes na capacidade de predição? A resposta obtém-se pelo número de instâncias que compõem as classes (+) e (-) no grupo de teste. Se nele constar mais instâncias pertencentes a (+), então, tendencialmente *Classif1* é inferior em taxa de erro e portanto melhor em performance, caso contrário, tendencialmente, será o *Classif2*. Assim sendo conclui-se que o peso de elementos pertencentes a uma classe poderá influir a taxa de erro, desvirtuando a sua propriedade de medir o acerto, salvo quando o peso das classes no grupo de teste é equivalente. Michie [5] comenta o absurdo dos “default classifiers”, em que se considerarmos que 5% das instâncias pertencem a uma classe num problema de 2 classes, então se classificarmos todas as instâncias como pertencentes à classe maioritária o erro será de aproximadamente, no conjunto de teste 5%, dificilmente igualável por outro classificador. Neste caso o classificador só será útil se obtiver uma taxa de erro inferior aos 5%.

Frank [4], aponta outras deficiências que se prendem com o tipo de linguagem de representação do classificador. Se sistematizarmos, existem 2 tipos de abordagem na construção de classificadores: A primeira, estrutura-se na construção de modelos probabilísticos em que a variável dependente é traduzida pela probabilidade condicional $P(C_i | i\text{-ésima instância})$, pressupondo que a população segue uma determinada lei de probabilidade; A segunda, não pressupondo qualquer lei de probabilidade, o classificador construído apenas prediz a classe, como o caso das árvores de decisão como o C4.5 [10]. Ora se um classificador nos indica $P(\text{Bom} | i\text{-ésima instância}) = 95\%$ de a instância pertencer à classe “Bom”, com certeza que será mais seguro afectar aquela instância como pertencente à classe “Bom”, comparativamente a outra instância em que o modelo estima 51%. Assim sendo a gradualidade nas estimativas, nos modelos probabilísticos, permitem que se possa medir, isso sim a sua discriminação, enquanto que a taxa de erro nos mede a performance da classificação. Acontece que um modelo

pode ser muito discriminante, mas a regra de classificação que lhe está associada pode ser pouco certa, sendo que, é certo, se o for pouco discriminante dificilmente será assertivo.

7.2 “ROC analysis”

Por tudo o que foi dito na Secção anterior, e porque no presente trabalho dispomos de dois modelos probabilísticos, importa elegermos a discriminação do modelo como o factor de performance a medir. Poderá não ser claro para o leitor a distinção entre discriminação e acerto, mas para melhor percepção relembro o exemplo do “default classifier”, onde o critério de classificação é de afectar às instâncias a classificar a classe maioritária. Neste caso o poder discriminatório do “modelo” é 0 (zero), no entanto o seu acerto é de 95%.

Assim sendo apresentamos o “ROC analysis”²⁴ como a abordagem mais indicada à aferição do poder discriminatório de um modelo e portanto da sua performance.

7.2.1 “Roc curves ou Lorentz diagram”

Consideremos um dado modelo probabilístico e um conjunto de teste constituído por n instâncias. Sobre as n instâncias, calculamos com base no modelo, a probabilidade condicionada $P(\text{Bom} | x_n)$, para cada instância x das n possíveis. Optou-se por tomar todas instâncias (do grupo de teste) com $P(\text{Bom} | x_n) > 75\%$ e considerá-las ou classificá-las como pertencentes à classe “Bom”. Uma vez que dispomos da verdadeira classe das instâncias, apresenta-se a “matriz de confusão” no Quadro 24, para as instâncias teste com $P(\text{Bom} | x_n) > 75\%$.

Quadro 24 – Exemplo de uma matriz de confusão e respectivos TP e FP rates

		Observado		Total
		Bons	Maus	
Predicto	Bons	11	4	15
	Maus	17	118	135
Total(a)		28	122	150

TP rate (> 75%)	39,3%
FP rate (> 75%)	3,3%

(a) Total de “Bons” e “Maus” no grupo de teste

Da matriz de confusão conseguimos obter o “TP rate (> 75%)” – “true positivite rate”, ou seja do total das instâncias, do grupo de teste, pertencentes à classe “Bom”, que totalizam 28, qual a percentagem que foi predita correctamente como “Bom” com $P(\text{Bom} | x_n) > 75\%$. Contrariamente temos o “FP rate (> 75%)” - “false positive rate”, ou seja

²⁴ A avaliação de modelos no âmbito do “Credit scoring”, a partir de curvas ROC é prática corrente.

do total das instâncias, do grupo de teste, pertencentes à classe “Mau”, que totalizam 122, qual a percentagem que foi predita erradamente como “Bom” com $P(\text{Bom} | x_n) > 75\%$. Passemos de seguida a representar num par de eixos cartesianos, constantes na Figura 15, em que “TP rate” figura no eixo das ordenadas, e o “FP rate” figura nos eixos das abcissas, o ponto A (39.3%,3.3%).

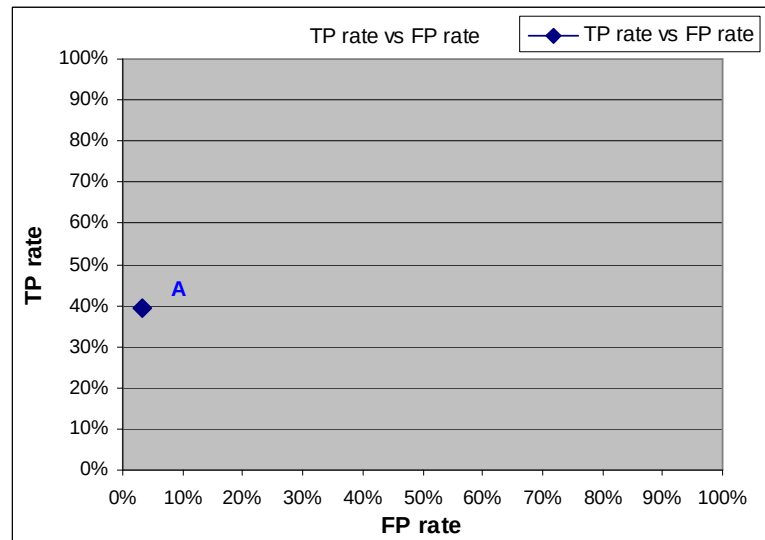


Figura 15 – representação do ponto A (39.3%, 3.3%)

O ponto A, interpreta-se como sendo o resultado de se optar por classificar todas as instâncias cuja $P(\text{Bom} | x_n) > 75\%$, como “Bom”. Tal resultado significa que classifico 39,3% dos “Bons” como “Bons” e erro em 3,93% dos “Maus” que classifico como “Bom”.

Imaginemos agora que pretendemos ensaiar e classificar como “Bom”, todas as instâncias para vários valores de $P(\text{Bom} | x_n) > 62,5\%, > 50\%, > 37.5\%$, tal daria origem a um conjunto de pontos semelhantes a “A”, que uma vez unidos dariam origem a uma curva denominada “ROC curve”, tal como se representa na figura 16, resultante dos elementos contidos no Quadro 25.

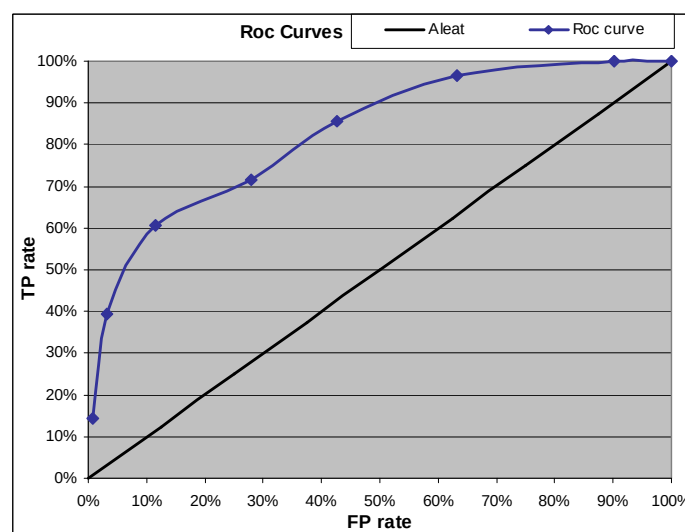


Figura 16 – Sucessões de pontos idênticos ao ponto A perfazem uma curva ROC

Quadro 25 – detalhe dos dados referentes à curva ROC da Figura 16

P(Bom Xn)*	Grand Total	Bons	Maus	Acum Bons	TP rate	Acum Maus	FP rate	Aleat
0%-12,5%	12	0	12	28	100,0%	122	100,0%	100,0%
12,51%-25%	34	1	33	28	100,0%	110	90,2%	87,5%
25,01%-37,5%	28	3	25	27	96,4%	77	63,1%	75,0%
37,51%-50%	22	4	18	24	85,7%	52	42,6%	62,5%
50%-62,5%	23	3	20	20	71,4%	34	27,9%	50,0%
62,5%-75%	16	6	10	17	60,7%	14	11,5%	37,5%
75%-87,5%	10	7	3	11	39,3%	4	3,3%	25,0%
>87,5%	5	4	1	4	14,3%	1	0,8%	12,5%
Grand Total	150	28	122					0,0%

* Optou-se por discretizar por intervalos ou "tranches" para facilitar a suavização do "Roc curve"

Repara-se que na figura 16 estão desenhadas duas 2 linhas: a curva ROC propriamente dita constituída por uma linha que suaviza 8 pontos, tantos quantos os intervalos previstos no Quadro 25 na coluna P (Bom|X_n); uma diagonal que pretende significar um modelo que classifica aleatoriamente, traduzindo-se em TP rate iguais a FP rate à medida que vamos cumulativamente percorrendo os intervalos da coluna "P(Bom|X_n)". À medida que vamos percorrendo os pontos do curva ROC, necessariamente as instâncias que pertencem a intervalos de P (Bom|X_n) elevados, deverão possuir a maior quantidade de "bons" e a menor quantidade "Maus", contrariamente acontecendo, para intervalos de P (Bom|X_n) reduzidos. O modelo será tanto mais perfeito no seu poder discriminante quanto mais se afastar da diagonal, significando que à medida que vamos progredindo nos intervalos, começando no mais elevado, todas as instâncias que classificamos na classe "Bom", estando os bons concentrados nos intervalos maiores, necessariamente a TP rate será consideravelmente superior ao FP rate revelando assim o seu poder discriminante.

Apresentamos de seguida a figura 17. Esta é constituída por duas curvas ROC, sendo a Roc curve1 a mesma retratada na figura anterior.

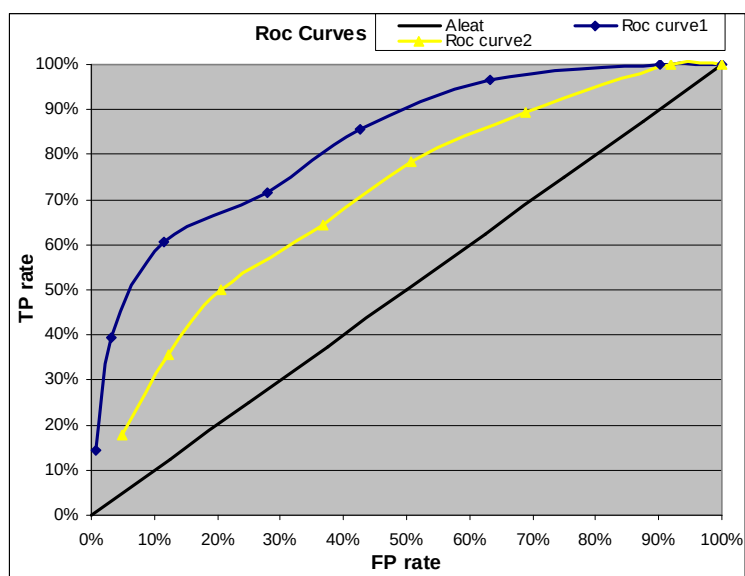


Figura 17 – Representação de duas curvas ROC em que uma domina a outra

As curvas ROC ajudam a descrever as diferenças entre 2 modelos, pois se a Roc curve1, for constituída por pontos, em que todos eles se distanciam ou “cobrem” a Roc curve2 então podemos afirmar que o modelo traduzido pela Roc curve1 é melhor na sua capacidade discriminante, em todos os intervalos de probabilidade, que o modelo traduzido pela Roc curve2. Ora na comparação de performances entre 2 modelos esta situação nem sempre ocorre, podendo haver intervalos de probabilidade em que um modelo cobre o outro e vice versa, tal significa que ambas curvas ROC se interceptam e não existe uma dominância de um sobre outro, bem pelo contrário podendo-se complementar. Enquanto que uma análise de performance orientada para o erro da classificação, forçosamente elege o melhor, a “Roc analysis”, permite visualizar, o modelo dominante em todos os intervalos de probabilidade, ou, caso contrário quais os melhores modelos para cada intervalo de probabilidade.

Conforme se pode observar no Quadro 26, é evidente a menor concentração de bons nos intervalos de probabilidade mais elevados e, necessariamente maior concentração de “maus” na Roc curve2.

Quadro 26 - detalhe dos dados referentes à Roc curve2 da Figura 17

Roc curve2								
P(Bom/Xn)*	Grand Total	Bons	Maus	Acum Bons	TP rate	Acum Maus	FP rate	Aleat
0%-12,5%	10	0	10	28	100,0%	122	100,0%	100,0%
12,51%-25%	31	3	28	28	100,0%	112	91,8%	87,5%
25,01%-37,5%	25	3	22	25	89,3%	84	68,9%	75,0%
37,51%-50%	21	4	17	22	78,6%	62	50,8%	62,5%
50%-62,5%	24	4	20	18	64,3%	45	36,9%	50,0%
62,5%-75%	14	4	10	14	50,0%	25	20,5%	37,5%
75%-87,5%	14	5	9	10	35,7%	15	12,3%	25,0%
>87,5%	11	5	6	5	17,9%	6	4,9%	12,5%
Grand Total	150	28	122					0,0%

7.3 Medidas de avaliação do poder discriminatório

Ao longo do presente trabalho temos evocado métricas que avaliam o poder discriminatório de um atributo, de um grupo de factores e de uma variável contínua discretizada. As mesmas podem ser aplicadas no âmbito desta Secção, no entanto a literatura [1], [2] e [7] menciona frequentemente as medidas que de seguida apresentamos.

7.3.1 Coeficiente de Gini (“Gini Coefficient”) ou “Efficiency Index”

Tal como explicado na Secção anterior, quanto mais afastado estiver a curva ROC da diagonal, onde o classificador será indiscriminante, então tanto menor será a área apurada, acima da curva ROC (área B na figura 18) e portanto mais discriminante será o modelo. O cálculo do *coeficiente de Gini* resulta da utilização das curvas ROC. Tomemos o exemplo ilustrado no gráfico da Figura 18 e passemos a descrever cálculo do *coeficiente de Gini* (G) [7]:

- a) Define-se como sendo o rácio da área entre a Diagonal e a curva (área A) ÷ área total acima da diagonal (soma da área A com a área B). Tal como foi dito, quanto mais a curva se distanciar da diagonal, maior será o rácio e maior será a medida de separação entre “Bons” e “Maus”.

$$G = \text{área A} / (\text{área A} + \text{área B}) \quad (7.1)$$

- b) Somando as áreas A e B obtemos toda área acima da diagonal sendo igual a metade da área de um quadrado, ou seja igual a $\frac{1}{2}$. Assim, $G = 2 \times \text{área A}$, ou seja 2 vezes a área entre a diagonal e a curva.

- c) A expressão descrita anteriormente, pode ser pensada da seguinte forma:

Sendo A = área acima da diagonal – área acima da curva, então:

$$G = 2 \times (\text{área acima da diagonal} - \text{área acima da curva}) \quad (7.2)$$

- d) A área acima da curva, área B, pode ser calculada pelo o somatório das áreas de todos os sucessivos trapézios que ocupam a área B. A área B contém tantos trapézios quantos os intervalos de probabilidade P (Bom|Xn), em que de acordo com a Figura 18, a área do trapézio²⁵ ilustrado é igual a,

$$(B_{j+1} - B_j) \times (M_{j+1} + M_j) / 2. \quad (7.3)$$

- e) Bastando substituir na expressão (7.2) a área acima da curva pela expressão (7.3) e considerar que a área acima da diagonal é igual a $\frac{1}{2}$, chegamos à expressão final do *coeficiente de Gini* :

$$G = 2 \times (1/2 - \sum_j (B_{j+1} - B_j) \times (M_{j+1} + M_j) / 2) = 1 - \sum_j (B_{j+1} - B_j) \times (M_{j+1} + M_j) \quad (7.4)$$

²⁵ A área de um trapézio é formulada pela expressão: $h \times (\text{base}_1 + \text{base}_2) \times \frac{1}{2}$. Aplicando ao trapézio ilustrado na Figura 18, h (altura) é calculado pela seguinte diferença $B_{j+1} - B_j$, base_1 igual a M_j e base_2 igual a M_{j+1} .

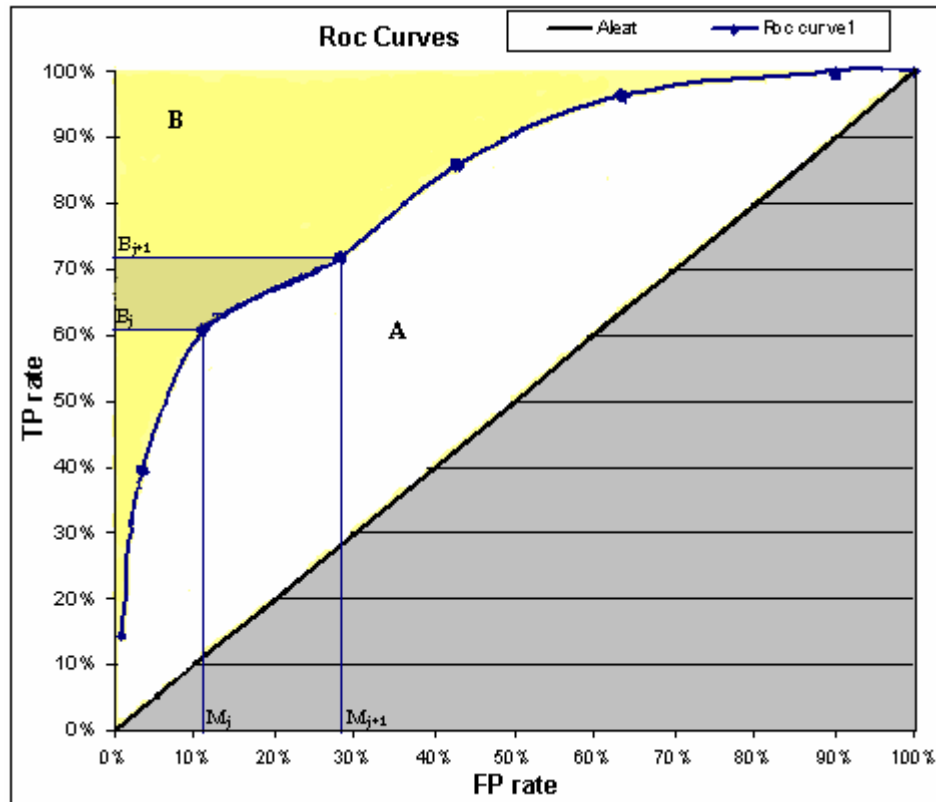


Figura 18 – As áreas que intervêm no cálculo do *Coeficiente de Gini*

7.3.2 “K-S distance”

Consiste numa métrica simples, bastando para efeito representar a distribuição acumulada, atendendo ao score, dos “Bons” e dos “Maus”, e aplicar a Kolmogrov-Smirnov distance entre as distribuições, ou seja:

$$KS = \text{Max} | \sum_i P(\text{score}_i | \text{Bom}) - \sum_i P(\text{score}_i | \text{Mau}) |$$

Abaixo encontramos um exemplo de cálculo (Quadro 27) e respectiva representação gráfica (figura 19)

Quadro 27 – A distância K-S e respectivo cálculo

Tranches	%Acum Bons	%Acum Maus	Max P(score/Bom)-P(score/Mau)
(0.0282,0.152]	2,17%	21,57%	0,19
(0.152,0.227]	4,35%	35,95%	0,32
(0.227,0.304]	7,97%	47,71%	0,40
(0.304,0.391]	13,04%	64,05%	0,51
(0.391,0.479]	22,46%	76,47%	0,54
(0.479,0.569]	34,06%	84,31%	0,50
(0.569,0.665]	44,93%	89,54%	0,45
(0.665,0.764]	60,14%	96,73%	0,37
(0.764,0.88]	76,09%	99,35%	0,23
(0.88,0.999]	100,00%	100,00%	0,00
Grand Total			

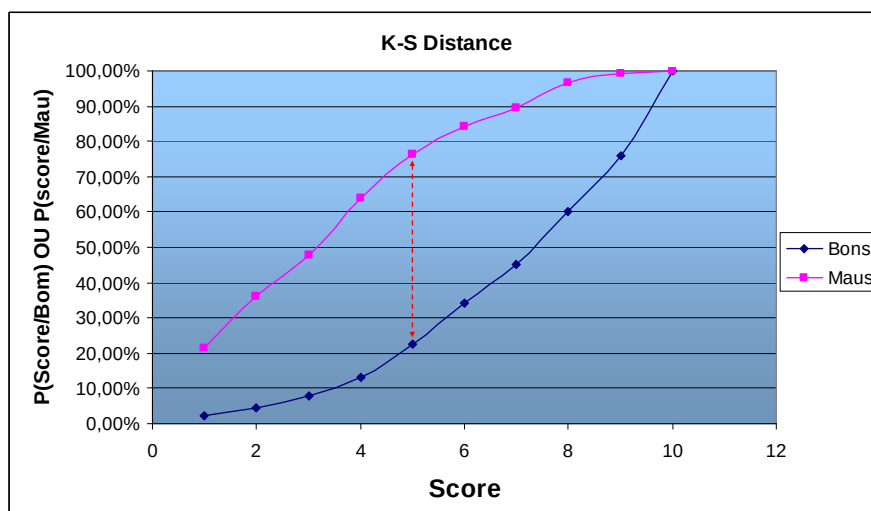


Figura 19 – A distância K-S representada graficamente a tracejado

7.4 “Roc analysis” do caso prático

Pretendemos nesta Secção, medir a performance dos classificadores gerados no âmbito do caso prático, utilizando a “Roc analysis”, contrapondo com a metodologia seguida na Secção 6.4. Consideremos, também, os classificadores Discrima0, Logis0, Discrima1, Logis1, Discrima2 e Logis2, descritos anteriormente (Secção 6.4)

Na figura 20 apresentamos quatro curvas ROC, sendo Logis2 e Discrima2 as que mais se distanciam da diagonal, concluindo-se que aqueles detêm de maior capacidade discriminatória comparativamente com as restantes. De referir que tanto os modelos de Regressão Logística e Discriminante Linear, independentemente do tipo de pré-processamento realizado, não apresentam diferenças nas curvas, pois praticamente se ajustam, revelando que estes dois modelos não se distinguem, no presente caso prático, no poder discriminatório. A diferença, essa sim é entre os Modelos com diferentes tipos de Pré-processamento. Se atentarmos à Figura 20 constatamos que os classificadores com o pré-processamento utilizado no Capítulo 6 (Discrima2 e Logis2), são dominantes face aos restantes, em quase todos os intervalos de probabilidade. As curvas ROC ilustram-nos, como já foi dito, qual o modelo mais discriminante, ou se não houver dominância de nenhum, quais os intervalos de probabilidade em que cada modelo se revela mais discriminante através da relação TP rates vs FP rates. No entanto precisamos de medir a discriminação pelo que apresentamos de seguida, no Quadro 28, o Coeficiente Gini dos quatro modelos representada numa escala de 0 a 1.

Quadro 28-Coeficientes de Gini dos Classificadores

Classificadores	Coeficiente Gini
Logis0	0,582
Logis2	0,641
Discrima0	0,593
Discima2	0,637

Conforme foi explicado na Secção 7.3.1 o Coeficiente de Gini, mede-nos a proporção que ocupa a área entre a curve Roc e diagonal da área total acima da diagonal (ou seja igual a $\frac{1}{2}$). Quanto mais a curva ROC se afastar da diagonal, maior é a área compreendida entre aquelas duas linhas e portanto maior o coeficiente de Gini e a capacidade discriminante do modelo. Pela leitura do Quadro 28, se conclui que os Gini de Discrima2 e Logis2 são superiores que Discrima0 e Logis0. Assim sendo, é importante lembrarmos que estes mesmos modelos avaliados à luz da taxa de erro, na Secção 6.4, produziram resultados idênticos, no que diz respeito à supremacia do Discrima2 e Logis2 sobre o Discrima0 e Logis0. A taxa de erro mede o acerto do modelo ou da sua regra de classificação, o coeficiente de Gini mede, independentemente da regra de classificação, a discriminação do modelo.

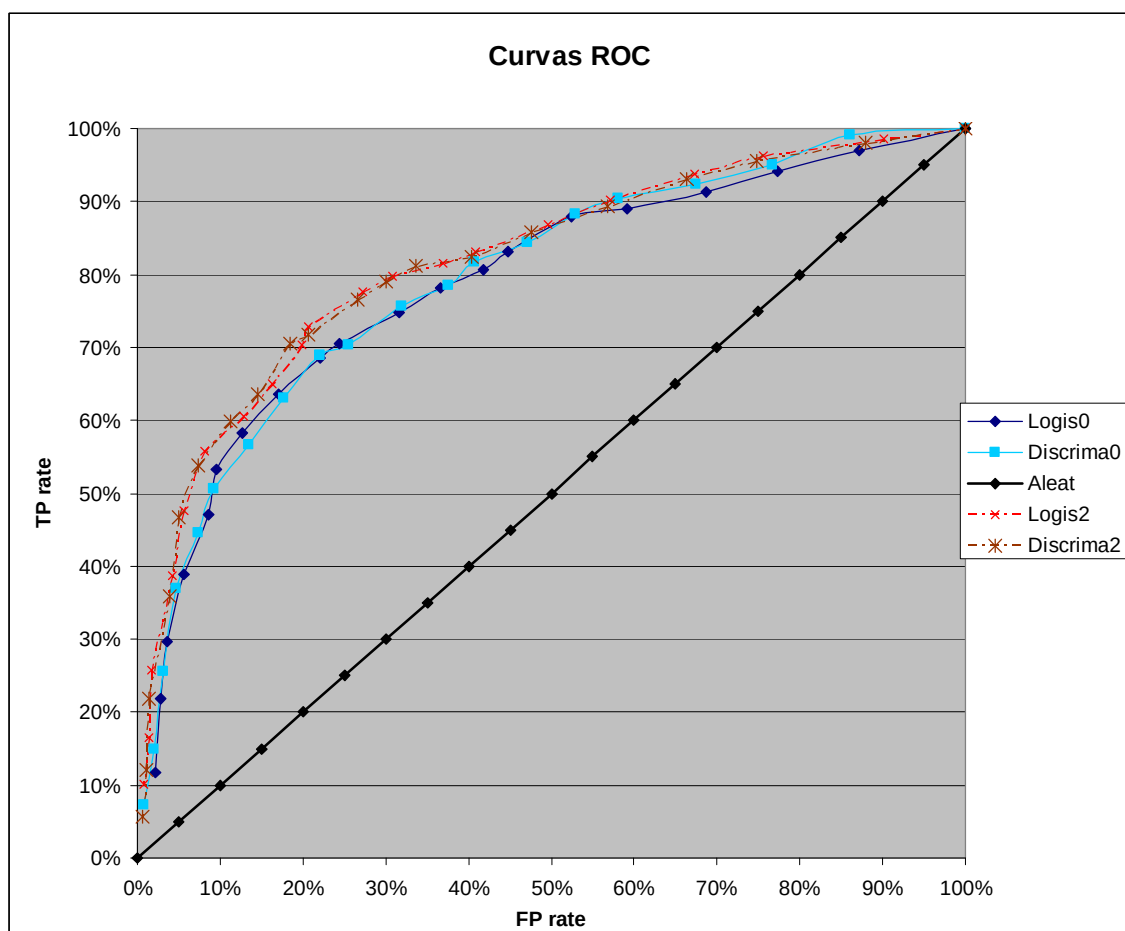


Figura 20 – “Roc Analysis” entre Discrima0, Logis0 e Discrima2, Logis2

Resta-nos à semelhança do que foi efectuado na Secção 6.4 completar a “Roc analysis” sobre o Discrima1, Logis1 e Discrima2, Logis2. Ao compararmos Discrima2 e Logis2 com Discrima1 e Logis1, sendo estes últimos construídos a partir das técnicas de pré-processamento, binarização de atributos nominais e selecção de atributos (“stepwise forward”), pretendemos isolar o efeito da eficácia das melhorias apresentadas nas Secções 6.1 e 6.2.

A Figura 21 ilustra-nos que os classificadores Discrima2 e Logis2 detêm de maior poder discriminante, pois são as que mais se distanciam da diagonal. Os *coeficientes de Gini* dos respectivos classificadores, apresentados no Quadro 29, confirmam a supremacia daqueles dois classificadores. Apesar de Discrima2 e Logis2 se distanciarem comparativamente mais da diagonal face às restantes curvas ROC, no entanto nos sítios assinalados a tracejado (na Figura 21) tal facto não sucede. Tal significa que existem intervalos de probabilidade, $P(Bom|X_n)$, cujo poder discriminante de Discrima1 e Logis1 é superior. Quando tal sucede diz-se que Discrima2 e Logis2, genericamente são superiores no poder discriminante, mas não são dominantes. Assim sendo se combinarmos os classificadores Logis2 e Logis1, aplicando-os nos intervalos de probabilidade onde o respectivo poder discriminante é superior, obteríamos um classificador com um poder discriminante superior ao Logis2.

Assim sendo, é importante lembrarmos que estes mesmos modelos avaliados à luz da taxa de erro, na Secção 6.4, produziram resultados idênticos, no que diz respeito à supremacia do Discrima2 e Logis2 sobre o Discrima1 e Logis1. No entanto se avaliarmos a performance de Discrima1 e Logis1 comparativamente ao Discrima0 e Logis0, pelo erro (taxa de erro) e pela discriminação (Coeficiente Gini), diverge-se nas conclusões. Constatou-se na Secção 6.4 que as taxas de erro dos Quadros 22 e 23 respeitantes aos classificadores Discrima1 e Logis1 não apresentam melhorias face a Discrima0 e Logis0. Constata-se, na presente Secção, que o poder discriminante medido pelo coeficiente de Gini indicado nos Quadros 28 e 29, de Discrima1 e de Logis1 apresenta melhorias face a Discrima0 e Logis0. Assim sendo conclui-se que a selecção de atributos (técnica de pré-processamento que diferencia Discrima1 e Logis1 de Discrima0 e Logis0), imprime maior poder discriminante aos classificadores, mas dada a regra de classificação em vigor não incrementa o acerto.

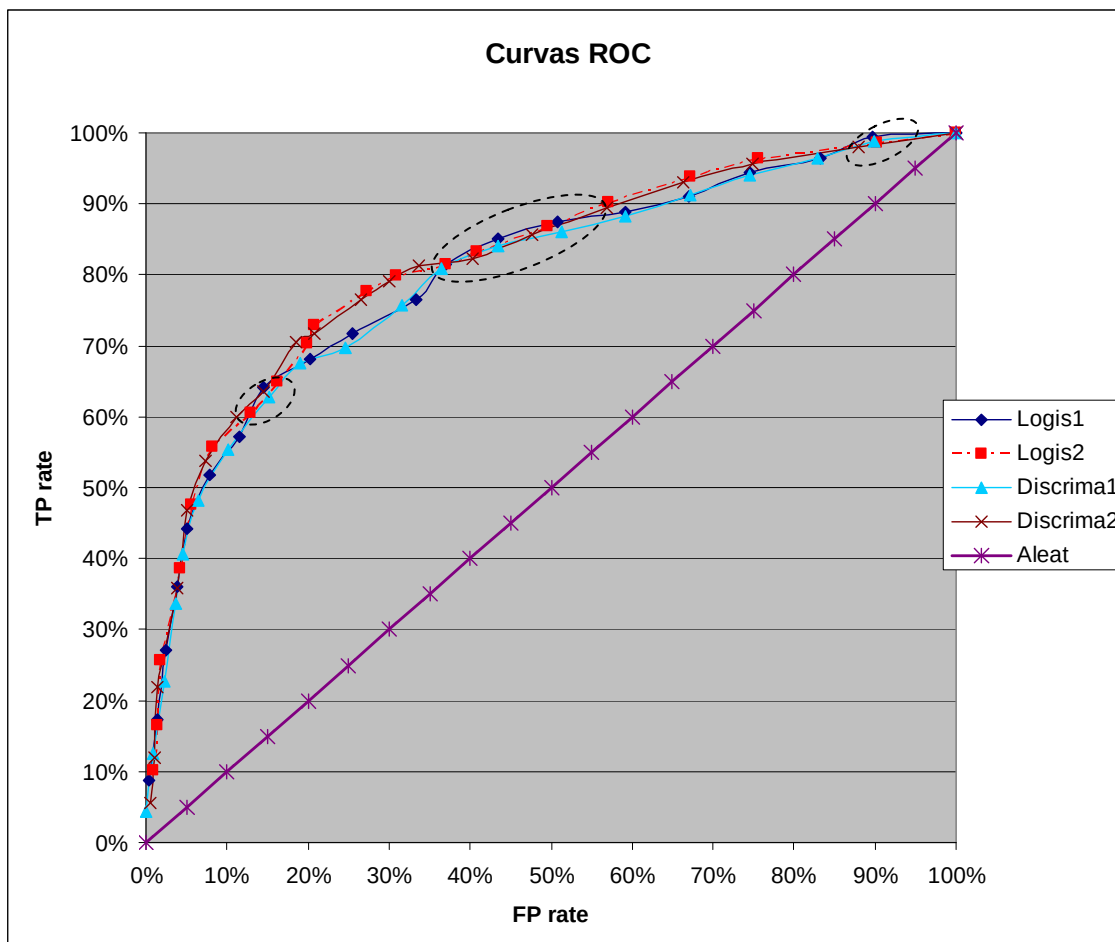


Figura 21 – “Roc Analysis” entre Discrima1,Logis1 e Discrima2, Logis2

Quadro 29-Coefficientes de Gini dos Classificadores

Classificadores	Coefficiente Gini
Logis1	0,613
Logis2	0,641
Discrima1	0,603
Discima2	0,637

8. Conclusões

O objectivo do presente trabalho consiste em avaliar se as propostas referentes ao agrupamento de factores e à discretização de variáveis numéricas beneficiam a performance dos modelos de Regressão Logística e do Discriminante Linear, comparativamente com outras duas estratégias de pré-processamento em uso.

A prossecução deste objectivo foi efectuada no âmbito de um problema de concessão de crédito, tradicionalmente objecto de tratamento dos técnicos de “Credit scoring”. A tomada de conhecimento de algumas técnicas ou práticas da área do “Credit scoring”, em grande medida enriqueceu este trabalho, complementando os conhecimentos obtidos na parte curricular.

Os dados foram extraídos a 31 de Outubro de 2003, tendo sido aleatoriamente seleccionados os contratos de crédito a termo, cujo início contratual se verificou entre 1 de Abril de 2002 e 31 de Dezembro de 2002. O conjunto de dados consta de 2.380 registos, cujos grupos (“Bom” e “Mau” pagador) da variável-classe se encontram igualmente repartidos, indo de encontro à exigência da instituição financeira.

Os 2.380 registos foram seleccionados após se ter verificado que o tempo médio de exposição dos contratos (tempo que medeia o momento do início do contrato e a data de observação a 31 Outubro de 2003) se aproximava da maturidade da população (momento a partir do qual a Taxa de sinistralidade de um conjunto de contraentes, não evolui mais) de onde estes registos foram extraídos. Assim poderemos definir o “Bom” ou “Mau” pagador com maior segurança zelando pela qualidade da definição da variável-classe.

Foi constituído um “holdout group test” para medir a performance dos classificadores e portanto avaliar as três estratégias de pré-processamento (descritas no Quadro 30) e concluir sobre a eficácia das propostas por nós apresentadas sobre o agrupamento de factores e a discretização de atributos numéricos. Composto por 714 registos extraídos, aleatoriamente, de forma a deterem igual proporção de “Bons” e “Maus”, o “holdout group test” equivale a 30% do conjunto de dados, restando para treino 1.666 registos.

Optamos por avaliar os classificadores segundo duas métricas. A primeira, frequente em “Machine Learning”, a taxa de erro de generalização. A segunda, frequente em “Credit scoring” mede a discriminação do classificador. A grande diferença entre uma e outra é que a primeira mede o acerto de uma regra de classificação concretizada num cut-off, ao passo que a segunda simula todos os cut-off’s possíveis, e para cada um deles quantifica o TP rate (a taxa de verdadeiros positivos) e o FP rate (a taxa de falsos positivos), pelo que quanto maior for a diferença entre aquelas duas taxas mais discriminante é o modelo. Fazendo representar o FP rate no eixo das abcissas e o TP rate nas ordenadas, dará origem a um conjunto de pontos (tantos quantos os cut-off) que uma vez unidos formam uma curva ROC. A necessidade de avaliar o erro e a discriminação, parte do facto que quando avaliamos o modelo pelo acerto, não estamos a avaliar o modelo mas sim a regra de classificação. Avaliando a discriminação, nos termos de “ROC analysis” estamos a avaliar o modelo independentemente da regra de classificação.

O Quadro 30 identifica para cada uma das três estratégias de pré-processamento e para cada um dos algoritmos de “Data Mining”, os seis classificadores sobre os quais se produziram resultados e importa tirar conclusões.

Quadro 30 – Identificação dos classificadores experimentados

Classificadores	Técnicas de pré-processamento	Algoritmo de “Data Mining”
<i>Logis0</i>	Binarização de variáveis nominais (1ª Estratégia de pré-processamento)	Regressão Logística
<i>Logis1</i>	Binarização de variáveis nominais e Selecção de atributos (2ª Estratégia de pré-processamento)	Regressão Logística
<i>Logis2</i>	Agrupamento de Factores, Discretização de variáveis, Binarização de variáveis nominais e Selecção de atributos (3ª Estratégia de pré-processamento)	Regressão Logística
<i>Discrima0</i>	Binarização de variáveis nominais (1ª Estratégia de pré-processamento)	Discriminante Linear
<i>Discrima1</i>	Binarização de variáveis nominais e Selecção de atributos (2ª Estratégia de pré-processamento)	Discriminante Linear
<i>Discrima2</i>	Agrupamento de Factores, Discretização de variáveis, Binarização de variáveis nominais e Selecção de atributos (3ª Estratégia de pré-processamento)	Discriminante Linear

Começamos por comparar os classificadores Logis0,Discrima0 com Logis2,Discrima2. Quanto ao erro, Logis2 e Discrima2 registam, respectivamente, 0.239 ± 0.0313 e 0.245 ± 0.0315 , contra 0.269 ± 0.0325 de Logis0 e 0.276 ± 0.0328 de Discrima0. Quanto ao poder discriminatório medido pelo o *coeficiente de Gini*, Logis2 e Discrima2 registam respectivamente, 0.641 e 0.637, contra 0.582 de Logis0 e 0.593 de Discrima0. Independentemente do algoritmo de “Data Mining”, a terceira estratégia de pré-processamento, comparativamente com a primeira, regista menos erro e mais discriminação. É importante constatar que a diminuição significativa da complexidade do modelo, impressa pela terceira estratégia, fez reduzir o número de variáveis de 153 (resultante da primeira estratégia) para 18, evitando o sobre-ajustamento (“overfitting”) e promovendo a interpretabilidade do modelo. Conclui-se também que em ambos algoritmos de “Data Mining” comparados entre si, não apresentam diferenças significativas tanto no acerto como na discriminação, podendo isto significar que os pressupostos assumidos no discriminante no tocante à multi-normalidade dos dados e da equicovariância dos grupos, não foram significativamente quebrados.

Comparando a segunda estratégia com a terceira, conseguimos isolar o efeito do agrupamento de factores e da discretização (nos termos por nós proposto) permitindo assim avaliar o impacto que aquele efeito terá nas performances, face às técnicas de pré-processamento usuais representadas na segunda estratégia.

Quanto ao erro, Logis1 e Discrimal registam, respectivamente, 0.269 ± 0.0325 e 0.275 ± 0.0327 . Quanto ao poder discriminatório medido pelo o *coeficiente de Gini*, Logis1 e Discrimal registam respectivamente, 0.613 e 0.603. Independentemente do algoritmo de “Data Mining”, a terceira estratégia de pré-processamento, comparativamente com a segunda, regista menos erro e mais discriminação. Igualmente se constata que a terceira estratégia mantém os modelos menos complexo (18 atributos contra 26 da segunda estratégia) e que os algoritmos de “Data Mining”, independentemente da estratégia de pré-processamento, se aproximam tanto no erro como na discriminação.

Portanto é inequívoco que o agrupamento de factores e discretização de variáveis numéricas, efectuado nos termos por nós propostos, constitui no presente caso prático uma melhoria na performance dos classificadores, observado no erro e na discriminação.

Outros factores deverão ser tidos em conta, para além da mera avaliação da performance produzidas pelas três estratégias. Um deles prende-se com a maior interpretabilidade que a terceira estratégia promove aos modelos. A agregação (dos factores em grupos e dos números em intervalos) em grupos ou intervalos homogêneos e generalizáveis tendencialmente reduz as hipóteses daqueles serem eliminados, no processo de selecção multi-variado de atributos, enquanto variáveis binárias. Das variáveis nominais existentes no caso prático, constata-se que o processo de selecção de atributos, após agrupamento de factores, ou elimina a integralidade dos factores de uma variável ou não elimina nenhum. Assim sendo, a terceira estratégia promove que todos (e não apenas uma parte) os factores originais de uma determinada variável nominal se preservem, donde portanto poderemos ter uma leitura mais completa do modelo, da importância atribuída a cada factor e a respectiva contribuição para variável a explicar.

Outro facto importante é que a segunda estratégia produz classificadores mais discriminantes do que assertivos, significando que o poder discriminante produzido pela segunda estratégia se aproxima mais da terceira comparativamente ao erro. Tal significa que a regra de classificação, igual para todos os classificadores, é na segunda estratégia onde se regista menos eficácia tendo em conta o poder discriminante do modelo. Quanto à discriminação não existem dúvidas que a terceira estratégia é aquela que mais promove a discriminação, no entanto o “ROC analysis” descreve que apesar das curvas ROC respeitantes à terceira estratégia se distanciarem mais das outras, não é dominante. Isto significa que as curvas ROC respeitantes à segunda estratégia, em alguns intervalos de probabilidade se sobrepõem às curvas ROC respeitantes à terceira estratégia, sendo nesses mais discriminante.

Idealmente se produzirmos um classificador a partir de um algoritmo de “Data Mining” (Regressão Logística) e resultante da combinação da segunda e da terceira estratégia, afectando a uma instância de teste a probabilidade maior atribuída pelas duas estratégias, necessariamente iríamos obter um classificador ligeiramente mais discriminante que o Logis2.

Com o presente trabalho, para além de testarmos a eficácia das nossas propostas, pretendemos avaliar o impacto que as três estratégias de pré-processamento provocam nas performances dos classificadores. Assim procuramos evidenciar a importância, por vezes menosprezada, das técnicas de pré-processamento no “KDD process”. Nas propostas apresentadas e contrariamente ao que é preconizado no “Credit scoring” nas tarefas de “coarse classifying” [1], procuramos propor um método que à boa maneira de “Machine Learning”, seja objectivo e que promova um pré-processamento automatizado, contrariamente ao apanágio do “coarse classifying” que preconiza o “tanto arte como de ciência”. Não podemos falar, nem praticar extracção automatizada de conhecimento a partir de técnicas que necessitem frequentemente da intervenção humana e sensível. No entanto, tal não significa que o técnico, durante o “KDD process” esteja isento de intervir, pelo contrário terá que intervir decidindo quais os algoritmos que pretende experimentar, qual o método de avaliação dos classificadores, qual o método de estimação do erro, qual a dimensão do conjunto de teste, e por aí fora. Tanto no agrupamento de factores como na discretização das variáveis numéricas, no método por nós proposto, o técnico é chamado a intervir na determinação do P-value que fixará o critério de paragem reflectindo o grau de generalização que aquele pretende. O problema do método de Quinlan [10] para o agrupamento de factores é precisamente não possibilitar a decisão do técnico, fazendo com que o número de grupos seja sistematicamente dois.

Tudo o que atrás referimos foi experimentado utilizando a Regressão Logística e o Discriminante. Apesar de se recear que a quebra dos pressupostos de normalidade e de equicovariância perturbassem a classificação, a verdade é que neste caso prático não aconteceu. Muitos outros algoritmos de “Data Mining” existem e que não foram alvo de experimentação e que seria interessante fazê-lo. Qual será o impacto das três estratégias de pré-processamento em classificadores obtidos a partir de árvores de decisão uni e multi-variadas (C4.5, CART e LTREE), Redes Neurais, “Support vector Machines”, algoritmos de aprendizagem baseada em instâncias (“instance-based learning”) como o KNN e muitos outros? Algo que o presente trabalho não responde e que seria interessante experimentar e obter conclusões.

ANEXOS

Anexo 1

1. Ordenar os n valores do atributo A de forma crescente.
2. Apurar todos os pontos médios entre dois valores sucessivos de A .
3. Assumir que cada ponto médio é um potencial ponto de corte e calcular “information gain” do atributo A para cada candidato a ponto de corte – c .

3.1 Cálculo de Entropia de A para o ponto de corte c quando $A > c$:

$$\text{Info}_c(A > c) = -P(\text{Bons}|A > c) \times \log_2(P(\text{Bons}|A > c)) - P(\text{Maus}|A > c) \times \log_2(P(\text{Maus}|A > c)) \quad (\text{A1})$$

3.2 Cálculo de Entropia de A para o ponto de corte c quando $A < c$:

$$\text{Info}_c(A < c) = -P(\text{Bons}|A < c) \times \log_2(P(\text{Bons}|A < c)) - P(\text{Maus}|A < c) \times \log_2(P(\text{Maus}|A < c)) \quad (\text{A2})$$

3.3 Cálculo de entropia de A para o corte c :

$$\text{Info}_c(A) = P(A > c | A) \times \text{Info}_c(A > c) + P(A < c | A) \times \text{Info}_c(A < c) \quad (\text{A3})$$

3.4 Cálculo do “information gain” de A para o corte c :

$\text{Gain}_c(A) = \text{Info}(\text{Classe}) - \text{Info}_c(A)$, tal que

$$\text{Info}(\text{Classe}) = -P(\text{Bons}|\text{Classe}) \times \log_2(P(\text{Bons}|\text{Classe})) - P(\text{Maus}|A < c) \times \log_2(P(\text{Maus}|\text{Classe})). \quad (\text{A4})$$

4. De todos os pontos de corte seleccionar o **Max Gain_c(A)**.
5. Eliminar as instâncias cujos valores de A são superiores ao ponto de corte seleccionado em 4.
- 6- Critério de paragem MDL: se sim continua em 3,4, e 5, se não pára.

$$\text{Gain} > \frac{\log_2(N-1)}{N} + \frac{\log_2(3^N - 2) - k \cdot E + K_1 E_1 + K_2 E_2}{N} \quad (\text{A5})$$

Tal que:

N – número de instâncias.

K – Número de classes.

E – Entropia do atributo A .

K_1 – Número de classes representadas no subintervalo₁.

K_2 – Número de classes representadas no subintervalo₂.

E_1 – Entropia do subintervalo₁, sendo equivalente à expressão (A1)

E_2 – Entropia do subintervalo₂, sendo equivalente à expressão (A2)

A Figura 1A exemplifica a aplicação do “Entropy-based discretizador” (sem critério de paragem). A sequência de cortes demonstra o apuramento cronológico dos cortes eleitos, sendo perfeitamente normal e intuitivo que o primeiro ponto de corte a ser eleito esteja localizado numa das extremidades. De registar que os intervalos criados conseguem, quase, conter apenas uma classe.

Sequência de cortes	→	6					5			4		3		2		1
Temperature	→	64	65	68	69	70	71	72	75	80	81	83	85			
Classe	→	Yes	No	Yes	yes	yes	No	Yes	Yes	No	Yes	Yes	No			

Golf Data set (Fonte: Ian Witten e Eibe Frank)

Figura 1A – Sequência de seis cortes utilizando o “Entropy-based discretizor” sem paragem

Anexo 2

Pretendeu-se testar a aderência da variável aleatória – ganho de informação de uma variável independente num determinado conjunto de instâncias, á lei Normal de densidade de probabilidade. Para esse efeito dividiu-se o conjunto de treino (composto por 1.666 instâncias) em quinze subconjuntos de igual dimensão, de forma a obtermos quinze valores independentes, para cada variável, de ganho de informação. Os valores do ganho de informação respeitantes aos quinze subconjuntos, encontram-se identificados no Quadro 1A.

Quadro 1A – Valores de ganho de informação por subconjunto e por variável independente

Variáveis	Sub1	Sub2	Sub3	Sub4	Sub5	Sub6	Sub7	Sub8	Sub9	Sub10	Sub11	Sub12	Sub13	Sub14	Sub15
Ndepend	0,036	0,061	0,017	0,003	0,013	0,035	0,007	0,028	0,027	0,025	0,008	0,033	0,023	0,008	0,007
Pais	0,051	0,009	0,025	0,009	0,038	0,018	0,036	0,036	0,028	0,036	0,005	0,023	0,019	0,026	0,001
Bem	0,048	0,039	0,087	0,068	0,050	0,043	0,013	0,061	0,106	0,031	0,085	0,043	0,063	0,156	0,003
Canal	0,005	0,005	0,010	0,001	0,013	0,002	0,001	0,014	0,015	0,030	0,009	0,001	0,001	0,001	0,008
Vprest	0,501	0,555	0,537	0,501	0,505	0,515	0,547	0,628	0,644	0,528	0,496	0,527	0,469	0,608	0,512
Vpvp	0,853	0,878	0,769	0,849	0,840	0,797	0,914	0,825	0,867	0,849	0,874	0,724	0,829	0,874	0,822
Vei	0,139	0,139	0,130	0,134	0,080	0,102	0,129	0,186	0,215	0,193	0,175	0,174	0,179	0,158	0,147
Vf	0,839	0,834	0,776	0,764	0,827	0,826	0,894	0,797	0,843	0,823	0,807	0,681	0,774	0,845	0,816
Prazo	0,112	0,172	0,116	0,104	0,039	0,133	0,076	0,121	0,160	0,118	0,102	0,123	0,105	0,172	0,059
idade1	0,273	0,269	0,275	0,379	0,343	0,327	0,351	0,296	0,374	0,421	0,330	0,316	0,282	0,311	0,311
estcivil1	0,055	0,063	0,068	0,050	0,032	0,037	0,035	0,049	0,053	0,047	0,086	0,114	0,063	0,059	0,074
profl	0,331	0,276	0,284	0,290	0,207	0,412	0,306	0,206	0,428	0,236	0,278	0,258	0,463	0,356	0,283
origtelpe	0,219	0,089	0,112	0,049	0,071	0,101	0,044	0,029	0,140	0,095	0,056	0,045	0,063	0,061	0,046
Estatuto1	0,017	0,027	0,030	0,059	0,022	0,052	0,018	0,037	0,008	0,017	0,012	0,015	0,028	0,047	0,023
vrend1	0,577	0,563	0,550	0,688	0,488	0,599	0,574	0,683	0,588	0,641	0,510	0,599	0,546	0,610	0,527
tipohab1	0,079	0,030	0,134	0,035	0,089	0,032	0,068	0,076	0,073	0,133	0,057	0,093	0,056	0,042	0,054
antemp1	0,152	0,261	0,289	0,302	0,171	0,293	0,186	0,183	0,218	0,270	0,196	0,192	0,224	0,286	0,256
Antbanc1	0,160	0,195	0,292	0,168	0,166	0,139	0,248	0,264	0,218	0,254	0,187	0,190	0,204	0,214	0,158
anthab1	0,166	0,295	0,191	0,266	0,244	0,386	0,216	0,259	0,190	0,372	0,198	0,217	0,264	0,325	0,226
sex1	0,002	0,000	0,003	0,000	0,001	0,000	0,017	0,008	0,018	0,000	0,003	0,009	0,005	0,008	0,006
Distrito	0,144	0,060	0,111	0,153	0,182	0,093	0,050	0,101	0,075	0,086	0,159	0,137	0,053	0,156	0,155

Procedeu-se ao teste K-S (Kolmogorov-Sminov) de ajustamento à Normal, com correcção de Lilliefors, cujos resultados obtidos em SPSS constam no Quadro 2A. Em todas as variáveis a hipótese nula não foi rejeitada (para um nível de significância de 5%), logo o ganho de informação, no caso prático, ajusta-se à distribuição Normal.

Quadro 2A – Resultados do teste de ajustamento à Distribuição Normal

Tests of Normality

	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
NDEPEND	,148	15	,200*	,908	15	,158
PAIS	,130	15	,200*	,971	15	,835
BEM	,148	15	,200*	,933	15	,364
CANAL	,198	15	,118	,815	15	,010**
VPREST	,176	15	,200*	,883	15	,056
VPVP	,173	15	,200*	,935	15	,378
VEI	,135	15	,200*	,976	15	,909
VF	,169	15	,200*	,904	15	,123
PRAZO	,169	15	,200*	,948	15	,483
IDADE1	,115	15	,200*	,944	15	,450
estcivil1	,166	15	,200*	,910	15	,171
PROF1	,191	15	,145	,922	15	,271
origtelpe	,184	15	,185	,832	15	,010**
estatuto1	,164	15	,200*	,910	15	,176
VREND1	,121	15	,200*	,966	15	,759
TIPOHAB1	,125	15	,200*	,913	15	,201
ANTEMP1	,163	15	,200*	,924	15	,290
ANTBANC1	,123	15	,200*	,956	15	,593
ANTHAB1	,162	15	,200*	,926	15	,301
SEX1	,180	15	,200*	,843	15	,014
DISTRITO	,163	15	,200*	,926	15	,308

*. This is a lower bound of the true significance.

**. This is an upper bound of the true significance.

a. Lilliefors Significance Correction

Anexo 3

O Quadro 3A demonstra que o algoritmo descrito no ANEXO 1 (“Entropy-based discretizador”) com o critério de paragem MDL (“Minimum description length”), não discretizou a variável *Atemp1*.

Quadro 3A – resultado de aplicação do “Entropy-based discretizador” com MDL, como critério de paragem, na variável *Atemp1*.

				info						
	Corte	Info	Gain	(classe)	K1	K2	E1	E2	Stop.Criteria	
1	5.5	0.958	6.028e-04	0.999	2	2	0.659	0.298	-0.008793524	
2	10.5	0.903	1.013e-04	0.918	2	2	0.418	0.484	-0.009226929	
3	19.5	0.842	1.102e-05	0.847	2	2	0.557	0.284	-0.009242502	
4	21.5	0.769	8.474e-04	0.774	2	2	0.232	0.537	-0.008315641	
5	22.5	0.799	4.019e-04	0.816	2	2	0.118	0.680	-0.008797978	
6	23.5	0.737	1.587e-03	0.773	1	2	0.000	0.737	-0.007536715	
7	25.5	0.800	6.316e-04	0.811	2	2	0.173	0.627	-0.008570043	
8	27.5	0.737	2.468e-03	0.768	2	2	0.101	0.635	-0.006654864	
9	28.5	0.821	2.017e-04	0.841	2	2	0.186	0.635	-0.009025938	
10	29.5	0.757	3.260e-03	0.783	1	2	0.000	0.757	-0.005886966	
11	31.5	0.806	1.604e-03	0.811	2	2	0.431	0.374	-0.007604332	
12	32.5	0.664	2.219e-02	0.749	1	2	0.000	0.664	0.013165208	
13	34.0	0.807	1.082e-03	0.845	2	2	0.181	0.625	-0.008127269	
14	35.5	0.671	9.288e-02	0.764	1	2	0.000	0.671	0.083846369	

O Quadro 3A foi construído com os resultados obtidos por cada corte eleito, registando os valores da coluna “Stop.Criteria”. O “Stop.Criteria” foi obtido a partir da expressão (A5), com uma ligeira transformação resultando a expressão (A6).

$$\text{Gain} = \frac{\log_2(N-1)}{N} + \frac{\log_2(3^k - 2) - k \cdot E + K_1 E_1 + K_2 E_2}{N} > 0 \quad (\text{A6})$$

Pela leitura da expressão (A6) se conclui que o algoritmo deverá parar sempre que daquela expressão resulte um valor negativo. Assim sendo conforme se constata no Quadro 3A, o algoritmo pára no primeiro corte eleito, logo não há discretização.

Referências Bibliográficas

- [1] Thomas, Lyn C.; Edelman, David B.; Crook, Jonathan N., *“Credit scoring and its Applications”* SIAM – Society for industrial and applied Mathematics (2002).
- [2] Bailey, Murray , *“Credit Scoring – The principles and Practicalities”*, White Box Publishing (2001).
- [3] Hosmer, David W.; Lemeshow, Stanley, *“Applied Logistic Regression”*, Jonh Wiley & sons (1989).
- [4] Witten, Ian H.; Frank, Eibe, *“Data Mining, Practical Machine, Learning tools and Techniques With Java Implementations”* , Morgan Kaufmann Publishers (1999).
- [5] Michie, D.; Spiegelhalter,D.J.; Taylor, C.C.; *“Machine Learning, Neural and Statistical Classification”* Publicação efectuada no âmbito do “Statlog Project”(1994).
- [6] Reis, Elizabeth, *“Estatística Multivariada Aplicada”*, Edições Silabo (2001).
- [7] Thomas, Lyn C.; Edelman, David B.; Crook, Jonathan N., *“Credit Scoring and Credit Control”*, Based on the proceedings of a conference on Credit Scoring and Credit Control, organized by the institute of Mathematics and its applications and held at the University of Edinburgh in August 1989.
- [8] Guimarães, Rui Campos; Cabral, José A. Sarsfield, *“Estatística”* Mcgraw-Hill (1999).
- [9] Dash, M.; Liu, H., *“Feature Selection for Classification”*, IDA-Elsevier B.V. (1997).
- [10] Quinlan, J.Ross, *“C4.5 Programs for Machine Learning”*, Morgan Kaufman Publishers (1988).
- [11] Flach, Peter, *“Roc Analysis in Machine Learning”* Intelligent systems Seminar (2002).
- [12] Domingos, Pedro; Provost, Foster, *“Tree Induction for Probability-based Ranking”*, Kluwer Academic Publishers (2002).
- [13] Margineantu, Dragos D.; Dietterich, Thomas G., *“Improved Class Probability Estimates from Decision Tree Models”*, Department of Computer Science, Oregon State University

- [14] Michalski, R., J. Carbonell and T. Mitchell. *Machine Learning – Artificial Intelligence approach* (volume 2), Morgan Kaufman, 1986.
- [15] Hall, Mark A.; Smith, Lloyd A., *“Practical Feature Subset Solution For Machine Learning”*, Department of Computer science, University of Waikato.
- [16] Fayyad, Piatetsky-Shapiro, Smyth, "From Data Mining to Knowledge Discovery: An Overview", in Fayyad, Piatetsky-Shapiro, Smyth, Uthurusamy, *Advances in Knowledge Discovery and Data Mining*, AAAI Press / The MIT Press, Menlo Park, CA, 1996, pp.1-34
- [17] J.C.W. Deuse, B. de la Iglesia, C.M. Howard and V.J. Rayward-Smith. "A methodology for Knowledge Discovery: The KDD Roadmap. University of East Anglia, School of Information Systems Technical Report SYS-C99-01.
- [18] Leo Breiman, Jerome H. Friedman, Richard A. Olshen, Charles J. Stone Belmont. "Classification and regression trees". Wadsworth International Group, c1984
- [19] Edite Fernandes. "Estatística Aplicada", Universidade do Minho, 1999
- [20] Tom Mitchell. "Machine Learning", McGraw Hill, 1997
- [21] Peter Dalgaard, "Statistics and computing – Introductory Statistics with R". Springer-Verlag, 2002.