Cross-Selling na Banca de Retalho – Estudo de Caso

Anabela Maria Teixeira Nogueira
Mestrado em Engenharia Matemática
Departamento de Matemática
2014

Orientadores
Prof.ª Doutora Ana Rita Pires Gaio, Professor Auxiliar, FCUP
Prof.º Doutor Joaquim F. Pinto da Costa, Professor Auxiliar, FCUP

Orientador de Estágio
Dr. José Carlos Azevedo, Banif S.A.

Todas as correções determinadas pelo júri, e só essas, foram efetuadas.

O Presidente do Júri,

Porto, _____/_____/_____

Anabela Maria Teixeira Nogueira

Cross-Selling na Banca de Retalho – Estudo de Caso

Dissertação submetida à Faculdade de Ciências da Universidade do Porto para obtenção do grau de Mestre em Engenharia Matemática

Orientação Científica FCUP:
Orientador: Prof.ª Doutora Ana Rita Gaio
Orientador: Prof.º Doutor Joaquim Pinto da Costa

Orientação Banif, S.A.:
Orientador: Dr. José Carlos Azevedo

Departamento de Matemática
Faculdade de Ciências da Universidade do Porto
2014

Agradecimentos

A presente dissertação fruto de um trabalho individual só foi possível graças aos apoios e contributos de diversas pessoas, às quais quero deixar o meu sincero agradecimento.

Ao Dr. José Carlos Azevedo pela partilha do saber, pelos conselhos para a vida profissional e pela oportunidade de estágio.

A todos os colaboradores do Departamento de Marketing do Banif, que se disponibilizaram a partilhar o conhecimento necessário para a realização do estágio e que contribuiu para a presente dissertação.

À Prof.ª Doutora Ana Rita Gaio e ao Prof.º Doutor Joaquim Pinto da Costa pelo tempo disponibilizado com sugestões e ensinamentos que valorizaram este trabalho.

À minha família por todo o apoio disponibilizado. Em especial aos meus pais, pelo incentivo e apoio incondicional na minha formação académica que permitiu que esta jornada acontecesse. À minha irmã Raquel, pelo companheirismo.

Ao Nuno, por ser um ouvinte atento. Agradeço toda a compreensão e palavras de incentivo.

Aos amigos, que me proporcionaram momentos de companheirismo, distração e de motivação.

Resumo

A banca de retalho bancário tem evoluído, prestando atenção aos mercados e aos seus clientes. Alguns produtos e serviços são mais padronizados, destinados a determinados grupos de clientes, mas por vezes são flexíveis, permitindo uma adaptação às necessidades do cliente. No entanto, as ofertas praticadas pelas instituições bancárias são similares, deixando de ser atrativo para os clientes estarem fidelizados a uma instituição. A fidelização permite um relacionamento duradouro com o cliente que permitirá uma maior sustentabilidade para a banco. A estratégia de marketing *cross-selling* permite não só a possibilidade de maiores lucros com a aquisição de produtos complementares, mas também aumentar o valor do cliente para a instituição.
Neste trabalho foi desenvolvida uma estratégia de *cross-selling* para os clientes particulares ativos da instituição bancária Banif. As metodologias utilizadas foram regras de associação, árvores de decisão e modelos de misturas finitas. As regras de associação apresentaram a existência de poucas associações entre produtos, mesmo tendo sida adicionada informação sócio-económica dos clientes. Apenas os clientes que adquiriram produtos considerados de *cross-selling* foram considerados para as análises posteriores. As árvores de decisão permitem estruturar regras de vendas para a maior parte dos produtos de *cross-selling*. Os modelos de misturas finitas utilizaram as variáveis sócio-económicas dos clientes como concomitantes, e os produtos foram as variáveis de entrada.
Para os clientes do Banif que possuem produtos de *cross-selling*, o modelo de misturas finitas obtido ajusta-se bem aos dados dos clientes, tendo permitido segmentar os clientes de acordo com a aquisição de produtos. Para uma maior eficácia na concretização das vendas de produtos complementares, a informação recolhida das árvores de decisão revelaram-se úteis.

Palavras-chave: CROSS-SELLING, RETALHO BANCÁRIO, REGRAS DE ASSOCIAÇÃO, ÁRVORES DE DECISÃO, MODELOS DE MISTURAS FINITAS.

Abstract

The retail banking has evolved, paying attention to markets and customers. Some products and services are more standardized, targeted at particular groups of customers, but sometimes are flexible, allowing for an adaptation to the customers needs. However, as banks are practicing very similar offerings to their clients, these are starting to see very few advantages in staying loyal to a single bank. The loyalty allow for a longer relationship with the client that will enable greater sustainability for the bank. The marketing strategy cross-selling allows not only for the possibility of greater profits, through the acquisition of complementary products, but also for the increase of the the customer's value to the institution.
In this work a strategy for cross-selling to active private clients from the Banif bank was developed. The used methodologies were association rules, decision trees and finite mixture models. Association rules showed the existence of few associations between products, even with the socio-economic customer information added. Only customers who purchased cross-selling products were considered for further analysis. Decision trees allow for the creation of sales rules for most of the cross-selling products. Finite mixtures models used socio-economic variables as concomitants variables and the products as input variables.
For customers of Banif bank that own cross-selling products, the obtained finite mixtures model fited well to customer data and enabled to segment customers according to their purchased products. For greater effectiveness in achieving sales of complementary products, information gathered from decision trees proved to be helpful.

Keywords: CROSS-SELLING, RETAIL BANKING, ASSOCIATION RULES, DECISION TREES, FINITE MIXTURE MODELS.

Conteúdo

Lista de Tabelas

Lista de Figuras

Capítulo 1

Introdução

No passado as vendas de retalho consistiam em colocar à venda produtos ao público. Se os produtos fossem vendidos, mais produtos eram adquiridos, e caso não se vendessem, os produtos deixavam de ser adquiridos. Era portanto um negócio orientado para o produto, necessitando da sensibilidade dos retalhistas para apostar no produto certo. Nos dias de hoje, os negócios tendem a ser orientados para os clientes. Resultando num melhor serviço prestado ao cliente, esta nova abordagem requer um melhor conhecimento do cliente, especialmente das suas necessidades e preferências. Na área da banca tal conhecimento pode ser adquirido através do comportamento de compra revelado através das transações.
A banca de retalho à semelhança de outras áreas de retalho também está sujeita à competitividade, competindo com a qualidade dos seus produtos/serviços, preços praticados e a sua reputação. Torna-se cada vez mais necessário vender o produto certo através do canal certo, ao cliente certo. Para tal, deve-se ter em conta, que quando um cliente opta por determinada instituição bancária devido à sua oferta de produtos, poderá adquirir produtos complementares, ao longo do tempo.

1.1 Tema da Dissertação

Esta dissertação resulta do trabalho desenvolvido numa primeira fase dum estágio realizado na entidade bancária Banif – Banco Internacional do Funchal, SA, e seguido de uma fase de aperfeiçoamento das técnicas utilizadas e de apreensão de novos conhecimentos. Inicialmente foi proposto como tema de estágio, o desenho de um modelo para *cross-selling* e *upselling* na banca de retalho, ambas estratégias de marketing relacional.
Cross-selling ou *venda cruzada* consiste em oferecer a clientes existentes produtos complementares ou vender produtos relacionados com os produtos anteriormente adquiridos. No retalho bancário importa oferecer produtos complementares aos clientes permitindo à entidade bancária aumentar a fidelização com estes clientes. Este processo torna mais difícil a um cliente com vários produtos bancários mudar de entidade bancária, devido aos custos de mobilidade. Outra vantagem para a entidade bancária, para além dos eventuais lucros com a aquisição de produtos,

consiste no facto de ser menos dispendioso vender produtos ou serviços a um cliente existente, do que a um novo cliente. De facto, é possível utilizar a informação adquirida até ao momento sobre o cliente existente, para saber mais sobre as suas preferências, de modo a direcionar a estratégia de *cross-selling* bem como campanhas de *marketing* para serem eficazes. Esta vantagem de informação, adicionada aos elevados custos de mobilidade bancária, produz um monopólio local virtual na instituição bancária, que fica melhor capacitada para competir pelos seus clientes do que outras firmas que não têm um relacionamento estabelecido ou acesso à mesma informação sobre as suas necessidades e preferências (Kamakura, 2008). A outra estratégia é *upselling*, que consiste em oferecer a clientes existentes *upgrades* de produtos anteriormente adquiridos oferecendo assim um produto de gama superior que satisfaça as suas novas necessidades.

Depois de realizado um pré-processamento dos dados disponibilizados concluiu-se que não seria possível estudar o *upselling* de produtos, devido ao modo como a base de dados foi construída. Assim sendo, o objetivo principal deste trabalho consiste em desenvolver uma estratégia de *cross-selling* para os clientes do banco Banif.

Este estudo por cliente está inserido na área de *Customer Relationship Management* (CRM), que tem vindo a ser utilizada desde 1990. Segundo Buttle (2009), CRM é uma estratégia de negócio que integra processos internos e funções, e redes externas, para criar e dar valor aos clientes-alvo num dado lucro. Fundamenta-se em dados relacionados de alta qualidade de clientes, e habilitados pela tecnologia de informação.

Quando o volume de dados é muito grande, as análises de dados tradicionais não podem ser usadas, contudo recorre-se a técnicas de *data mining* para processar grandes volumes de dados para além de outros tipos de análises. Fayyad et al. (1996) refere que *data mining* é uma parte de *Knowledge Discovery in Databases* (KDD), este último termo surge em 1989 e contempla uma série de passos: seleção de dados, pré-processamento, transformação de dados, *data mining* e interpretação/validação. Aqui, *data mining* consiste na aplicação de algoritmos específicos para extrair padrões dos dados. As tarefas de *data mining* utilizadas nesta dissertação são: regras de associação, classificação e *clustering*.

Este trabalho encontra-se segmentado em vários capítulos. O presente capítulo permite contextualizar a temática desta dissertação. O capítulo 2 aborda a base teórica de análise de associações, com foco nas regras de associação. O capítulo 3 inicia-se com uma abordagem sobre métodos de classificação, com destaque na construção de árvores de decisão. O capítulo 4 aborda a base teórica da análise de *clusters*, focando-se nos modelos de mistura finita. No capítulo 5 são apresentados os resultados da aplicação das diversas metodologias bem como uma análise crítica. E por fim, no capítulo 6 são apresentadas as principais conclusões, as limitações e considerações finais relevantes.

1.2 Entidades Envolvidas

O Banif – Banco Internacional do Funchal, SA, é uma sociedade anónima que integra a Banif Comercial SGPS, SA, que por sua vez está integrada na BANIF SGPS, *holding* do Banif – Grupo Financeiro, SA. A origem do Banif – Banco Internacional do Funchal, SA remonta a 1988, tendo sido integrado todo o ativo e passivo da extinta Caixa Económica do Funchal, uma pequena instituição financeira de âmbito regional em dificuldades. O Banif é um banco de relação, expresso na sua atitude sistemática de servir o cliente através de uma gama completa de produtos e serviços especificamente desenhados para ir ao encontro das necessidades dos seus clientes. A estrutura organizacional de suporte assenta na existência de uma *holding* geral, a BANIF SGPS, SA, que organiza as suas quatro *sub-holdings* por diferentes segmentos de negócio:

- Banif – para atuar na área da banca comercial, e desde a sua fundação dá passos decisivos para afirmar a qualidade e extensão dos seus produtos e serviços;
- Companhia de Seguros Açoreana – criada em 1892, rapidamente se expandiu para território continental e ainda no Funchal, contudo em 1975 em resultado da sua nacionalização, todo o património e pessoal do continente e Madeira foram transferidos para outra companhia, enquanto a Açoreana ficava restringida aos Açores. Em 1990 reinicia a sua expansão para o continente, em 1993 já possuía balcões no Porto, Lisboa, Braga e Setúbal e com agentes por todo o território. No ano de 1996 foi integrada no Grupo Banif para atuar na área de seguros;
- Banif Investment Bank – criado no ano de 2000 para atuar na área da banca de investimento. É através desta instituição que o grupo passa a ter uma atuação privilegiada e altamente especializada nas áreas de gestão de ativos, mercado de capitais, *corporate finance*, corretagem e *private banking*;
- Banif Mais - em 2009 o Banif – Grupo Financeiro passou a incorporar integralmente o Grupo Tecnicrédito, que detém a totalidade do Banco Mais, SA, instituição de crédito que opera no setor financeiro automóvel em Portugal, Espanha, Eslováquia e Polónia, e do Bank Plus Bank, Zrt. (atualmente designado Banif Plus Bank, Zrt.), que opera no mesmo setor, na Hungria. Esta incorporação reveste-se de um conjunto de vantagens e sinergias nomeadamente ao nível do alargamento e diversificação da base acionista e de clientes do Banif – Grupo Financeiro. Desta integração resulta uma mudança de marca de Banco Mais para Banif Mais, que atua na área de crédito especializado.

Compreende-se assim que, por razões históricas, algumas gamas de produtos e serviços foram disponibilizados aos seus clientes à medida que o grupo Banif ia integrando as diversas *sub-holdings*. O grupo detém atualmente uma oferta de produtos e serviços financeiros que percorre todas as necessidades, interesses e expectativas desde o individual e particular até à grande empresa ou organismo público: banca comercial de retalho, crédito especializado, corretagem,

seguros, serviços de banca de investimento, *corporate finance*, consultoria financeira, gestão de ativos, *project finance*, *private banking* e mercado de capitais são alguns dos principais produtos e serviços da sua oferta permanente.

1.3 Processamento dos dados

Os dados utilizados nesta dissertação foram recolhidos no mês de Setembro de 2013 contendo informação dos clientes da instituição bancária Banif – Banco Internacional do Funchal, SA. Os dados estão organizados de acordo com as contas de depósitos à ordem contendo todos os clientes existentes até à data. Estes últimos são de dois tipos: empresas e particulares, sendo os particulares os clientes de interesse para esta dissertação, pois são os clientes em maior número e com elevado interesse para com a sua fidelização ao banco. Os dados incluem quais os produtos ou serviços que o cliente possuía até ao mês de recolha da informação, bem como valores patrimoniais, observações mensais de saldos médios e informações sobre a abertura de conta. Foram também disponibilizados dados sócio-demográficos dos clientes constituídos como primeiro titulares da conta. Do total das observações foram retiradas as informações sobre as contas que estavam ativas, isto é, contas que possuem produtos. Ainda sobre estes dados, foram escolhidos os clientes que possuíam atualmente uma só conta ativa, independentemente de possuir outras mas que estejam inativas, representando assim 82,67% sobre os dados das contas ativas.
Numa pré-análise, os dados revelaram que em muitas variáveis existia um grande número de *outliers* devido a valores monetários elevados em alguns clientes. As variáveis que indicam o estado civil, habilitações e profissão do cliente apresentaram valores omissos, em cerca de 11%, 46% e 37% dos dados, respetivamente. Estes valores omissos são justificáveis devido à política de obrigatoriedade no preenchimento da informação dos clientes ser recente. Os dados convergidos de várias *sub-holdings* por vezes possuíam diferentes tipos de informação, tendo sido necessário generalizar a informação sobre cada produto; nomeadamente, indicar se o cliente era possuidor do produto discriminando o número total de produtos.
Com o intuito de realizar uma segmentação de clientes, o algoritmo K-Médias foi utilizado para determinar grupos. A maioria das variáveis permaneceram como numéricas e foram estandardizadas. Numa segunda abordagem, pretendeu-se estudar a venda efetiva de produtos de *cross-selling*, excluindo todos os produtos vinculados. Inicialmente foi criado um modelo logístico, em que se pretendia distinguir os clientes com e sem produtos de *cross-selling*. Contudo, dado que os clientes sem produtos de *cross-selling* são os clientes que dominam, importa observar os clientes com apenas um produto dos clientes com mais produtos.
Estas abordagens iniciais não permitiram estruturar uma estratégia que possibilitasse vendas de *cross-selling*, tendo sido obtidas ideias gerais sobre os clientes e que o baixo número de clientes com produtos de cross-selling dificulta o visionamento de possibilidades de venda. Por estas razões, foram utilizadas outras metodologias com a base de dados discretizada. Uma breve descrição das variáveis utilizadas está disponível na Tabela 1.1. Ao longo deste trabalho, surgiu

a necessidade de proteger o nome dos produtos bem como alguma informação concreta sobre os clientes, de modo a proteger a confidencialidade da informação sobre os clientes, colaboradores e fornecedores. Pela Tabela 1.1 pode observar-se que os produtos estão numerados de 1 a 30, são produtos pertencentes às diversas *sub-holdings* apresentadas anteriormente.

Tabela 1.1: Especificação das variáveis.

Variáveis	Definição
Sexo	Género do cliente: 0 - Feminino; 1 - Masculino.
Idade	Idade do cliente com categorização ordinal: 0, 1, 2, 3 e 4.
Residente	Informação sobre se o cliente reside em Portugal; 0 - não; 1 - sim.
EstCivil	Estado civil do cliente: C/U - casado/a ou em união de facto; Solt - Solteiro/a; e D/S/V - Divorciado/a, separado/a ou viúvo/a.
Habilit	Habilitações literárias do cliente: Bas - ensino básico; Sec - ensino secundário; e Sup - ensino superior.
Profissao	Profissão do cliente: 0 - doméstica ou estudante; 1 - quadro médio de empresa; e 2 - quadro superior de empresa.
AnosCliente	Número de anos do indivíduo como cliente do Banif com categorização ordinal: 0, 1, 2 e 3.
PatFin	Património financeiro do cliente em setembro de 2013, com categorização ordinal: 0, 1 e 2.
PatFinAA	Informação sobre o património do ano anterior: 0 - o património do ano anterior é inferior ao atual; 1 - caso contrário.
SldMdSem	Saldo médio semestral do cliente relativamente ao mês de setembro de 2013, com categorização ordinal: 0, 1 e 2.
TRecursos	Total de recursos que o cliente posui no mês de setembro de 2013, com categorização ordinal: 0, 1, e 2.
P1, . . . , P30	Posse do produto: 0 - não possui; 1 - possui o produto.

Capítulo 2

Análise de Associações

As regras de associação permitem encontrar padrões frequentes, associações ou correlações entre conjuntos de itens em bases de dados de transações, relacionais ou outras. As regras de associação foram introduzidas por Agrawal et al. (1993) e a sua aplicação clássica tem sido em análises de dados em *market basket* (ou cestos de compras), que procura descobrir entre os itens comprados pelos clientes, quais são os que estão associados. É precisamente com esta visão que é pretendido usar este tipo de regras para identificar novas oportunidades de *cross-selling* de produtos dos clientes do banco.
Existem estudos que utilizam regras de associação na área de marketing e de *cross-selling*, nomeadamente: Anand et al. (1998) que utilizam as regras de associação numa amostra de clientes do sector financeiro; Wong et al. (2005) propõem um método para obter recomendações acionáveis para maximizar o lucro da venda de produtos, descobrindo o conjunto de produtos que devem ser descontinuados; Lee et al. (2013) utilizam regras de associação não baseadas na frequência mas na utilidade, de modo a refletir não só a correlação estatística mas a significância semântica, isto é, o preço e a quantidade. Embora Berry and Linoff (2004) tenham afirmado que as regras de associação não são uma boa escolha para a criação de modelos de *cross-selling* em indústrias como o retalho bancário, porque descrevem promoções anteriores de *marketing*. Existem, ainda, vários estudos noutras áreas: *web mining* (Pei et al., 2000; Tan and Kumar, 2002), análise de documentos (Holt and Chung, 1999), bioinformática (Satou et al., 1997; Xiong et al., 2005), diagnóstico de alarme de telecomunicação (Klemettinen, 1999), deteção de intrusos numa *network* (Barbara et al., 2001; Lee et al., 2000) e em previsão de *stock movement* (Lu et al., 1998).

2.1 Conceitos Básicos de Regras de Associação

Hahsler et al. (2009b) apresenta uma descrição formal da tarefa da obtenção de regras de associação do seguinte modo:
Seja $I = \{i_1, i_2, ..., i_r\}$ um conjunto de atributos binários, chamados itens. Seja $T = \{t_1, t_2, ..., t_s\}$ um conjunto de transações conhecido por base de dados. Cada transação em T possui um

único ID de transação e contém um subconjunto de itens de I. Uma regra é definida como uma implicação da forma $A \rightarrow C$ onde $A, C \subseteq I$ e $A \cap C = \emptyset$. Os conjuntos de itens A e C são chamados de antecedente (ou *left-hand-side*, LHS) e consequente (ou *right-hand-side*, RHS), respetivamente, da regra.
Para selecionar regras interessantes do conjunto de todas as regras possíveis, são utilizadas medidas de significância e de interesse, respetivamente suporte e confiança. A regra $A \rightarrow C$ possui um suporte sup no conjunto de transações T se sup% das transações em T contêm A e C e sendo definido como:

$$sup(A \rightarrow C) = \frac{\text{Total de transações que contêm A e C}}{\text{Total de transações}}. \tag{2.1}$$

O suporte para um conjunto de itens A é definido do seguinte modo:

$$sup(A) = \frac{\text{Total de transações que contêm A}}{\text{Total de transações}}. \tag{2.2}$$

Segundo Liu (2011), o suporte é uma medida útil porque se for muito baixo, a regra deve ocorrer apenas por acaso. Além disso, num contexto empresarial, uma regra que cobre poucos casos (ou transações) pode não ser útil porque não deverá ser lucrativo atuar usando essa regra. A regra $A \rightarrow C$ possui uma confiança $conf$ no conjunto de transações T se $conf$% das transações em T contêm A e C e é definida como:

$$conf(A \rightarrow C) = \frac{sup(A \rightarrow C)}{sup(A)}. \tag{2.3}$$

A confiança pode ser interpretada como uma probabilidade de encontrar o consequente da regra nas transações sob a condição de que a transação também contém o antecedente. Segundo Liu (2011), a confiança determina a preditabilidade da regra, se a confiança de uma regra for muito baixa, não é possível inferir ou prever C através de A, sendo uma regra de pouca preditabilidade e de uso limitado. Assim é possível obter uma regra em que um cliente que compra os produtos a_1 e a_2 também comprará o produto c com uma probabilidade de $conf$% (Hipp et al., 2000).
O problema da obtenção de regras de associação pode ser definido formalmente como o seguinte: dado um conjunto de transações T, encontrar todas as regras que possuam um suporte $\geq minsup$ e confiança $\geq minconf$, onde $minsup$ e $minconf$ são os *thresholds* de suporte e confiança, respetivamente.
Tendo sido obtidas as regras é possível filtrá-las ou classificá-las usando a medida *lift*. Esta última é definida como

$$lift(A \rightarrow C) = \frac{sup(A \rightarrow C)}{sup(A)sup(C)}, \tag{2.4}$$

e pode ser interpretada como o desvio do suporte de toda a regra em relação ao suporte esperado sob a hipótese de independência dado o suporte dos antecedentes e do consequente da regra. Quanto maior o valor de *lift*, maior é a força da associação. Se o valor da medida *lift* for exatamente 1, significa que A e C são independentes; caso seja superior a 1, são

correlacionados positivamente; e caso seja inferior a 1, são correlacionados negativamente. Existem vários algoritmos capazes de encontrar um conjunto de regras de associação, que sejam computacionalmente eficientes e com requisitos de memória diferentes. O melhor algoritmo para obtenção de regras de associação é o algoritmo Apriori que é descrito a seguir.

2.2 Algoritmo Apriori

O Algoritmo Apriori foi desenvolvido por Agrawal and Srikant (1994). Este algoritmo é pioneiro no uso do corte baseado no suporte para controlar sistematicamente o crescimento exponencial dos conjuntos de itens candidatos. E decompõe o problema em duas tarefas:

1. Geração de todos os conjuntos de itens frequentes: tem como objetivo encontrar todos os conjuntos de itens que satisfazem o *threshold* $minsup$.
2. Geração de regras: tem como objetivo extrair todas as regras de confiança elevada (não inferior a $minconf$) encontradas na tarefa anterior. Estas são as chamadas regras fortes.

Segundo Tan et al. (2014), para realizar a primeira tarefa, o algoritmo Apriori segue o seguinte princípio:

Teorema 2.2.1 (Princípio Apriori) *Se um conjunto de itens é frequente, então todos os seus subconjuntos também são frequentes.*

Segundo Agrawal and Srikant (1994), os algoritmos para descobrir grandes conjuntos de itens realizam múltiplas passagens pelos dados. Na primeira passagem, o algoritmo faz uma única passagem por todos os dados para determinar o suporte de cada item e determinar quais deles são $fortes$, isto é, que possuem $minsup$. Em cada passagem subsequente, começa-se com o conjunto de sementes de conjuntos de itens encontrados como sendo fortes na passagem anterior. E usa-se este conjunto para gerar novos conjuntos de itens potencialmente fortes chamados de conjuntos de itens $candidatos$, e contar o seu suporte atual destes últimos durante a passagem pelos dados. No final da passagem, determina-se quais dos conjuntos de itens candidatos são na verdade fortes, e serão o próximo conjunto de sementes para a próxima passagem. Este processo continua até que não sejam encontrados mais conjuntos de itens fortes.

2.3 Avaliação de Regras de Associação

O algoritmo Apriori bem como outros algoritmos utilizados na obtenção de regras de associação têm poder para gerar um grande número de regras, e quanto maior for a base de dados, maior será o número de regras obtidas através destes algoritmos. Por isso, importa discriminar de entre centenas ou mesmo milhares de regras quais as interessantes através de critérios que possam avaliar a qualidade dos padrões de associação. Existem dois tipos de critérios: os estatísticos e

os subjetivos. Os critérios estatísticos utilizam medidas de interesse para determinar se um determinado padrão é interessante. Exemplos dessas medidas que foram abordadas anteriormente são: suporte, confiança e *lift*. Neste tipo de critérios, as regras que usam itens mutuamente independentes ou que cobrem muito poucas transações são consideradas desinteressantes e por isso mesmo são eliminadas utilizando as medidas de interesse. Os critérios subjetivos são critérios teóricos que requerem um certo *know-how* sobre o problema em si. Um padrão é considerado subjetivamente interessante se revelar informação inesperada acerca dos dados ou se gerar conhecimento útil que possa originar ações lucrativas. No caso de terem sido obtidas regras inesperadas, estas poderão evidenciar uma nova oportunidade de *cross-selling* ou levar à toma de medidas que visam explorar novos nichos de mercado.

No entanto, a utilização das medidas de interesse também possui algumas limitações. No caso da medida de suporte ser baixa, algumas regras potencialmente interessantes poderão ser eliminadas pelo *threshold* de suporte, e por consequência alguns itens poderão nem aparecer nas regras. A confiança também pode originar que a informação obtida seja traiçoeira. Considere-se por a situação de retalho não bancário: pretende-se analisar o relacionamento entre as pessoas que bebem chá e café, estando a informação disposta numa tabela de contingências, na Figura 2.1a. Esta última pode ser usada para avaliar a regra $\{\text{Chá}\} \rightarrow \{\text{Café}\}$. À primeira vista parece

Tabela 2.1: Exemplo de preferências alimentares, num grupo de 1000 pessoas.

(a)

	Café	$\overline{\text{Café}}$	Total
Chá	150	50	200
$\overline{\text{Chá}}$	650	150	800
Total	800	200	1000

(b)

	Cupcake	$\overline{\text{Cupcake}}$	Total
Éclair	150	100	250
$\overline{\text{Éclair}}$	100	650	750
Total	250	750	1000

razoável afirmar que as pessoas que bebem chá tendem também a beber café porque o suporte da regra é de 15% (i.e. $sup(\{\text{Chá}\} \rightarrow \{\text{Café}\}) = 150/1000 = 0,15$) e com uma confiança de 75% (i.e. $conf(\{\text{Chá}\} \rightarrow \{\text{Café}\}) = sup(\{\text{Chá}\} \rightarrow \{\text{Café}\})/sup(\{\text{Chá}\}) = 0,15/0,20 = 0,75$). Todavia, a percentagem de pessoas que bebem café, independentemente de beberem chá, é 80% (i.e. $P(\text{Café}) = 800/1000 = 0,80$), enquanto a percentagem de pessoas que bebem café, sabendo que bebem chá, é de 75% (i.e. $P(\text{Café}|\text{Chá}) = 150/200 = 0,75$). Então, uma pessoa que beba chá diminui a probabilidade de beber café de 80% para 75%. A regra, apesar do seu alto valor de confiança, ignora o suporte do consequente do conjunto de itens da regra. Se o suporte das pessoas que bebem café fosse tido em conta, não haveria surpresa de que muitas pessoas que bebem chá também bebem café.

Para ajudar a avaliar as regras de associação obtidas, apesar das limitações das medidas de suporte e confiança, a medida *lift* pode ser utilizada com esse propósito, medindo a força da associação. Todavia, possui pequenas desvantagens como ser suscetível ao ruído em base de dados pequenas e ainda, os conjuntos de itens que ocorrem com pouca frequência podem pro-

duzir valores elevados da medida *lift*. Considere-se agora como exemplo o estudo da frequência de ocorrência de {Chá,Café} e {Éclair,Cupcake} nos hábitos consumistas utilizando as Tabelas 2.1a e 2.1b. Os valores de *lift* para ambos os hábitos consumistas são:
$lift(\{\text{Chá}\} \rightarrow \{\text{Café}\}) = \frac{sup(\{\text{Chá}\} \rightarrow \{\text{Café}\})}{sup(\{\text{Chá}\}) * sup(\{\text{Café}\})} = 0,15/(0,20 * 0,80) = 0,938$;
$lift(\{\text{Éclair}\} \rightarrow \{\text{Cupcake}\}) = \frac{sup(\{\text{Éclair}\} \rightarrow \{\text{Cupcake}\})}{sup(\{\text{Éclair}\}) * sup(\{\text{Cupcake}\})} = 0,15/(0,25 * 0,25) = 2,40$;
sugerindo uma ligeira correlação negativa para {Chá,Café}; e para {Éclair,Cupcake} uma correlação positiva. Apesar das probabilidades das pessoas que bebem chá e café ser 15%, a mesma probabilidade do que as pessoas que consomem *éclairs* e *cupcakes*; neste caso a medida de confiança seria a melhor medida, porque considera que a associação entre Chá e Café (i.e. $conf(\{\text{Chá}\} \rightarrow \{\text{Café}\}) = 0,75$) é mais forte do que a associação de *Éclair* e *Cupcakes* (i.e. $conf(\{\text{Éclair}\} \rightarrow \{\text{Cupcake}\}) = 0,60$).

Capítulo 3

Classificação

A classificação é também conhecida como aprendizagem supervisionada ou aprendizagem indutiva. Este tipo de aprendizagem é semelhante à aprendizagem humana sobre experiências passadas para ganhar novos conhecimentos, de modo a melhorar a nossa habilidade para realizar as tarefas na vida real. Dado não ser possível os computadores terem experiência, estes podem aprender através da análise de dados adquiridos anteriormente.
Há vários métodos de aprendizagem supervisionada, neste capítulo são abordadas as árvores de decisão, uma metodologia não-paramétrica capaz de lidar com falta de informação. E por estas razões é uma metodologia utilizada para reconhecimento de padrões em diversas áreas como na medicina, por exemplo na gestão das *guidelines* do tratamento da doença de Parkinson (Olanow and Koller, 1998); no reconhecimento de caracteres, (Wang and Suen, 1984) utilizaram um classificador de árvore de decisão baseado num processo recursivo de redução de entropia, para reconhecimento de caracteres chineses, utilizando um grande número de classes. E também pode ser aplicado a *cross-selling*; Salazar et al. (2007) utilizam árvores de aquisição de compra para cada segmento de clientes do ramo financeiro, sugerem qual será a sequência de compras após a aquisição de determinado produto.

3.1 Conceitos Básicos

Segundo Tan et al. (2014), para a tarefa de classificação tendo como dados de *input* uma coleção de registos, cada registo ou instância é caracterizado por um tuplo $(\mathbf{x}, y)$, onde **x** é o conjunto de variáveis e y uma variável categórica. As variáveis em **x** podem ser discretas ou contínuas. Caso a variável y seja uma variável contínua trata-se de um problema de regressão. O modelo de classificação é útil na modelação descritiva como uma ferramenta exploratória para distinguir os objetos de classes diferentes; e ainda na modelação preditiva para prever a classe de registos, cuja classe seja desconhecida.
As técnicas de classificação são mais adequadas para prever ou descrever conjuntos de dados com categorização binária ou nominal. São menos eficientes para variáveis categóricas ordinais porque não consideram a ordem implícita entre as categorias.

A abordagem geral para a resolução de um problema de classificação utiliza validação cruzada para avaliar a capacidade de generalização do modelo. Os dados são divididos em duas amostras aleatórias e mutuamente exclusivas: uma amostra de treino utilizada para a estimação do modelo; e uma amostra de teste utilizada para a validação do modelo. Para um conjunto de dados de tamanho *suficientemente* grande o método *holdout* de validação cruzada pode ser utilizado, e o conjunto de dados pode ser separado em quantidades iguais ou desiguais, como por exemplo 2/3 dos dados para a amostra de treino e o restante para amostra de teste. Após a obtenção das duas amostras, a estimação do modelo é realizada com a amostra de treino, e posteriormente, o modelo é aplicado aos dados de teste e o erro de predição é calculado.
A avaliação da performance do modelo de classificação é baseada na contagem de registos cuja classe foi correta e incorretamente prevista pelo modelo. Estas contagens são apresentadas numa tabela de contingências. A Tabela 3.1 apresenta a tabela de contingências para um problema de classificação binário. Cada entrada m_{ij} nesta tabela denota o número de registos da classe i prevista como sendo da classe j. Baseado nas entradas da tabela de contingências, o número total de previsões corretas pelo modelo é $(m_{11} + m_{00})$ e o número total de previsões incorretas é $(m_{10} + m_{01})$.
Embora a tabela de contingências providencie a informação necessária para determinar se o

Tabela 3.1: Tabela de contingências de um problema com 2 classes.

		Classe Prevista	
		Classe = 1	Classe = 0
Classe Atual	Classe = 1	m_{11}	m_{10}
	Classe = 0	m_{01}	m_{00}

modelo de classificação é bom, sumariar a informação com um único número tornaria mais conveniente comparar o desempenho de modelos diferentes. Isto pode ser feito usando uma métrica de performance tal como a taxa de precisão definida como:

$$\text{Taxa de precisão} = \frac{\text{Número de previsões correctas}}{\text{Número total de previsões}} = \frac{m_{11} + m_{00}}{m_{11} + m_{10} + m_{01} + m_{00}} \tag{3.1}$$

De modo equivalente, a performance do modelo pode ser expressa em termos da taxa de erro, que é dada pela equação seguinte:

$$\text{Taxa de erro} = \frac{\text{Número de previsões incorrectas}}{\text{Número total de previsões}} = \frac{m_{10} + m_{01}}{m_{11} + m_{10} + m_{01} + m_{00}} \tag{3.2}$$

A maioria dos algoritmos de classificação procura modelos que possuam uma elevada taxa de precisão, ou de modo equivalente, uma taxa de erro baixa quando aplicada ao conjunto de teste.

3.2 Árvores de decisão

Os métodos baseados em árvores particionam o espaço de variáveis num conjunto de regiões, e de seguida encontram um modelo simples que se ajuste a cada uma das regiões. Sendo uma abordagem não-paramétrica para criar modelos de classificação, não requer suposições prévias relativamente ao tipo de distribuições de probabilidade satisfeitas pela classe e restantes variáveis. Esta secção introduz a análise CART (*Classification and Regression Trees*) desenvolvido por Breiman et al. (1984), que pode ser aplicada como uma árvore de classificação ou uma árvore de regressão dependendo se a variável a predizer é discreta ou contínua. Esta secção é baseada nos trabalhos de Duda et al. (2001), Hastie et al. (2009) e Tan et al. (2014).

Os classificadores de árvores de decisão são construídos através de sucessivas divisões do espaço, as quais podem ser múltiplas ou binárias. As divisões binárias serão utilizadas nesta dissertação devido à sua simplicidade. O classificador realiza uma série de questões, em que a próxima questão depende da resposta da atual questão. Esta abordagem além de intuitiva é útil para variáveis categóricas, em que as respostas podem ser do tipo ”sim/não” ou ainda um determinado valor de entre um conjunto de valores. Estas questões podem ser representadas como uma árvore, que se trata de uma estrutura hierárquica representada através de nós e ramos. Por convenção, o nó no topo é chamado de raiz e está ligado por ramos aos nós descendentes. Estes estão de modo similar ligados a outros nós, até atingir um nó terminal ou também chamado de nó folha, pois não possuem descendentes. Numa árvore de decisão, a cada nó folha está atribuída uma classe. Os nós não terminais contêm condições de teste sobre variáveis para separar os registos que contêm diferentes características. Obtida a árvore de decisão facilmente se classifica uma instância da amostra de teste. Começando pelo nó raiz, aplicam-se sucessivamente as condições de teste e segue-se o ramo apropriado baseado na resposta da condição de teste.

A grande vantagem na utilização das árvores de decisão é a facilidade de interpretação. O caminho desde o nó raiz até um nó folha permite traçar uma decisão sobre um padrão. Contudo se a árvore for muito grande pode elevar o erro, e por isso surge a necessidade, durante a construção de uma árvore, decidir se é necessário parar a sua construção para não ficar extensa (Secção 3.2.2) ou após a sua construção podar a árvore (Secção 3.2.3).

Um número exponencial de árvores de decisão podem ser construídas a partir de um dado conjunto de variáveis. Enquanto algumas árvores são mais precisas do que outras, encontrar a árvore ótima é computacionalmente inviável devido ao número exponencial do espaço de procura. No entanto, algoritmos eficientes têm sido desenvolvidos para induzir uma árvore de decisão razoavelmente precisa, num razoável período de tempo. Estes algoritmos usualmente utilizam uma estratégia gulosa que constrói uma árvore de decisão utilizando decisões localmente ótimas sobre as variáveis, para particionar os dados. Em suma, a construção da árvore depende de três elementos: a seleção das divisões; as decisões para declarar um nó folha ou continuar a dividir; e a atribuição de uma classe a cada nó folha (Breiman et al., 1984).

3.2.1 Condições de Teste e Impureza do Nó

Os algoritmos de árvores de decisão devem providenciar um método para expressar as condições de teste do conjunto das variáveis e dos seus resultados correspondentes para diferentes tipos de variáveis. Para variáveis binárias, a condição de teste gera dois potenciais resultados. Nas variáveis nominais, como podem possuir vários valores, a condição de teste pode ser expressada de duas formas: divisão binária, em que existe $2^{v-1} - 1$ formas de criar uma partição binária com v valores da variável; ou uma divisão múltipla, em que o número de resultados depende do número de valores distintos da variável utilizada na condição de teste. As variáveis ordinais podem possuir divisões binárias ou múltiplas, embora os valores da variável ordinal tenham que ser agrupados de forma a não violar a propriedade de ordem dos valores da variável. E relativamente às variáveis contínuas, a condição de teste T deve ser expressa como um teste de comparação $(T < d)$ ou $(T \geq d)$ com resultados binários ou múltiplos intervalos na forma $d_j \leq T \leq d_{j+1}$, para $j = 1, ..., v$. Para o caso binário, o algoritmo da árvore de decisão deve considerar todas as divisões possíveis d, e selecionar a que produz a melhor partição. Para uma divisão múltipla, o algoritmo deve considerar todos os possíveis intervalos dos valores contínuos. Uma possível abordagem seria discretizar a variável e atribuir um novo valor a cada intervalo criado. Intervalos adjacentes podem ser agregados em intervalos maiores, para que a ordem seja preservada.
O princípio fundamental subjacente à criação da árvore é a simplicidade; sendo preferível a tomada de decisões que dão origem a uma árvore simples, compacta e com poucos nós. Procura-se então a condição de teste T para cada nó N que permita que a pureza dos dados atinja os seus nós descendentes de forma o mais "pura" possível. Para formalizar esta noção, é conveniente definir a impureza de um nó. Diversas medidas de impureza têm sido propostas, e apesar de diferentes, todas possuem o mesmo comportamento. Seja $i(N)$ a impureza de um nó N. Em todos os casos, pretende-se que $i(N)$ seja zero se todos os padrões que atingem o nó forem da mesma categoria, e que seja máximo se as categorias forem diferentes.
A medida mais popular é a *impureza de entropia*:

$$i(N) = -\sum_j P(z_j)\, log_2\, P(z_j), \tag{3.3}$$

onde $P(z_j)$ é a fração de padrões no nó N que são da categoria z_j. Se todos os padrões são da mesma categoria, então a impureza é zero, caso contrário é positiva, com o maior valor a ocorrer quando as diferentes classes são igualmente prováveis.
Outra definição de impureza particularmente útil num caso de duas categorias é a seguinte. Dado o desejo de ter zero de impureza quando o nó representa apenas padrões de uma só categoria, a forma polinomial mais simples é:

$$i(N) = P(z_1)P(z_2). \tag{3.4}$$

Isto pode ser interpretado como a *impureza da variância*, dado que sob as suposições razoáveis está relacionado com a variância da distribuição associada às duas categorias. Uma generalização de impureza da variância, aplicável a duas ou mais categorias, é a *impureza de Gini*:

$$i(N) = \sum_{i \neq j} P(z_i)P(z_j) = 1 - \sum_{j} P^2(z_j). \tag{3.5}$$

Esta é apenas a taxa de erro esperada no nó N se a categoria é selecionada aleatoriamente da distribuição da classe presente em N.
A *impureza de má classificação* pode ser escrita como:

$$i(N) = 1 - max_j P(z_j), \tag{3.6}$$

e mede a probabilidade mínima de que um padrão de treino seja mal classificado em N.
Dada uma sub-árvore a partir do nó N, o valor utilizado na divisão d para a condição de teste T é escolhido de acordo com o teste que diminui a impureza tanto quanto o possível. A diminuição da impureza é definida como

$$\Delta i(N) = i(N) - P_E i(N_E) - (1 - P_E) i(N_D), \tag{3.7}$$

onde N_E e N_D são os nós descendentes à esquerda e direita, respetivamente, $i(N_E)$ e $i(N_D)$ as suas impurezas, e P_E a fração de padrões no nó N que vão para N_E quando a condição de teste T é usada. Então o "melhor" valor d utilizado no teste é a escolha para T que maximiza $\Delta i(T)$.
As medidas de impureza como a entropia e Gini tendem a favorecer as variáveis que possuem um grande número de valores distintos. Mesmo em situações extremas em que uma condição de teste, que resulta num grande número de partições pode não ser desejável, porque o número de registos associados a cada partição pode ser muito pequeno para permitir realizar uma previsão de confiança. CART possui uma estratégia para ultrapassar este problema, restringir as condições de teste a divisões binárias.
Contudo, a escolha da medida de impureza afeta pouco a performance da árvore de decisão, dado que as medidas são consistentes entre elas. Geralmente, a medida de entropia é frequentemente usada devido à sua simplicidade computacional, embora a impureza de Gini também seja por vezes usada. De facto, o critério de paragem e o método para podar a árvore são mais importantes do que a escolha da medida de impureza.

3.2.2 Critérios de Paragem

O crescimento da árvore binária faz-se de cima para baixo; o número de registos torna-se mais pequeno à medida que a árvore cresce. Se a árvore continuar a crescer até que cada nó folha corresponda à impureza mais baixa, então a árvore estará a sofrer de *overfitting*, pois o número de registos pode ser muito pequeno para tomar uma decisão estatisticamente significativa. Mas se a paragem de divisões ocorrer demasiado cedo, então o erro nos dados de treino não é

suficientemente baixo e consequentemente a performance sofre as consequências.
Uma abordagem tradicional consiste no uso de validação cruzada, em que os nós são divididos continuamente em sucessivos níveis até que o erro na amostra de teste seja minimizado. Outro método é definir um pequeno valor de *threshold* na redução da impureza; a divisão termina se o melhor candidato a uma divisão num dado nó reduzir a impureza num valor inferior há quantidade pré-definida, i.e., se $max_d \Delta i(d) \leq \beta$. Este método tem dois benefícios: ao contrário da validação cruzada, a árvore é treinada diretamente usando todos os dados da amostra de treino; e ainda o facto dos nós folha poderem estar em diferentes níveis da árvore, o que é desejável sempre que a complexidade dos dados varia. Contudo, este método possui uma desvantagem: o facto de ser frequentemente difícil saber como escolher o *threshold* porque existe apenas um relacionamento simples entre β e a performance. Um método muito simples consiste em parar quando um nó representa menos do que um certo número de registos, ou uma percentagem do total da amostra de treino, por exemplo 5%.

3.2.3 Poda da Árvore

Ocasionalmente, a paragem de divisões sofre da falta de informação, o chamado efeito horizonte; e a determinação da divisão ótima num nó N não é influenciada pelas decisões nos nós descendentes de N, i.e., aqueles em nível subsequentes. Numa paragem de divisões, o nó N pode ser declarado como nó folha, eliminando a possibilidade de divisões benéficas nos nós subsequentes; como tal uma condição de paragem pode ocorrer demasiado cedo para a precisão geral ótima de reconhecimento.
A principal abordagem alternativa é a poda da árvore. Considere-se, que uma árvore se encontra totalmente construída, isto é, todos os nós folhas têm um mínimo de impureza. Todos os pares de nós folhas circundantes (i.e., aqueles que estão ligados a um nó antecedente comum, um nível acima) são considerados para eliminação. Qualquer par cuja eliminação produz um aumento satisfatório na impureza é eliminado, e o nó comum antecedente é declarado como nó folha. Claramente, tal fusão ou união de dois nós folhas é o inverso da divisão de nós. Não é pouco habitual que, após a poda, os nós folhas existam num grande número de níveis e que a árvore seja desequilibrada.
Os benefícios da poda da árvore são evitar o efeito horizonte. Contudo adicionam um maior custo computacional do que a divisão, e para amostras de treino grandes, o custo pode ser proibitivo. Para amostras pequenas, o custo computacional é pequeno e a poda da árvore é geralmente preferível do que a paragem de divisões.

3.2.4 Antecedentes e Custos

Até agora assumiu-se que a categoria z_i é representada com a mesma frequência na amostra de treino e teste. Se não for esse o caso, é necessário um método para controlar a construção da árvore de modo a obter um erro mais baixo na tarefa de classificação final quando as frequências são diferentes. O método mais direto é ”pesar” as amostras para as frequências antecedentes

também conhecidas como *priors*. Além disso, pode-se procurar minimizar o custo geral, em vez de restringir o custo de má classificação. E ainda representar tal informação numa matriz de custo λ_{ij}, o custo de classificar um padrão w_i quando na verdade é z_j. O custo de informação é facilmente incorporado na impureza de Gini, dando a seguinte expressão de impureza de Gini pesada,

$$i(N) = i(N) = \sum_{ij} \lambda_{ij} P(z_i) P(z_j), \tag{3.8}$$

que deverá ser utilizada na amostra de treino. De modo análogo os custos podem ser incorporados nas outras medidas de impureza.

3.2.5 Variáveis com valores em falta

Os problemas de classificação podem possuir falta de informação nalgumas variáveis, quer durante o treino do classificador, quer na classificação ou mesmo em ambas as fases. Considere-se primeiro a situação de treino de um classificador de árvore em que alguns padrões de treino são na verdade variáveis em falta. Uma abordagem seria apagar os padrões deficientes, mas seria um desperdício, e esta abordagem só deve ser utilizada se existirem muitos padrões completos. Uma técnica melhor é prosseguir como na Secção 3.2.1, mas em vez de calcular as impurezas no nó N é usada apenas a informação da variável presente. Suponhamos que existem n pontos de treino no nó N e que cada um tem três variáveis, exceto um padrão para o qual não se possui informação sobre uma variável x_3. Para encontrar a melhor divisão em N, calculam-se as divisões possíveis usando todos os n pontos, usando a variável x_1, depois todos os n pontos para a variável x_2, e depois $n-1$ pontos não-deficientes para a variável x_3. Cada divisão tem uma redução associada em impureza, calculada como anteriormente, embora agora sejam utilizados números diferentes de padrões. Como sempre, a divisão desejada é aquela com maior diminuição na impureza.

Considere-se agora criar e usar árvores que possam classificar um padrão com falta de informação. As árvores descritas anteriormente não podiam lidar diretamente com condições de teste com falta de variáveis e apesar de se suspeitar que condições de teste deficientes possam ocorrer, deve-se modificar o procedimento de treino discutido na Secção 3.2.1. A abordagem básica durante a classificação é usar a decisão (”primária”) tradicional num nó sempre que possível mas usar condições de teste alternativas sempre que a informação para usar em teste está em falta nessa variável.

Durante o treino da árvore, juntamente com a divisão primária, a cada nó não terminal N é dado um conjunto ordenado de divisões suplentes. A divisão suplente maximiza a associação preditiva com a divisão primária. Uma medida simples da associação preditiva de duas divisões d_1 e d_2 é meramente uma contagem numérica de padrões que são enviados para a esquerda por d_1 e d_2 mais a contagem de padrões que são enviados para a direita por ambas as divisões. A segunda divisão suplente é definida de modo similar; usa outra variável e é a divisão que melhor se aproxima da divisão primária da forma definida acima. Claro que durante a classificação de condições de teste deficientes, usa-se a primeira divisão suplente que não envolve a condição

de teste com falta de informação na variável. Esta estratégia de valor em falta corresponde a um modelo linear que substitui o valor em falta do padrão por um valor de uma variável que exista mais fortemente correlacionada. Esta estratégia usa ao máximo a vantagem das associações entre as variáveis para decidir a melhor divisão, quando os valores da variável não existem. Um método muito relacionado com as divisões suplentes é a utilização de valores virtuais; a cada valor em falta é atribuído o valor mais provável.

Capítulo 4

Análise de *clusters*

A análise de *clusters* ou *clustering*, é uma técnica de aprendizagem não supervisionada. Ao contrário da aprendizagem supervisionada não se possui informação *a priori* sobre o valor das classes particionadas ou dos grupos nos dados. De acordo com esta definição poder-se-ia dizer que as regras de associação são uma tarefa de aprendizagem não supervisionada, contudo, por razões históricas, *clustering* está associado a aprendizagem não supervisionada enquanto as regras de associação não estão.
Um outro modo de promover o *cross-selling* de produtos consiste em segmentar os clientes em pequenos grupos de acordo com as suas similaridades e selecionar os produtos alvo para cada grupo. Para esta tarefa é muito comum serem usados algoritmos de *clustering*, criando partições de clientes. Esta tarefa na área de pesquisa de marketing é intitulada de segmentação.
Dos diversos métodos de *clustering* disponíveis, é abordado o método de *clustering* baseado em modelos, nomeadamente modelos de misturas finitas. As áreas de aplicação de modelos de misturas variam entre medicina Fahey et al. (2007), física Akpinar and Akpinar (2009), economia Ferrall (2005) e marketing Wedel and DeSarbo (2002).

4.1 Clustering usando Modelos de Misturas Finitas

Clustering usando modelos de misturas finitas é um tipo de *clustering* baseado em modelos estatísticos. Por vezes é eficiente assumir que os dados foram gerados como resultado de um processo estatístico e descrever os dados através de um modelo estatístico que melhor se ajusta aos dados, em que este modelo é descrito pela sua distribuição e por um conjunto de parâmetros para essa distribuição. Esta secção descreve um tipo particular de modelos estatísticos, modelos de misturas finitas, que permitem modelar os dados usando várias distribuições estatísticas. Cada mistura está associada a um *cluster* e os parâmetros das várias distribuições nesse *cluster* providenciam uma descrição desse *cluster*, tipicamente em termos do seu centro e amplitude. Os modelos de mistura são adequados para conjuntos de observações que são uma misturas de diferentes distribuições de probabilidade. Esta secção baseou-se nos trabalhos de Tan et al. (2014) e Leisch and Grun (2008).

4.1.1 Modelos de Misturas Finitas

Formalmente, assume-se que existem K componentes e que cada componente segue uma distribuição paramétrica. As componentes têm um peso atribuído que indica a probabilidade a priori de uma observação pertencer a uma componente, e a distribuição da mistura é dada pela soma pesada das K componentes. Se os pesos dependerem de outras variáveis, estas últimas são chamadas de variáveis concomitantes. Exemplos destas variáveis são as variáveis sócio-demográficas como a idade ou género que por vezes estão relacionadas com os diferentes segmentos. O modelo é descrito da seguinte forma

$$\begin{aligned} h(y|x,w,\psi) &= \sum_{k=1}^{K} \pi_k(w,\alpha) f_k(y|x,\theta_k) \\ &= \sum_{k=1}^{K} \pi_k(w,\alpha) \prod_{d=1}^{D} f_{kd}(y_d|x_d,\theta_{kd}), \end{aligned}$$

onde: ψ denota o vetor de parâmetros para as densidades da mistura $h()$ e é dado por $(\alpha,(\theta_k)_{k=1,...,K})$; a resposta é denotada por y, o preditor por x e as variáveis concomitantes por w; f_k é a função de densidade específica da componente k; as variáveis multivariadas y são assumidas como sendo particionadas em D subconjuntos onde as densidades das componentes são independentes entre os subconjuntos, i.e. a densidade de componente f_k é dada pelo produto sobre D densidades que são definidas como o subconjunto de variáveis y_d e x_d para $d = 1, ..., D$. Os parâmetros específicos da componente são dados por $\theta_k = (\theta_{kd})_{d=1,...,D}$. Sobre a suposição de que M observações estão disponíveis, as dimensões das variáveis são dadas por $y = (y_d)_{d=1,...,D} \in \mathbb{R}^{M \times \sum_{d=1}^{D} k_{yd}}$, $x = (x_d)_{d=1,...,D} \in \mathbb{R}^{M \times \sum_{d=1}^{D} k_{xd}}$ e $w \in \mathbb{R}^{M \times k_{xd}}$. Nesta notação k_{yd} denota a dimensão da d-ésima resposta, k_{xd} a dimensão do d-ésimo preditor e k_w a dimensão das variáveis concomitantes. Os pesos das componentes π_k estão restringidos para todo w de modo a que

$$\sum_{k=1}^{K} \pi_k(w,\alpha) = 1 \quad \text{e} \quad \pi_k(w,\alpha) > 0, \quad \forall\, k, \tag{4.1}$$

onde α é o conjunto dos parâmetros associados às variáveis concomitantes. O modelo logístico multinomial dado por

$$\pi_k(w,\alpha) = \frac{e^{w'\alpha_k}}{\sum_{u=1}^{K} e^{w'\alpha_u}} \quad \forall\, k,$$

é assumido para π_k, com $\alpha = (\alpha_k^t)_{k=1,...,K}^t$ e $\alpha_1 \equiv 0$.
Para os modelos de misturas, cada distribuição descreve um grupo diferente, i.e., um *cluster* diferente. É possível identificar quais os objetos que pertencem aos *clusters*, contudo a modelação de misturas não produz uma atribuição dos objetos nos *clusters*, mas providencia a probabilidade de um objeto específico pertencer a um *cluster*.

4.1.2 Estimação dos Parâmetros

Dado um modelo estatístico para os dados, é necessário estimar os parâmetros desse modelo. Nos casos mais simples, sabe-se quais os objetos que vêm de determinadas distribuições, e a situação reduz-se a uma estimação de parâmetros de uma única distribuição através dos dados dessa distribuição. Para as distribuições mais comuns, a estimação dos parâmetros pelo método da máxima verosimilhança é calculada por fórmulas simples envolvendo os dados. Contudo numa situação mais realista, não se sabe que objetos foram gerados por quais distribuições.
Uma abordagem comum para esta tarefa é a estimação por máxima verosimilhança, conseguida através do algoritmo Expectativa–Maximização (EM), para um número fixo K de componentes. Este algoritmo foi assim denominado por Dempster et al. (1977). Dada uma estimativa para os valores dos parâmetros, o algoritmo EM calcula a probabilidade de cada ponto pertencer a cada distribuição e depois usa essas probabilidades para calcular uma nova estimativa de parâmetros. As iterações continuam até que as estimativas dos parâmetros não sejam alteradas ou mudem *muito pouco*. Por conseguinte, a estimação por máxima verosimilhança é aplicada, mas por procura iterativa.
O algoritmo EM aplica um esquema de extensão de dados omissos. Assume-se que uma variável latente $l_m \in \{0,1\}^K$ existe para cada observação m e indica de que componente m é membro, i.e., l_{mk} é 1 se a observação m vem da componente k e 0 caso contrário. Tem-se claramente, $\sum_{k=1}^{K} l_{mk} = 1$ para qualquer m. No algoritmo EM, as observações cujas componentes não são observadas l_{mk}, são tratadas como valores omissos e os dados são aumentados por estimativas de pertença das componentes, i.e., pelas probabilidades a posteriori estimadas $\widehat{p}_{mk}$. Para uma amostra com M observações $\{(y_1, x_1, w_1), ..., (y_M, x_M, w_M)\}$ o algoritmo EM é dado por:

Passo-E: Dadas as estimativas dos parâmetros atuais $\psi^{(i)}$ na i-ésima iteração, substitui os dados em falta l_{mk} pelas probabilidade a posteriori estimadas

$$\widehat{p}_{mk} = \frac{\pi_k(w_m, \alpha^{(i)}) f(y_m|x_m, \theta_k^{(i)})}{\sum_{u=1}^{K} \pi_u(w_m, \alpha^{(i)}) f(y_m|x_m, \theta_u^{(i)})}.$$

Passo-M: Dadas as estimativas para as probabilidade a posteriori $\widehat{p}_{mk}$, que são funções de $\psi^{(i)}$, obtém novas estimativas $\psi^{(i+1)}$ dos parâmetros por maximizar

$$Q(\psi^{(i+1)}|\psi^{(i)}) = Q_1(\theta^{(i+1)}|\psi^{(i)}) + Q_2(\alpha^{(i+1)}|\psi^{(i)})$$

onde

$$Q_1(\theta^{(i+1)}|\psi^{(i)}) = \sum_{m=1}^{M} \sum_{k=1}^{K} \widehat{p}_{mk}\, log(f(y_m|x_m, \theta_k^{(i+1)}))$$

e

$$Q_2(\alpha^{(i+1)}|\psi^{(i)}) = \sum_{m=1}^{M} \sum_{k=1}^{K} \widehat{p}_{mk}\, log(\pi_k(w_m, \alpha^{(i+1)})).$$

Q_1 e Q_2 podem ser maximizados separadamente. A maximização de Q_1 dá novas estima-

tivas $\theta^{(i+1)}$ e a maximização de Q_2 dá $\alpha^{(i+1)}$. Q_1 usa a estimação em modelos lineares generalizados e Q_2 usa a estimação de modelos *logit* multinominais.

O algoritmo K-Médias para dados euclidianos é um caso especial do algoritmo EM para distribuições gaussianas esféricas, com matrizes de covariância iguais, mas com diferentes médias. O passo E (Expectativa) corresponde ao passo do K-Médias de atribuição de cada objeto a um *cluster*. Em vez disso, cada objeto é atribuído a todos os *clusters* (distribuição) com uma determinada probabilidade. O passo M (Maximização) corresponde a calcular os centróides do *cluster*. Em vez disso, todos os parâmetros das distribuições, bem como os pesos dos parâmetros, são selecionados para maximizar a verosimilhança.

4.1.3 Avaliação do Ajustamento do Modelo

Não existe uma só forma, que seja a melhor, para avaliar o ajustamento de um modelo de misturas finitas. As técnicas utilizadas nesta dissertação são de duas categorias: critérios de informação e qualidade de classificação. Relativamente aos critérios de informação foram utilizados: *Akaike Information Criterion* (AIC) e o *Bayesian Information Criterion* (BIC). AIC é uma medida da qualidade do ajustamento de um modelo dada por:

$$AIC = 2\,q - 2\,log\,L,$$

sendo q o número total de parâmetros do modelo e L a estimativa de verosimilhança máxima. BIC é uma medida da qualidade do ajustamento de um modelo dada por:

$$BIC = 2\,log(N) - 2\,log\,L,$$

sendo N o número total de observações N e L a estimativa de verosimilhança máxima. O modelo com melhor ajustamento aos dados é o modelo com menor AIC ou BIC. Estas medidas foram bem estudadas na teoria estatística, e não são suficientes para decidir qual o modelo que deverá ser escolhido. E por esta razão foi utilizada uma técnica de qualidade de classificação, a entropia relativa. A entropia de um modelo é definida como uma medida de incerteza de classificação, dada por:

$$EN(\widehat{p}) = -\sum_{m=1}^{M}\sum_{k=1}^{K}\widehat{p}_{mk}\,log\,\widehat{p}_{mk},$$

em que $\widehat{p}_{mk}$ é a probabilidade a posteriori da observação m pertencer à classe k. Esta medida está delimitada para os valores $[0, \infty[$, em que os valores mais elevados indicam uma maior quantidade de incerteza na classificação. O software *MPLUS* apenas apresenta a medida de entropia relativa de um modelo, que é na verdade a medida de entropia mas definida em $[0, 1]$, dada por:

$$E = 1 - \frac{EN(\widehat{p})}{N\,log\,J}.$$

O modelo com melhor ajustamento aos dados possui um valor próximo de um, pois indica elevada certeza na classificação, enquanto valores próximos de zero indicam baixa certeza na classificação.

4.1.4 Vantagens e Limitações

Encontrar *clusters* através da modelação de dados usando modelos de mistura e aplicando o algoritmo EM para estimar os parâmetros destes modelos possui uma variedade de vantagens e desvantagens. Pelo lado negativo, o algoritmo EM pode ser lento, e não é prático para modelos com um grande número de componentes, e não se comporta bem quando os *clusters* contêm apenas alguns pontos ou se os pontos são quase colineares. Existe também um problema na estimação do número de *clusters* ou, mais genericamente, na escolha da forma exata do modelo a usar. Este problema tem sido tratado através da aplicação de uma abordagem bayesiana, a qual, falando grosseiramente, dá o *odds* de um modelo contra outro, baseado na estimativa derivada dos dados. Os modelos de misturas têm dificuldades com ruído e *outliers*, embora algum trabalho tenha vindo a ser desenvolvido para lidar com este problema.
Pelo lado positivo, os modelos de misturas são mais gerais do que algoritmos como o K-Médias porque podem usar distribuições de vários tipos. Como resultado, os modelos de misturas (baseados em distribuições gaussianas) podem encontrar *clusters* de diferentes tamanhos e formas elípticas. Além disso, uma abordagem baseada num modelo providencia um modo disciplinado de eliminar alguma complexidade associada aos dados. De modo a que os *clusters* produzidos são facilmente caracterizados, uma vez que podem ser descritos por um pequeno número de parâmetros. Finalmente, muitos conjuntos de dados são na verdade o resultado de processos aleatórios, e assim devem satisfazer as suposições estatísticas destes modelos.

Capítulo 5

Resultados

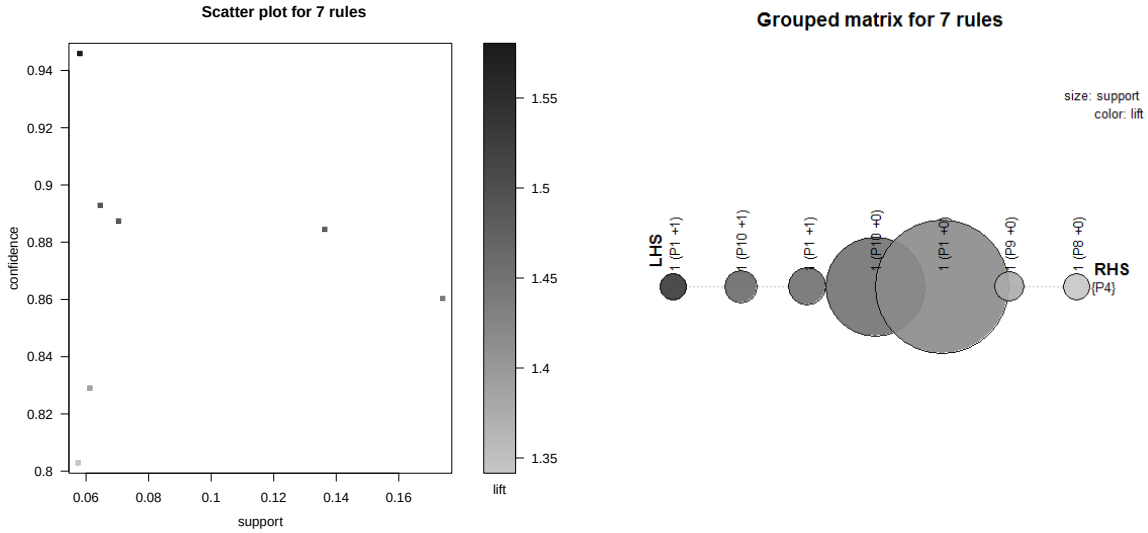
Neste capítulo são apresentados e analisados todos os resultados obtidos da aplicação das diversas metodologias descritas anteriormente. O nosso objetivo é que estes resultados possam contribuir para a definição de novas estratégias de marketing no Banif, se possível mais lucrativas. Os produtos foram nomeados por P1, P2, ..., P30. Os onze produtos com numeração de 20 a 30 são produtos gerais que pertencem a diversas *sub-holdings*, e tal como referido no capítulo inicial, são considerados produtos de *cross-selling*, pela entidade bancária.

5.1 Regras de Associação

As análises desta secção foram realizadas usando o ambiente estatístico *R* (R Core Team, 2013). Em particular, foram utilizados os pacotes: *arules* (Hahsler et al., 2009a) para a mineração das regras de associação e *arulesViz* (Hahsler and Chelluboina, 2013) para visualização das regras de associação. Os dados utilizados são referentes aos produtos de todos os clientes particulares ativos, com exceção dos produtos vinculados. Não é de interesse serem obtidas regras sobre produtos que existem devido à compra e existência de outros produtos específicos, encontrando-se assim dependentes da existência de outros. Para a obtenção das regras de associação, a base de dados foi transformada numa base de dados transacional, e optou-se pelo formato *single* da forma $<$CIF, Item$>$, onde CIF corresponde ao número de identificação de cliente e Item o produto que o cliente possui. Deste modo, cada registo representa um item e cada item está associado a um CIF.
Os conjuntos de regras sobre os produtos dos clientes bancários particulares ativos apresentados nesta dissertação consistem de três conjuntos:

A1 - Regras obtidas com um suporte mínimo de 0.05 e uma confiança mínima de 0.80. Foram obtidas 7 regras que tiveram como produto consequente apenas um único produto, P4.

A2 - Regras obtidas com um suporte mínimo de 0.01 e uma confiança mínima de 0.60. Foram obtidas 60 regras e que tiveram mais do que um produto como produto consequente: P1, P2 e P4.



(a) Diagrama de dispersão.

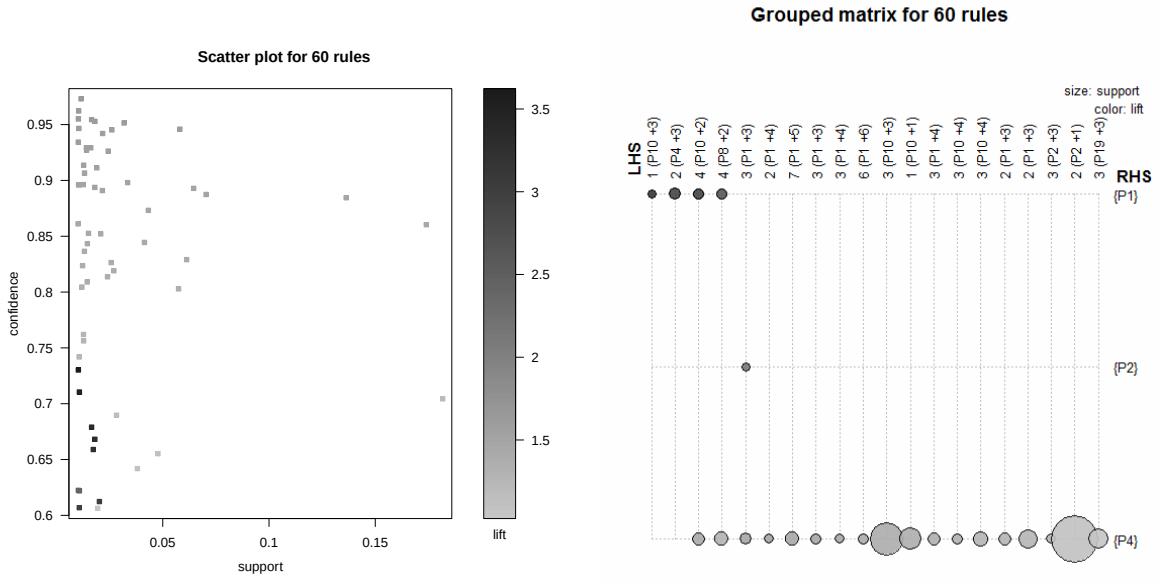
(b) Representação agrupada.

Figura 5.1: Representação gráfica do conjunto de regras A1.

A3 - Regras obtidas com um suporte mínimo de 0.01, uma confiança mínima de 0.10 e com uma restrição relativamente aos produtos consequentes, que consideram apenas produtos de interesse para o banco. Foram obtidas 2 regras com apenas um único produto consequente, P30.

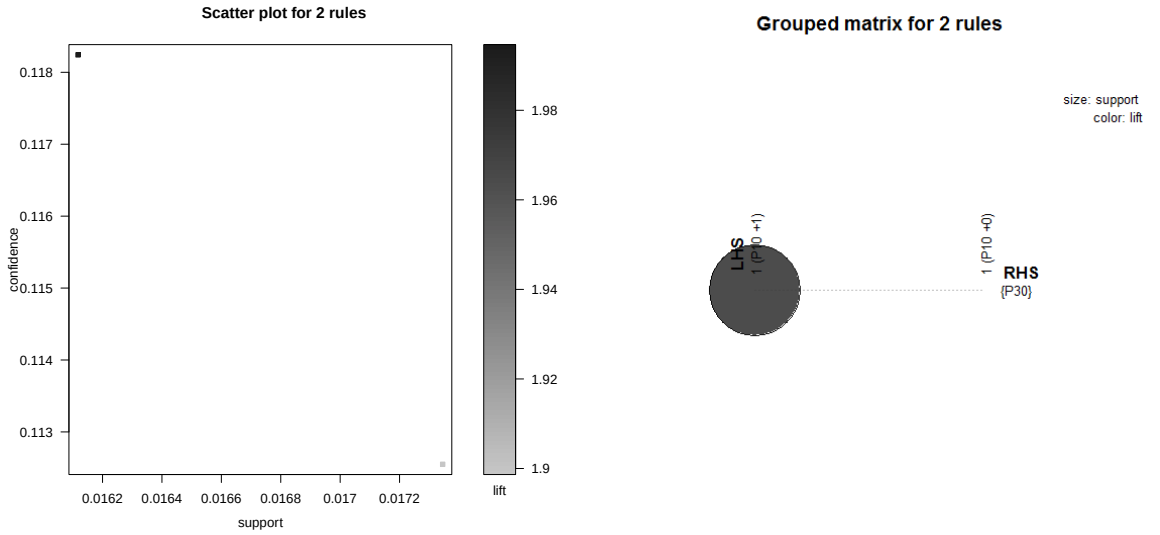
O conjunto de regras A1 foi obtido com um nível escolhido de confiança elevado, acima ou igual a 80%, em que os produtos do conjunto antecedente (LHS) ocorreram em pelo menos 5% dos clientes. A Figura 5.1a apresenta as regras do conjunto A1. Cada regra apresenta os conjuntos antecedentes e consequentes correlacionados positivamente, devido aos valores da medida *lift*. Todas as regras têm como produto consequente P4, que é um produto que não produz lucro contudo permite aumentar a fidelização do cliente (Figura 5.1b). Nesta última figura, a representação agrupada permite relacionar as medidas suporte e *lift*, bem como agrupar as regras, especialmente quando são em grande número pelos produtos que possuem em comum. Os conjuntos P_i $+0$ significam apenas P_i; os conjuntos P_i $+m$ sendo m um número superior a 0, são constituídos por $m + 1$ produtos, com $i = 1, ..., 30$. As duas regras do conjunto A1 com maior força de associação são as que possuem um suporte mais baixo, e com mais do que um produto no conjunto LHS, nomeadamente $\{$P1,P10$\}$ e $\{$P10,P2$\}$. Pelo facto do nível de confiança ser elevado, estas regras não devem ser ignoradas. Na maioria das restantes regras, o seu suporte é superior, mas este está relacionado com o facto das regras possuírem apenas um único produto no conjunto LHS.

Para a obtenção do conjunto de regras A2, foi alargado o nível de confiança e suporte, de modo a permitir observar maior variabilidade de padrões apesar de pouco frequentes. Por isso, este último conjunto inclui as regras do conjunto A1, tendo sido verificado que as regras com menor confiança eram as regras com maior força de associação (Figura 5.2a), que tinham como



(a) Diagrama de dispersão. (b) Representação agrupada.

Figura 5.2: Representação gráfica do conjunto de regras A2.



(a) Diagrama de dispersão. (b) Representação agrupada.

Figura 5.3: Representação gráfica do conjunto de regras A3.

produtos consequentes P1 e P2. Os produtos P1 e P2 têm em comum com P4 o facto de não gerarem lucros para a entidade bancária. E tendo em conta, que para os itens pouco frequentes produzem valores de *lift* elevados. Apesar da medida *lift* considerar o suporte do conjunto consequente (RHS), deve-se considerar a confiança como a melhor medida para avaliar estas regras. E por esta razão, as regras com maior *lift* não são mais importantes do que as restantes. No geral, o conjunto A2 não gerou muito mais informação sobre as associações de produtos do que A1. E dado o suporte ser baixo, o conjunto de regras A2 não poderá ser considerado para a elaboração de estratégias de *cross-selling*.

Na realidade, o banco pretende aumentar a venda de outro tipo de produtos, nomeadamente os produtos entre P20 a P30. Assim sendo, a obtenção dos produtos consequentes (RHS) foi restringida aos produtos P20 a P30, tendo sido obtido o conjunto de regras A3. O suporte e a confiança precisaram de ser alargados, devido à existência de regras quando a confiança é muito baixa. Foram obtidas duas regras (Figura 5.3a). Curiosamente, as regras obtidas foram $\{\text{P10,P4}\} \Rightarrow \{\text{P30}\}$ e $\{\text{P10}\} \Rightarrow \{\text{P30}\}$. A regra cujo LHS é constituído por dois produtos possui um suporte menor, mas uma confiança e *lift* maiores relativamente à regra com um único produto no conjunto LHS. Sendo assim, $\{\text{P10,P4}\} \Rightarrow \{\text{P30}\}$ é a regra com maior força de associação, apesar de ser constituída por dois produtos no conjunto LHS.

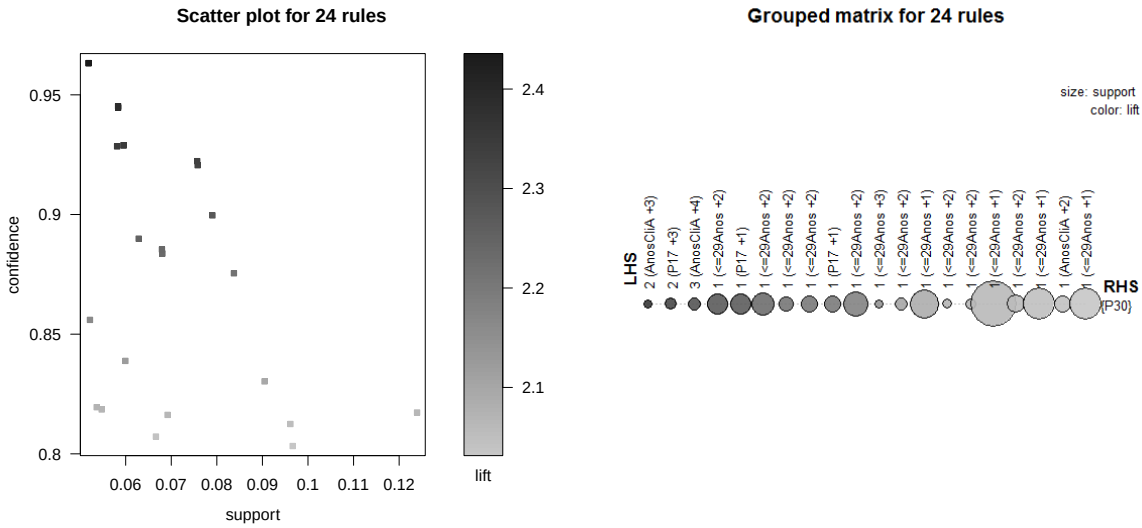
Os conjuntos A1, A2 e especialmente A3 evidenciaram uma fraca adesão por parte dos clientes, isto é, dado o suporte utilizado ter sido baixo, o número de clientes que possui produtos considerados de *cross-selling* é muito baixo. Eventualmente a conjetura económica também produziu efeito sobre o poder económico dos clientes.

Nos resultados anteriores desta secção, cada item foi tratado como uma variável binária assimétrica, i.e. possui dois estados mas um é mais valioso do que o outro. Para a obtenção dos próximos resultados, o conjunto de dados foi alargado para conter variáveis categóricas e binárias simétricas (ambos os estados têm o mesmo peso), contendo informação sobre o cliente. Considerou-se apenas a informação dos clientes que de facto adquiriram produtos de *cross-selling*. Os conjuntos de regras considerados foram os seguintes:

B1 - Regras obtidas com um suporte mínimo de 0.05 e uma confiança mínima de 0.80. Foram obtidas 24 regras e um único produto consequente, P30.

B2 - Regras obtidas com um suporte mínimo de 0.05 e uma confiança mínima de 0.55. Foram obtidas 75 regras e dois produtos consequentes: P23 e P30.

O conjunto de regras B1 foi obtido com um nível escolhido de confiança elevado, acima dos 80%, e com um suporte mínimo de 5%. Em cada uma das regras, o valor da medida *lift* indica que os conjuntos antecedente e consequente estão correlacionados positivamente (Figura 5.4a). O facto de um grande número de regras ter sido obtido, deveu-se à informação sobre o cliente que foi adicionada. Todavia o único produto no conjunto RHS foi P30 (Figura 5.4b). Mais uma vez, o nível de confiança foi alargado, para permitir que outros produtos surgissem nas regras de associação, tendo sido obtido o conjunto B2, com um nível de confiança mínimo de 55%. Este conjunto de regras inclui as 24 regras de B1 e outras para os produtos P30 e P23 (Figura 5.5b).



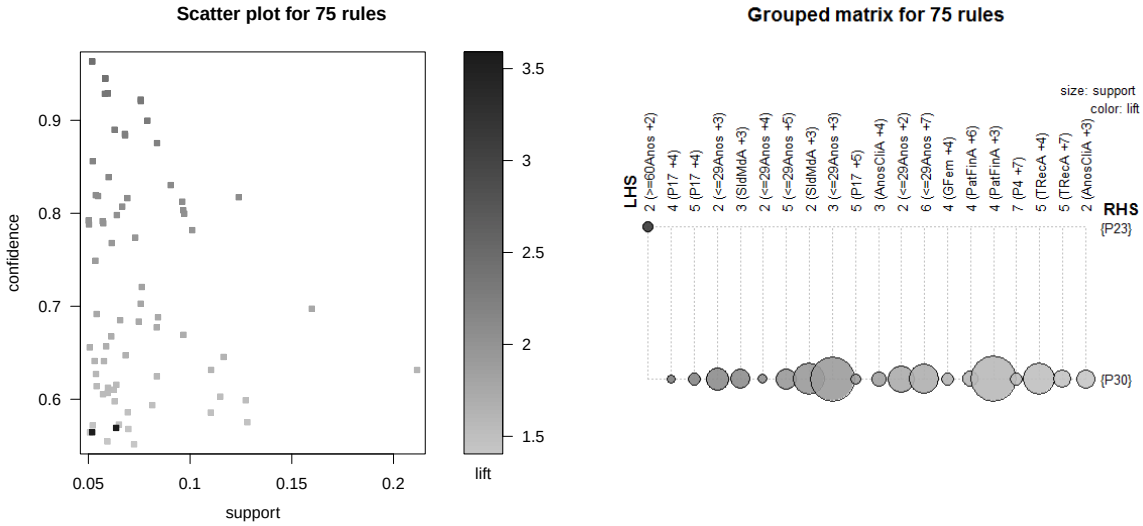
(a) Diagrama de dispersão.

(b) Representação agrupada.

Figura 5.4: Representação gráfica do conjunto de regras B1.

A Figura 5.5a evidencia que as regras sobre o produto consequente P23 são desinteressantes. Na verdade, apesar de possuírem um valor da medida *lift* elevado, têm suporte baixo, tendo em conta que a confiança para estas regras é baixa. Sendo então o conjunto de regras B1 mais interessante, este é analisado com maior pormenor de seguida.

As regras do conjunto B1 incidiram na sua maioria sobre: idade $\leq$29 anos, classe A dos anos como cliente, classe A do património financeiro, classe A do saldo médio semestral, classe A do total de recursos e o produto P17. A Tabela 5.1 sugere que na maioria das regras obtidas, o facto do cliente possuir uma idade $\leq$29 anos não influencia muito. De facto, considerando o par de regras 1 e 2, os valores das três medidas não se alteram significativamente. Resultados análogos para os pares de regras: 3 e 4; 5 e 6; 7 e 23; 8 e 9; 12 e 13. O género do cliente também não parece ser muito relevante, apenas a regra 16 especifica que um cliente com idade inferior a 30 anos, do género feminino e de classe A do património financeiro poderá comprar o produto P30 com um nível de confiança de cerca de 84%. O facto da posse da informação sobre o cliente ser da classe A relativamente ao seu total de recursos diminui o nível de confiança da regra bem como o suporte da mesma, visível entre os pares de regras: 14 e 15; 20 e 21. As regras com maior confiança e medida *lift* geralmente possuem o produto P17. Contudo a regra 24 tem um nível de confiança relativamente baixo, e inclui o produto P17.



(a) Diagrama de dispersão.

(b) Representação agrupada.

Figura 5.5: Representação gráfica do conjunto de regras B2.

Tabela 5.1: Regras de associação do conjunto B1 ordenadas pela medida de confiança.

Id	LHS							Suporte	Confiança	Lift
	Idade	Sexo	AnosCli	PatFin	SldMd	TRec	Produto			
1	-	-	A	A	-	-	P17	0,052	0,963	2,434
2	≤29	-	A	A	-	-	P17	0,052	0,963	2,434
3	≤29	-	-	A	A	-	P17	0,058	0,945	2,388
4	-	-	-	A	A	-	P17	0,058	0,945	2,387
5	-	-	A	-	-	-	P17	0,060	0,929	2,347
6	≤29	-	A	-	-	-	P17	0,060	0,929	2,347
7	≤29	-	A	A	A	-	-	0,058	0,928	2,346
8	≤29	-	-	A	-	-	P17	0,076	0,922	2,330
9	-	-	-	A	-	-	P17	0,076	0,921	2,326
10	≤29	-	A	A	-	-	-	0,079	0,900	2,273
11	≤29	-	A	-	A	-	-	0,063	0,890	2,248
12	≤29	-	-	-	A	-	P17	0,068	0,885	2,237
13	-	-	-	-	A	-	P17	0,068	0,884	2,233
14	≤29	-	-	A	A	-	-	0,084	0,875	2,212
15	≤29	-	-	A	A	A	-	0,052	0,856	2,163
16	≤29	F	-	A	-	-	-	0,060	0,839	2,120
17	≤29	-	A	-	-	-	-	0,090	0,830	2,098
18	≤29	-	-	-	A	A	-	0,054	0,819	2,070
19	≤29	-	-	A	-	B	-	0,055	0,819	2,068
20	≤29	-	-	A	-	-	-	0,124	0,817	2,065
21	≤29	-	-	A	-	A	-	0,069	0,816	2,062
22	≤29	-	-	-	A	-	-	0,096	0,812	2,053
23	-	-	A	A	A	-	-	0,067	0,807	2,039
24	≤29	-	-	-	-	-	P17	0,097	0,803	2,029

5.2 Árvores de Decisão

As análises desta secção foram realizadas usando o ambiente estatístico *R* (R Core Team, 2013). Em particular, foram utilizados os pacotes: *rpart* (Therneau et al., 2013) para a construção de árvores de decisão e *rpart.plot* (Milborrow, 2014) para visualização das mesmas. Os dados utilizados são referentes a todos os clientes particulares ativos, que possuem produtos de *cross-selling*. Mais uma vez, os produtos vinculados foram retirados. As variáveis disponibilizadas para a construção das árvores foram todos os produtos de P1 a P30, idade, sexo, residência, estado civil, profissão, habilitações, saldo médio semestral, anos como cliente e património financeiro (2013). As árvores de decisão foram construídas para cada um dos produtos a realizar *cross-selling*, nomeadamente do produto P20 ao P30. As árvores possuem como variável resposta uma variável categórica, tendo sido utilizadas árvores de classificação. A medida de Gini foi utilizada para escolher qual a melhor divisão e as probabilidades a priori são proporcionais à frequência dos dados observados. Não foi aplicado nenhum critério de paragem mas as árvores foram podadas.

A Figura 5.6 apresenta as etiquetas de cada uma das divisões, bem como a probabilidade por classe das observações em cada nó (a soma dessas probabilidade por nó é 1). O algarismo 0 significa a não venda do produto P20, enquanto 1 representa o oposto. A árvore de decisão sugere que, se um cliente não tem os produtos P21, P23, P24, P25, P28 e P30, então é possível concretizar a venda do produto P20 mas com pouca segurança. Mas se o cliente também não possuir o produto P29 ou P29 e P26 então deverá ser possível concretizar a venda do produto P20 com maior probabilidade. O erro de má classificação utilizando validação cruzada para o produto P20 é de aproximadamente 1%. De forma análoga foram construídas as árvores de decisão para os restantes produtos.

Para os produtos P21, P22 e P27 não foi possível obter árvores de decisão devido à posse destes produtos por parte dos clientes ser pequena. A Figura 5.7 sugere que, aos clientes com a categoria mais alta de património financeiro que não possuam os produtos P8 e P24 mas que possuam P28, é possível concretizar a venda do produto P23. E ainda, se os clientes de património financeiro elevado não possuírem P8 e P28 é possível concretizar a venda do produto P23. Também é possível realizar a venda de P23 se o cliente não possuir os produtos P30, P24, P20, P29 e P21. O erro de má classificação utilizando validação cruzada é de aproximadamente 1.1%.

Relativamente ao produto P24, a árvore foi construída apenas com informação de outros produtos. A Figura 5.8 sugere que aos clientes que não possuam os produtos P30, P28 e P25 é possível concretizar a venda do produto P24. Esta venda também é possível de concretizar se o cliente não tiver os produtos P23, P20, P29 e P21. O erro de má classificação utilizando a validação cruzada é de aproximadamente 3.3%.

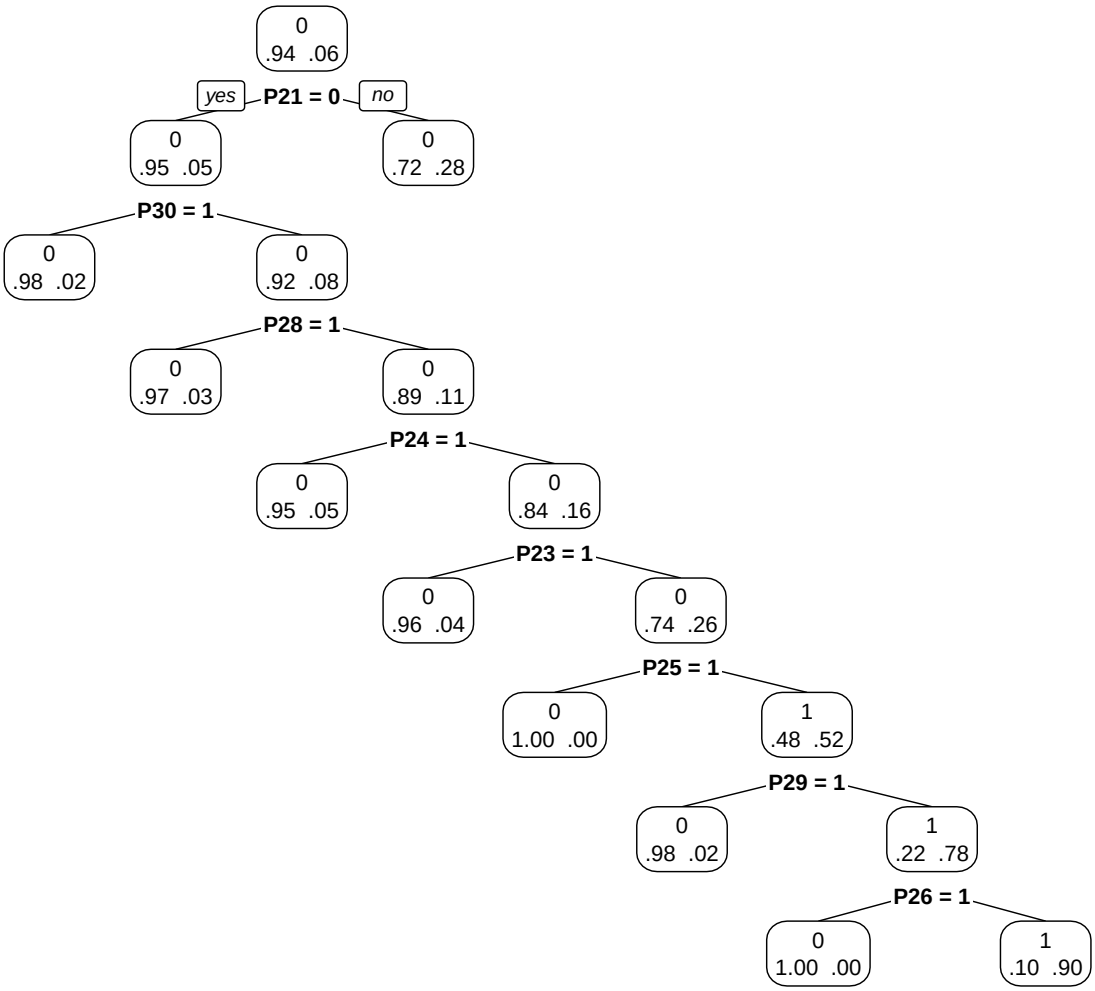


Figura 5.6: Árvore de decisão podada para o produto P20.

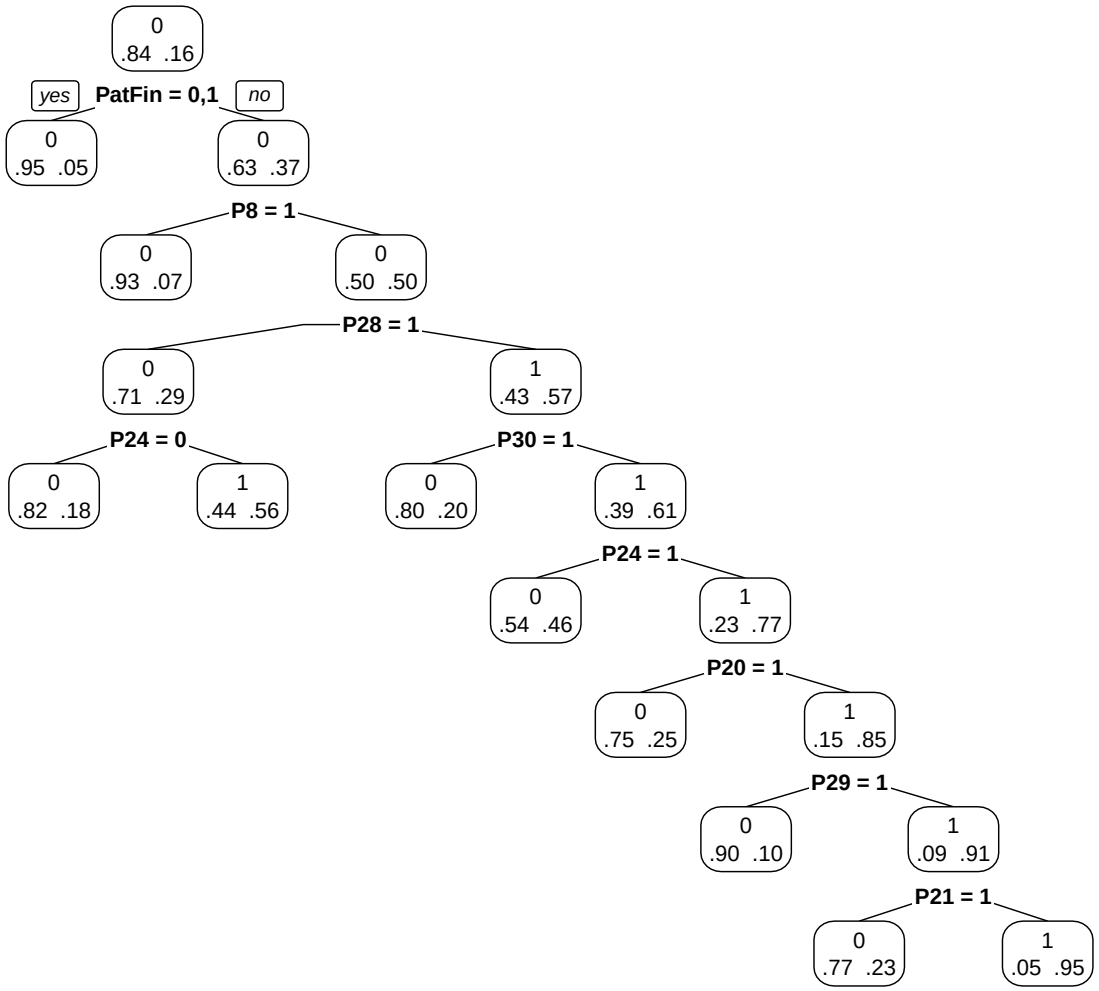


Figura 5.7: Árvore de decisão podada para o produto P23.

Em relação ao produto P25, a Figura 5.9 sugere dois grupos de decisões, ambos incluindo o produto P1. Num primeiro grupo a categoria do património financeiro do cliente é 1 ou 2, e para a venda do produto P25 poder ser equacionada, o cliente não deve possuir os produtos P28, P24, P30 e P23; também é possível realizar a venda a clientes que não possuam os produtos P20, P29 e P21. Num segundo grupo os cliente possuem a categoria mais baixa relativamente ao património financeiro, sendo possível concretizar a venda do produto P25 se o cliente não possuir o produto P30, embora se o cliente também não possuir os produtos P28, P24 e P26, por esta ordem, também ainda será possível concretizar a venda do produto. O erro de má classificação utilizando validação cruzada para o produto P25 é de aproximadamente 1%.

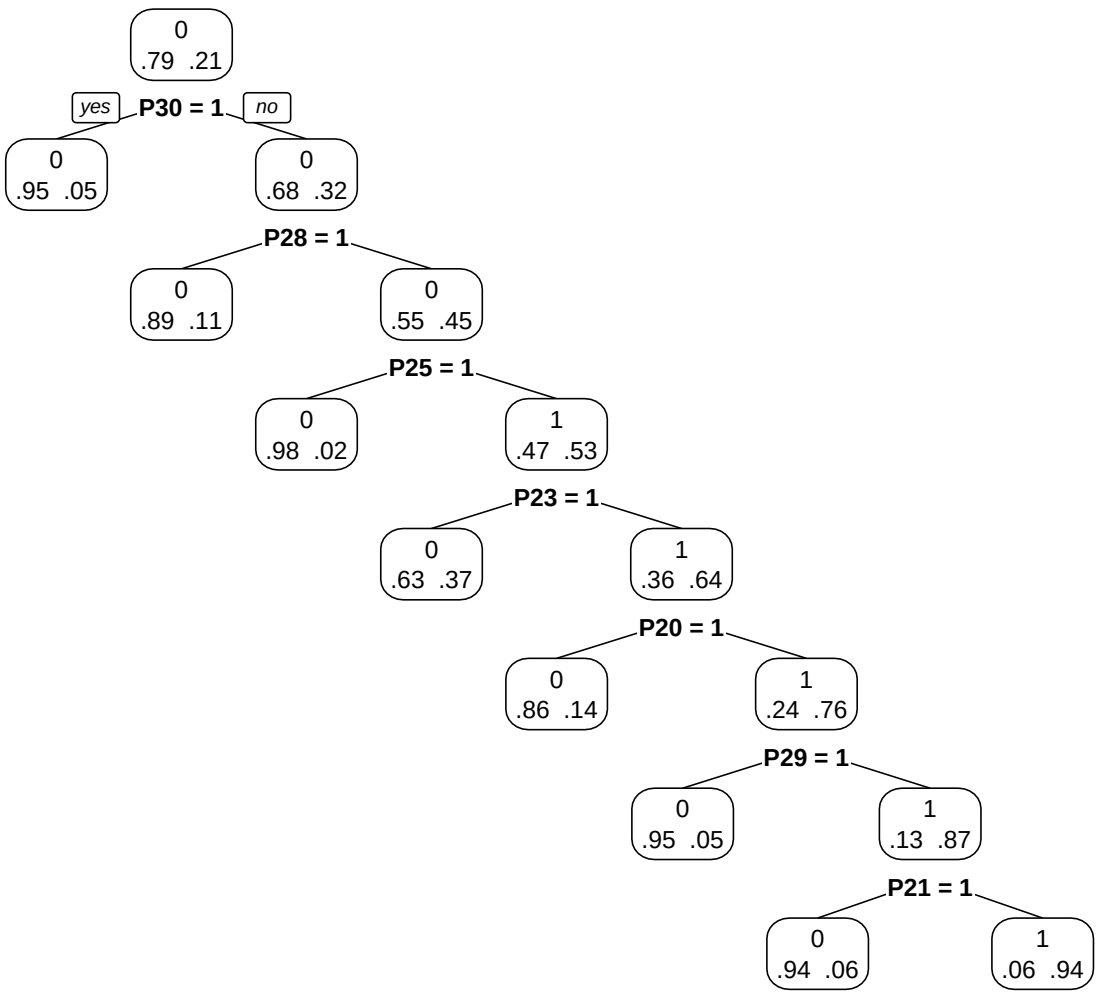


Figura 5.8: Árvore de decisão podada para o produto P24.

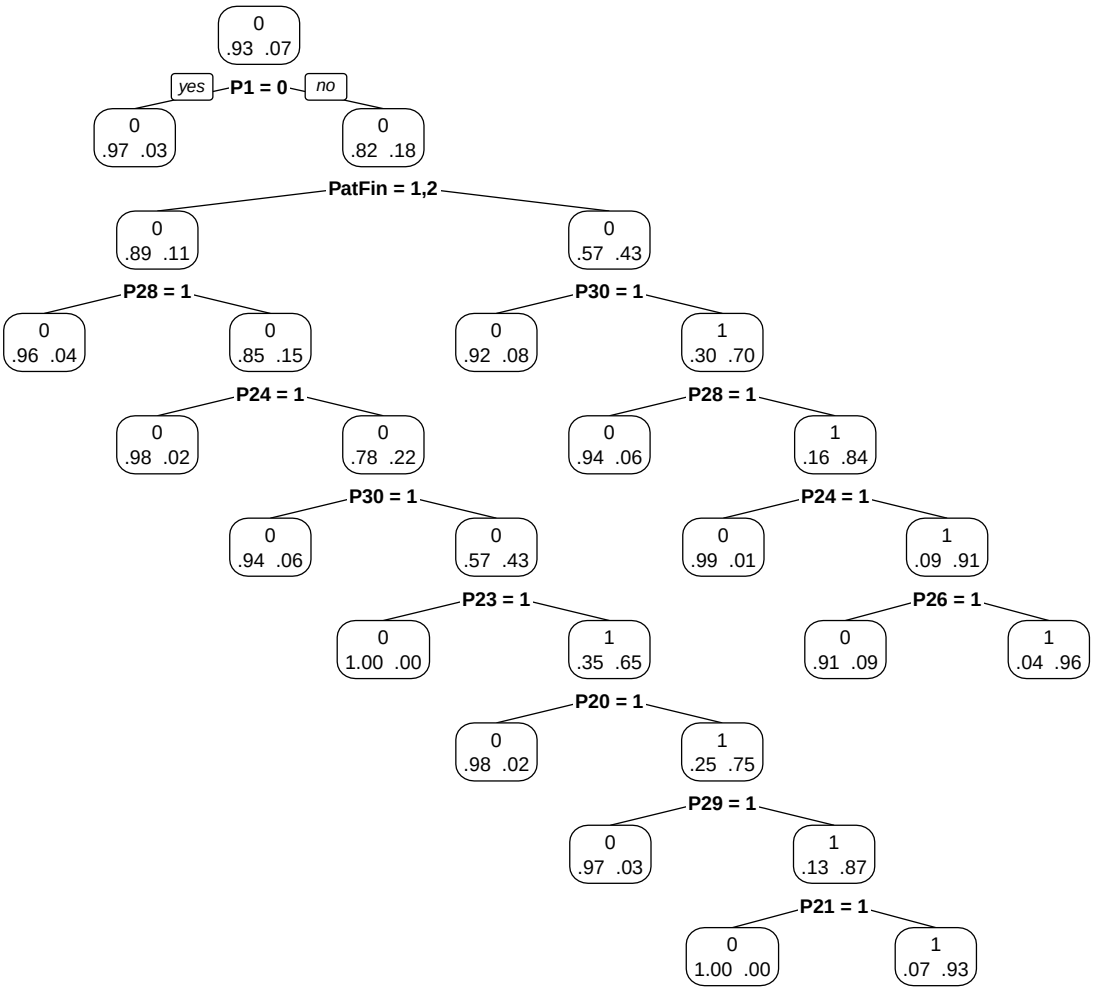


Figura 5.9: Árvore de decisão podada para o produto P25.

O produto P26 é geralmente adquirido por clientes com património financeiro na categoria mais baixa e que não possuem os seguintes produtos: P30, P28, P25, P24, P20, P29 e P21. E tal como sugere a Figura 5.10, se o cliente também não possuir o produto P23 então tem uma probabilidade alta de adquirir o produto P26. O erro de má classificação utilizando validação cruzada para a árvore de decisão sobre o produto P26 é de aproximadamente 0.5%.

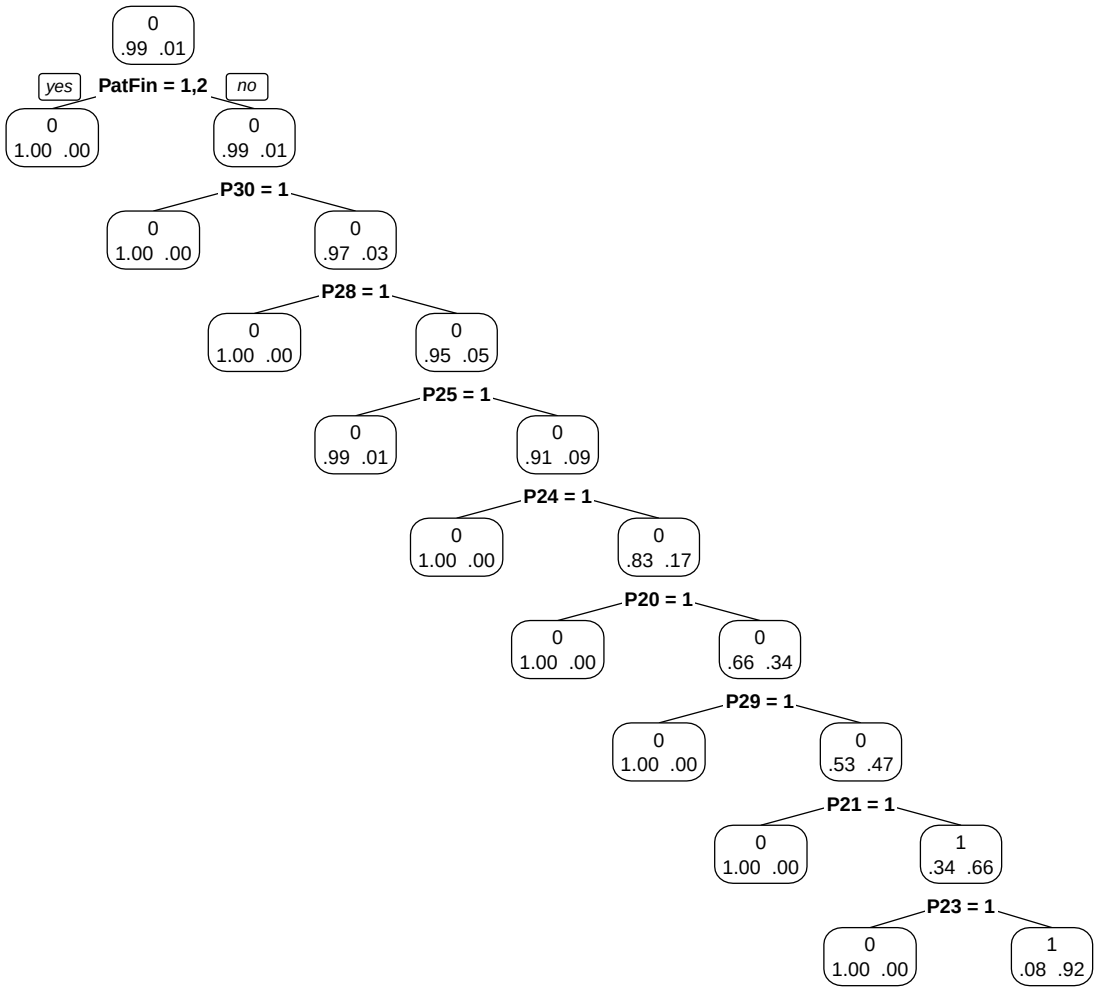


Figura 5.10: Árvore de decisão podada para o produto P26.

A árvore de decisão para o produto P28 foi construída com informação sobre os outros produtos, como mostra a Figura 5.11. É possível concretizar a venda do produto P28 se o cliente não possuir o produto P30 nem P24. Também é possível realizar a venda deste produto se o cliente não possuir os produtos P23, P25, P20 e P29. O erro de má classificação utilizando validação cruzada para o produto P28 foi de aproximadamente 3%.

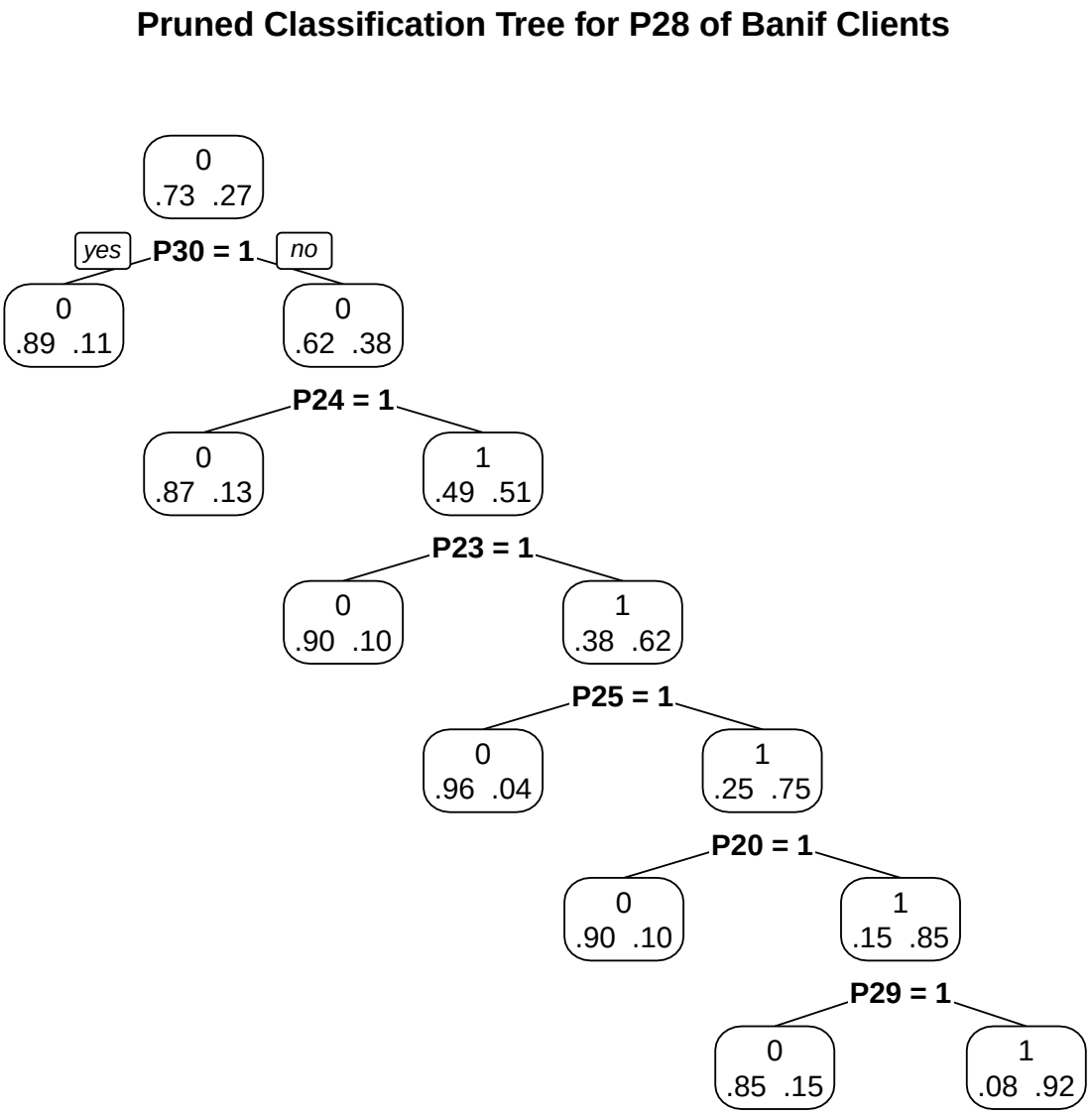


Figura 5.11: Árvore de decisão podada para o produto P28.

Relativamente ao produto P29, a árvore de decisão construída, mais uma vez, deve-se inteiramente à informação sobre a compra de outros produtos. A Figura 5.12 sugere que se um cliente não possuir os produtos P30, P28, P24, P23, P25 e P20 pode concretizar a compra do produto P29. Também existe a possibilidade de venda do mesmo produto se para além dos produtos já mencionados também não possuir o produto P21; ou P21 e P26; ou ainda P21, P26 e P22. O erro de má classificação utilizada validação cruzada para a construção da árvore podada foi de aproximadamente 0.4%.

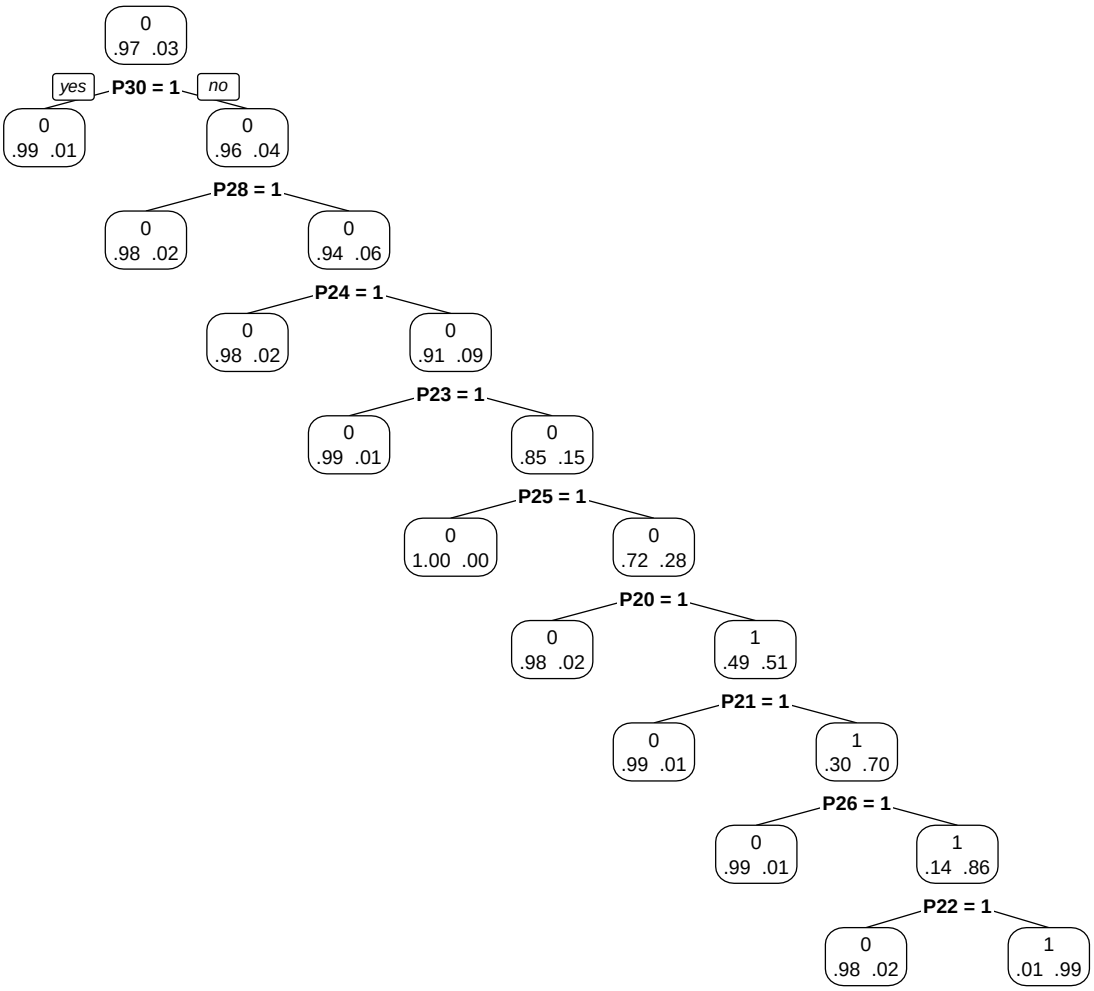


Figura 5.12: Árvore de decisão podada para o produto P29.

E por fim, o produto P30, cuja árvore de decisão utiliza informação sobre outros produtos e sobre o património financeiro, e divide as decisões em dois grupos. A Figura 5.13 sugere que num primeiro grupo para a compra de P30 encontram-se os clientes com património financeiro das categorias 1 e 2, que não possuam os produtos P9, P23, P24 e P28 ou que não possuam P20, ou ainda P20 e P29. Ainda no grupo dos clientes com património financeiro nas categorias 1 ou 2, caso possuíssem o produto P9 ou se ainda não possuíssem o produto P28, seria possível realizar a venda de P30. Num segundo grupo, se os clientes são da categoria mais baixa do património financeiro é possível concretizar a venda do produto P30. Contudo também é possível concretizar esta venda, sabendo que o cliente não possui P28; ou P28 e P25; ou P28,

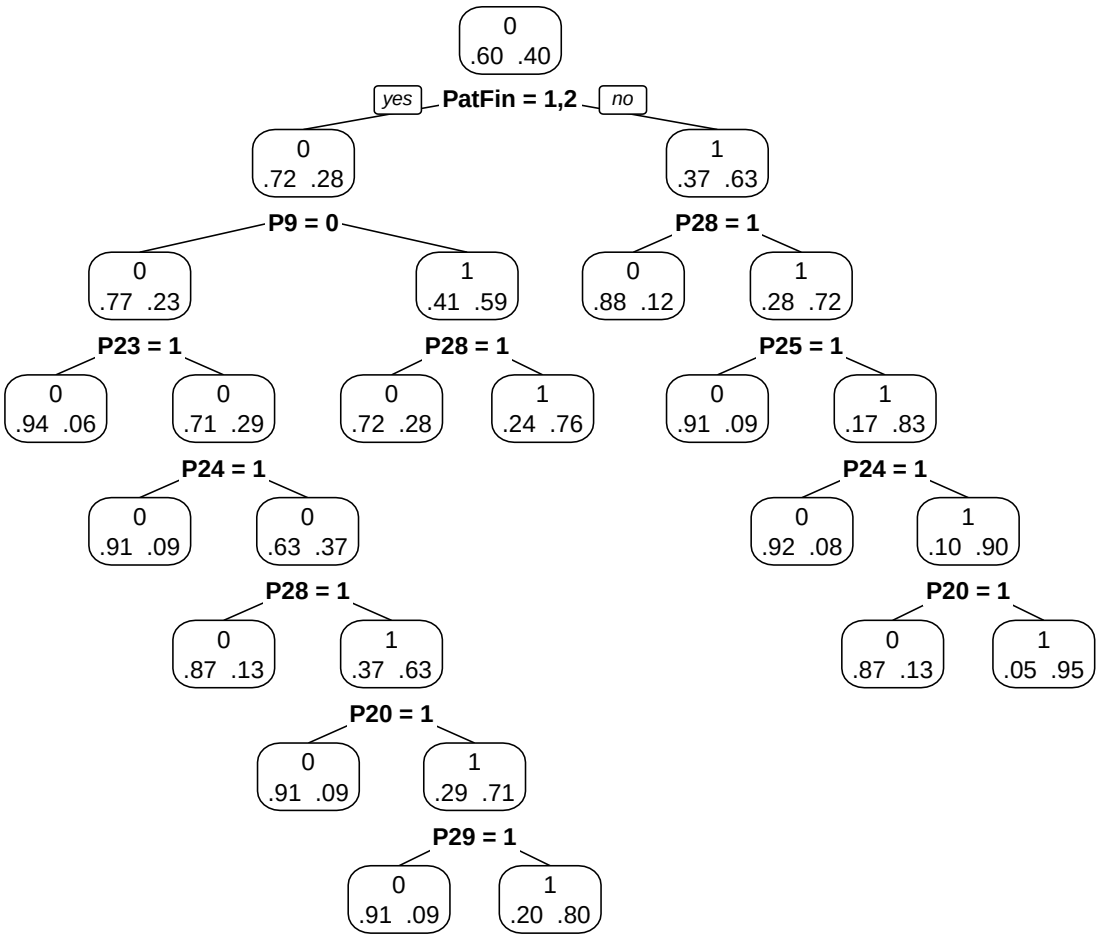


Figura 5.13: Árvore de decisão podada para o produto P30.

P25 e P24; ou ainda P28, P25, P24 e P20. O erro de classificação utilizando validação cruzada foi de aproximadamente 6.5%.

5.3 Modelos de Misturas Finitas

As análises desta secção foram realizadas usando o ambiente estatístico *MPLUS* versão 6.12 (Muthén and Muthén, 2011). Os dados utilizados para análise são referentes a todos os clientes particulares ativos que possuem produtos de *cross-selling*. Consideraram-se modelos de misturas finitas tendo como variáveis de entrada os produtos P1 a P30. As variáveis concomitantes consideradas foram: sexo, idade, património financeiro, saldo médio semestral, total de recursos e anos como cliente. Estas variáveis concomitantes permitem determinar os pesos das componentes do modelo. Foram ajustados modelos com diferentes números de classes. As medidas BIC e AIC foram diminuindo à medida que o número de classes aumentava (Tabela 5.2). Devido à complexidade computacional dos modelos e ao elevado número de parâmetros, estes demoraram entre 22 a 91 horas a correr. Cada modelo de misturas providencia a probabilidade de um dado objeto pertencer a determinada classe, e através destas probabilidades é possível atribuir o objeto a uma classe. Analisando o tamanho (em percentagem) dos indivíduos, em cada uma das classes, a Tabela 5.3 sugere que o modelo com quatro classes consegue explicar e separar melhor os indivíduos. Isto porque, nos modelos com duas e três classes, um destes últimos possui cerca de metade dos indivíduos, e também não se pretende muitas classes e com poucos indivíduos tal como acontece no modelo com cinco classes. A escolha do modelo com quatro classes vai de encontro ao melhor valor da medida de entropia relativa da Tabela 5.2, o mais alto valor entre 0 e 1.

De seguida, o modelo com quatro classes é então avaliado através das probabilidades a posteriori de pertença a cada classe e das diferenças existentes entre as classes para cada variável considerada. Analisando os histogramas da Figura 5.14 verificou-se que os indivíduos estão razoavelmente bem atribuídos a cada uma das classes, dado o maior número de observações nas caudas dos histogramas. De seguida, o modelo com quatro classes é avaliado através

Tabela 5.2: Medidas de classificação e duração do processamento dos modelos.

N.º classes	AIC	BIC	Entropia Relativa	Tempo
2	1450268,417	1450982,852	0,876	22h
3	1405031,818	1406072,599	0,894	40h
4	1382467,482	1383834,610	0,901	61h
5	1366690,334	1368383,808	0,894	91h

Tabela 5.3: Tamanho em percentagem de cada classe, para os diversos modelos.

Classe n.º	2 Classes	3 Classes	4 Classes	5 Classes
1	52%	49%	13%	31%
2	48%	36%	33%	11%
3		15%	40%	31%
4			14%	13%
5				14%

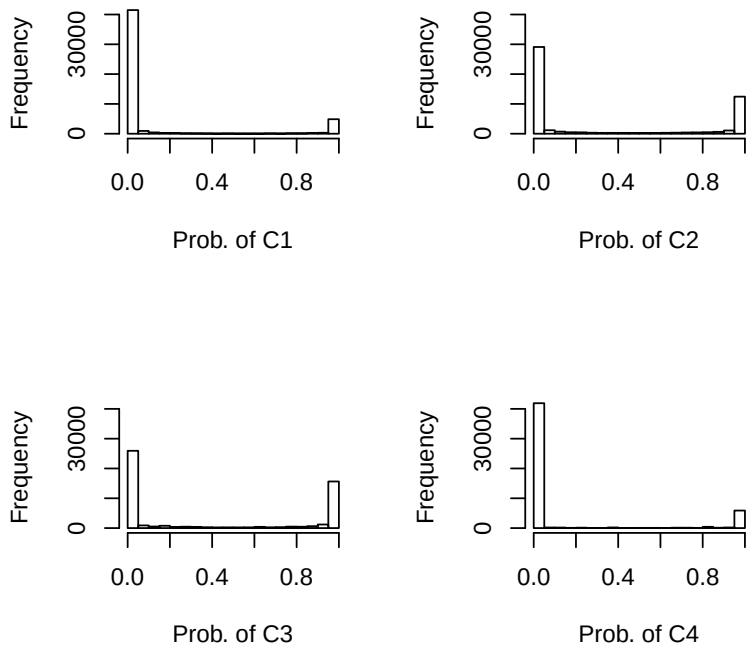


Figura 5.14: Histogramas das probabilidades de pertença dos indivíduos a cada *cluster*.

da determinação das diferenças entre os *clusters* obtidos para cada variável. Com esse fim, foi utilizado o teste paramétrico ANOVA e o teste de Tukey para comparação de populações. O teste ANOVA teve como hipótese nula de que todos os grupos são da mesma população, i.e., as médias dos grupos são iguais. Os resultados do teste ANOVA indicaram que nas variáveis TRec, P6, P13 a P16, P20, P22 e P27, a hipótese nula teve que ser rejeitada, dado a sua significância ser superior a 0.05. A comparação do teste de Tukey é realizada através de um teste de hipóteses, cuja hipótese nula assume que as médias são da mesma população, contra a hipótese alternativa de que pelo menos duas das médias são de populações diferentes. O teste de Tukey identificou diferenças em cada um dos *clusters*, nos produtos P3, P9, P11, P13, P17 e P30; nos anos como cliente, entre as categorias 2 e 3; no saldo médio semestral, entre as categorias 2 e 1; e na idade, entre as categorias 3 e 2. Os restantes pares de categorias das restantes variáveis não apresentaram diferenças significativas.

As estimativas do modelo de misturas foram calculadas numa escala de probabilidade, estando a informação disposta numa escala mais interpretável. A utilização desta escala permite indicar qual a probabilidade de um indivíduo de uma classe possui sobre pertencer a uma categoria de uma variável. Em todas as variáveis foi omitida a categoria 0, cuja estimativa para uma variável é a diferença entre 1 e a soma das probabilidades das restantes categorias. As estimativas do modelo para os produtos permite diferenciar as classes. A Figura 5.15 indica que um cliente da classe 1 possui uma probabilidade superior a 50% de possuir os produtos P1, P2, P4, P8 ou P10. Um cliente da classe 2 possui uma probabilidade superior a 50% de adquirir os produtos

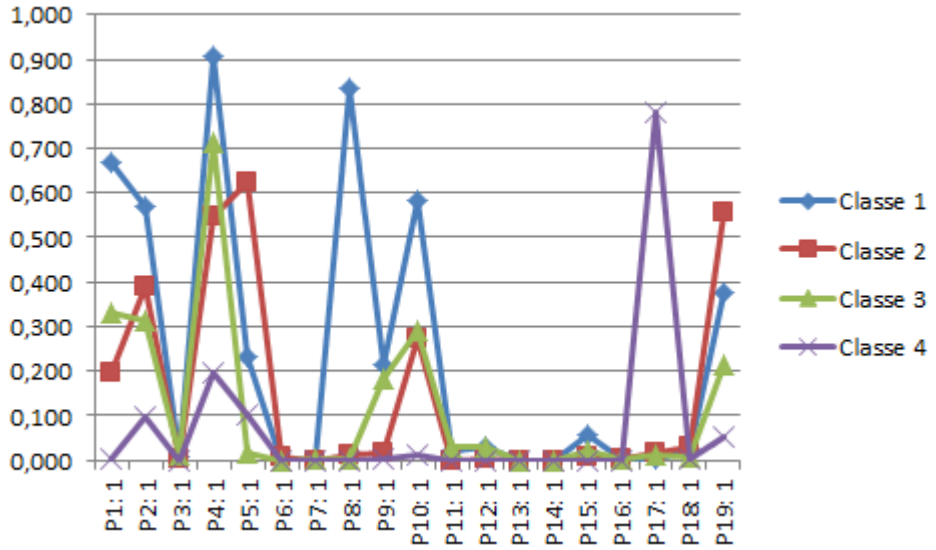


Figura 5.15: Estimativas do modelo sobre os produtos P1 a P19, numa escala de probabilidade.

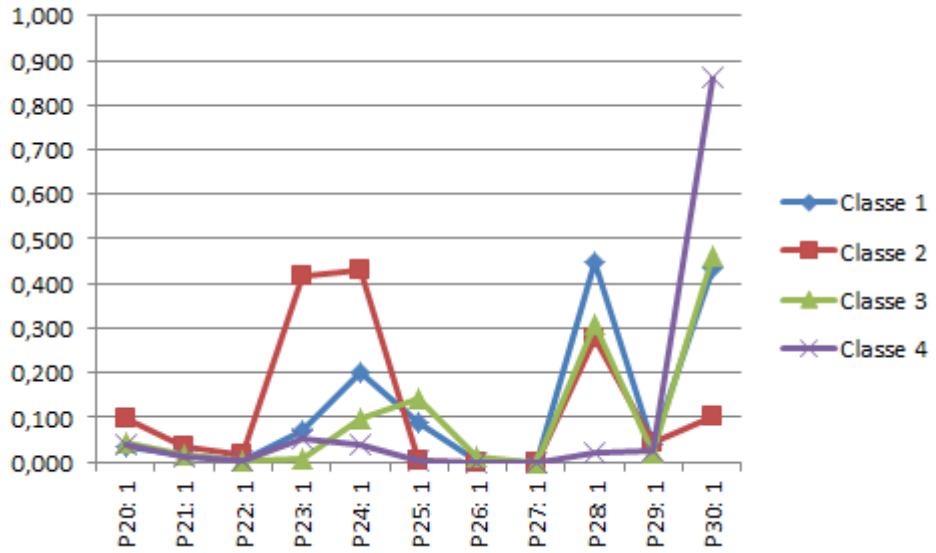


Figura 5.16: Estimativas do modelo sobre os produtos P20 a P30, numa escala de probabilidade.

P4, P5 ou P19. Os clientes da classe 3 não possuem muitos produtos, e apenas o produto P4 possui uma probabilidade superior a 50% de ser adquirido. O produto P17 ganha destaque para os clientes da classe 4, com uma probabilidade superior a 70% de ser adquirido. As estimativas sobre os produtos de *cross-selling* são apresentadas na Figura 5.16. A maioria destes produtos apresenta estimativas muito próximas; embora os clientes da classe 1 adquirem mais P2, na classe 2 os produtos P23 e P24 são os mais comprados, e na classe 4 o produto mais adquirido é P30. Os clientes da classe 3 são muito similares aos clientes da classe 1 sobre os produtos de *cross-selling*.

As classes dos clientes com produtos de *cross-selling* estão descritas numericamente na Tabela

5.4, em todas as variáveis a categoria 0 foi omitida pois o seu valor é a diferença entre 100% e a soma da percentagem das restantes categorias. Esta tabela não apresenta frequências absolutas devido à confidencialidade dos dados, para não divulgar o número total de clientes que possui produtos de *cross-selling*. Reunida toda a informação de avaliação do modelo foi possível segmentar os clientes em quatro grupos distintos: os mais antigos, os consolidados, os medianos e os jovens. Em todos os grupos há uma predominância do género masculino; todavia o género refere-se ao primeiro titular da conta e se esta última possuir outros titulares do género feminino estes não são considerados.

Os clientes mais antigos correspondem à classe 1 sendo 13% dos clientes considerados nesta análise. Os clientes mais jovens (categoria 0 da variável Idade) não estão incluídos neste grupo. Possuem elevado património financeiro e um saldo médio semestral e um total de recursos mediano. A maioria dos clientes possui os produtos P1, P2, P4, P8, P10 e relativamente aos produtos de *cross-selling*, 45% dos clientes possui P28. Na sua generalidade, os clientes estão fidelizados e utilizam vários serviços disponibilizados pelo banco além de possuírem algum crédito.

Os clientes consolidados correspondem à classe 2 constituindo 33% dos clientes considerados nesta análise. Os clientes mais jovens não estão incluídos neste grupo. Estes clientes possuem património financeiro médio-alto e um saldo médio semestral e um total de recursos em média elevados. A maioria dos clientes possui os produtos: P4, P5 e P19 e relativamente aos produtos de *cross-selling* P20, P23 e P24 foram os produtos que apesar de não ser possuído pela maioria, é possuído em maior quantidade relativamente aos restantes grupos. No geral, os clientes já são clientes há algum tempo, e utilizam produtos de poupanças ou investimentos.

Os clientes medianos correspondem à classe 3 sendo o maior segmento de clientes com 40% dos clientes considerados na análise. As idades dos clientes está em média na categoria 2. O património financeiro, saldo médio semestral e o total de recursos são relativamente baixos. A maioria dos clientes possui o produto P4 e relativamente aos produtos de *cross-selling*, o produto P25 é possuído por cerca de 14% dos clientes, contudo este segmento é o que mais contém clientes com este produto. Na sua generalidade, os clientes utilizam serviços disponibilizados pelo banco e são também possuidores de pequenos créditos.

Por último, o segmento dos clientes jovens corresponde à classe 4 constituído por 14% dos clientes considerados na análise. A denominação deste segmento deve-se precisamente ao facto de todos os clientes deste segmento estarem na categoria mais baixa da idade e não serem clientes há muito tempo. O património financeiro, saldo médio semestral e o total de recursos são relativamente baixos. A maioria dos clientes possui o produto P17 e relativamente aos produtos de *cross-selling* o produto P30. No geral, estes clientes possuem alguns recursos e possuem produtos de poupança.

Tabela 5.4: Descrição das classes.

		Classe 1	Classe 2	Classe 3	Classe 4
Tamanho		13%	33%	40%	14%
Sexo:	1	71%	65%	61%	52%
Idade:	1	30%	14%	25%	0%
	2	36%	18%	25%	0%
	3	24%	22%	21%	0%
	4	6%	40%	13%	0%
PatFin:	1	4%	33%	43%	23%
	2	96%	66%	1%	1%
SldMdSem:	1	30%	23%	44%	27%
	2	49%	65%	19%	3%
TRec:	1	40%	17%	38%	52%
	2	32%	79%	4%	6%
AnosCli:	1	20%	23%	26%	29%
	2	38%	27%	32%	12%
	3	38%	33%	23%	1%
P1:	1	67%	20%	33%	0%
P2:	1	56%	39%	32%	9%
P3:	1	1%	0%	1%	0%
P4:	1	91%	55%	72%	19%
P5:	1	22%	63%	1%	10%
P6:	1	0%	1%	0%	0%
P7:	1	0%	0%	0%	0%
P8:	1	88%	1%	0%	0%
P9:	1	21%	2%	18%	0%
P10:	1	58%	28%	29%	1%
P11:	1	2%	0%	3%	0%
P12:	1	3%	0%	3%	0%
P13:	1	0%	0%	0%	0%
P14:	1	0%	0%	0%	0%
P15:	1	6%	1%	2%	0%
P16:	1	0%	0%	0%	0%
P17:	1	0%	2%	1%	78%
P18:	1	2%	3%	1%	0%
P19:	1	37%	56%	22%	5%
P20:	1	3%	10%	4%	4%
P21:	1	2%	4%	2%	1%
P22:	1	0%	2%	0%	0%
P23:	1	7%	42%	1%	5%
P24:	1	20%	43%	10%	4%
P25:	1	9%	1%	14%	0%
P26:	1	0%	0%	1%	0%
P27:	1	0%	0%	0%	0%
P28:	1	45%	28%	31%	2%
P29:	1	4%	4%	2%	3%
P30:	1	44%	10%	46%	86%

Capítulo 6

Conclusão

O propósito desta dissertação centrou-se no estudo de *cross-selling* de produtos da instituição bancária Banif. Foram utilizados dados de clientes obtidos em setembro de 2013. Numa análise cuidada sobre os dados, e dado o grande volume de clientes, foi necessário utilizar metodologias que pudessem ser aplicadas a grandes volumes de dados. Os dados apresentaram uma grande quantidade de valores omissos em algumas variáveis, devido ao facto da não obrigatoriedade do seu preenchimento desde a abertura da instituição. Os dados convergidos de várias *sub-holdings* por vezes possuíam diferentes tipos de informação, tendo sido necessário generalizar a informação sobre cada produto; nomeadamente, indicar se o cliente era possuidor do produto discriminando, em alguns produtos, o número total de produtos.

Uma estratégia para uma maior eficiência de *cross-selling* necessita de informação sobre o produto a vender, mais precisamente sobre quando e que canal deverá ser utilizado. Contudo os dados disponibilizados só possuíam informação sobre os clientes. Por essa razão, definiu-se uma estratégia para saber que produtos deverão ser vendidos a que clientes, no sentido de concretizar vendas de *cross-selling*. De futuro, durante a venda efetiva de produtos de *cross-selling*, seria útil guardar informação sobre quando é que se deu a venda e por que canal (*e-mail*, telefone ou outros), de forma a complementar as análises feitas neste trabalho.

A aplicação de regras de associação permitiu evidenciar que não existem muitos padrões de associações sobre os produtos considerados de *cross-selling*. O facto de poucos clientes possuírem tais produtos influenciou bastante este resultado. Mostrando que as regras de associação não são uma boa escolha para a criação de um modelo de *cross-selling*, porque descrevem promoções anteriores de *marketing*.

As árvores de decisão obtidas não incluíram informação do cliente relativamente à idade, sexo, situação profissional ou informação do cliente bancário, tendo sido na sua maioria incluída informação sobre a não posse de determinados produtos sabendo que estes não estão correlacionados entre si. Este facto deve-se à maioria dos clientes não possuir muitos produtos. Os erros de má classificação mostraram-se baixos, sendo possível estruturarem-se regras de venda para cada um dos produtos considerados de *cross-selling*.

O modelo de misturas finitas possui uma grande vantagem sobre os restantes métodos utilizados;

considerando as informações do cliente como variáveis concomitantes, foi possível observar o impacto que estas exercem sobre a compra dos produtos. Nas metodologias anteriores isto não aconteceu, pois as informações dos clientes não sobressaiam perante os produtos, pouco adquiridos pelos clientes.

A melhor estratégia de *cross-selling* para este caso de estudo seria utilizar a segmentação de clientes obtida através do modelo de misturas finitas e, perante os produtos de *cross-selling* mais vendidos, serem utilizadas as árvores de decisão para melhor direcionar a venda para o produto certo. Contudo, deve-se ter presente que o estudo poderá não ser representativo da população dos clientes do Banif, uma vez que foram utilizados apenas os clientes que efetivamente possuíam produtos de *cross-selling*.

Referências

Agrawal, R., Imieliski, T., and Swami, A. (1993). Mining association rules between sets of items in large databases. In *Proceedings of 1993 ACM SIGMOD International Conference on Management of Data*, pages 207–216.

Agrawal, R. and Srikant, R. (1994). Fast algorithms for mining association rules. In *Proc. of 20th Intl. Conf. on VLDB*, pages 487–499.

Akpinar, S. and Akpinar, E. K. (2009). Estimation of wind energy potential using finite mixture distribution models. *Energy Conversion and Management*, 50(4):877–884.

Anand, S., Patrick, A., Hughes, J., and Bell, D. (1998). A data mining methodology for cross-sales. *Knowledge-based systems*, 10:449–461.

Barbara, D., Couto, J., Jajodia, S., and Wu, N. (2001). Adam: A testbed for exploring the use of data mining in intrusion detection. *SIGMOD Record*, 30(4):15–24.

Berry, M. J. and Linoff, G. (2004). *Data Mining Techniques: For Marketing, Sales, and Customer Support*. John Wiley & Sons, Inc., New York, NY, USA, second edition.

Breiman, L., Friedman, J., Stone, C., and Olshen, R. (1984). *Classification and Regression Trees*. The Wadsworth and Brooks-Cole statistics-probability series. Taylor & Francis.

Buttle, F. (2009). *Customer Relationship Management: Concepts and Technologies*. Butterworth-Heinemann.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of The Royal Statistical Society, Series B*, 39(1):1–38.

Duda, R., Hart, P., and Stork, D. (2001). *Pattern classification*. Pattern Classification and Scene Analysis: Pattern Classification. Wiley.

Fahey, M. T., Thane, C. W., Bramwell, G. D., and Coward, W. A. (2007). Conditional gaussian mixture modelling for dietary pattern analysis. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 170(1):149–166.

Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). From data mining to knowledge discovery in databases. *American Association of Artificial Intelligence*.

Ferrall, C. (2005). Solving finite mixture models: Efficient computation in economics under serial and parallel execution. *Computational Economics*, 25(4):343–379.

Hahsler, M., Buchta, C., Gruen, B., and Hornik, K. (2009a). *arules: Mining Association Rules and Frequent Itemsets*. R package version 0.6-8.

Hahsler, M. and Chelluboina, S. (2013). *arulesViz: Visualizing Association Rules and Frequent Itemsets*. R package version 0.1-7.

Hahsler, M., Grün, B., Hornik, K., and Buchta, C. (2009b). Introduction to arules – A computational environment for mining association rules and frequent item sets. *The Comprehensive R Archive Network*.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. Springer Series in Statistics. Springer.

Hipp, J., Güntzer, U., and Nakhaeizadeh, G. (2000). Algorithms for association rule mining – a general survey and comparison. *SIGKDD Explorations*, 2(2):1–58.

Holt, J. D. and Chung, S. M. (1999). Efficient mining of association rules in text databases. In *Proc. of the 8th Intl. Conf. on Informatics and Knowledge Management*, pages 234–242.

Kamakura, W. (2008). Cross-selling: offering the right product to the right customer at the right time. *Journal of Relationship Marketing*, pages 41–58.

Klemettinen, M. (1999). *A Knowledge Discovery Methodology for Telecommunication Network Alarm Databases*. PhD thesis, University Of Helsinki.

Lee, D., Park, S.-H., and Moon, S. (2013). Utility-based association rule mining: A marketing solution for cross-selling. *Expert Systems with Applications*, 40(7):2715–2725.

Lee, W., Stolfo, S. J., and Mok, K. M. (2000). Adaptive intrusion detection: A data mining approach. *Artificial Intelligence Review*, 14(6):533–567.

Leisch, F. and Grun, B. (2008). FlexMix Version 2 : Finite Mixtures with Concomitant Variables and Varying and Constant Parameters. *Journal of Statistical Software*, 28(4).

Liu, B. (2011). *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data*. Data-Centric Systems and Applications. Springer.

Lu, H., Han, J., and Feng, L. (1998). Stock movement and n-dimensional inter-transaction association rules. In *Proc. 1998 SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery*.

Milborrow, S. (2014). *rpart.plot: Plot rpart models. An enhanced version of plot.rpart.* R package version 1.4-4.

Muthén, L. and Muthén, B. (1998-2011). Mplus user's guide. Sixth Edition. Los Angeles, CA: Muthén & Muthén.

Olanow, C. W. and Koller, W. C. (1998). An algorithm (decision tree) for the management of parkinson's disease treatment guidelines. *Neurology*, 50(3 Suppl 3):S1–S1.

Pei, J., Han, J., Mortazavi-asl, B., and Zhu, H. (2000). Mining access patterns efficiently from web logs. In *Proc. of 4th Pacific-Asia Conf. on Knowledge Discovery and Data Mining*, pages 396–407.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Salazar, M. T., Harrison, T., and Ansell, J. (2007). An approach for the identification of cross-sell and up-sell opportunities using a financial services customer database. *Journal of Financial Services Marketing*, 12(2):115–131.

Satou, K., Shibayama, G., Ono, T., Yamamura, Y., E. Furuichi, S. K., and Takagi, T. (1997). Finding association rules on heterogeneous genome data. In *Proc. of the Pacific Symp. on Biocomputing*, pages 397–408.

Tan, P., Steinbach, M., and Kumar, V. (2014). *Introduction to data mining*. Always learning. Pearson Education, Limited.

Tan, P.-N. and Kumar, V. (2002). Mining association patterns in web usage data. In *Proc. of the Intl. Conf. on Advances in Infrastructure for e-Business, e-Education, e-Science and e-Medicine on the Internet*.

Therneau, T., Atkinson, B., and Ripley, B. (2013). *rpart: Recursive Partitioning*. R package version 4.1-3.

Wang, Q. R. and Suen, C. Y. (1984). Analysis and design of a decision tree based on entropy reduction and its application to large character set recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, PAMI-6(4):406–417.

Wedel, M. and DeSarbo, W. S. (2002). Market segment derivation and profiling via a finite mixture model framework. *Marketing Letters*, 13(1):17–25.

Wong, R., Fu, A., and Wang, K. (2005). Data mining for inventory item selection with cross-selling considerations. *Data mining and knowledge discovery*, 11(1):81–112.

Xiong, H., Shekhar, S., Tan, P. N., Kumar, V., and Holbrook, S. R. (2005). Identification of functional modules in protein complexes via hyperclique pattern discovery. In *Proc. of the Pacific Symp. on Biocomputing*.